Estimation and Optimization of Information Measures with Applications to Fairness and Differential Privacy

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Estimation and Optimization of Information Measures with Applications to Fairness and Differential Privacy"

presented by: Wael Mohammed A Alghamdi

Signature ______________________________
Typed name: Professor F. Calmon

Signature Cynthia Dwork ______________________________
Typed name: Professor C. Dwork

Signature ______________________________
Typed name: Professor Y. Lu

Signature ______________________________
Typed name: Professor S. Vadhan

April 27, 2023

Estimation and Optimization of Information Measures with Applications to Fairness and Differential Privacy

A dissertation presented

by

Wael Mohammed A Alghamdi

to

The Harvard John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Mathematics

Harvard University

Cambridge, Massachusetts

April 2023

Dissertation Advisor: Prof. Flavio Calmon Wael Mohammed A Alghamdi

Estimation and Optimization of Information Measures with Applications to Fairness and Differential Privacy

Abstract

My dissertation solves three theoretical problems on optimizing and estimating information measures, and it also builds on this theory to introduce novel practical algorithms for: 1) Optimal mechanism design for differential privacy (DP); 2) Optimal group-fair enhancement in machine learning; and 3) Estimation of information measures from data using sample moments. Information measures (in particular, f-divergences) provide a rigorous way to tackle several real-world problems. Examples include: 1) Quantifying the degree of *privacy* afforded by data releasing mechanisms—using the *hockey-stick divergence*; 2) Correcting machine learning (ML) trained classifiers for *group-fairness*—via optimizing *cross-entropy*; and 3) Detecting new dependencies between pairs of natural phenomena—via estimating *mutual information* from data. Herein, we put forth mathematically grounded approaches for the above three practical problems. In the first third of the dissertation, we design optimal DP mechanisms in the large-composition regime, and we also derive a fast and accurate DP accountant for the large-composition regime via the method of steepest descent from mathematical physics. We prove that the privacy parameter is equivalent to a KL-divergence term, then we provide solutions to the ensuing minmax KL-divergence problem. In the second third of the dissertation, we generalize the ubiquitous concept of *information projection* to the case of conditional distributions—which we term *model projection*. We derive explicit formulas for model projection, as well as a parallelizable algorithm to compute it efficiently and at scale. We instantiate our model projection theory to the domain of group-fair ML, thereby obtaining an optimal multi-class fairness enhancement method that runs in the order of seconds on datasets of size more than 1 million samples. In the last third of the dissertation, we derive the functional form of the relationship between information measures and the underlying moments. Plugging in the sample moments of data into our new moments-based formulas, we are able to estimate mutual information and differential entropy efficiently and robustly against affine-transformations of the samples.

Contents

List of Tables

List of Figures

Acknowledgments

I have been privileged throughout my journey at Harvard to be surrounded by great people, whom I hold in the highest regard and would like to sincerely thank.

To my awesome advisor, Professor Flavio Calmon, I cannot thank you enough. I could not have asked for a better advisor. You have made me the researcher that I am, and I will forever be grateful for that. Your continuous technical help, vision, and intuition facilitated my academic discoveries. And your kindness and patience helped me achieve that in a healthy environment. Your faith in me was a driving force in my determination and perseverance to do more. The list of qualities I admire in my advisor runs very long, and I know for sure that it will serve as my guideline when becoming an advisor myself.

My sincere thanks are also to my thesis committee: Professor Cynthia Dwork, Professor Salil Vadhan, and Professor Yue Lu. Your feedback greatly helped me realize important potential avenues for future work. I also thank my qualifying exam committee: Professor Cynthia Dwork, Professor Li Na, and Professor Madhu Sudan. You feedback was immensely helpful in the beginning of my Ph.D. for me to focus on important problems.

The researchers I had the pleasure to collaborate with deserve special thanks, and I credit them for much of the intellectual outcome of this dissertation. Professor Oliver Kosut had a great impact on my academic growth, from whom I learned how not to settle for a theorem with many assumptions. I am also thankful to Shahab Asoodeh, Hsiang Hsu, Haewon Jeong, Hao Wang, Felipe Gomez, and Professor Lalitha Sankar for the extensive collaborations we had.

Life was colorful during the Ph.D. years because of my friends. Thanks to Madeleine Barowsky, the most pleasant person one could hope to interact with. I greatly enjoy our fun random walks, and because of your heavenly cookies I cannot now rate other cookies above 7/10. You have also been always there for me and your perspective has helped me overcome many obstacles in my life, so truly thank you. I also thank again my collaborators Shahab, Hsiang, Haewon, Hao, and Felipe for being great friends outside the lab too. I will miss many of the activities we did together: the deep and helpful chats with my great friend Hsiang during our invented breaks; the fun times with Shahab and others; truly celebrating when accomplishments are done with Haewon and others; and of course I will miss the 2:00am crunch meetings with Hsiang and Haewon. Thanks to my other lab mates Lucas Monteiro, Carol Long, and Alex Oesterling for the joyful and productive times we spent together. I also thank my close group of friends in the first couple of years at Harvard for the fun times we had together: Donghun Lee, Oussama Dhifallah, and Ismail Ben Atitallah.

I am grateful to my old friends, including Louai Wafa, Abdullah Almogren, Abdulrahman Alzahrani, Sultan Aboalela, Meshal Alawein, Badr Alghamdi, Tariq Alturkestani, Nasir Alfaraj, Mansour Kamal, John Mikhael, Talal Al-Mulla, and Mina Nassif for the fun activities and interesting conversations we had together. Special thanks to Abdulrahman and Sultan for giving me great company throughout the Ph.D. years. I am also grateful to the new friends I made, including Bogdan Kulynych, Behrooz Razeghi, J. C. S. Santos Filho (Candido), Polina Kirichenko, Monica Welfert, Atefeh Gilani, and Serdar Aslan for the fun times we had in Boston and around the country at conferences.

To all my friends, knowing you has been and will continue to be joyful, and I will cherish these memories forever.

I am also thankful to my other collaborators, Wangyu Luo, Winston Michalak, Dennis Wei, Karthikeyan Ramamurthy, and Fei Wei; to people who helped me grow academically: Professor Yury Polyanskiy, Mario Diaz, Professor Mohamed-Slim Alouini, and Walid Abediseid; and to people who helped me grow professionally: Professor John Girash, Ann Greaney-Williams, Lisa Frazier Zezze, Allison Choat, and Rob Lorino.

Most importantly, I am forever indebted to my family for their unconditional love. I would not be where I am without the unwavering support of my mom, Nora, and dad, Mohammed, and I could not thank you enough for that. Thanks to my sisters Lujain and Mayada, and brothers Ayman and Haytham, for your love and support and for championing my success. I also thank my uncles Saeed and Hamed, grandma Fikrah, and cousin Turki for their love and support. I also thank my nephew Sahm who was born one week before my thesis defense and brought joy to the family, and the furball Bo who likes to steal my seat whenever I have important Zoom meetings.

To my beloved family.

Chapter 1

Introduction

Automated decision making has been increasingly integrated in practice, and that has been observed to sometimes come at a cost on the individual. Examples of such unwanted consequences include facing discrimination from predictions made by machine learning (ML) models [ALMK16, BDH+18], and having private data exposed [DMNS06, ACG+16]. Several algorithmic interventions for these problems have been introduced in the literature and deployed in practice, e.g., methods for training group-fairness aware classifiers [FSV+19] and optimized private data-release mechanisms [DMNS06]. Our approach is a disciplined framework to tackle the underlying mathematical problems via solving a variety of information-measures based optimization problems.

Information measures lie at the heart of the field of information theory. Similar to how norms in the usual Euclidean spaces quantify sizes of vectors, so do information measures provide a mathematically rigorous description of how much information a random event contains. For example, it is intuitive that one is more uncertain about the outcome of the roll of a fair die than they would be about the outcome of the toss of a fair coin. In this sense, rolling a die contains more *information* than tossing a coin. *Entropy* is an information measure that captures exactly this intuition, and it tells us that the former random event has $\log 6$ bits of information, which is greater than $\log 2$ bits, the entropy of the latter random event. The *Kullback-Leibler (KL) divergence* goes a step further, as it measures how *dissimilar* two random events are. For example, in coding theory, the KL-divergence measures the additional number of bits when encoding randomly sampled data using a codebook optimized for their underlying distribution.

Generalizing the concepts of entropy and KL-divergence, one arrives in information theory to a class of information measures that concretely quantify dissimilarity between probability measures: f-*divergences* [AS66, Csi67]. With $f : (0, \infty) \to \mathbb{R}$ being a convex function, one defines the f-divergence of a measure μ from another measure ν by

$$D_f(\mu \parallel \nu) := \mathbb{E}_\nu \left[f\left(\frac{d\mu}{d\nu} \right) \right] - f(1), \tag{1.1}$$

where $d\mu/d\nu$ is the Radon-Nikodym derivative. The KL-divergence is a concrete example of f-divergences, where it is defined by $D_{\mathrm{KL}}(P \parallel Q) := D_f(P \parallel Q)$ for the convex function $f(t) := t \log t$. Explicitly, if P and Q are discrete probability measures supported over $\{1, \cdots, n\}$, then the KL-divergence is given by

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}. \tag{1.2}$$

It is true that $D_{\mathrm{KL}}(P \parallel Q) \geq 0$ with equality if and only if $P = Q$. Thus, the KL-divergence quantifies how *dissimilar* P is from Q: the larger $D_{\mathrm{KL}}(P \parallel Q)$ is, the further P is from Q.

The appeal of f-divergences is that they give us a handle on modeling real-world problems, thereby opening the door for solving those concrete problems. Such solutions aid in, e.g., revealing optimal design models and estimating desirable figures of merit. The expressability via f-divergences of many real-world phenomena is true in virtue of them being best described using probabilities.

Consider for example the problem of achieving private data release. Differential Privacy (DP) [DMNS06] is the widely adopted standard in the practice of privacy-preserving machine learning algorithms [EPK14, Dif17, KMR+20]. A mechanism, i.e., randomized algorithm (viewed as a channel $P_{Y|X}$), is considered differentially private if its output distributions $P_{Y|X=x}$ and $P_{Y|X=x'}$ do not vary significantly with small deviation of the inputs (i.e., small $d(x, x')$ for some metric d). The input X is considered the true response to a query on a dataset containing sensitive information, and Y is its privatized version. Whether $X = x$ or $X = x'$ could be determined, for example, by whether a certain fixed individual is included in the queried dataset. When Y is a DP version of the query, an observer of Y cannot determine with high confidence whether Y came from x or x', thus they cannot associate any fixed individual to the publicly released data Y. Mathematically, DP can be expressed using the maximum of an f-divergence: $P_{Y|X}$ is (ε, δ)-DP for $\varepsilon \geq 0$ and $\delta \in [0, 1]$ if

$$\sup_{d(x,x') \leq s} D_{f_\varepsilon}(P_{Y|X=x} \parallel P_{Y|X=x'}) \leq \delta \tag{1.3}$$

where $f_\varepsilon(t) = \max(0, t - e^\varepsilon)$, D_f is as defined in (1.1), and $d(x, x') \leq s$ is some preset condition capturing "neighboring" inputs. This special case of f-divergence is known as the *hockey-stick* *divergence* [PPV10]. Hence, the DP condition (1.3) puts a uniform bound on how dissimilar a pair of publicly released data are when they come from similar true queries. From this definition of DP, one sees that a smaller value of the bound δ corresponds to more privacy. The definition of DP in (1.3) not only gives us a rigorous nd practical way to measure privacy, but this f-divergence expression also allows us to study the important problem of optimal DP mechanism design. Specifically, as lower values of δ are desirable, the optimal DP mechanism (for the *single-shot* setting) is the one that minimizes the maximal hockey-stick divergence in (1.3). Trivially, one could pick Y to be independent of X and attain a value of $\delta = 0$ for every ε. However, this choice is evidently impractical since it ignores the utility side: the released data Y tell us nothing about the desired query X. Generally, any given random mechanism $P_{Y|X}$ will decrease the utility. Thus, a sensible DP mechanism-design optimization problem would take the minmax form

$$\inf_{P_{Y|X}\in\mathcal{U}} \ \sup_{d(x,x')\le s} \ D_{f_\varepsilon}(P_{Y|X=x} \parallel P_{Y|X=x'}), \tag{1.4}$$

where belonging to $\mathcal{U}$ means that a certain prescribed utility measure is attained. In words, solving (1.4) amounts to finding the mechanism $P^\star_{Y|X}$ that optimizes the privacy-utility trade-off. A number of recent works tackle this optimization problem in a variety of settings [GRS12, GS10, GV15, GKOV15, SCDF13, GV16, GDGK20, GDGK19], but they all consider only the settings of single-shot private data release and scalar-valued queries. In this dissertation, we address the more complicated setting of optimal DP mechanism design under composition and with vector-valued queries. In other words, we consider X to be $\mathbb{R}^m$-valued, and we apply the mechanism $P_{Y|X}$ to $k > 1$ different queries. In this case, the probability measure $P_{Y|X=x}$ is replaced in (1.4) by the product of measures $P_{Y|X=x_1} \times \cdots \times P_{Y|X=x_k}$. The composition setting yields an intractable expression for the DP parameters, and our work in Chapter 2 addresses this challenge.

Challenge 1. *Given the intractability of the value* $\sup_{d(x_j,x_j')\le s; 1\le j\le k} \ D_{f_\varepsilon}(\prod_{j=1}^k P_{Y|X=x_j} \parallel \prod_{j=1}^k P_{Y|X=x_j'})$ *of the privacy parameter* δ*, how can we find mechanisms* $P_{Y|X}$ *with favorable privacy-utility trade-off after a large number of compositions? In addition, how can we quantify privacy in the large-composition regime accurately and efficiently?*

Another f-divergence based optimization considered in this thesis is *model projection*. In contrast to the maximal f-divergence considered in (1.3)–(1.4), we now consider minimizing the *average* f-divergence of an arbitrary conditional distribution from another fixed one. Consider the following "abstract" experiment. We have a random event X, which is inaccessible to us, and that we believe causes another random event Y, which is observable by us. We collect many samples of Y and we build a model $P_{Y|X}$ that captures the transition probabilities from X to Y using a desired method $\mathscr{M}$. However, we observe, after the fact, that this model $P_{Y|X}$ violates some *other* necessary properties. For example, we might know—or postulate—via some other physical law $\mathscr{L}$ that we must have $P_{Y|X} \in \mathcal{F}$ for some feasible set of models $\mathcal{F}$ but it so happens that $P_{Y|X} \notin \mathcal{F}$. It is natural then to ask: what is the closest member of $\mathcal{F}$ to the fitted model $P_{Y|X}$? Mathematically, closeness of a candidate $Q_{Y|X} \in \mathcal{F}$ to $P_{Y|X}$ can be captured via an f-divergence. Taking an averaged penalty, i.e., the penalty $D_f(Q_{Y|X=x} \,\|\, P_{Y|X=x})$ is weighted by $P(dx)$, we arrive thus at the optimization problem

$$\begin{aligned} \underset{Q_{Y|X}}{\text{minimize}} \quad & \mathbb{E}_{Z\sim P_X}\left[D_f(Q_{Y|X=Z} \,\|\, P_{Y|X=Z})\right] \\ \text{subject to} \quad & Q_{Y|X} \in \mathcal{F}. \end{aligned} \tag{1.5}$$

In words, we want to change the fitted model from data, $P_{Y|X}$, as least as possible so that the new model—the $Q^{\star}_{Y|X}$ that minimizes (1.5)—satisfies the new information (here, belonging to the feasible set $\mathcal{F}$). Thus, in an f-divergence sense, we are *projecting* the model $P_{Y|X}$ onto the set $\mathcal{F}$, hence the name *model projection*. A concrete example of this model projection setup is correcting ML models for group-fairness. Here, the method $\mathscr{M}$ to train the model $P_{Y|X}$ could be ML (e.g., applying logistic-regression on the collected data); the desired property $\mathscr{L}$ is a group-fairness constraint (e.g., the accuracy of $P_{Y|X}$ is independent of, say, the race of individuals). The model projection formulation (1.5) is naturally inscribed within the *information projection* literature [Che68, Csi75]. However, the practical prior results on information projection (e.g., explicit formulas and numerical methods) are inapplicable to the model projection setup when the input random variable X has an infinite support (as this amounts to a constrained information projection with infinitely many constraints). We address this challenge in Chapter 3.

Challenge 2. *Given the model projection formulation in* (1.5)*, what is the explicit formula for the projected model $Q^{\star}_{Y|X}$ (i.e., the minimizer in* (1.5)*)? And can we compute such a formula efficiently?*

The final chapter of this dissertation reveals the functional relation between information measures and the underlying moments, then uses these new formulas to introduce a new moments-based estimator of information measures that is robust to affine transformations of the samples. As the degree of dependence between random variables X and Y is conceptually equivalent to the dissimilarity between the joint probability measure $P_{X,Y}$ and the product measure $P_X \times P_Y$, the *mutual information* $I(X;Y) := D_{\mathrm{KL}}(P_{X,Y} \parallel P_X \times P_Y)$ is a popular measure of independence. In view of this definition, we have $I(X;Y) \geq 0$ (whenever the integration is well-defined), X and Y are independent if and only if $I(X;Y) = 0$, and the larger $I(X;Y)$ is the more dependence there is between X and Y. Furthermore, the mutual information is invariant to one-to-one transformations, because it is determined by the underlying probability measures rather than the values the induced random variables take. A closely related information measure is *differential entropy*, defined as $h(Z) := -D_{\mathrm{KL}}(P_Z \parallel \lambda)$, where λ denotes the Lebesgue measure. In very general situations, one has the equality $I(X;Y) = h(X) + h(Y) - h(X,Y)$. The appealing properties of mutual information and differential entropy have led to their adoption of in practice as metrics for quantifying associations between data [GNO+12, CLA+10, Fle04]. However, in a practice, one has access to only finitely many samples drawn from the underlying distributions, which makes the task of reliably estimating these information measures a difficult task. Several estimators of mutual information and differential entropy have recently been proposed within the information theory and computer science communities [KSG04, VV11, JVHW15, WY16, GKOV17]. Still, the state-of-the-art k-nearest-neighbors (k-NN) based estimators [GKOV17] do not capture some of the desirable properties of the estimated information measures, such as invariance to affine transformations. In addition, it is well-known that a probability measure is completely determined by its moments if, e.g., it has a finite moment-generating function around the origin. In such case, information measures that are defined in terms of distributions should, in principle, be expressible in terms of moments. Such a functional relation between information measures and moments are potentially helpful in the estimation task, as estimation of moments is a far easier task than that of probability measures (e.g., faster convergence). We address these challenges in Chapter 4.

Challenge 3. *Given probability measures whose moment-generating functions are finite, what is the exact functional form of the relationship between information measures and the moments of the underlying distributions? Also, how can we estimate information measures from data using sample moments?*

We address Challenges 1–3 in Chapters 2–4, respectively, of this dissertation. In the remainder of this introduction below, we give a brief overview of our contributions.

1.1 Optimal Differential Privacy in the Large-Composition Regime

In practical application, privacy mechanisms are composed (i.e., repeatedly applied) hundreds of times on private data, and it is known that DP guarantees degrade under composition [ACG+16]. In addition, the expression for the DP parameters in (1.3) becomes unwieldy after composition of mechanisms. This is not surprising in view of the important result in [MV16, Theorem 1.5] showing the #P-completeness of the general problem of computing optimal privacy parameters under composition. This leads to a central question in DP:

How can one optimize and quantify privacy under composition?

We tackle this problem in Chapter 2 [AAC+a, AAC+b, AAC+c, AAC+22].

In contrast to previous works on optimal DP mechanism design in the literature [GRS12, GS10, GV15, GKOV15, SCDF13, GV16, GDGK20, GDGK19], we consider the large-composition regime instead of the single-shot setting. The starting point of our approach is reducing the mechanism design problem into one about minimizing the KL-divergence. Specifically, we derive the following limit as the number of compositions k grows without bound: for each fixed $\delta \in (0, 1/2)$, under mild conditions on the mechanism $P_{Y|X}$, we have that

$$\frac{1}{k} \cdot \inf\{\varepsilon \geq 0 \ : \ \text{the } k\text{-fold composition of } P_{Y|X} \text{ is } (\varepsilon, \delta)\text{-DP}\} \to \sup_{d(x,x') \leq s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}) \quad (1.6)$$

As smaller values of ε correspond higher DP, it suffices by the above limit to minimize the maximal KL-divergence term on the right-hand side (subject to a utility constraint). The first part of Chapter 2 is devoted to solving this constrained minmax KL-divergence problem in the settings of scalar queries, vector queries, and in the small-sensitivity regime (i.e., $s \to 0^+$). Our result lead to new DP mechanisms, namely, the Cactus, isotropic, and Schrödinger mechanisms, that are optimal in the large-composition regime in some precise senses.

The other part of Chapter 2 is devoted to the problem of quantifying DP in the large-composition regime in an efficient and accurate way. Several privacy accountants have been introduced in the literature, e.g., [DRS22, KJH20, KH21, KJPH21, GLW21, GKKM22, DGK+22], which compute upper bounds on the privacy budget parameters (ε, δ) in DP (see (1.3)). However, there are still limitations to those accountants. First, the closed-form accountants, while attaining the theoretically optimal runtime, suffer from either overestimating or underestimating the privacy parameters. In addition, FFT-based approaches, while working well in practice, do not have constant runtimes, cannot generate the full privacy curve, and their current implementations cannot estimate the privacy parameters for very small δ (e.g., below 10^{-10}). To circumvent these issues, we introduce the saddle-point accountant (SPA), which is a lightweight and accurate DP accountant for the large-composition regime. Specifically, using Parseval's identity we derive a new formula for the privacy curve (i.e., the function $\varepsilon \mapsto \delta(\varepsilon)$ where $\delta(\varepsilon)$ is the least value of δ for which the mechanism is (ε, δ)-DP), expressing it as a contour integral over a line running parallel to the imaginary axis and with a free positive real-intercept. Then, the saddle-point method from mathematical physics yields a preferable choice of this real-intercept, referred to as the saddle-point.

Our main contributions in Chapter 2 include: (Here, we assume an ℓ^2 sensitivity, i.e., $d(x, x') = \|x - x'\|$ is the ℓ^2 norm in (1.3).)

1. We prove a tight composition theorem for the large-composition regime showing the asymptotic equivalence between the privacy parameter ε and the following maximal KL-divergence term $\sup_{\|x-x'\|\le s} D(P_{Y|X=x} \parallel P_{Y|X=x'})$. Therefore, we reduce the problem of optimizing a DP mechanism in the large-composition regime to the following minmax KL-divergence problem:
$$\begin{aligned} \inf_{P_{Y|X}} \quad & \sup_{\|x-x'\|\le s} \ D(P_{Y|X=x} \parallel P_{Y|X=x'}) \\ \text{subject to} \quad & \sup_{x\in\mathbb{R}^m} \ \mathbb{E}[c(Y-x) \mid X = x] \le C, \end{aligned} \tag{1.7}$$
where $c : \mathbb{R}^m \to \mathbb{R}_+$ is a preset cost function (e.g., a variance cost).

2. We prove that additive, continuous, spherically-symmetric mechanisms are optimal in the KL-divergence problem (1.7).

3. We derive a finite-dimensional convex program whose solutions are arbitrarily close to optimal for (1.7), thereby introducing the Cactus mechanism (for scalar queries) and the isotropic mechanisms (for vector queries and monotone mechanisms).

4. In the small-sensitivity regime, i.e., $s \to 0^+$, we show that the square of the ground-state eigenfunction of the Schrödinger operator yields optimal mechanisms. We call those the Schrödinger mechanisms.

5. We introduce the saddle-point (SPA), a closed-form DP accountant that has the theoretically optimal runtime (e.g., constant runtime for self-composition), estimates the privacy parameters accurately with provable error bounds, and works for arbitrarily small values of δ.

The results of Chapter 2 are based on the following papers:

- [AAC$^+$b]: **Wael Alghamdi**, Shahab Asoodeh, Flavio P. Calmon, Juan Felipe Gomez, Oliver Kosut, and Lalitha Sankar. Optimal Multidimensional Differentially Private Mechanisms in the Large-Composition Regime. Accepted in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023.

- [AAC$^+$a]: **Wael Alghamdi**, Shahab Asoodeh, Flavio P. Calmon, Juan Felipe Gomez, Oliver Kosut, and Lalitha Sankar. Schrödinger Mechanisms: Optimal Differential Privacy Mechanisms for Small Sensitivity. Accepted in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023.

- [AAC$^+$c]: **Wael Alghamdi**, Shahab Asoodeh, Flavio P. Calmon, Juan Felipe Gomez, Oliver Kosut, and Lalitha Sankar. The Saddle-Point Method in Differential Privacy. *Under review*.

- [AAC$^+$22]: **Wael Alghamdi**, Shahab Asoodeh, Flavio P. Calmon, Oliver Kosut, Lalitha Sankar, and Fei Wei. Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1838–1843, 2022.

1.2 Model Projection and Optimal Group-Fairness Intervention

In Chapter 3, we generalize the ubiquitous concept of information projection [Che68, Csi75, CM03, DZ96, YB17, Csi84, Top79, Csi84, Bar00, Slo02, BC80, AS16, KS15a, KS15b, Csi95a, Csi95b] to the case of conditional distributions. Specifically, we derive the explicit solution to the model projection problem (1.5) for any $\mathbb{R}^m$-valued input X (under mild regularity assumptions on the other given quantities). In addition, we derive an efficient and provably convergent numerical procedure to compute the solution to the model projection problem in the presence of only finitely many samples from the underlying distributions. Thus, the main question we are considering here is the following:

How can we solve model projection (1.5) explicitly and efficiently?

Chapter 3 provides theoretical and practical solutions to the above question [AHJ$^+$22, AAW$^+$20].

We also showcase the practicality of our theoretical results on model projection by applying them to the domain of group-fair machine-learning (ML). In this setting, the relevant quantities in the model projection formulation (1.5) are a given biased model $P_{Y|X}$ and a set of group-fair models $\mathcal{F}$. For example, $P_{Y|X}$ could be a model predicting the recidivism risk Y of an individual with prior criminal record X, but it exhibits very different accuracies in its prediction when assessed on different protected groups (e.g., determined by race). Here, $\mathcal{F}$ would be the set of models that have very similar accuracies across the different protected groups. Then, the solution to the model projection problem (1.5) is the unique group-fair model that is constructed via the least possible amount of changes (in the sense of a preset f-divergence) to the score assignments of the given biased model $P_{Y|X}$. This makes model projection a principled method for achieving *optimal* group-fairness in wide range of practical scenarios.

Several group-fairness intervention methods have been introduced in recent literature, e.g., [WRC20, WRC21, JN20, ABD+18, CHKV19, YCK20, CJG+19, ZVRG17, CDPF+17, MW18, KGZ19, LPB+21, BNBR19, PQC+19], and extensive comparisons between such group-fairness intervention methods can be found in [BDH+18, FSV+19, WRC21]. One can see from these studies that we do not have a universally optimal group-fairness intervention method. More importantly, almost all available implementations are tailored to binary classification, whereas there are many cases where the predicted variable is not binary. For example, in education, grading algorithms assign one out of several grades to students; in healthcare, predicted outcomes are frequently not binary (e.g., severity of disease). In addition, group-fairness intervention methods are often benchmarked on overused and small datasets, such as UCI Adult [Lic13] and COMPAS [ALMK16]. We address these issues by using our theoretical results on model projection to introduce `FairProjection`, a group-fairness enhancement method that works for any number of prediction classes or protected groups. Further, we benchmark `FairProjection` on a new dataset containing more than one million samples.

In summary, our contributions in Chapter 3 include:

1. We derive the explicit solution to the model projection problem (1.5).

2. We derive a parallelizable algorithm that computes the projected model in the presence of only finitely many samples, and we show convergence and sample-complexity guarantees.

3. We introduce `FairProjection`, the application of model projection to the group-fairness domain. `FairProjection` is applicable for multi-class prediction and for any number of protected groups.

4. We show in extensive comparisons on real-world datasets that `FairProjection` can achieve

competitive performance while running in a fraction of the time required for other state-of-the-art group-fairness intervention methods.

5. We benchmark `FairProjection` on a new dataset of size more than 1 million samples. This dataset is derived from open and anonymized data from Brazil's national high school exam, and benchmarking on it answer recent calls [BZZ+21, DHMS21] for moving away from overused datasets including UCI Adult [Lic13] and COMPAS [ALMK16].

Chapter 3 is based on work that appeared in the following papers:

- [AHJ+22]: **Wael Alghamdi**,* Hsiang Hsu,* Haewon Jeong,* Hao Wang, Peter Winston Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and COMPAS: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*, 2022. (**Selected as Oral Presentation**; * = equal contribution.)
- [AAW+20]: **Wael Alghamdi**, Shahab Asoodeh, Hao Wang, Flavio P. Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2711–2716, 2020.

1.3 Measuring Information from Moments

We consider the problem of expressing information measures in terms of moments. Suppose X is a random variable whose moment-generating function exists, i.e., $\mathbb{E}[e^{tX}] < \infty$ over some nontrivial interval $t \in (-\delta, \delta)$. Then, it is well-known that the moments of X determine its distribution uniquely, i.e., if Y is a random variable with $\mathbb{E}[Y^k] = \mathbb{E}[X^k]$ for all $k \in \mathbb{N}$ then we must have $P_Y = P_X$. Then, in principle, it should be possible to express any distribution functional with such input X using only its moments. However, no such functional relationship between information measures and moments existed before. In addition too being a theoretically intriguing problem, finding moments-based formulas for information measures are potentially helpful in estimating those information measures from data, since estimating moments is a task that we know how to do well. Thus, we tackle the following question:

What is the functional relationship between information measures and moments?

We derive in Chapter 4 the functional relationship between information measures and moments [AC22, AC21c, AC19].

We show, for example, that the differential entropy of a random vector X whose moment-generating function is finite is given by

$$h(X) = \lim_{n\to\infty} \int_0^\infty \rho_X^{(n)}(t)\, dt, \tag{1.8}$$

where each $t \mapsto \rho_X^{(n)}(t)$ is a rational function whose coefficients are multivariate polynomials in the moments of X (and we give explicit formulas for those coefficients). We develop our moments-based formulas by first studying polynomial approximations of the conditional expectation.

Viewing the conditional expectation $\mathbb{E}[X \mid Y]$ as the minimum mean-square error (MMSE) in estimating X using Y, we introduce the best-polynomial approximations. The n-th polynomial MMSE (PMMSE) is defined as the best degree-n polynomial in Y when estimating X. We study when the PMMSE converges to the MMSE, and quantify the convergence rate. We also develop the PMMSE formulas further for the special case of a Gaussian channel $Y = \sqrt{t}X + N$ for standard normal N and constant $t \geq 0$. This helps us derive the moments-based formulas for information measures with the aid of the I-MMSE relationship [GSV05], which expresses information measures in terms of the MMSE in Gaussian channels. Thus, our work is inscribed within literature on the I-MMSE relation, its extensions, and its applications [GSV05, Zak05, Guo09, Ver10, GWSV11, WV12, AVW14, HJW15, DBPS17, DV20, LTV06, TV06]

The moments-based formulas, such as (1.8), can help when estimating information measures from data. Indeed, truncating the limit in (1.8) at a fixed n, we define the n-th order approximation of differential entropy as

$$h_n(X) = \int_0^\infty \rho_X^{(n)}(t)\, dt. \tag{1.9}$$

This formula can be computed using only the first $2n$ moments of X. Further, $h_1(X) \geq h_2(X) \geq \cdots \geq h(X)$ and $h_n(X) \searrow h(X)$ as $n \to \infty$. Then, in the presence of only samples of X, we may replace the moments in (1.9) by *sample moments*, thereby introducing an approximation $\widehat{h}_n(X)$ of $h(X)$ from data. Writing the mutual information $I(X;Y)$ in terms of differential entropy, we also obtain moments-based formula, approximations I_n, and estimators $\widehat{I}_n$.

The moments-based estimators we introduce have desirable physical properties. They behave under affine transformations exactly like the information measures they estimate, e.g., $\widehat{h}_n$ and h both are shifted by $\log |a|$ is the underlying random variable is scaled by a nonzero constant a. This stems from the fact that the PMMSE we introduce behaves like the MMSE under affine transformations. Also, the mutual information estimator $\widehat{I}_n$ detects independence exactly: if X and Y are independent, then $\widehat{I}_n(X;Y) = 0$.

We summarize our contributions in Chapter 4 below:

1. We introduce the polynomial MMSE and show its convergence to the MMSE.

2. We derive the functional form of the relationship between information measures and moments.

3. We introduce new moments-based estimators of information measures from data.

4. We show experimentally that the proposed estimators can outperform the state-of-the-art estimators of information measures. We also prove that the proposed estimators are consistent and derive their sample complexity.

The presented work in Chapter 4 is based on the following papers:

- [AC22]: **Wael Alghamdi** and Flavio P. Calmon. Measuring information from moments. In *IEEE Transactions on Information Theory*, 2022, doi: 10.1109/TIT.2022.3202492.

- [AC21c]: **Wael Alghamdi** and Flavio P. Calmon. Polynomial approximations of conditional expectations in scalar gaussian channels. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 420–425, 2021.

- [AC19]: **Wael Alghamdi** and Flavio P. Calmon. Mutual information as a function of moments. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 3122—3126, 2019.

Chapter 2

Optimal Differential Privacy in the Large-Composition Regime

Differential Privacy (DP) [DMNS06] has become the de-facto standard for designing privacy-preserving machine learning algorithms (including in practice [EPK14, Dif17, KMR+20]). Intuitively, a randomized algorithm (or mechanism), viewed as a channel $P_{Y|X}$, is said to be differentially private if its output does not vary significantly with small perturbations of the input. Here, $P_{Y|X}$ models a randomized query response mechanism in which the input $X = x$ is the true response to a query on a dataset containing sensitive information and Y is its privatized version. There are several variants of DP that formalize this intuition. For instance, consider the original variant of DP, defined in [DMNS06, DKM+06] as follows.

Definition 2.1 ((ε, δ)-DP [DMNS06, DKM+06])**.** A mechanism $P_{Y|X}$ is said to be (ε, δ)-*differentially private* (or (ε, δ)-*DP* for short) for $\varepsilon \geq 0$ and $\delta \in [0, 1]$ if

$$\sup_{\|x-x'\|\leq s} \sup_{A\subset\mathcal{Y}} P_{Y|X=x}(A) - e^{\varepsilon} P_{Y|X=x'}(A) \leq \delta, \tag{2.1}$$

where $\mathcal{Y} := \mathrm{supp}(Y)$, A varies over measurable subsets of $\mathcal{Y}$, $\|\cdot\|$ is some norm, and s is the corresponding *sensitivity* of the query, i.e., the maximum change of the query's response over all pairs of "neighboring" datasets (e.g., differing in one entry). The case $\delta = 0$ is typically referred to as *pure* DP and denoted by ε-DP.

Any random mechanism will therefore introduce distortion on the query's output, reducing utility. Thus, it is natural to ask how to design mechanisms that achieve the optimal trade-off between privacy and utility. In addition, it is known that DP guarantees degrade if one composes (i.e., repeatedly applies) a privacy mechanism a large number of times [ACG$^+$16]. Hence, it is important to have efficient and accurate procedures to keep track of the afforded privacy budget. This chapter tackles both of these questions, where we propose DP mechanisms (dubbed the Cactus, isotropic, and Schrödinger mechanisms) that are optimal in the large-composition regime, and we also introduce the saddle-point method which is a lightweight and accurate DP accountant for the large-composition regime.

2.1 Optimizing Differentially-Private Mechanisms

A number of works [GRS12, GS10, GV15, GKOV15, SCDF13, GV16, GDGK20, GDGK19] have sought optimal DP mechanisms in a variety of settings. However, these works all focus on the *single-shot* setting, in which a single mechanism is applied to a single query. This chapter departs from previous work in that we focus on the large-composition regime instead of optimizing (2.1).

Most practical differentially-private mechanisms are applied several times on sensitive data. In this case, quantifying privacy guarantees turns out to be a challenging problem. We tackle the problem of designing DP mechanisms in the large-composition regime in this work by reducing the DP problem to a KL-divergence minmax problem. We start by proving a tight composition theorem for the large-composition regime showing the asymptotic equivalence between the privacy parameter ε and a maximal KL-divergence term. Then, we introduce new mechanisms (both for the scalar and vector query cases, and also for the small-sensitivity regime in the scalar case) that optimize such maximal KL-divergence terms. We also present numerical experiments showing that the new mechanisms outperform the Gaussian and Laplace mechanisms.

2.2 Accounting for Differential Privacy

Quantifying the privacy loss after a large number of compositions of DP mechanisms is a central challenge in privacy-preserving ML. A key result by [MV16, Theorem 1.5] states that computing exact privacy parameters under composition is in general #P-complete, hence infeasible. This challenge has spurred several follow-up works on *privacy accounting*, e.g., [DRS22, KJH20, KH21, KJPH21, GLW21, GKKM22, DGK+22], which compute upper bounds on the privacy budget parameters (ε, δ) in DP (see (2.1)).

The currently available accountants have several limitations. The accountants that have closed-form formulas—thereby attaining constant (in composition) runtimes—such as the moments accountant [ACG+16, Mir17] and the CLT-based Gaussian-DP accountant [BDLS20], suffer from either overestimating or underestimating, respectively, the privacy parameters. On the other hand, convolution-based accountants, such as FFT-based approaches [KJH20, GLW21], while working well in practice, do not have constant runtimes, cannot generate the full privacy curve, and are limited by machine precision due to their purely numerical nature.[1] For example, FFT-based approaches fail to estimate values of δ below 10^{-10} [GLW21, Appendix B] or 10^{-12} [DGK+22, Appendix C].

We overcome these challenges by introducing a new approach for estimating DP parameters using complex analysis. Our approach is based on the method of steepest descent for integral approximation—a well-known method in mathematical physics [JJ99]. We derive the *saddle-point accountant* (SPA), which:

1) has a computable closed-form formula, hence enjoys constant runtime complexity in the number of compositions for self-composition;

2) estimates the privacy parameters accurately and with provable error bounds; and

3) works for any value of δ, however small, thus describing the full range of (ε, δ) guarantees.

2.3 Chapter Organization

This chapter is organized as follows. The next section gives a brief overview of how the DP optimization problem reduces to a KL-divergence optimization problem in the large-composition regime. Then, we give a brief overview of our main contributions, review the relevant literature, and state our notation and assumptions.

We present our large-composition theorem in Section 2.9, reducing the DP problem to one about KL-divergence. Then, we show that designing optimal DP mechanisms for the large-composition regime can be significantly reduced to the case of additive, continuous, and spherically-symmetric

[1] Of course, this limitation can be alleviated by using custom implementations and arbitrary float-point precision libraries. Our point is that closed-form formulas do not have this limitation.

mechanisms in Section 2.10. After that, we give the mathematical construction of our three proposed families of mechanisms in Section 2.11 (Cactus, isotropic, and Schrödinger mechanisms), followed by proofs of their respective optimalities in the next three sections. Specifically, we treat the 1-dimensional Cactus mechanism in Section 2.12, where we show that this mechanism can get arbitrarily close to optimal; the multidimensional case is treated in Section 2.13, where we prove optimality of the proposed monotone isotropic mechanisms; and in Section 2.14, we go back to the 1-dimensional setting, where we show that eigenfunctions of the Schrödinger operator yield the optimal mechanisms in the small-sensitivity regime. We demonstrate via numerical experiments in Section 2.15 how our proposed mechanisms outperform the Gaussian or Laplace mechanisms.

In Sections 2.16–2.19, we turn to the DP accounting problem, where we introduce and analyze the saddle-point-accountant. The method of steepest descent is recalled in Section 2.16.3. We derive a new contour-integral formula and an asymptotic expansion for the privacy curve in Section 2.17. This asymptotic expansion gives rise to heuristics for approximating the privacy curve, which leads to the the SPA-MSD method in Section 2.17.3. Then, we derive a tight composition theorem and the decay rate of the saddle-point in Section 2.18. In Section 2.19, we introduce the SPA-CLT (the second version of the SPA) and apply the results from Section 2.18 to derive rigorous bounds on the privacy curve.

2.4 Concentration of DP: From DP to KL-Divergence

The definition of (approximate) DP can be cast in terms of properties of the *privacy loss random variable* (PLRV), defined as

$$L_{x,x'} := \log \frac{dP_{Y|X=x}}{dP_{Y|X=x'}}(Y), \tag{2.2}$$

where $Y \sim P_{Y|X=x}$ and $x, x' \in \mathbb{R}^m$. Namely, it can be shown that (2.1) is equivalently expressed as

$$\sup_{\|x-x'\|\leq s} \mathbb{E}\left[\left(1-e^{\varepsilon-L_{x,x'}}\right)^+\right] \leq \delta, \tag{2.3}$$

where $a^+ := \max\{0, a\}$. In the simplest case of composition in DP, where the same mechanism $P_{Y|X}$ is independently applied k times on data X generating output Y^k, i.e., $P_{Y^k|X} = \prod_{i=1}^k P_{Y_i|X}$, the PLRV is given by

$$L_{x,x'}^k := \sum_{i=1}^k \log \frac{dP_{Y_i|X=x}}{dP_{Y_i|X=x'}}(Y_i), \tag{2.4}$$

where $Y_i \sim P_{Y_i|X=x}$. Thus, analogous to (2.1)–(2.3), the composed mechanism is (ε,δ)-DP if

$$\sup_{\|x-x'\|\leq s} \mathbb{E}\left[\left(1-e^{\varepsilon-L^k_{x,x'}}\right)^+\right] \leq \delta. \tag{2.5}$$

From the law of large numbers, the distribution of $L^k_{x,x'}/k$ will concentrate around its mean, the KL-divergence, as

$$\frac{1}{k}\mathbb{E}\left[L^k_{x,x'}\right] = D\left(P_{Y|X=x} \parallel P_{Y|X=x'}\right). \tag{2.6}$$

Since the function $f(u) := (1-e^{\varepsilon-ku})^+$ is non-decreasing, in the limit of large compositions, privacy mechanisms with lower values of $D(P_{Y|X=x} \parallel P_{Y|X=x'})$ will enjoy stronger (ε,δ)-DP guarantees. Thus, regardless of the exact distribution of the privacy loss random variable, its mean (2.6) plays a central role in the privacy guarantees offered after many compositions. In applications such as privacy-ensuring machine learning, the number of compositions frequently exceeds $k = 10^3$.

Inspired by the above observation on concentration of the PLRV, our first main contribution is proving the following limit:

$$\frac{1}{k}\cdot\inf\{\varepsilon\geq 0 \;:\; k \text{ compositions of } P_{Y|X} \text{ are } (\varepsilon,\delta)\text{-DP}\} \to \sup_{\|x-x'\|\leq s} D(P_{Y|X=x} \parallel P_{Y|X=x'}) \tag{2.7}$$

as the number of compositions k grows without bound. We prove three versions of this limit in Theorem 2.1 under mild regularity assumptions on the mechanism $P_{Y|X}$, where we also quantify the rate of convergence as $-\Phi^{-1}(\delta)\sqrt{\mathrm{V}}/\sqrt{k}$, where Φ is the standard-normal cumulative distribution function and V is a constant related to the variance of the PLRV of $P_{Y|X}$.

Equipped with the limit (2.7), we dedicate the bulk of this chapter to designing privacy mechanisms with favorable (ε,δ)-DP guarantees under a large number of compositions via minimizing the maximal KL-divergence term on the right-hand side of (2.7). Since after many compositions, privacy will be mostly determined by the mean of the privacy loss random variable (2.6), we solve the optimization problem

$$\begin{aligned} \inf_{P_{Y|X}\in\mathscr{R}} \quad & \sup_{\|x-x'\|\leq s} D(P_{Y|X=x} \parallel P_{Y|X=x'}) \\ \text{subject to} \quad & \sup_{x\in\mathbb{R}^m} \mathbb{E}[c(Y-x) \mid X=x] \leq C, \end{aligned} \tag{2.8}$$

where $c:\mathbb{R}^m \to [0,\infty)$ is a pre-specified cost function, $s, C > 0$ are constants, and $\mathscr{R}$ is the set of all Markov kernels on $\mathbb{R}^m$. Note that the cost function is critical: without the constraint, (2.8) can be trivially solved by any mechanism that is independent of X. In this chapter, we introduce new

DP mechanisms that can solve (2.8) to arbitrary accuracy under various setups, namely, (i) *Cactus mechanisms* for the scalar case ($m = 1$), (ii) *isotropic mechanisms* for the vector ($m > 1$) and monotone case, and (iii) *Schrödinger mechanisms* for the scalar and small-sensitivity regime ($s \to 0^+$).

2.5 Main Contributions

The main contributions of the work underlying this chapter are as follows:

1. We prove the limit (2.7) in Theorem 2.1, thereby showing the asymptotic equivalence between the ε parameter of DP and the associated maximal KL-divergence term.

2. We show (Theorem 2.3) that additive, continuous, spherically-symmetric mechanisms—i.e., where $Y = X + Z$ for a noise vector Z independent of X and with probability density function p_Z that is constant on every sphere centered around the origin—suffice to solve (2.8).

3. Even restricting to additive mechanisms, (2.8) is an infinite-dimensional optimization problem, so it cannot be solved directly. Instead, we formulate an approximate problem that is finite dimensional and can be solved efficiently. We prove (Theorem 2.8) that this approximate problem can get arbitrarily close to optimal for solving (2.8) in the scalar case ($m = 1$).

4. We solve the approximate scalar problem to derive (near) optimal mechanisms for the quadratic cost function, i.e., $c(x) = x^2$. We dub the resulting mechanism the "Cactus mechanism" due to the shape of the distribution (see Figure 2.1). Surprisingly, the Gaussian distribution is strictly sub-optimal for (2.8), as the Cactus mechanism achieves a smaller KL-divergence for the same variance.

5. Similarly, for the vector-valued case ($m > 1$), we formulate an approximate problem that is finite dimensional and can be solved efficiently. We restrict attention here to monotone mechanisms, as these are the ones for which current DP accounting can be extended. We prove (Theorem 2.10) that this approximate problem can get arbitrarily close to optimal for solving (2.8) in the vector case ($m > 1$) when restricted to monotone mechanisms.

6. We fully characterize the small-sensitivity regime, i.e., where $s \ll 1$, in which case we derive closed-form optimal distributions that we call the Schrödinger mechanism. In this case, the minimax optimization of KL-divergence in (2.8) reduces to finding a *unique* minimizer of the

Fisher information $I(P)$ over all probability measures P satisfying the utility constraint (see Section 2.14). This reduced formulation reveals a remarkable characterization of the optimizer $P^\star$: its square root is the eigenfunction of the Schrödinger operator corresponding to the smallest eigenvalue (Theorem 2.14).

7. In the small-sensitivity and scalar case, we identify closed-form DP mechanisms with the optimal privacy-utility trade-off where the utility is measured via the cost function c. In particular, we show that $P^\star$ is the Gaussian measure for the L^2 cost function (Proposition 2.4), thereby proving that the Gaussian mechanism is optimal in this sense in the small-sensitivity regime. Our results also show that $P^\star$ for the L^1 cost is given by the Airy function, leading to the introduction of a new optimal DP mechanism, which we call the *Airy mechanism* (see Definition 2.7).

8. We provide numerical benchmarks that demonstrate that the proposed isotropic and Schrödinger mechanisms achieve a favorable privacy-utility trade-off under a large number of compositions when compared to the Gaussian or Laplace mechanisms (Section 2.15).

9. We derive a new formula for the DP curve, expressing it as a contour integral with integrand expressible in terms of the cumulant-generating function of the PLRV and with a contour running parallel to the imaginary axis and of free positive real-axis intercept (Theorem 2.15).

10. We apply the method of steepest descent from mathematical physics to choose the real-axis intercept in our new DP formula as the saddle-point, thereby giving rise to the saddle-point accountant (SPA), a lightweight and accurate closed-form DP accountant that works for even vanishingly small values of δ.

2.6 Related Work

Identifying optimal mechanisms is a fundamental and challenging problem in the domain of differential privacy. There have been several works in the literature that have attempted to address this problem. For instance, within the class of additive noise mechanisms and under the single shot setting (i.e., no composition), Ghosh et al. [GRS12] showed that the geometric mechanism is universally optimal for $(\varepsilon, 0)$-DP in a Bayesian framework, and Gupte and Sundararajan [GS10] derived the optimal noise distribution in a minimax cost framework. For a rather general cost function, the optimal noise distribution was shown to have a staircase-shaped density function [GV15, GKOV15, SCDF13].

Geng and Viswanath [GV16] showed that for (ε, δ)-DP and integer-valued query functions, in the single-shot setting, the discrete uniform noise distribution and the discrete Laplacian noise distribution are asymptotically optimal (for L^1 and L^2 costs) within a constant multiplicative gap in the high privacy regime (i.e., both ε and δ approach zero). Geng et al. [GDGK20] studied the same setting except for real-valued query functions and identified truncated Laplace distribution is asymptotically optimal in various high privacy regimes. Finally, Geng et al. [GDGK19] showed that the optimal noise distribution for real-valued query and $(0, \delta)$-DP is uniform with probability mass at the origin. Our work differs from these works in that we focus on the optimal mechanisms under a large number of compositions, rather than the single shot setting.

When considering a composition of n mechanisms, an important line of research has been to derive tighter composition results: relationships between the DP parameters of the composed mechanism and the parameters of each constituent mechanism. There are several composition results in the literature, such as [DRV10, MV16, KOV15, ACG$^+$16, ALC$^+$21, MM18]. More recently, Dong et al. [DRS22] have proposed a composition result for large n and for a new variant of DP, called Gaussian-DP, that leverages the central limit theorem. These results can be sub-optimal (see, for example, [GLW21, Fig. 1]). Consequently, numerical composition results have gained increasing traction as they lead to easier, yet powerful, methods for accounting the privacy loss in composition [KJH20, GLW21, KJPH21, ZDW21]. In particular, Koskela et al. [KJH20] obtained a numerical composition result based on a numerical approximation of an integral that gives the DP parameters of the composed mechanism. The approximation is carried out by discretizing the integral and by evaluating discrete convolutions via the fast Fourier transform algorithm. The running time and memory needed for this approximation were subsequently improved [GLW21]. While our work shares the focus on the large composition regime, we are primarily interested in synthesizing optimal mechanisms rather than analyzing existing mechanisms.

The connection we put forth starting from DP and leading to the Schrödinger equation is new. The component connections, however, have been noted in some form in the literature. Nevertheless, our work serves to make the existing results into one coherent unit, fills some existing gaps rigorously, and extends existing setups. Specifically, for the problem of minimizing the Fisher information, we:

1. work with a larger class of cost functions,

2. do not restrict the support of the PDFs we optimize over,

3. do not require any regularity assumptions whatsoever on the PDFs we optimize over.

We circumvent imposing any assumptions on the PDFs, and are able to extend the class of cost constraints, by introducing a novel proof technique for minimizing Fisher information that does not depend on the theory of calculus of variation, and also by deriving an estimate of the logarithmic derivative of the ground-state eigenfunction of the Schrödinger operator (which is of independent interest).

The statistics literature is rife with results on Fisher-information-minimizing distributions. The Cramér-Rao bound implies that Gaussian measures have the smallest Fisher information among all densities with a given variance. The minimizer over compactly-supported distributions or over those supported on $\mathbb{R}^+$ were characterized in [UK95] and [BV09], respectively. Kagan [Kag86] studied the same problem for densities on $\mathbb{R}$ with fixed first and second moments, which was later extended to other moments by Ernst [Ern17]. A connection between minimizing Fisher information and the Schrödinger equation has been established in [HR09, Example 5.1]. Formulating a privacy problem in terms of minimizing Fisher information has appeared in [FS18, FS19], but not in a DP sense; rather, the analyses therein pertain to privacy-preserving battery charging methods to obfuscate household information, and the Fisher information itself is proposed as a privacy metric. Fisher information minimization in [FS18] is done for PDFs of compact support, and that is extended to unbounded support in [FS19] but for only a quadratic cost. Further, the PDFs considered in [FS18, FS19] are assumed *a priori* to be twice continuously differentiable. Therefore, none of these previous works has a setup encompassing ours, namely, they minimize Fisher information: over PDFs supported over a compact set [UK95, FS18] or over $\mathbb{R}^+$ [BV09]; assuming regularity of the PDFs [HR09, FS18, FS19]; or under a strictly smaller or different class of constraint functions [Kag86, Ern17, FS19].

We discuss in more detail how our work differs from the existing literature closest to ours [HR09, Ern17, FS18, FS19] regarding the Fisher information minimization problem in Appendix A.1.

2.7 Notation

We fix a Euclidean space $\mathbb{R}^m$ throughout, and an m-dimensional random vector X, whose induced Borel probability measure is denoted by P_X. Denote by λ and $\|\cdot\|$ the Lebesgue measure and ℓ^2

norm, respectively, on $\mathbb{R}^m$. The open ball around $x \in \mathbb{R}^m$ of radius r is denoted by $B_r(x)$. The shift operator is denoted by T_x, i.e., $(T_x r)(A) := r(A - x)$.

For a probability measure P on $\mathbb{R}^m$ and $c : \mathbb{R}^m \to \mathbb{R}$, the expectation is denoted by $\mathbb{E}_P[c] := \int_{\mathbb{R}^m} c(x)\, dP(x)$. For probability measures P, Q over $\mathbb{R}^m$, the KL-divergence is denoted by $D(P \,\|\, Q)$, the variance of the information density is denoted by

$$\mathrm{V}(P \,\|\, Q) := \mathbb{E}_P\left[\left(\log \frac{dP}{dQ} - D(P \,\|\, Q)\right)^2\right], \tag{2.9}$$

and the E_γ-divergence is defined for $\gamma \geq 0$ as

$$\mathsf{E}_\gamma(P \,\|\, Q) := \sup_{A \text{ Borel}} P(A) - \gamma Q(A) = \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - e^{\varepsilon}\right)^+\right],$$

where $a^+ := \max(0, a)$. We write $F(p \,\|\, q)$ or $F(X \,\|\, Y)$ if $P, Q \ll \lambda$ with densities p and q or $X \sim P$ and $Y \sim Q$, where $F \in \{D, \mathrm{V}, \mathsf{E}_\gamma\}$. We also denote the expectation by $\mathbb{E}_p[g] := \mathbb{E}_P[g]$ if $P \ll \lambda$ has probability density function (PDF) p. A probability measure P over $\mathbb{R}^m$ is said to be spherically-symmetric if $P(\{Ux : x \in B\}) = P(B)$ for any Borel $B \subset \mathbb{R}^m$ and every orthogonal matrix $U \in \mathbb{R}^{m \times m}$.

We denote by $\mathscr{R}$ the set of all Markov kernels[2] on $\mathbb{R}^m$, i.e., conditional distributions $P_{Y|X}$ for $\mathbb{R}^m$-valued X and Y such that $x \mapsto P_{Y|X=x}(B)$ is a Borel function for all Borel sets $B \subset \mathbb{R}^m$. The set $\mathscr{B}$ denotes all Borel probability measures on $\mathbb{R}^m$. The set of all probability density functions on $\mathbb{R}$ is denoted by $\mathcal{P}$. The Fisher information of $p \in \mathcal{P}$ is denoted by $I(p)$, i.e., if p is absolutely continuous then

$$I(p) := \int_{\{x \in \mathbb{R}\,;\, p(x) > 0\}} \frac{p'(x)^2}{p(x)}\, dx, \tag{2.10}$$

and $I(p) = \infty$ otherwise.

For a random variable L, the moment-generating function (MGF) is denoted by $M_L(t) := \mathbb{E}[e^{tL}]$, and the cumulant-generating function (CGF) by $K_L(t) := \log M_L(t)$. The standard normal cumulative density function is denoted by Φ. The Q function is defined by $Q(x) := 1 - \Phi(x)$. We also denote the function $q : \mathbb{R} \to (0, \infty)$ by

$$q(z) := Q(z) \cdot \sqrt{2\pi}\, e^{z^2/2}. \tag{2.11}$$

[2] It is true that any conditional distribution from $\mathbb{R}^m$ into $\mathbb{R}^m$ has a version that is a Markov kernel [Ç11, Chapter 4, Theorem 2.10].

The (m, k)-th partial Bell polynomial is denoted by (with $\boldsymbol{x} = (x_1, \cdots, x_m)$)

$$B_{m,k}(\boldsymbol{x}) := \sum_{\substack{k_1+\cdots+k_m=k \\ 1\cdot k_1+\cdots+m\cdot k_m=m}} \binom{m}{k_1, \cdots, k_m} \prod_{j=1}^{m} \left(\frac{x_j}{j!}\right)^{k_j} \tag{2.12}$$

and the m-th complete Bell polynomial by $B_m(\boldsymbol{x}) := \sum_{k=1}^{m} B_{m,k}(\boldsymbol{x})$. We will use the standard Bachmann-Landau notations $O, \Omega, \Theta, o, \omega$. The notation $a_k \sim b_k$ means $a_k/b_k \to 1$ as $k \to \infty$.

2.8 Assumptions

Any assumption required for a particular result to hold will be explicitly invoked in the statement of the same result. We collect the various assumptions needed for this chapter in this section for reference.

2.8.1 Assumption for Optimal Mechanism Design

Sensitivity

Throughout this chapter, we consider *only* the ℓ^2 sensitivity.

Cost Function

For the results of this chapter we will require the cost function c to satisfy some assumptions, and we will always explicitly invoke such assumptions clearly in the relevant context. For Sections 2.10 and 2.12–2.13, we will require the following assumption on the cost function c.

Assumption 2.1. *The cost function $c : \mathbb{R}^m \to \mathbb{R}$ satisfies:*

(a) Growing from 0 to ∞: *$c(0) = 0$; $c(u) \leq c(v)$ if $\|u\| \leq \|v\|$; and $c(x) \to \infty$ as $\|x\| \to \infty$.*

(b) Spherical symmetry: *there is a function $\widetilde{c} : \mathbb{R}_+ \to \mathbb{R}$ such that $c(x) = \widetilde{c}(\|x\|)$ for all $x \in \mathbb{R}^m$.*

(c) Lower-semicontinuity: *c is continuous at the origin, and it is lower semicontinuous over $\mathbb{R}^m$.*

A natural choice of cost function satisfying Assumption 2.1 is positive multiples and powers of the the quadratic cost $c(x) = \beta\|x\|^\alpha$ for $\alpha, \beta > 0$, but we allow $c(x)$ to be any function that satisfies the above assumptions. For the optimality results in Sections 2.12–2.13, we also require $c(x) \sim \beta\|x\|^\alpha$ as $\|x\| \to \infty$, which will be explicitly invoked when needed.

The results of Section 2.14 hold for the class of cost functions c satisfying the following properties. We note that this class includes functions such as $c(x) = \beta|x|^{\alpha}$ and $c(x) = \beta \log(|x|+1)^{\alpha}$ for any $\alpha, \beta > 0$.

Assumption 2.2. *The cost function $c : \mathbb{R} \to \mathbb{R}$ satisfies:*

(a) Growing from 0 to ∞: *$c(0) = 0$; $c(u) \leq c(v)$ if $|u| \leq |v|$; and $c(x) \to \infty$ as $|x| \to \infty$.*

(b) Evenness: *$c(x) = c(-x)$ for all $x \in \mathbb{R}$.*

(c) Continuity: *c is continuous over $\mathbb{R}$.*

(d) Controlled derivative: *$c'(x) = o\left(c(x)^{3/2}\right)$ as $x \to \infty$.*

(e) Tail regularity: *$\int_{x_0}^{\infty} |c'|^2/|c|^{5/2}, \int_{x_0}^{\infty} |c''|/|c|^{3/2} < \infty$ for some $x_0 \in \mathbb{R}$.*

(f) Moderate growth: *$x \mapsto \sqrt{c(x)}/\exp(\gamma \int_0^{|x|} \sqrt{c(t)}\, dt)$ is integrable for all $\gamma > 0$.*

(g) Additive/Multiplicative regularity: *there is a locally bounded strictly positive function ρ on $\mathbb{R}$ such that $c(x-t), c(tx) \leq \rho(t)(c(x)+1)$ for all $x, t \in \mathbb{R}$.*

Remark 2.1. In the assumptions involving c' or c'', it is to be understood that c is required to be differentiable (or twice differentiable) *only* at large enough values.

Mechanism

We prove optimality of additive, continuous, spherically-symmetric mechanisms in Section 2.10 (Theorems 2.3–2.6) *without* any assumption at all on the considered mechanisms $P_{Y|X}$. Similarly, for the Fisher information minimization result in Theorem 2.13, we do not impose any restriction at all on the considered PDFs p.

For proving the limit (2.7) in Theorem 2.1, we use the CLT, so we will impose a subset of the following properties on the variance of the PLRV.

Assumption 2.3. *The Markov kernel $P_{Y|X}$ satisfies:*

(a) *Bounded variance:* $\sup_{\|x-x'\|\leq s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}), \sup_{\|x-x'\|\leq s} \mathrm{V}(P_{Y|X=x} \,\|\, P_{Y|X=x'}) < \infty$.

(b) *If $x_\ell, x_\ell' \in \mathbb{R}^m$ are such that $\mathrm{V}(P_{Y|X=x_\ell} \,\|\, P_{Y|X=x_\ell'}) \to 0$ as $\ell \to \infty$, then $D(P_{Y|X=x_\ell} \,\|\, P_{Y|X=x_\ell'}) \to 0$.*

Finally, to extract optimal DP mechanisms from our results on the global minimization of Fisher information in Section 2.14, we will require that the considered PDFs satisfy the expansion of KL-divergence in terms of Fisher information and that they have uniformly bounded information-density variance. We define this class, denoted $\mathcal{F}$, in Definition 2.8.

Assumption 2.4. *The PDF p on $\mathbb{R}$ satisfies:*

(a) Fisher information expansion: $D(p \,\|\, T_a p) = a^2 I(p)/2 + o(a^2)$ as $a \to 0$.

(b) Bounded variance: there is an $s > 0$ such that $\sup_{|a| \le s} D(p \,\|\, T_a p)$, $\sup_{|a| \le s} \mathrm{V}(p \,\|\, T_a p) < \infty$.

2.8.2 Assumptions for the Saddle-Point Accountant

We will require the privacy-loss random variable (PLRV)—see Definition 2.11—to have a finite MGF.

Assumption 2.5. *The MGF $M_L(t)$ of the PLRV L is finite for every $t > 0$.*

Under Assumption 2.5, both the MGF and CGF can be extended to be holomorphic functions over the half-plane $z \in (0, \infty) + i\mathbb{R} \subset \mathbb{C}$.

We impose the following technical assumption on the distribution of the PLRV so that Parseval's identity applies.

Assumption 2.6. *The induced probability measure P_L by the PLRV L decomposes as a sum $P_L = Q_L + R_L$ for Q_L absolutely continuous with respect to the Lebesgue measure and discrete R_L. Further, with q_L denoting the PDF of Q_L, we assume that $x \mapsto e^{tx} q_L(x)^2$ is integrable for each $t > 0$.*

For our error analysis, we will assume the following on the growth of the first three moments of a PLRV.

Assumption 2.7. *With $\widetilde{L} = \widetilde{L}_1 + \cdots + \widetilde{L}_n$ being the exponential tilting with parameter $t > 0$ (see Definition 2.12), and denoting*

$$\mathrm{P}_t := \sum_{j=1}^{n} \mathbb{E}\left[\left| \widetilde{L}_j - \mathbb{E}[\widetilde{L}_j] \right|^3 \right], \tag{2.13}$$

we assume that there are constants $\mathrm{KL}, \mathrm{V} > 0$, and P such that $t = o(n^{-1/3})$ yields the limit (as $n \to \infty$)

$$\frac{1}{n} \cdot (\mathbb{E}[\widetilde{L}], \sigma_{\widetilde{L}}^2, \mathrm{P}_t) \to (\mathrm{KL}, \mathrm{V}, \mathrm{P}). \tag{2.14}$$

Remark 2.2. Assumption 2.7 is automatically satisfied under Assumption 2.5 for *self-composition*.

It is worth noting that all of the above three assumptions are satisfied by both the subsampled Gaussian mechanism (because it is continuous with a PDF that decays super-exponentially) and the subsampled Laplace mechanism (because its continuous part is bounded). See Appendix A.15 for more details.

2.9 Large-Composition Theorem

We start by delineating how designing optimal DP mechanisms in the high-composition regime leads to the information-theoretic KL-divergence optimization problem we consider in (2.8). We emphasize that our analysis in this section is focused in the setting of sufficiently large number of compositions. Thus, our focus is solving the resulting information-theoretic problem presented in (2.8). In a nutshell, we show in Theorem 2.1 that: for a wide class of mechanisms $P_{Y|X}$, with $P_{Y|X}^{\circ k}$ denoting the k-fold self-composition, and $\varepsilon \mapsto \delta_{P_{Y|X}^{\circ k}}(\varepsilon)$ the worst-case privacy curve after k compositions, we have the limit

$$\lim_{k\to\infty} \frac{1}{k} \cdot \inf\left\{ \varepsilon \geq 0 \ : \ \delta_{P_{Y|X}^{\circ k}}(\varepsilon) \leq \delta \right\} = \sup_{\|x-x'\|\leq s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}). \tag{2.15}$$

Thus, designing optimal DP mechanisms in the large-composition regime is equivalent, to a "first-order" approximation, to minimizing the maximal KL-divergence (i.e., solving (2.8)).

2.9.1 The Privacy Curve After Composition

We first set up some useful notation. We let $\mathcal{D}$ be some set containing the datasets. Let $P_{Y|X}$ be a mechanism, i.e., a Markov kernel on $\mathbb{R}^m$, and $f : \mathcal{D} \to \mathbb{R}^m$ a query function. Let $d \in \mathcal{D}$ and write $x = f(d)$. Then, the result of applying $P_{Y|X}$ for the query function f to the dataset d is denoted by $Y \sim P_{Y|X=x}$. Note that Y is a random variable.

We consider next the k-fold adaptive composition of $\mathcal{M}$ with k different queries. Fix k query functions $f_1 : \mathcal{D} \to \mathbb{R}^m$ and $f_j : \mathcal{D} \times (\mathbb{R}^m)^{j-1} \to \mathbb{R}^m$ for $2 \leq j \leq k$. Let $d \in \mathcal{D}$, $x_1 := f_1(d)$, and $Y_1 \sim P_{Y|X=x_1}$ the output of the first application of the mechanism $P_{Y|X}$. Define $x_j := f_j(d, Y_1, \cdots, Y_{j-1})$ for $2 \leq j \leq k$. Then, the k-fold composition of $P_{Y|X}$ is defined by $P_{Y|X}^{\circ k}[x_1, \cdots, x_k] := (P_{Y|X=x_1}, \cdots, P_{Y|X=x_k})$. Let $\mathcal{Z}_1 := \mathcal{D}$, and, for each $j \geq 2$, let $\mathcal{Z}_j \subset \mathcal{D} \times (\mathbb{R}^m)^{j-1}$ denote the set of all possible sequences $z_j = (d, y_1, \cdots, y_{j-1})$ as d ranges over $\mathcal{D}$. We use the shorthands $\tau^{(\ell)} := (\tau_1, \cdots, \tau_\ell)$. We assume that the f_j have the same sensitivity, defined as

$$s := \sup_{\substack{z_j=(d,u^{(j-1)}),z_j'=(d',v^{(j-1)})\in\mathcal{Z}_j \\ d,d'\in\mathcal{D},d\simeq d'}} |f_j(z_j) - f_j(z_j')|, \tag{2.16}$$

where $d \simeq d'$ is used to indicate that the datasets d and d' are neighboring. Let $\mathcal{S}_d^{(k)} \subset (\mathbb{R}^m)^k$ denote the set of all possible outcomes $x^{(k)}$ generated from $d \in \mathcal{D}$. Then, according to the DP definition in (2.1), the mechanism $P_{Y|X}^{\circ k}$ is (ε, δ)-DP if and only if

$$\sup_{d\simeq d'} \sup_{(u^{(k)},v^{(k)})\in\mathcal{S}_d^{(k)}\times\mathcal{S}_{d'}^{(k)}} \sup_{A \text{ Borel}} P_{Y^{(k)}|X^{(k)}=u^{(k)}}(A) - e^{\varepsilon} P_{Y^{(k)}|X^{(k)}=v^{(k)}}(A) \le \delta. \tag{2.17}$$

We will rewrite the DP definition (2.1) in terms of the E_γ-divergence.

Definition 2.2 (Hockey-stick divergence)**.** The *hockey-stick divergence* with parameter $\gamma \ge 0$ (or, the E_γ-divergence) of P from Q, with (P, Q) being a pair of Borel probability measures on $\mathbb{R}^m$, is defined as

$$\mathsf{E}_\gamma(P \parallel Q) := (P - \gamma Q)^+(\mathbb{R}^m) = \sup_{A \text{ Borel}} P(A) - \gamma Q(A). \tag{2.18}$$

It is immediate that the DP definition (2.1) may be rewritten in terms of the E_γ-divergence [BO13]. Specifically, a mechanism $P_{Y|X}$ is (ε, δ)-DP if and only if

$$\sup_{\|x-x'\|\le s} \mathsf{E}_{e^\varepsilon}(P_{Y|X=x} \parallel P_{Y|X=x'}) \le \delta. \tag{2.19}$$

Rewriting DP in terms of the hockey-stick divergence facilitates mathematical reasoning about composition of mechanisms, as we briefly review next. Note that inequality (2.17) may be rewritten as

$$\sup_{d\simeq d'} \sup_{(u^{(k)},v^{(k)})\in\mathcal{S}_d^{(k)}\times\mathcal{S}_{d'}^{(k)}} \mathsf{E}_{e^\varepsilon}\left(P_{Y^{(k)}|X^{(k)}=u^{(k)}} \parallel P_{Y^{(k)}|X^{(k)}=v^{(k)}}\right) \le \delta. \tag{2.20}$$

The left-hand side may be upper bounded, by definition of sensitivity, as

$$\begin{aligned}
&\sup_{d\simeq d'} \sup_{(u^{(k)},v^{(k)})\in\mathcal{S}_d^{(k)}\times\mathcal{S}_{d'}^{(k)}} \mathsf{E}_{e^\varepsilon}\left(P_{Y^{(k)}|X^{(k)}=u^{(k)}} \parallel P_{Y^{(k)}|X^{(k)}=v^{(k)}}\right) \\
&\qquad \le \sup_{\substack{u^{(k)},v^{(k)}\in(\mathbb{R}^m)^k \\ \|u_j-v_j\|\le s,\, j\in[k]}} \mathsf{E}_{e^\varepsilon}\left(P_{Y^{(k)}|X^{(k)}=u^{(k)}} \parallel P_{Y^{(k)}|X^{(k)}=v^{(k)}}\right) && (2.21) \\
&\qquad = \sup_{\substack{u^{(k)},v^{(k)}\in(\mathbb{R}^m)^k \\ \|u_j-v_j\|\le s,\, j\in[k]}} \mathsf{E}_{e^\varepsilon}\left(\prod_{j\in[k]} P_{Y|X=u_j} \middle\| \prod_{j\in[k]} P_{Y|X=v_j}\right) && (2.22)
\end{aligned}$$

Further, this upper bound is at least as tight as the bound considered by all existing PLRV-based DP accountants. Indeed, let (P, Q) be a tightly dominating pair, i.e.,

$$\sup_{\|x-x'\|\le s} \mathsf{E}_\gamma(P_{Y|X=x} \parallel P_{Y|X=x'}) = \mathsf{E}_\gamma(P \parallel Q) \tag{2.23}$$

for every $\gamma \ge 0$. Then,

$$\mathsf{E}_\gamma(P_{Y|X=x} \parallel P_{Y|X=x'}) \le \mathsf{E}_\gamma(P \parallel Q) \tag{2.24}$$

whenever $\|x - x'\| \le s$ and $\gamma \ge 0$. Then, we have that, for every $\gamma \ge 0$, (see, e.g., [DRS22, Theorem 4])

$$\mathsf{E}_\gamma\left(\prod_{j\in[k]} P_{Y|X=u_j} \,\middle\|\, \prod_{j\in[k]} P_{Y|X=v_j}\right) \le \mathsf{E}_\gamma(P^{\otimes k} \parallel Q^{\otimes k}) \tag{2.25}$$

whenever $\|u_j - v_j\| \le s$ for every $j \in [k]$. In other words,

$$\sup_{\|u_j-v_j\|\le s,\, j\in[k]} \mathsf{E}_\gamma\left(\prod_{j\in[k]} P_{Y|X=u_j} \,\middle\|\, \prod_{j\in[k]} P_{Y|X=v_j}\right) \le \mathsf{E}_\gamma(P^{\otimes k} \parallel Q^{\otimes k}). \tag{2.26}$$

Further, the upper bound in (2.22) is tight in general; indeed, equality is attained if the f_j depend only on d. Thus, for the mechanism design part of this chapter, we focus on the privacy curve

$$\delta_{P_{Y|X}^{\circ k}}(\varepsilon) := \sup_{\|u_j-v_j\|\le s,\, j\in[k]} \mathsf{E}_{e^\varepsilon}\left(\prod_{j\in[k]} P_{Y|X=u_j} \,\middle\|\, \prod_{j\in[k]} P_{Y|X=v_j}\right). \tag{2.27}$$

We show in Theorem 2.1 below that, under mild conditions on $P_{Y|X}$, the behavior of $\delta_{P_{Y|X}^{\circ k}}(\varepsilon)$ for large k is governed by the KL-divergence.

The DP mechanism-design problem aims at finding a mechanism $P_{Y|X}$ for which, given $\varepsilon \ge 0$, the number $\delta_{P_{Y|X}^{\circ k}}(\varepsilon)$ is as small as possible. Naturally, one would also impose some cost constraint on $P_{Y|X}$ so that the optimization problem is nontrivial. Dually, we may fix $\delta \in [0, 1]$ and look for a mechanism $P_{Y|X}$ that minimizes $\varepsilon \ge 0$, i.e., one that minimizes the left-inverse function

$$\varepsilon_{P_{Y|X}^{\circ k}}(\delta) := \inf\left\{\varepsilon \ge 0 \;:\; \delta_{P_{Y|X}^{\circ k}}(\varepsilon) \le \delta\right\}. \tag{2.28}$$

In other words, we are considering the following problem.

Problem 1. *For $\varepsilon \ge 0$ and $k \in \mathbb{N}$, minimize $\delta_{P_{Y|X}^{\circ k}}(\varepsilon)$ over $P_{Y|X}$. Dually, for $\delta \in [0, 1]$ and $k \in \mathbb{N}$, minimize $\varepsilon_{P_{Y|X}^{\circ k}}(\delta)$ over $P_{Y|X}$.*

2.9.2 From DP to KL-divergence: A Large-Composition Theorem

We derive in Theorem 2.1 a sense in which Problem 1 about DP mechanism-design under composition reduces to the following KL-divergence optimization.

Problem 2. *Minimize* $\sup_{\|x-x'\|\le s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}) + \Theta(1/\sqrt{k})$ *over* $P_{Y|X}$.

Optimizing the maximal KL-divergence (2.8) is, therefore, a "first-order" approximation of Problem 2. We note that the implicit constant in the $\Theta(1/\sqrt{k})$ term in Problem 2 is within a multiplicative factor of $1+o(1)$ from belonging to the interval $\left[-\Phi^{-1}(\delta)\underline{\sigma}, -\Phi^{-1}(\delta)\sigma_{\max}\right]$ where: $\sigma_{\max}^2$ is the maximal variance of the random variables $L_{x,x'}$ defined in (2.2); $\underline{\sigma}$ is the infimal variance of $L_{x,x'}$ whose induced KL-divergence gets arbitrarily close to maximal; and Φ is the standard-normal CDF. Importantly, this term vanishes in k, so the KL-divergence term dominates the objective function.

The crux of our technical approach in proving the reduction from Problem 1 to Problem 2 in Theorem 2.1 is showing that δ is sandwiched between two values $\delta_{P_{Y|X}^{\circ k}}(\underline{\varepsilon}_k(\delta))$ and $\delta_{P_{Y|X}^{\circ k}}(\overline{\varepsilon}_k(\delta))$, where for our choice of values we have both $\underline{\varepsilon}_k(\delta)/k$ and $\overline{\varepsilon}_k(\delta)/k$ of order $\sup_{\|x-x'\|\le s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}) + \Theta(1/\sqrt{k})$, from which one may conclude that the true value $\varepsilon_{P_{Y|X}^{\circ k}}(\delta)/k$ has this order too. We introduce next some useful notation. Denote

$$\mathrm{KL}_{x,x'} := D(P_{Y|X=x} \,\|\, P_{Y|X=x'}), \tag{2.29}$$

and set

$$\mathrm{KL}_{\max} := \sup_{\|x-x'\|\le s} \mathrm{KL}_{x,x'}. \tag{2.30}$$

We suppress the dependence on $P_{Y|X}$ in the above notation for readability. Note that the proposed problem in (2.8) aims to minimize $\mathrm{KL}_{\max}$ subject to a cost constraint:

$$\begin{aligned} &\inf_{P_{Y|X}\in\mathscr{R}} && \mathrm{KL}_{\max} \\ &\text{subject to} && \sup_{x\in\mathbb{R}^m} \mathbb{E}_{P_{Y|X=x}}[T_x c] \le C. \end{aligned} \tag{2.31}$$

Note that $\mathrm{KL}_{x,x'} \ge 0$. It suffices to consider the case $\mathrm{KL}_{\max} > 0$, since the degenerate case $\mathrm{KL}_{\max} = 0$ yields an impractical mechanism that outputs only noise independent of the input data. We shall assume that $\mathrm{KL}_{\max} < \infty$, since otherwise the mechanism $P_{Y|X}$ would be infeasible for the problem (2.8).

Note that $\mathrm{KL}_{x,x'}$ is the mean $\mathrm{KL}_{x,x'} = \mathbb{E}[L_{x,x'}]$ (see (2.2) for the definition of the PLRV $L_{x,x'}$). We will also consider the variance of $L_{x,x'}$, which we denote by

$$\mathrm{V}_{x,x'} := \mathbb{E}\left[\left(L_{x,x'} - \mathrm{KL}_{x,x'}\right)^2\right]. \tag{2.32}$$

For tuples of vectors $u^{(k)}, v^{(k)} \in (\mathbb{R}^m)^k$, we will denote $\mathrm{KL}_{u^{(k)},v^{(k)}} := \sum_{j\in[k]} \mathrm{KL}_{u_j,v_j}$, and $\mathrm{V}_{u^{(k)},v^{(k)}}$ is defined similarly. We will assume that $P_{Y|X}$ satisfies $\mathrm{V}_{\max} := \sup_{\|x-x'\|\le s} \mathrm{V}_{x,x'} < \infty$. Note that the values of $\mathrm{KL}_{\max}$ and $\mathrm{V}_{\max}$ are obtained by element-wise maximization:

$$\sup_{\|u_j-v_j\|\le s,\ j\in[k]} \mathrm{KL}_{u^{(k)},v^{(k)}} = k\cdot \mathrm{KL}_{\max}, \tag{2.33}$$

$$\sup_{\|u_j-v_j\|\le s,\ j\in[k]} \mathrm{V}_{u^{(k)},v^{(k)}} = k\cdot \mathrm{V}_{\max}. \tag{2.34}$$

The following result shows that $\varepsilon_{P_{Y|X}^{\circ k}}(\delta) \sim k\cdot \mathrm{KL}_{\max}$, thereby giving an asymptotic sense in which the DP mechanism-design problem reduces to the KL-divergence optimization (2.8) in the large-composition regime.

Theorem 2.1. *Let $P_{Y|X}$ be a Markov kernel on $\mathbb{R}^m$ satisfying item (a) of Assumption 2.3, i.e., its induced information densities have uniformly bounded means and variances. Then, for any $\delta \in (0,1/2)$, we have the bounds*

$$k\cdot(\mathrm{KL}_{\max} - o(1)) \le \varepsilon_{P_{Y|X}^{\circ k}}(\delta) \le k\cdot \mathrm{KL}_{\max} + \left(-\Phi^{-1}(\delta) + o(1)\right)\sqrt{k\cdot \mathrm{V}_{\max}}, \tag{2.35}$$

where $o(1)$ denotes a function that vanishes as $k\to\infty$. If, in addition, item (b) of Assumption 2.3 holds, then we have the refined expansion

$$\varepsilon_{P_{Y|X}^{\circ k}}(\delta) = k\cdot \mathrm{KL}_{\max} - \Phi^{-1}(\delta)\sqrt{k\cdot \mathrm{V}_k^\star}, \tag{2.36}$$

where the constants $\mathrm{V}_k^\star \ge 0$ satisfy the inequalities $\underline{\mathrm{V}} \le \liminf_{k\to\infty} \mathrm{V}_k^\star \le \limsup_{k\to\infty} \mathrm{V}_k^\star \le \mathrm{V}_{\max}$, with the constant in the lower bound $\underline{\mathrm{V}} \ge 0$ defined by

$$\underline{\mathrm{V}} := \inf\left\{ \liminf_{\ell\to\infty} \mathrm{V}_{x_\ell,x'_\ell} \ :\ x_\ell, x'_\ell \in \mathbb{R}^m,\ \sup_{\ell\in\mathbb{N}} \|x_\ell - x'_\ell\| \le s,\ \lim_{\ell\to\infty} \mathrm{KL}_{x_\ell,x'_\ell} = \mathrm{KL}_{\max} \right\}. \tag{2.37}$$

Proof. See Appendix A.2. □

Remark 2.3. The mechanisms we propose in later sections all satisfy the premises of Theorem 2.1. Since the proposed distributions can perform arbitrarily close to optimal for the *unrestricted* problem (2.8), there is no loss in generality in imposing the restrictions in Theorem 2.1.

Remark 2.4. We derive in Theorem 2.16 an alternative large-composition theorem for a mechanism

$P_{Y|X}$ in terms of its dominating pairs (see (2.23)).

An related result to Theorem 2.1 gives a more precise asymptotic when one has access to tightly-dominating pairs. Namely, let (P, Q) be a tightly-dominating pair for $P_{Y|X}$, i.e.,

$$\sup_{\|x-x'\| \le s} \mathsf{E}_\gamma(P_{Y|X=x} \,\|\, P_{Y|X=x'}) = \mathsf{E}_\gamma(P \,\|\, Q) \tag{2.38}$$

for every $\gamma \ge 0$, and consider the privacy curve

$$\delta^{\mathrm{TD}}_{P^{\circ k}_{Y|X}}(\varepsilon) := \mathsf{E}_\gamma(P^{\otimes k} \,\|\, Q^{\otimes k}). \tag{2.39}$$

This curve is well-defined and it gives a privacy guarantee [DRS22, Theorem 4]: the k-fold composition of $P_{Y|X}$ satisfies $(\varepsilon, \delta^{\mathrm{TD}}_{P^{\circ k}_{Y|X}}(\varepsilon))$-DP for every $\varepsilon \ge 0$. Similarly to (2.28), consider the inverse curve

$$\varepsilon^{\mathrm{TD}}_{P^{\circ k}_{Y|X}}(\delta) := \inf\left\{ \varepsilon \ge 0 \;:\; \delta^{\mathrm{TD}}_{P^{\circ k}_{Y|X}}(\varepsilon) \le \delta \right\}, \tag{2.40}$$

for which we prove the following asymptotic.

Theorem 2.2. *Let $P_{Y|X}$ be a Markov kernel on $\mathbb{R}^m$. Assume that (P, Q) is a tightly dominating pair for $P_{Y|X}$ (see (2.38), and suppose that $\mathbb{E}_P[|\log \frac{dP}{dQ}|^3] < \infty$. Then, for any fixed $\delta \in (0, 1/2)$, we have the asymptotic (see (2.40))*

$$\varepsilon^{\mathrm{TD}}_{P^{\circ k}_{Y|X}}(\delta) = k \cdot D(P \,\|\, Q) - \Phi^{-1}(\delta)\sqrt{k \cdot \mathrm{V}(P \,\|\, Q)} + o(\sqrt{k}). \tag{2.41}$$

as $k \to \infty$.

Proof. This follows from the stronger result we prove later in Theorem 2.16 by instantiating it to the case of fixed δ. □

2.10 Optimality of Additive, Continuous, Spherically-Symmetric Mechanisms

We start by deriving characterizations of solutions to the optimization problem (2.8). The difficulty of this problem lies in the fact that we are optimizing over all conditional distributions. This not only makes the problem infinite-dimensional, but it also renders direct approaches ineffective. The main result of this section, shown in Theorem 2.3 below, is that it suffices to consider continuous additive channels. In other words, the optimization in (2.8) may be restricted to conditional distributions of the form $P_{Y|X=x} = T_x P$ for some Borel probability measure P on $\mathbb{R}$ that is absolutely continuous with respect to the Lebesgue measure and which is spherically symmetric. Equipped with this reduction, we build in the next two sections an explicit family of finitely-parametrized distributions that are also optimal in (2.8). Thus, the main result of this section is stated as follows.

Theorem 2.3. *If the cost function c satisfies Assumption 2.1, then there is an additive, continuous, spherically-symmetric mechanism solving the optimization problem* (2.8).

Proof. We break down the proof of Theorem 2.3 into the three parts in Theorems 2.4–2.6 which we prove throughout this section. By Assumption 2.1, we have both continuity and the vanishing of c at the origin. Hence, choosing an additive mechanism with a sufficiently rapidly decaying PDF, we see that problem (2.8) is feasible, i.e., its optimal value is not ∞. By Theorem 2.4, there is an additive mechanism $P^\star$ achieving this optimal value. By Theorem 2.5, $P^\star$ must be continuous. By Theorem 2.6, we may assume that $P^\star$ is spherically-symmetric, and the proof is complete. □

Let $\mathscr{P} \subset \mathscr{R}$ be the set of conditional distributions $P_{Y|X}$ satisfying the cost constraint in (2.8), i.e.,

$$\mathscr{P} := \left\{ P_{Y|X} \in \mathscr{R} \, ; \, \sup_{x\in\mathbb{R}^m} \mathbb{E}[c(Y-x) \mid X = x] \le C \right\}. \tag{2.42}$$

The infimal value in (2.8) is then

$$\mathrm{KL}^\star := \inf_{P_{Y|X}\in\mathscr{P}} \sup_{x,x'\in\mathbb{R}^m:\|x-x'\|\le s} D(P_{Y|X=x} \,\|\, P_{Y|X=x'}), \tag{2.43}$$

where we are considering the ℓ_2-sensitivity here. We are interested in computing $\mathrm{KL}^\star$, as well as mechanisms $P_{Y|X}$ that approach this optimal value. Note that, for clarity of presentation, we suppress the dependence on (s, c, C) in the notations $\mathscr{P}$ and $\mathrm{KL}^\star$.

2.10.1 Additive Mechanisms are Optimal

In the main problem (2.8), we allow $P_{Y|X}$ to be any mechanism that produces Y given X. A more restrictive but natural and easy-to-implement class of mechanisms is the *additive* mechanism class. An additive mechanism is given by $P_{Y|X=x}(B) = T_x P(B)$ where P is a Borel probability measure on $\mathbb{R}^m$. In other words, an additive mechanism $P_{Y|X}$ has Y of the form $Y = X + Z$ for some noise random variable $Z \sim P \in \mathscr{B}$ that is independent of the input X. Let $\mathscr{P}_{\mathrm{add}} \subset \mathscr{B}$ be the set of additive mechanisms satisfying the cost constraint in (2.8),

$$\mathscr{P}_{\text{add}} := \{P \in \mathscr{B} \,;\, \mathbb{E}_P[c] \le C\}\,. \tag{2.44}$$

Since the KL-divergence is shift-invariant, restricting the optimization (2.8) to additive mechanisms amounts to considering the simplified optimization problem

$$\mathrm{KL}^\star_{\text{add}} := \inf_{P \in \mathscr{P}_{\text{add}}} \sup_{a \in \mathbb{R}^m : \|a\| \le s} D(P \,\|\, T_a P). \tag{2.45}$$

Of course, it is immediate that $\mathrm{KL}^\star \le \mathrm{KL}^\star_{\text{add}}$. In fact, we show below that these quantities are the same, meaning that there is no loss in restricting to additive mechanisms.

The optimization problem in (2.8) is a convex problem, but the fact that the feasible set $\mathscr{P}$ is of infinite dimension means it cannot be solved directly, nor do the tractable properties one expects of a convex optimization problem necessarily follow. For example, in any finite dimensional convex optimization problem, a symmetry in the problem leads to the same symmetry in the solution. In this problem, and under Assumption 2.1, one can see that shifting the mechanism—i.e., given $P_{Y|X}$, construct $Q_{Y|X=x}(B) = P_{Y|X=x+z}(B+z)$ for some z—does not change the cost constraint nor the objective value in (2.8). Thus, one might be inclined to conclude that the optimal mechanism is invariant to a shift (i.e., is an additive mechanism). Unfortunately, the infinite-dimensional nature of the problem means that this conclusion is not immediate. We resolve this issue in the following theorem which states that additive mechanisms are in fact optimal in (2.8).

Theorem 2.4. *If the cost function c satisfies Assumption 2.1, then $\mathrm{KL}^\star = \mathrm{KL}^\star_{\text{add}}$, and there exists an additive mechanism achieving this optimal value.*

Proof sketch. The proof is given in Appendix A.3. We give here only a high level description of the approach. Let $P^{(k)}_{Y|X}$ be a sequence achieving $\mathrm{KL}^\star$. We make these mechanisms increasingly closer to being additive, while sacrificing neither feasibility nor utility, by considering the convex combinations

$$\overline{P}^{(k)}_{Y|X=x}(A) := \mathbb{E}\left[P^{(k)}_{Y|X=x+Z_k}(A + Z_k)\right] \tag{2.46}$$

where $Z_k \sim \mathrm{Unif}(B_k(0))$. Specifically, one can invoke Prokhorov's theorem on the $\overline{P}^{(k)}_{Y|X}$, thereby extracting a probability measure $P^\star$ such that $\overline{P}^{(k)}_{Y|X=x} \to T_x P^\star$ weakly for each fixed x. Finally, we show that the additive mechanism $P^\star$ is optimal by invoking joint convexity and lower-semicontinuity of the KL-divergence. □

2.10.2 Feasible Additive Mechanisms Must Be Continuous

Recall that the finiteness of the KL-divergence $D(P \parallel Q) < \infty$ necessarily implies the absolute continuity $P \ll Q$. This fact can be used to conclude that an additive mechanisms can yield a finite objective value in (2.8) only if it is continuous.

Theorem 2.5. *If $\mu \in \mathscr{B}$ satisfies $\sup_{\|x\| \le s} D(\mu \parallel T_x\mu) < \infty$, then we necessarily have $\mu \ll \lambda$. In particular, any feasible additive mechanism in (2.8) must be continuous.*

Proof. We show that the relation $\mu \ll T_x\mu$ for every $\|x\| \le s$ is enough to conclude that $\mu \ll \lambda$. Fix a Borel set $A \subset \mathbb{R}^m$ such that $\lambda(A) = 0$, and we will show that $\mu(A) = 0$. Note that the function $x \mapsto (T_x\mu)(A)$ is Borel as it is given by the convolution $1_A * \eta$ where $\eta(E) := \mu(-E)$. Then, by Tonelli's theorem and translation-invariance of the Lebesgue measure,

$$\int_{\mathbb{R}^m} (T_x\mu)(A)\, d\lambda(x) = \int_{\mathbb{R}^{2m}} 1_{A-x}(b)\, d\mu(b)\, d\lambda(x) \tag{2.47}$$

$$= \int_{\mathbb{R}^{2m}} 1_{A-x}(b)\, d\lambda(x)\, d\mu(b) \tag{2.48}$$

$$= \int_{\mathbb{R}^{2m}} 1_{A-b}(x)\, d\lambda(x)\, d\mu(b) \tag{2.49}$$

$$= \int_{\mathbb{R}^m} (T_b\lambda)(A)\, d\mu(b) \tag{2.50}$$

$$= \int_{\mathbb{R}^m} \lambda(A)\, d\mu(b) = 0. \tag{2.51}$$

Thus, $(T_x\mu)(A) = 0$ for λ-almost every x. In particular, $(T_x\mu)(A) = 0$ for at least one $x \in B_s(0)$. Thus, $\mu \ll T_x\mu$ implies $\mu(A) = 0$, and the proof is complete. □

2.10.3 Spherically-Symmetric Additive Mechanisms are Optimal

We show in the following theorem that any mechanism can be replaced with another spherically-symmetric mechanism without reducing either of the objective value or the cost incurred in problem (2.8). Note that spherical symmetry amounts to evenness in the single-dimensional case (i.e., when $m = 1$). In this simpler case, we would take $q(z)$ to be the average $\frac{p(z)+p(-z)}{2}$, and joint convexity of the KL-divergence would finish the proof. For $m > 1$, however, we need to average over all possible rotations of p. In $\mathbb{R}$, these are just multiplications of the input by an element in $\{\pm 1\}$, but in higher dimensions we need to consider multiplication by infinitely many orthogonal matrices. Thus, we need to utilize the existence of the Haar measure in general.

Theorem 2.6. *Suppose that the cost function c satisfies Assumption 2.1. For any continuous additive mechanism, there is a corresponding spherically-symmetric, continuous, additive mechanism that increases neither the objective value nor the cost constraint in problem (2.8).*

Proof. Fix a continuous additive mechanism P and let p be its PDF. We will use p to define a new PDF q that is isotropic, satisfies the cost constraint, and does not increase the maximal KL-divergence.

Specifically, let $\mathrm{O}(m)$ denote the orthogonal group, i.e., the topological group of orthogonal $m \times m$ real matrices under multiplication and with the subspace topology inherited from $\mathbb{R}^{m\times m}$. Let μ denote the right Haar measure on $\mathrm{O}(m)$ with the normalization $\mu(\mathrm{O}(m)) = 1$; as $\mathrm{O}(m)$ is a compact topological group, μ is well-defined. In particular, μ is a Borel probability measure, and $\mu(\mathcal{S}U) = \mu(\mathcal{S})$ for any element $U \in \mathrm{O}(m)$ and Borel subset $\mathcal{S} \subset \mathrm{O}(m)$ (i.e., μ is invariant under multiplication on the right). We will define the PDF $q : \mathbb{R}^m \to \mathbb{R}_+$ by

$$q(z) := \int_{\mathrm{O}(m)} p(Uz)\ d\mu(U). \tag{2.52}$$

We will check that q is well-defined, is isotropic, satisfies the cost constraint, and has a maximal KL-divergence upper bounded by that of p. Then, the mechanism $Y = X + Z$ with Z having PDF q and independent of X would verify the claim of the theorem.

To see that q is well-defined, note that the mapping $(U, z) \mapsto Uz$ is continuous and p is Borel, hence $(U, z) \mapsto p(Uz)$ is Borel, and Fubini's theorem yields that $z \mapsto q(z)$ is Borel. For isotropy of q, note that for every $V \in \mathrm{O}(m)$, right-invariance of μ yields that

$$q(Vz) = \int_{\mathrm{O}(m)} p(UVz)\ d\mu(U) = \int_{\mathrm{O}(m)} p(Uz)\ d\mu(U) = q(z). \tag{2.53}$$

That q satisfies the cost constraint can be seen via Tonelli's theorem and isotropy of c:

$$\begin{aligned}
\int_{\mathbb{R}^m} c(z)q(z)\,dz &= \int_{\mathbb{R}^m} c(z) \int_{\mathrm{O}(m)} p(Uz)\,d\mu(U)\,dz && (2.54)\\
&= \int_{\mathrm{O}(m)} \int_{\mathbb{R}^m} c(z)p(Uz)\,dz\,d\mu(U) && (2.55)\\
&= \int_{\mathrm{O}(m)} \int_{\mathbb{R}^m} c(U^T w)p(w)\,dw\,d\mu(U) && (2.56)\\
&= \int_{\mathrm{O}(m)} \int_{\mathbb{R}^m} c(w)p(w)\,dw\,d\mu(U) && (2.57)\\
&\le \int_{\mathrm{O}(m)} C\,d\mu(U) = C. && (2.58)
\end{aligned}$$

Finally, that q does not increase the maximal KL-divergence can be deduced from joint convexity of the KL-divergence. Indeed, consider the PDFs $r_{x,V}(z) := (T_{Vx}p)(Vz)$ for $(x,V,z) \in \mathbb{R}^m \times \mathrm{O}(m) \times \mathbb{R}^m$. Then, with $U \sim \mu$ and $\|x\| \le s$ (so $\|Ux\| = \|x\| \le s$), we have that

$$\begin{aligned} D(q \,\|\, T_x q) &= D\left(z \mapsto \mathbb{E}[r_{0,U}(z)] \,\|\, z \mapsto \mathbb{E}[r_{x,U}(z)]\right) && (2.59) \\ &\le \mathbb{E}\left[D(r_{0,U} \,\|\, r_{x,U})\right] && (2.60) \\ &= \mathbb{E}\left[D(p \,\|\, T_{Ux} p)\right] && (2.61) \\ &\le \sup_{\|a\| \le s} D(p \,\|\, T_a p). && (2.62) \end{aligned}$$

The proof is thus complete. □

2.11 Proposed Mechanisms

For clarity of presentation, we include separate section for the construction of our proposed mechanisms and for respective results regarding their optimality. This section is devoted to only *defining* our proposed mechanisms and illustrating the shapes of their PDFs. Proofs of their respective optimalities occupy Sections 2.12–2.14.

We introduce next new mechanisms that are optimal or can get arbitrarily close to optimal for the main KL-divergence problem (2.8). Namely, we introduce: (i) *Cactus mechanisms* for scalar queries and fixed sensitivity, (ii) *isotropic mechanisms* for vector queries and fixed sensitivity, and (iii) *Schrödinger mechanisms* for scalar queries and small sensitivity ($s \to 0^+$). We give the mathematical construction of these mechanisms in this section, and we prove optimality in the next three sections. Later, in Section 2.15, we demonstrate the DP performance of these mechanisms.

In Theorem 2.3 we reduced the main KL-divergence problem (2.8) to the case of additive, continuous, spherically-symmetric mechanisms. Symbolically, it suffices to solve the problem

$$\begin{aligned} \inf_{p} \quad & \sup_{\|a\| \le s} D(p \,\|\, T_a p) \\ \text{subject to} \quad & \mathbb{E}_p[c] \le C, \end{aligned} \qquad (2.63)$$

where p ranges over spherically-symmetric PDFs on $\mathbb{R}^m$. Still, the optimization problem over additive mechanisms in (2.63) is infinite-dimensional, so it cannot be solved numerically as-is, and it appears to have no closed-form solution for non-trivial cost functions and fixed sensitivity s. The lack of closed-form solution is true even for the simple 1-dimensional variance case, i.e., $m = 1$ and $c(x) = x^2$:

to our surprise, as will be illustrated later, the Gaussian mechanism is not optimal![3] Therefore, in the regime of fixed positive s, and for arbitrary dimension m, to find practically achievable near-optimal mechanisms, we resort to numerical approximation of (2.45). In Section 2.14, we explore the regime where $s \to 0^+$; in this limit, we show that the optimal distribution can be determined exactly, and in fact for quadratic cost the limiting optimal distribution *is* Gaussian—although for other costs the optimal distribution is much more surprising.

In what follows, we introduce new mechanisms that perform arbitrarily close to optimal for the problem (2.63) (hence for the main problem (2.8) too by Theorem 2.3). We start in this section by giving a brief overview of the construction of our proposed mechanisms.

Remark 2.5. In the fixed-sensitivity regime, we set $s = 1$. We can do this without loss of generality simply by scaling: that is, the optimization problem in (2.63) with sensitivity s and cost function $c(x)$ is equivalent to the same problem with sensitivity 1 and cost function $c(sx)$.

2.11.1 The Cactus Mechanisms

We consider scalar mechanisms first, so we set $m = 1$ in this subsection. We are also considering a fixed sensitivity, so we set $s = 1$ (see Remark 2.5). To approximate (2.63) by a numerically tractable problem, we: (i) quantize the distribution, and (ii) only explicitly parameterize the distribution in a certain interval. Specifically, we construct a mapping from finite-length vectors to continuous measures as follows.

Definition 2.3 (Cactus mechanism)**.** Fix two positive integers n and N, and a constant $r \in (0, 1)$. Consider the partition of $\mathbb{R}$ by intervals $\{\mathcal{J}_{n,i}\}_{i\in\mathbb{Z}}$ defined by: $\mathcal{J}_{n,0} := [-1/(2n), 1/(2n)]$ and

$$\mathcal{J}_{n,i} := \begin{cases} \left(\frac{i-1/2}{n}, \frac{i+1/2}{n}\right], & \text{if } i > 0, \\ \left[\frac{i-1/2}{n}, \frac{i+1/2}{n}\right), & \text{if } i < 0. \end{cases} \tag{2.64}$$

We associate to each vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_N) \in [0, 1]^{N+1}$ a piecewise constant function that is defined by

$$f_{n,r,\boldsymbol{p}}(x) = \begin{cases} np_{|i|}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| < N, \\ np_N r^{|i|-N}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| \geq N. \end{cases} \tag{2.65}$$

[3] Of course, simply because Gaussian is not optimal does not imply that there is no closed-form solution. It is possible to write a set of KKT conditions for (2.45). This set of KKT conditions cannot be solved in closed-form.

We also associate with $f_{n,r,\boldsymbol{p}}$ the Borel measure $P_{n,r,\boldsymbol{p}}$, where

$$P_{n,r,\boldsymbol{p}}(B) := \int_B f_{n,r,\boldsymbol{p}}(x)\, dx. \tag{2.66}$$

For any fixed triplet (n, N, r), and for any $\boldsymbol{p}^\star$ that solves the restriction of problem (2.63) to the class $\{f_{n,r,\boldsymbol{p}} : f_{n,r,\boldsymbol{p}} \text{ is a PDF}\}_{\boldsymbol{p}\in[0,1]^{N+1}}$, we call the additive mechanism $P_{n,r,\boldsymbol{p}^\star}$ a *Cactus mechanism*.

Remark 2.6. Note that

$$\int_{\mathbb{R}} f_{n,r,\boldsymbol{p}}(x)\, dx = p_0 + \sum_{i=1}^{N-1} 2p_i + \frac{2p_N}{1-r} =: S_{r,\boldsymbol{p}}. \tag{2.67}$$

If $S_{r,\boldsymbol{p}} = 1$, then $P_{n,r,\boldsymbol{p}}$ is a probability measure with density $f_{n,r,\boldsymbol{p}}$. This distribution is symmetric around the origin, i.e., $f_{n,r,\boldsymbol{p}}(x) = f_{n,r,\boldsymbol{p}}(-x)$. Further, its tails decay almost geometrically with base r^n: for $(N + 1/2)/n < x_1 < x_2$ one has $f_{n,r,\boldsymbol{p}}(x_2) = r^{nk} \cdot f_{n,r,\boldsymbol{p}}(x_1)$ where $k = (\lceil nx_2 - 1/2\rceil - \lceil nx_1 - 1/2\rceil)/n \approx x_2 - x_1$.

Figure 2.1 shows an example of a Cactus mechanism for a quadratic cost. This plot shows the Cactus PDF $f_{200,0.9,\boldsymbol{p}^\star}$, where $\boldsymbol{p}^\star \in [0,1]^{1601}$ is obtained by solving the restriction of problem (2.63) to the class of potential Cactus mechanisms (i.e., PDFs of the form $f_{200,0.9,\boldsymbol{p}}$ for some $\boldsymbol{p} \in [0,1]^{1601}$). We explicitly give the numerical procedure for finding $\boldsymbol{p}^\star$ in Theorem 2.7 in the next section. The shape of this distribution[4] has inspired the name the "Cactus distribution." We note that the Cactus PDF is only *piece-wise* continuous, but that the number of quantization bins $N = 1600$ is large is the reason it appears continuous in Figure 2.1.

We investigate the Cactus mechanisms in detail in Section 2.12. Specifically, we show the explicit form of the optimization problem that is the restriction of problem (2.63) to the construction in Definition 2.3 in Theorem 2.7. We also show that the Cactus mechanisms perform arbitrarily close to optimal for the main problem (2.8) in Theorem 2.8. Numerical experiments for the Cactus mechanism are also presented in Section 2.15.

2.11.2 Isotropic Mechanisms

Next, we consider vector-valued mechanisms in this subsection (so the dimension m is free). We generalize our approach for constructing the Cactus mechanism in the previous subsection. However, in this multidimensional setting, we only consider *monotone* PDFs, defined as follows.

[4] In addition to the state of Arizona being home of several of the authors.

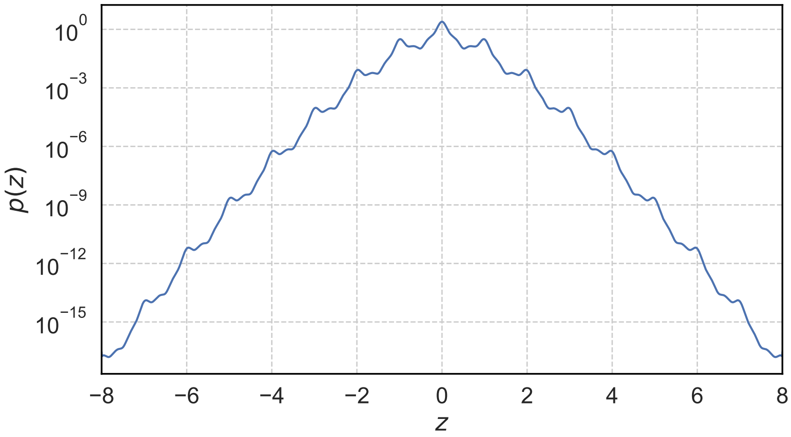


Figure 2.1: *The* Cactus distribution *plotted on a semi-log scale. The cost function is $c(z) = z^2$, and the parameters are: $s = 1$, $C = 0.25$, $n = 200$, $N = 1600$, and $r = 0.9$ (see Definition 2.3).*

Definition 2.4. We say that a continuous random vector Z is *monotone* if it has a PDF $p(z)$ such that for every $z \in \mathbb{R}^m$ and $t \in [0, 1)$, we have $p(tz) \geq p(z)$.

Remark 2.7. Note that a continuous random vector is monotone *and* spherically-symmetric if its PDF can be written as $p(z) = \widetilde{p}(\|z\|)$ such that $\widetilde{p} : \mathbb{R}_+ \to \mathbb{R}_+$ is non-increasing. One example is the Gaussian mechanism $Z \sim \mathcal{N}(0, \sigma^2 I_m)$.

Remark 2.8. Restricting attention to monotone mechanisms naturally leads to suboptimal solutions to the problem (2.63), unlike for the Cactus mechanism on the real line which is allowed to be non-monotone hence can be universally optimal. Nevertheless, we focus on monotone mechanisms since, for such mechanisms, it is tractable to both do DP accounting (see Lemma 2.1) as well as solve the KL-divergence optimization (2.8) (see Proposition 2.1). In particular, for general vector-valued mechanisms that are *not* monotone, it does not seem that current DP accounting techniques can be readily used to test such mechanisms' performance. The main difficulty here lies in the fact that one does not have in general a way to determine "worst shifts" if the mechanism is non-monotone; in contrast, we show in Lemma 2.1 that maximal shifts are worst shifts for monotone and spherically-symmetric mechanisms. We note that even with this restriction we will show that isotropic mechanisms are optimal for (2.8) among monotone mechanisms; see Remark 2.12 for more details.

As with the scalar case, the search space for the KL-divergence optimization (2.63) is infinite-dimensional, hence we resort to a quantization approach. We generalize the construction of the Cactus mechanism in Definition 2.3. We fix a large enough ball, which we divide into spherical shells of fixed small enough width. We require that the mechanism be constant over the individual spherical shells. Then, we impose geometric tails outside the fixed large ball. In addition, for the multidimensional case, we require the mechanism to be *a priori* monotone. Formally, we introduce the following construction.

Definition 2.5 (Isotropic mechanism)**.** Fix two positive integers n and N, a constant $r \in (0,1)$, and a vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_N) \in [0,\infty)^{N+1}$ with $p_0 \geq \cdots \geq p_N$. Consider the partition of $\mathbb{R}$ by intervals $\{\mathcal{J}_{i,n} := \left[\frac{i}{n}, \frac{i+1}{n}\right)\}_{i\in\mathbb{N}}$. We define the piecewise-constant function

$$\widetilde{f}_{n,r,\boldsymbol{p}}(\rho) := \begin{cases} p_i, & \text{if } \rho \in \mathcal{J}_{i,n}, \text{with } i < N, \\ p_N r^{i-N}, & \text{if } \rho \in \mathcal{J}_{i,n}, \text{with } i \geq N. \end{cases} \tag{2.68}$$

We also define the density $f_{n,r,\boldsymbol{p}} : \mathbb{R}^m \to [0,\infty)$ by

$$f_{n,r,\boldsymbol{p}}(x) := \widetilde{f}_{n,r,\boldsymbol{p}}(\|x\|), \tag{2.69}$$

and associate with $f_{n,r,\boldsymbol{p}}$ the Borel measure $P_{n,r,\boldsymbol{p}}$ given by

$$P_{n,r,\boldsymbol{p}}(B) := \int_B f_{n,r,\boldsymbol{p}}(x)\, dx. \tag{2.70}$$

For any fixed triplet (n,N,r), and for any $\boldsymbol{p}^\star$ that solves the restriction of problem (2.63) to the class $\{f_{n,r,\boldsymbol{p}} : f_{n,r,\boldsymbol{p}} \text{ is a PDF}\}_{\boldsymbol{p}\in[0,\infty)^{N+1}, p_0\geq\cdots\geq p_N}$, we call the mechanism $P_{n,r,\boldsymbol{p}^\star}$ an *isotropic mechanism*.

Visualizing an isotropic mechanism can be done via the distribution of its *radius*. It is not hard to see that any spherically symmetric random vector Z can be written in the form

$$Z = R \cdot U \tag{2.71}$$

where U is a uniformly distributed random vector over the unit $(m-1)$-sphere in $\mathbb{R}^m$, and R is a nonnegative scalar random variable (non necessarily independent of U). In fact, we may set $R = \|Z\|$ and $U = Z/\|Z\|$. We call $R = \|Z\|$ the radius of Z.

We plot in Figure 2.2 the distribution of the radius $R = \|Z\|$ with Z being an isotropic mechanism. Specifically, we fix the dimension to $m = 10$, use the quadratic cost with cost bound 2.5, and choose the construction parameters $(n,N,r) = (400, 1200, 0.9)$. Thus, $Z \sim P_{400,0.9,\boldsymbol{p}^\star}$, where $\boldsymbol{p}^\star$ is found

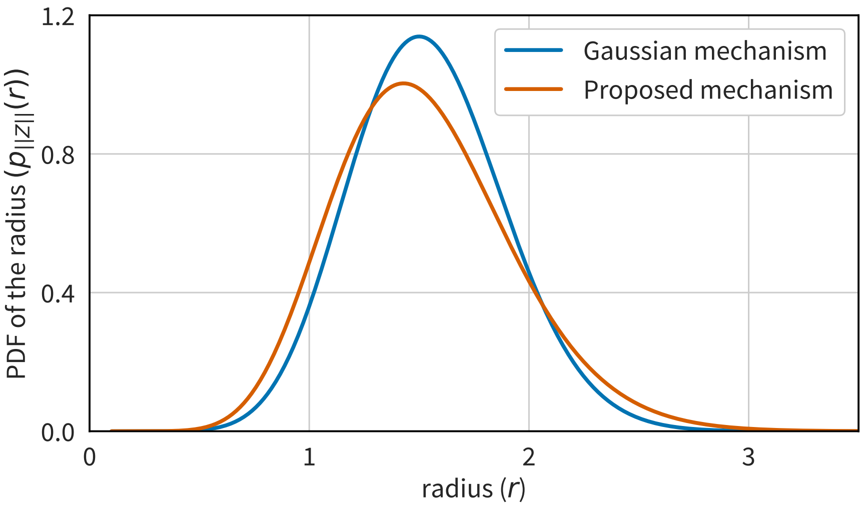


Figure 2.2: *The distributions of the radii of both the proposed* isotropic mechanism *and the Gaussian mechanism (for comparison), both in $m = 10$ dimensions and with a quadratic cost $\mathbb{E}[\|Z\|^2] = 2.5$. The construction parameters for the isotropic mechanism are $n = 400$, $N = 1200$, and $r = 0.9$.*

by solving (2.63) when restricted to the class of potential isotropic mechanisms (i.e., PDFs of the form $f_{400,0.9,\boldsymbol{p}}$ for $\boldsymbol{p} \in [0,\infty)^{1201}$ satisfying $p_0 \geq \cdots \geq p_N$). We also plot the PDF of the radius of a corresponding Gaussian vector, i.e., the mechanism adds the Gaussian vector $G \sim \mathcal{N}(0, 0.25 I_{10})$ whose radius satisfies $\|G\| \sim \frac{1}{2}\chi_{10}$. Note that both mechanisms in Figure 2.2 are monotone according to Definition 2.4, but this generally does not imply monotonicity of the PDF of the radial part of the random vectors.

We investigate the isotropic mechanisms in detail in Section 2.13. Specifically, we show the explicit form of the optimization problem that is the restriction of problem (2.63) to the construction in Definition 2.5 in Theorem 2.9. We also show that the isotropic mechanisms perform arbitrarily close to optimal for the main problem (2.8) among all monotone mechanisms in Theorem 2.10. Numerical experiments for the isotropic mechanism are also presented in Section 2.15.

2.11.3 Schrödinger Mechanisms

The third and final family of mechanisms we introduce are closed-form solutions to the scalar-query case in the regime of *small-sensitivity*. Thus, we fix the dimension to $m = 1$ in this subsection, and we will consider sensitivities in the regime $s \to 0^+$. We will introduce mechanisms that are the squares of the ground-state eigenfunctions of the Schrödinger operator with the potential function being a positive multiple of the cost function c. Formally, we define the Schrödinger mechanism as follows.

Definition 2.6 (Schrödinger mechanism)**.** The *Schrödinger mechanism* given the cost function c and parameter $\theta > 0$ is defined by $Y = X + Z$ for Z having the PDF $y^2_{\theta,c}$ where $y = y_{\theta,c}$ is the unique unit-L^2-norm and strictly positive solution to the Schrödinger equation

$$y'' = (\theta c - E)y, \tag{2.72}$$

with E an arbitrary constant. In addition, with $C = \mathbb{E}_{y^2_{\theta,c}}[c]$, we denote the PDF of the Schrödinger mechanism by $p^\star_{c,C} := y^2_{\theta,c}$.

Remark 2.9. As we will show in Lemma 2.2, there is a unique E for which the ODE (2.72) is uniquely solvable with the prescribed properties for the solution y. Further, this value of E is the minimal eigenvalue of the Schrödinger operator with potential θc.

For example, if $c(x) = x^2$ is the quadratic cost, then the Schrödinger eigenproblem treats what is known as the quantum harmonic oscillator in quantum physics. The ground-state eigenfunction is known to be the Gaussian function. Then, the Schrödinger mechanism is in fact the Gaussian mechanism, and we have

$$p^\star_{c,C}(x) = \frac{1}{\sqrt{2\pi C}} e^{-x^2/(2C)}, \tag{2.73}$$

i.e., $p^\star_{c,C}$ is the centered Gaussian PDF with variance C.

As another example, consider the absolute value cost $c(x) = |x|$. In this case, the Schrödinger mechanism can be described using the Airy function [NIS, Chapter 9], as follows. The differential equation

$$y''(x) = xy(x) \tag{2.74}$$

has two linearly independent solutions, called the Airy functions. They are denoted by Ai and Bi, where Ai is the solution such that $\mathsf{Ai}(x) \to 0$ as $x \to \infty$; specifically, Ai is approximated as $\mathsf{Ai}(x) \sim e^{-2x^{3/2}/3}/(2\sqrt{\pi}x^{1/4})$. This function can be expressed by the improper Riemann integral

$$\mathsf{Ai}(x) = \frac{1}{\pi} \lim_{N\to\infty} \int_0^N \cos\left(\frac{t^3}{3} + xt\right) dt. \tag{2.75}$$

This function is analytic, and there are countably many zeros of Ai and Ai' all falling on the negative half-line. As is customary, the zeros of Ai and Ai' are denoted by $a_1 > a_2 > \cdots$ and $a'_1 > a'_2 > \cdots$, respectively. It is also known that we have the values

$$a_1 = -2.33810\ldots, \quad a_1' = -1.01879\ldots, \quad \text{and} \quad \mathsf{Ai}(a_1') = 0.53565\ldots. \tag{2.76}$$

In particular, the function Ai is strictly positive and strictly decreasing over $[a_1', \infty)$. We use the Airy function to construct the following density, which we show in Section 2.14 to be equal to $p^\star_{c,C}$.

Definition 2.7 (Airy distribution)**.** For $C > 0$, we define the *Airy distribution* with first absolute moment C as the probability measure whose PDF $p_{\mathsf{Ai},C}$ is given by

$$p_{\mathsf{Ai},C}(x) := \frac{1}{3C\mathsf{Ai}(a_1')^2}\mathsf{Ai}\left(\frac{-2a_1'}{3C}|x| + a_1'\right)^2. \tag{2.77}$$

Remark 2.10. We show in Lemma 2.3 that $p_{\mathsf{Ai},C}$ is indeed a PDF with first absolute moment C, and we also derive its variance.

In Proposition 2.5, we show that $p^\star_{c,C} = p_{\mathsf{Ai},C}$ when $c(x) = |x|$. In Figure 2.3, we illustrate the Airy distribution and compare it with the Laplace distribution. We note that the Airy distribution has a lighter tail than that of the Laplace distribution, where the decay rate of the former is $e^{-\Theta(|x|^{3/2})}$ and that of the latter is $e^{-|x|}$. Further, since the Airy function Ai is strictly positive and strictly decreasing over $[a_1', \infty)$, we see that the Airy PDF $p_{\mathsf{Ai},C}$ is even, strictly positive everywhere, and strictly decreasing over $[0, \infty)$. Also, the Airy distribution is differentiable at the origin, unlike the Laplace distribution.

We investigate the Schrödinger mechanisms in more detail in Section 2.14. There, we show their optimality in the small sensitivity regime. This optimality is a byproduct of the stronger result that the Schrödinger PDF $p^\star_{c,C}$ is the unique *global* minimizer of the Fisher information—a result of independent interest. That is, we show that (under Assumption 2.2) the PDF $p^\star_{c,C}$ uniquely solves the minimization

$$p^\star_{c,C} = \underset{\substack{p \in \mathcal{P} \\ \mathbb{E}_p[c] \le C}}{\operatorname{argmin}}\, I(p). \tag{2.78}$$

2.12 Optimality of the Cactus Distribution on the Real Line

We show in this section that the Cactus distribution family introduced in Definition 2.3 is optimal for (2.8), and we show also that each Cactus distribution is obtainable via a tractable finite-dimensional convex optimization problem. Recall that the Cactus mechanism is scalar, so we are fixing $m = 1$ in this section.

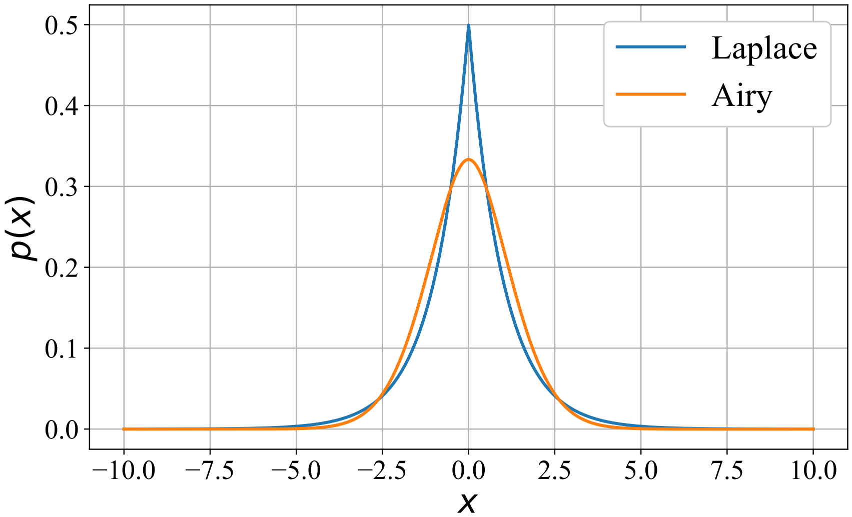


Figure 2.3: *The densities of the Laplace distribution and the Airy distribution,* $p_{\mathsf{Ai},C}(x)$ *(introduced in Definition 2.7). Both of these densities have absolute first moment equal to one.*

We use the following notation. Consider the restriction of (2.63) to the mechanisms constructible by Definition 2.3. For a fixed triplet $(n, N, r) \in \mathbb{N}^2 \times (0, 1)$, consider the set of mechanisms $\mathscr{C}_{n,N,r} \subset \mathscr{B}$,

$$\mathscr{C}_{n,N,r} := \left\{ P_{n,r,\boldsymbol{p}} \; ; \; \boldsymbol{p} \in [0,1]^{N+1}, S_{r,\boldsymbol{p}} = 1 \right\}. \tag{2.79}$$

(Recall that we define $S_{r,\boldsymbol{p}} = P_{n,r,\boldsymbol{p}}(\mathbb{R})$ in (2.67).) Denote the optimal value achievable by the class $\mathscr{C}_{n,N,r}$ with

$$\mathrm{KL}^{\star}_{n,N,r}(C) := \inf_{\substack{P \in \mathscr{C}_{n,N,r} \\ \mathbb{E}_P[c] \le C}} \sup_{|a| \le 1} D(P \,\|\, T_a P). \tag{2.80}$$

We show next that we may restrict the shift a in the supremum in (2.80) to take values over the finite set $\{1/n, 2/n, \cdots, 1\}$ (rather than varying over the whole interval $[-1, 1]$), thereby rendering (2.80) a finite-dimensional optimization problem amenable to standard numerical convex-programming methods.

For each $i \in \mathbb{Z}$, we denote the constants

$$c_{n,i} := \int_{\mathcal{I}_{n,i}} n c(x) \, dx. \tag{2.81}$$

Theorem 2.7. *Fix $r \in (0, 1)$, and positive integers $n < N$. The minimization (2.80) can be recast as the* *following convex program over the variable* $\boldsymbol{p} = (p_0, \cdots, p_N) \in \mathbb{R}^{N+1}$

$$\begin{aligned}
\underset{\boldsymbol{p}}{\text{minimize}} \quad & \max_{k \in \{1,\dots,n\}} \frac{1}{2} \sum_{i=-N+1}^{N-k-1} (p_{|i|} - p_{|i+k|}) \log \frac{p_{|i|}}{p_{|i+k|}} + \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} \\
& \quad + p_N \frac{1-r^k}{1-r} k \log r^{-1} \\
\text{subject to} \quad & p_0 c_{n,0} + \sum_{i=1}^{N-1} 2 p_i c_{n,i} + 2 p_N \sum_{i=N}^{\infty} c_{n,i} r^{i-N} \le C, \\
& p_0 + \sum_{i=1}^{N-1} 2 p_i + \frac{2 p_N}{1-r} = 1, \\
& p_i \ge 0 \text{ for all } i \in \{0, \dots, N\}.
\end{aligned} \tag{2.82}$$

Proof. See Appendix A.4. □

The main result regarding the Cactus mechanisms is the following theorem, showing that the Cactus mechanisms derived from the optimization problem (2.82) are in fact globally optimal for the main optimization problem (2.8).

Theorem 2.8. *Consider* $m = 1$*, and suppose the cost function* $c : \mathbb{R} \to \mathbb{R}$ *satisfies Assumption* 2.1*. Assume also that there are constants* $\alpha, \beta > 0$ *such that* $c(x) \sim \beta |x|^\alpha$ *as* $|x| \to \infty$*. Denote the optimal value a Cactus distribution (see Definition* 2.3*) can achieve by*

$$\mathrm{KL}^\star_{\text{Cactus}} := \lim_{\eta \to 0^+} \inf_{(n,N,r) \in \mathbb{N}^2 \times (0,1)} \mathrm{KL}^\star_{n,N,r}(C + \eta). \tag{2.83}$$

We have that $\mathrm{KL}^\star = \mathrm{KL}^\star_{\text{Cactus}}$*. In other words, Cactus mechanisms can get arbitrarily close to optimal for the problem* (2.8)*.*

Proof. See Appendix A.5. □

Remark 2.11. The proof of Theorem 2.8 gives some guidelines for choosing the parameters (n, N, r). For example, optimal Cactus distributions can be obtained by restricting the ratio N/n (chosen sufficiently large), and choosing $r = 1 - \Theta_\alpha(N^{-1})$.

2.13 Optimality of Multidimensional Isotropic Mechanisms

We turn our attention to the multidimensional setting in this section, where we show optimality of the isotropic mechanisms introduced in Definition 2.5. We fix an arbitrary dimension $m \ge 1$, and we set $s = 1$ (see Remark 2.5). We focus on monotone mechanisms, which is supported by the results of Lemma 2.1 and Proposition 2.1 below. In this section, we show three facts:

1. For monotone and spherically symmetric PDFs, maximal shifts are worst shifts when solving problems of the form $\sup_{\|a\|\le s} D_f(p \,\|\, T_a p)$ for *any* f-divergence. In particular, this holds for the E_γ divergence ($\gamma \ge 1$) and the KL-divergence.

2. The isotropic mechanism can be found via a tractable finite-dimensional convex program.

3. The isotropic mechanism is optimal among monotone mechanisms for the main problem (2.8).

2.13.1 Maximal Shifts and Worst Shifts

The following lemma shows that accounting for monotone spherically-symmetric DP mechanisms reduces to computing the E_γ divergence at the maximal shift. This property is known to hold for the Gaussian mechanism [ACG$^+$16].

Lemma 2.1. *If $Z \sim p$ is a monotone spherically-symmetric random vector (as in Definition 2.4), and $\gamma \ge 1$, then $a \mapsto \mathsf{E}_\gamma(p \,\|\, T_a p)$ is spherically symmetric and increasing in the norm $\|a\|$. In particular, for any $s > 0$ we have*

$$\max_{\|a\|\le s} \mathsf{E}_\gamma(p \,\|\, T_a p) = \mathsf{E}_\gamma(p \,\|\, T_{se_1} p). \tag{2.84}$$

Proof. See Appendix A.6.1. □

We generalize Lemma 2.1 in another dimension, namely, we show next that the same result holds for any f-divergence. Specializing this result to the KL-divergence will help simplify the numerical implementation we give later in this section for the isotropic DP mechanism we propose.

Proposition 2.1. *Let $f : (0,\infty) \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. For any monotone spherically-symmetric random vector $Z \sim p$, the mapping $a \mapsto D_f(p \,\|\, T_a p)$ is spherically symmetric and increasing in the norm $\|a\|$. In particular, for any $s > 0$ we have*

$$\max_{\|a\|\le s} D_f(p \,\|\, T_a p) = D_f(p \,\|\, T_{se_1} p). \tag{2.85}$$

Proof. See Appendix A.6.2. □

Remark 2.12. The above results show that monotonicity facilitates DP accounting—indeed, accounting for multidimensional non-monotonic mechanisms presents a significant challenge, since, as suggested by (2.27), one must maximize over all possible $u_i - v_i$ for $1 \leq i \leq k$. Thus, in this section we restrict attention to the subclass of monotone mechanisms. Recall that Theorem 2.6 shows that, among all additive mechanisms, spherically-symmetric ones are optimal. It can be seen that spherically-symmetric mechanisms would still be optimal among all monotone mechanisms for the KL-divergence problem (2.8). Indeed, as in the proof of Theorem 2.6, if p is the PDF of an optimal mechanism that is monotone but not necessarily spherically-symmetric, then constructing the PDF

$$q(z) := \int_{\mathrm{O}(m)} p(Uz)\, d\mu(z) \tag{2.86}$$

(where μ the Haar measure over the orthogonal group $\mathrm{O}(m)$, see equation (2.52)) we see that q is the PDF of a monotone spherically-symmetric mechanism that performs at least as well as p for the problem (2.8) (hence, optimally among monotone mechanisms). In the sequel, we will denote the optimal value achievable by a monotone mechanism for the problem (2.8) by $\mathrm{KL}^\star_{\mathrm{monotone}}$.

2.13.2 Computing the Isotropic Mechanism

We show next that the isotropic mechanism in Definition 2.5 can be found via a simple finite-dimensional convex optimization problem. For each $(n,N,r) \in \mathbb{N}^2 \times (0,1)$, let $\mathscr{F}_{n,N,r}$ denote the family of mechanisms

$$\mathscr{F}_{n,N,r} := \left\{ P_{n,r,\boldsymbol{p}} \ ; \ \boldsymbol{p} \in [0,\infty)^{N+1}, P_{n,r,\boldsymbol{p}}(\mathbb{R}^m) = 1 \right\}. \tag{2.87}$$

Denote also the optimal value

$$\mathrm{KL}^\star_{n,N,r}(C) := \inf_{\substack{P \in \mathscr{F}_{n,N,r} \\ \mathbb{E}_P[c] \leq C}} \ \sup_{\|a\| \leq 1} \ D(P \,\|\, T_a P). \tag{2.88}$$

Note that mechanisms $P_{n,r,\boldsymbol{p}}$ (for $\boldsymbol{p} \in [0,\infty)^{N+1}$) achieving $\mathrm{KL}^\star_{n,N,r}(C)$ are what we call isotropic mechanisms in Definition 2.5.

To state our next result more compactly, we introduce the following shorthands. For each $s, \rho, \theta \geq 0$, let $H(s,\rho,\theta)$ denote the area of the triangle with side lengths s, ρ, and θ, i.e., $H(s,\rho,\theta) = 0$ if there is no triangle with such side lengths, and otherwise

$$H(s,\rho,\theta) := \frac{1}{4}\sqrt{(s+\rho+\theta)(s+\rho-\theta)(s-\rho+\theta)(-s+\rho+\theta)}. \tag{2.89}$$

For each $i, j, n \in \mathbb{N}$, denote the constant

$$\gamma_{i,j,n} := \int_{\mathcal{J}_{i,n}} \int_{\mathcal{J}_{j,n}} \theta\rho \cdot H(1, \rho, \theta)^{m-3} \, d\theta \, d\rho. \tag{2.90}$$

Also, denote the constants

$$c_{i,n} := \int_{\|x\| \in \mathcal{J}_{i,n}} c(x) \, dx. \tag{2.91}$$

Denote the open balls

$$\mathcal{B}(\rho) := \{x \in \mathbb{R}^m \, : \, \|x\| < \rho\}. \tag{2.92}$$

For integers $i \geq 0$ and $n \geq 1$, denote the volume of the spherical shell

$$v_{i,n} = \lambda\left(\mathcal{B}\left(\frac{i+1}{n}\right) \setminus \mathcal{B}\left(\frac{i}{n}\right)\right). \tag{2.93}$$

Denote also the volume of the unit ball

$$V_m := \lambda\left(\mathcal{B}(1)\right) = \frac{\pi^{m/2}}{\Gamma\left(\frac{m}{2}+1\right)}. \tag{2.94}$$

The following result shows that the optimization (2.88) required to numerically construct our proposed isotropic mechanism (i.e., finding the vector $\boldsymbol{p} \in \mathbb{R}_+^{N+1}$ for a fixed choice of (n, N, r)) can be carried out as a finite-dimensional convex optimization problem.

Theorem 2.9. *The optimization* (2.88) *can be rewritten as*

$$\begin{aligned} \underset{\boldsymbol{p} \in (0,\infty)^{N+1}}{\text{minimize}} \quad & A_m \sum_{i,j \geq 0} \gamma_{i,j,n} \, p_i \log \frac{p_i}{p_j} && (2.95) \\ \text{subject to} \quad & \sum_{i \geq 0} p_i v_{i,n} = 1 && (2.96) \\ & \sum_{i \geq 0} p_i c_{i,n} \leq C, && (2.97) \end{aligned}$$

where $A_m = 2^{m-3}(m-1)V_{m-1}$ and $p_i = p_N r^{i-N}$ for $i > N$.

Proof. See Appendix A.7. □

2.13.3 Optimality of the Isotropic Mechanism

Finally, we prove optimality of our proposed mechanisms introduced in Definition 2.5 for the optimization problem (2.8) among monotone mechanisms (see Remark 2.12).

Theorem 2.10. *Suppose $c : \mathbb{R}^m \to \mathbb{R}$ satisfies Assumption 2.1, and suppose c is also continuous and that, for* *some $\alpha, \beta > 0$, $c(x) \sim \beta \|x\|^{\alpha}$ as $\|x\| \to \infty$. With the optimal value obtainable by isotropic mechanisms (as constructed in Definition 2.5) denoted by*

$$\mathrm{KL}^{\star}_{\text{isotropic}}(C) := \lim_{\theta \to 0^+} \inf_{(n,N,r) \in \mathbb{N}^2 \times (0,1)} \mathrm{KL}^{\star}_{n,N,r}(C + \theta), \tag{2.98}$$

we have the equality $\mathrm{KL}^{\star}_{\text{monotone}} = \mathrm{KL}^{\star}_{\text{isotropic}}$.

Proof. See Appendix A.8. □

2.14 Optimality of Schrödinger's Mechanism for Small Sensitivity

We return to the scalar-query case in this section (so $m = 1$), and we also focus here on the regime of small sensitivity ($s \to 0^+$). We will show optimality of the Schrödinger mechanisms introduced in Definition 2.6 in this regime. Further, we show the stronger result in Theorem 2.13 that the PDF of the Schrödinger mechanism is in fact the *unique global minimizer* of the Fisher information.

As we are consider varying sensitivity in this section, we make that explicit in the notation for ε. Thus, with $\varepsilon_{P^{\circ k}_{Y|X}}$ as defined in (2.28), we replace this notation by $\varepsilon_{p^{\circ k}, s}$ in this section, where p is the PDF of the independent noise Z in the additive mechanism $Y = X + Z$. A schematic for our approach is the following sequence of reductions:

$$\min_p \varepsilon_{p^{\circ k}} \leftrightsquigarrow \min_p \max_a D(p \,\|\, T_a p) \leftrightsquigarrow \operatorname*{argmin}_p I(p) \leftrightsquigarrow \mathcal{H}_{\theta c}(\sqrt{p}) = E\sqrt{p}, \ p > 0. \tag{2.99}$$

That is, we reduce the problem of minimizing the DP parameter ε to that of minimizing the maximal KL-divergence (thereby removing the composition number k), then to finding unique minimizers of Fisher information (thereby removing the shift a), and finally to solving the Schrödinger eigenproblem with positive eigenfunctions.

Note that in this section $\mathcal{P}$ denotes the set of all PDFs over $\mathbb{R}$.

2.14.1 Definition of Optimality

Our approach is based on the well-known (see, e.g., [Kul59, Section 2.6]) result that, under mild regularity conditions on a PDF p, one has the expansion

$$D(p \,\|\, T_a p) = \frac{a^2}{2} I(p) + o(a^2) \quad \text{as } a \to 0. \tag{2.100}$$

We restrict attention in this section to PDFs satisfying this expansion. We also impose a restriction on the variance of the information density so that the CLT applies to reduce the DP parameter ε to the KL-divergence. Thus, we restrict attention in this section to the following subset of PDFs on the real line.

Definition 2.8. Let $\mathcal{F} \subset \mathcal{P}$ be the subset of PDFs p on the real line that satisfy the expansion in (2.100) and which satisfy both $\sup_{|a|\le s} D(p \parallel T_a p) < \infty$ and $\sup_{|a|\le s} \mathrm{V}(p \parallel T_a p) < \infty$ for some $s > 0$.

The definition we use for optimality of a noise PDF for queries with small sensitivities is given below.

Definition 2.9. We say that a PDF $p \in \mathcal{F}$ is *optimal in the small-sensitivity regime* for the cost function c and the cost bound C if $\mathbb{E}_p[c] \le C$, and for every other PDF $q \in \mathcal{F}$ (i.e., $\lambda(\{p = q\}) = 0$) satisfying $\mathbb{E}_q[c] \le C$ there is a constant $s(q) > 0$ such that $0 < s < s(q)$ implies

$$\sup_{0<\delta<\frac{1}{2}} \lim_{n\to\infty} \frac{\varepsilon_{p^{\circ n},s}(\delta)}{\varepsilon_{q^{\circ n},s}(\delta)} < 1. \tag{2.101}$$

Remark 2.13. For the Gaussian density $\varphi^{\sigma}(x) = e^{-x^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$, we have $D(\varphi^{\sigma} \parallel T_a\varphi^{\sigma}) = a^2/(2\sigma^2)$. Thus, if one insists that the PDF p satisfies $D(p \parallel T_a p) \le D(\varphi^{\sigma} \parallel T_a\varphi^{\sigma})$ for all small a, then the mapping $a \mapsto D(p \parallel T_a p)$ is necessarily differentiable at $a = 0$ with vanishing derivative. In particular, one reasonably expects that desirable PDFs for the small-shift regime to satisfy the expansion (2.100).

2.14.2 From DP to KL-divergence

Specializing the composition result in Theorem 2.1 to additive continuous mechanisms, we immediately obtain the following asymptotic.

Theorem 2.11. *Fix a PDF $p \in \mathcal{P}$. Suppose that there is an $s > 0$ such that $\sup_{|a|\le s} D(p \parallel T_a p) < \infty$ and $\sup_{|a|\le s} \mathrm{V}(p \parallel T_a p) < \infty$. Then, for any $\delta \in (0, 1/2)$, we have the limit*

$$\lim_{n\to\infty} \frac{\varepsilon_{p^{\circ n},s}(\delta)}{n} = \sup_{|a|\le s} D(p \parallel T_a p). \tag{2.102}$$

Proof. We apply Theorem 2.1 for the mechanism $P_{Y|X}$ given by $Y = X + Z$, where Z is a continuous random variable that is independent of X and which has PDF p. From (2.35), we have the inequalities

$$n\left(\sup_{|a|\le s} D(p\|T_a p) - o(1)\right) \le \varepsilon_{p^{\circ n},s}(\delta) \le n \cdot \sup_{|a|\le s} D(p\|T_a p) + \left(-\Phi^{-1}(\delta) + o(1)\right)\sqrt{n \cdot \sup_{|a|\le s} \mathrm{V}(p\|T_a p)}, \tag{2.103}$$

where $o(1)$ denotes a function that vanishes as $n \to \infty$. Dividing by n and taking $n \to \infty$, we obtain the desired limit (2.102). □

According to this asymptotic, characterizing $\varepsilon_{p^{\circ n},s}(\delta)$ for sufficiently large n boils down to computing the maximum of $D(p \parallel T_a p)$ over all $|a| \le s$.

2.14.3 From KL-Divergence to Fisher Information

In light of (2.100), another corollary of Theorem 2.1 is that the unique minimizer of the Fisher information is automatically the optimal PDF in the small-sensitivity regime.

Theorem 2.12. *If $p \in \mathcal{F}$ is the unique minimizer*

$$p = \operatorname*{argmin}_{\substack{q\in\mathcal{F} \\ \mathbb{E}_q[c]\le C}} I(q), \tag{2.104}$$

then p is the optimal PDF in the small-sensitivity regime for the cost function c and the cost bound C as per Definition 2.9.

Proof. Let $q \in \mathcal{F}$ be another PDF (i.e., $\lambda(\{p = q\}) = 0$) satisfying $\mathbb{E}_q[c] \le C$. Then, by assumption, $I(p) < I(q)$. Let $s_0 > 0$ be small enough so that the maximal KL-divergences and variances of the information densities of p and q from $T_a p$ and $T_a q$ over $|a| \le s_0$ are finite, the expansion (2.100) holds for p and q for $|a| \le s_0$, and the inequalities $D(p \parallel T_a p) \le \frac{a^2}{2} I(p) + \beta a^2$ and $D(q \parallel T_a q) \ge \frac{a^2}{2} I(q) - \frac{\beta}{2} a^2$ hold with the constant $\beta := (I(q) - I(p))/4$ for all $|a| \le s_0$. Fix $s \in (0, s_0)$. Let $\{a_k\}_{k\in\mathbb{N}} \subset [-s, s]$ be a convergent sequence so that $D(p \parallel T_{a_k} p) \to \sup_{|a|\le s} D(p \parallel T_a p)$. We may assume $\alpha \neq 0$ since the KL-divergence is nonnegative and expansion (2.100) implies that $D(p \parallel T_{a_k} p) \to 0$ if we had $a_k \to 0$. Denote $\alpha = \lim_{k\to\infty} a_k$. Then,

$$\sup_{|a|\le s} D(p \parallel T_a p) = \lim_{k\to\infty} D(p \parallel T_{a_k} p) \tag{2.105}$$

$$\le \lim_{k\to\infty} \frac{a_k^2}{2} I(p) + \beta a_k^2 \tag{2.106}$$

$$= \alpha^2 \cdot \left(\frac{I(p)}{2} + \beta\right) \tag{2.107}$$

$$
\begin{aligned}
&< \frac{\alpha^2}{2} I(q) - \frac{\beta}{2}\alpha^2 && (2.108)\\
&\leq D(q \parallel T_\alpha q) && (2.109)\\
&\leq \sup_{|a|\leq s} \; D(q \parallel T_a q) && (2.110)
\end{aligned}
$$

Finally, applying the asymptotic in Theorem 2.11 on both p and q, we obtain that, for every $\delta \in (0, 1/2)$,

$$
\lim_{n\to\infty} \frac{\varepsilon_{p^{\circ n},s}(\delta)}{\varepsilon_{q^{\circ n},s}(\delta)} = \lim_{n\to\infty} \frac{\varepsilon_{p^{\circ n},s}(\delta)/n}{\varepsilon_{q^{\circ n},s}(\delta)/n} = \frac{\sup_{|a|\leq s} \; D(p \parallel T_a p)}{\sup_{|a|\leq s} \; D(q \parallel T_a q)} < 1. \tag{2.111}
$$

As this limit is independent of δ, it still holds after taking the supremum over $\delta \in (0, 1/2)$. Thus, the condition in (2.101) is satisfied for every $s \in (0, s_0)$, i.e., p is the optimal PDF in the small-sensitivity regime for the cost function c and the cost bound C as per Definition 2.9. □

We derive in the remainder of this section unique minimizers of Fisher information over *all* PDFs $\mathcal{P}$, then we also show that such minimizers in fact fall within the subset $\mathcal{F}$.

2.14.4 From Fisher Information to the Schrödinger Equation

Solving the Fisher information minimization problem reveals a bridge between differential privacy and the celebrated Schrödinger operator. This connection enables us to borrow tools from the rich theory of the Schrödinger equation and show that the global minimizers of Fisher information are fully characterized by the minimal-eigenvalue eigenfunctions of the Schrödinger operator (see Theorem 2.13) with the potential given by the cost function c. More specifically, the global minimizer of Fisher information is identical to the distribution of a particle that is subjected to an energy potential given by a positive multiple of c and is in the ground state.

We recall the definition of the Schrödinger operator and some of its known properties.

Definition 2.10 (Schrödinger operator, [BS91, Section 2.4])**.** Given a measurable $v : \mathbb{R} \to \mathbb{R}$, the Schrödinger operator $\mathcal{H}_v$ on $L^2(\mathbb{R})$ with potential v is defined as[5]

$$
\mathcal{H}_v(y) := -y'' + vy. \tag{2.112}
$$

We say $y \in L^2(\mathbb{R})$ is an eigenfunction of $\mathcal{H}_v$ if y is differentiable, y' is absolutely continuous, and

[5] One may define $\mathcal{H}_v$ initially on compactly-supported $\mathcal{C}^\infty$ functions, then show that its closure is self-adjoint if v satisfies mild conditions (see [BS91, Chapter 2, Theorem 1.1]). In particular, this extension goes through if v is nonnegative (and measurable).

there exists a constant E such that $\mathcal{H}_v(y) = Ey$ holds a.e.

The spectrum of $\mathcal{H}_v$ is discrete: if v is locally bounded and $\lim_{|x|\to\infty} v(x) = \infty$ then $L^2(\mathbb{R})$ has an orthonormal complete set consisting of eigenfunctions of $\mathcal{H}_v$ with eigenvalues $\{E_k\}_{k\in\mathbb{N}}$ such that $E_k \to \infty$ (see [BS91, Chapter 2, Theorem 3.1]). Moreover, one may order the E_k in an increasing fashion, and then the eigenfunction associated to E_k has exactly k zeros (see [BS91, Chapter 2, Theorem 3.5]). We are interested in the smallest eigenvalue E_0 and the associated eigenfunction, i.e., the ground-state eigenfunction.

Lemma 2.2. *For any $\theta > 0$, there exists a unique unit-L^2-norm eigenfunction $y_{\theta,c}$ of $\mathcal{H}_{\theta c}$ satisfying $y_{\theta,c}(x) > 0$ for all $x \in \mathbb{R}$. Further, $y_{\theta,c}$ is even, and its eigenvalue is the smallest eigenvalue of $\mathcal{H}_{\theta c}$.*

Proof. See Appendix A.9.1. □

Remark 2.14. This lemma validates the claim in Remark 2.9 that there is a unique E for which the ODE (2.72) is uniquely solvable with the prescribed properties for the solution y in Definition 2.6, and that then E is the minimal eigenvalue of the Schrödinger operator $\mathcal{H}_{\theta c}$. The notation $y_{\theta,c}$ as given by Lemma 2.2 will be used in the sequel.

The notation $y_{\theta,c}$ as given by Lemma 2.2 will be used in the sequel. In fact, as per Definition 2.6, the PDF of the Schrödinger mechanism we introduce herein is exactly $p^\star_{c,C} = y^2_{\theta,c}$, where $C = \mathbb{E}_{y^2_{\theta,c}}[c]$.

Recall the recipe we provide in Theorems 2.11–2.12 for showing that the Schrödinger PDF $p^\star_{c,C}$ is the unique optimal DP mechanisms in the small-sensitivity regime (as per Definition 2.9):

1. First, show that $p^\star_{c,C}$ globally minimizes Fisher information (i.e., over $\mathcal{P}$);
2. Then, show that the global minimizer $p^\star_{c,C}$ in fact falls within $\mathcal{F}$;
3. Finally, use Theorem 2.12 to conclude that the Fisher information global minimizer $p^\star_{c,C}$ (i.e., the Schrödinger mechanism) is the optimal DP mechanism in the small-sensitivity regime.

We carry out step 1 in Theorem 2.13 below, where we show that $p^\star_{c,C} = y^2_{\theta,c}$ is the unique global minimizer of the Fisher information. After that, we complete our general derivations in Proposition 2.3 by showing that step 2 holds, i.e., $p^\star_{c,C} = y^2_{\theta,c} \in \mathcal{F}$.

Theorem 2.13. *Suppose c satisfies Assumption 2.2, fix $\theta > 0$, set $C = \mathbb{E}_{y^2_{\theta,c}}[c]$, and consider the PDF $p^\star_{c,C} = y^2_{\theta,c}$. Then, the PDF $p^\star_{c,C}$ uniquely minimizes the Fisher information among all PDFs $p \in \mathcal{P}$ that* *satisfy* $\mathbb{E}_p[c] \leq C$*, i.e.,*

$$p^\star_{c,C} = \operatorname*{argmin}_{\substack{p \in \mathcal{P} \\ \mathbb{E}_p[c] \leq C}} I(p). \tag{2.113}$$

Proof. See Appendix A.10. □

Since Theorem 2.13 gives a general unconditional result, our work can be seen as a way to fill the gaps in [FS18, FS19, Ern17, HR09]. Later in this section, we also provide a new *explicit* solution for the absolute-value cost case. Our method of proof deviates from those in [FS18, FS19, Ern17, HR09], where we borrow results from the quantum mechanics literature (such as [BS91]) to show that the needed properties for p can be *derived* instead of assumed. For instance, we show that the unique eigenfunction $y_{\theta,c}$ as given by Lemma 2.2 satisfies the following bound.

Proposition 2.2. *For c satisfying Assumption 2.2 and any* $\theta > 0$*, we have the bound*

$$\limsup_{|x|\to\infty} \left| \frac{y'_{\theta,c}(x)}{y_{\theta,c}(x)\sqrt{c(x)}} \right| \leq \sqrt{\theta}. \tag{2.114}$$

Proof. See Appendix A.9.2. □

Finally, we show in the following result that the PDF $y^2_{\theta,c}$ falls within the set $\mathcal{F}$ introduced in Definition 2.8.

Proposition 2.3. *For any c satisfying Assumption 2.2 and any* $\theta > 0$*, we have that* $y^2_{\theta,c} \in \mathcal{F}$*.*

Proof. See Appendix A.11. □

Next, we combine Theorems 2.11–2.13 and Proposition 2.3 to show in Theorem 2.14 that the PDF $y^2_{\theta,c}$ is the optimal DP mechanism in the sense of Definition 2.9.

Combining Theorem 2.13, Proposition 2.3, and Corollary 2.12, we get that the Schrödinger mechanism is optimal in the small-sensitivity regime.

Theorem 2.14. *If the cost function c satisfies Assumption 2.2, then the Schrödinger mechanism (see Definition 2.6) is optimal in the small-sensitivity regime in the sense of Definition 2.9.*

Proof. The Schrödinger PDF $p^\star_{c,C}$ uniquely minimizes the Fisher information by Theorem 2.13, and it belongs to $\mathcal{F}$ by Proposition 2.3. Hence, by Theorem 2.12, $p^\star_{c,C}$ is optimal in the small-sensitivity regime. □

Remark 2.15. For the two examples we discuss next, we give a reversing procedure producing θ given C that takes the form $\theta = aC^{-b}$ for absolute constants a and b.

2.14.5 From the Schrödinger Equation to the Gaussian and Airy Mechanisms

Next, we instantiate Theorem 2.14 for two different cost functions, namely the quadratic and absolute-value cost functions.

Consider first the quadratic cost function $c(x) = x^2$. By particularizing Theorem 2.14 to this case, we show that the Gaussian distribution is optimal in the small-sensitivity regime in the sense of Definition 2.9. This is a direct consequence of the Cramér-Rao bound, but we derive it here using Theorem 2.14. The Schrödinger to be solved becomes

$$y''(x) = \left(\theta x^2 - E\right) y(x). \tag{2.115}$$

Proposition 2.4. *Let $c(x) = x^2$. For any $C > 0$, we have*

$$p^\star_{c,C}(x) = \frac{1}{\sqrt{2\pi C}} e^{-x^2/(2C)}, \tag{2.116}$$

i.e., the Gaussian distribution is optimal in the small-sensitivity regime under a variance cost in the sense of Definition 2.9.

Proof. See Appendix A.12.1. □

We next consider the absolute value cost function $c(x) = |x|$. In this case, the eigenvalue problem $\mathcal{H}_{\theta c}(y) = Ey$ becomes

$$y''(x) = (\theta|x| - E)\, y(x), \tag{2.117}$$

for some $\theta > 0$. It is useful to recall the definition of the Airy functions we give in Section 2.11.3. In particular, Ai is a solution to the ODE $y''(x) = xy(x)$ satisfying $\mathsf{Ai}(x) \to 0$ as $x \to \infty$ and which is explicitly given by (2.75). Recall that we define the Airy distribution in Definition 2.7 as

$$p_{\mathsf{Ai},C}(x) := \frac{1}{3C\mathsf{Ai}(a_1')^2} \mathsf{Ai}\left(\frac{-2a_1'}{3C}|x| + a_1'\right)^2, \tag{2.118}$$

where $a_1' < 0$ is the zero of Ai' closest to the origin.

The following lemma verifies the claim in Remark 2.10 that $p_{\mathsf{Ai},C}$ is a valid PDF having first absolute moment C, and it also computes its variance.

Lemma 2.3. *With $p_{\mathsf{Ai},C}$ as in Definition 2.7, we have that*

$$\int_{\mathbb{R}} p_{\mathsf{Ai},C} = 1, \quad \int_{\mathbb{R}} |x| p_{\mathsf{Ai},C}(x)\, dx = C, \quad \textit{and} \quad \int_{\mathbb{R}} x^2 p_{\mathsf{Ai},C}(x)\, dx = \frac{9}{4}\left(\frac{8}{15} + \frac{1}{5\cdot(-a_1')^3}\right)\cdot C^2. \tag{2.119}$$

Proof. See Appendix A.12.2. □

We show in Proposition 2.5 below that the Airy distribution solves the ODE (2.117). Hence, per Theorem 2.13, the Airy distribution uniquely minimizes the Fisher information subject to an absolute value cost. If the cost bound is set to $C = 1$, then we obtain the minimal value

$$I(p_{\mathsf{Ai},C}) \approx 0.6266 < 1 = I(q), \tag{2.120}$$

where $q(x) := e^{-|x|}/2$ is the Laplace distribution.

Proposition 2.5. *Let $c(x) = |x|$. For any $C > 0$, we have that*

$$p^{\star}_{c,C} = p_{\mathsf{Ai},C}, \tag{2.121}$$

i.e., the Airy distribution is optimal in the small-sensitivity regime for the absolute-value cost constraint in the sense of Definition 2.9.

Proof. See Appendix A.12.3. □

2.15 Numerical Comparison with the Gaussian and Laplace Mechanisms

We apply state-of-the-art accounting methods and privacy-amplification techniques to simulate a real-world application for the proposed vector-valued isotropic mechanism introduced in Definition 2.5 (see also Theorem 2.9). In particular, we subsample our mechanism, following standard practice in the DP machine learning literature for amplifying privacy guarantees [KLN$^+$11, BNS13, ACG$^+$16]. Moreover, we use the arbitrary-accuracy FFT-based numerical accountant introduced in [GLW21] to compute tight privacy bounds for finite compositions.[6] To find the isotropic mechanism, we solve the optimization problem in Theorem 2.9 using an interior-point method.

In Figure 2.4, we fix $\delta = 10^{-8}$ and compute ε under a varying number of compositions. Under this setup, the accountant in [GLW21] computes both upper and lower bounds on ε. This accountant allows one to set the additive error in ε and δ via the parameters $\varepsilon_{\text{error}}, \delta_{\text{error}}$. We choose $\varepsilon_{\text{error}} = 0.002$ and $\delta_{\text{error}} = 10^{-10}$, effectively making the upper and lower bounds indistinguishable (they are both plotted in Figure 2.4). We compare the resulting privacy curve for the proposed mechanism with

[6] This accountant uses the privacy loss random variable (PLRV) formulation of DP to compute the privacy parameters. We note that the PLRV formulation is a lossless reparameterization of the (ε, δ)-curve [SMM19], hence it can applied to our proposed mechanism.

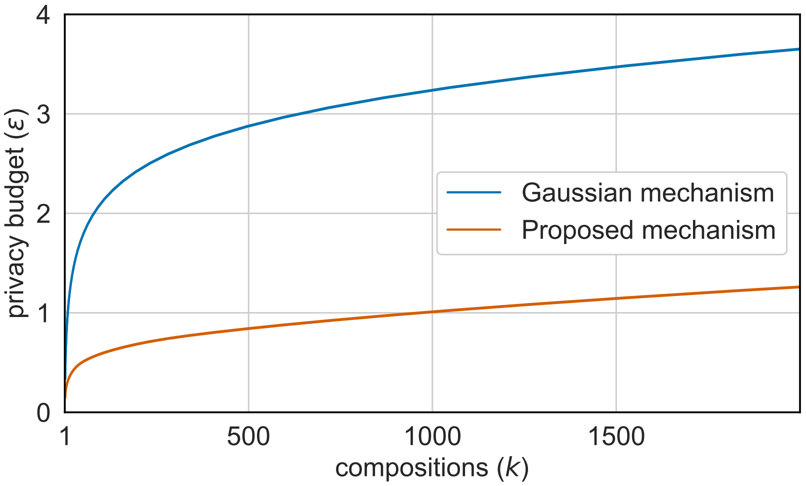


Figure 2.4: *Privacy budget ε versus number of the compositions, for $\mathbb{E}[\|Z\|^2] = 2.5$ (corresponding to $\sigma = 0.5$ for $\mathcal{N}(0, \sigma^2 I_{10})$), $\delta = 10^{-8}$, and subsampling rate $q = 0.001$. The proposed mechanism has 10 dimensions, and its construction parameters are $(n, N, r) = (400, 1200, 0.9)$, whereas its vector $\boldsymbol{p} \in (0, \infty)^{N+1}$ is computed numerically with the aid of Theorem 2.9.*

that of the subsampled Gaussian mechanism, for the same dimension $m = 10$ and variance cost $\mathbb{E}[\|Z\|^2] = 2.5$. Our proposed mechanism provides stronger privacy guarantees for all values of compositions $1 \leq k \leq 2000$.

We note that the proposed isotropic mechanism is not optimized for subsampling, though our numerical results imply that it still outperforms the subsampled Gaussian. An interesting future direction is to modify the optimization (2.80) (and Theorem 2.9) to explicitly optimize for subsampling, hence yielding a mechanism with better privacy guarantees than the one in Figure 2.4.

Next, we demonstrate that the Airy mechanism can achieve better DP parameters than the Laplace mechanism for the same fixed absolute-value cost in the small-sensitivity regime. In Figure 2.5, we fix $\delta = 10^{-8}$ and estimate ε under a varying number of compositions. We still use the accountant in [GLW21] with $\varepsilon_{\text{error}} = 0.002$ and $\delta_{\text{error}} = 10^{-10}$, effectively making the upper and lower bounds indistinguishable (they are both plotted in Figure 2.5). The Airy mechanism provides stronger privacy guarantees for all values of compositions ($1 \leq n \leq 2000$), and the gap increases with composition.

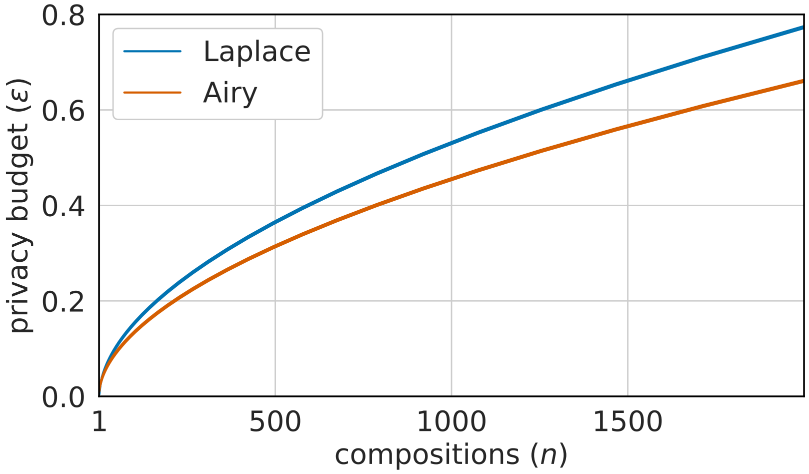


Figure 2.5: *The privacy budget ε versus the number of the compositions n, for the constraint $C = \mathbb{E}[|X|] = 2$, $s = 1$, fixed privacy parameter $\delta = 10^{-8}$, and subsampling rate $q = 0.01$.*

2.16 The Saddle-Point Accountant in the Large-Composition Regime

We now turn our attention to the DP accounting problem in the large-composition regime. We first illustrate the properties that the saddle-point accountant (SPA) enjoys using the experiment in Figure 2.6, which shows a comparison between the SPA and the state-of-the-art (SOTA) DP accountants when computing the (ε, δ) curve of a composition of 3000 subsampled Gaussian mechanisms. Only the moments accountant and the SPA are able to trace the whole privacy curve (see for example the region $\delta > 10^{-15}$). Further, the SPA upper and lower bounds have a narrow gap between them.

The SPA combines large-deviation and central-limit approaches for bounding expectations of sums of independent random variables, thereby attaining the best of both worlds. The large deviation approach uses the moment-generating function to approximate the probability of very unlikely events. The central limit theorem (CLT) approximates a random variable by a Gaussian with the same mean and variance. For DP accounting, the large deviation approach led to the moments accountant [ACG$^+$16]; the CLT approach led to Gaussian-DP [SMM19, DRS22]. Both these accountant methods can be computed in constant time, but their accuracy is far less than the SOTA FFT accountant [GLW21]. The saddle-point method can be viewed as a combination of two basic approaches: maintaining from large deviations the ability to handle very small values of δ, as well as the precise guarantees of the CLT. The resulting SPA achieves better accuracy than either approach

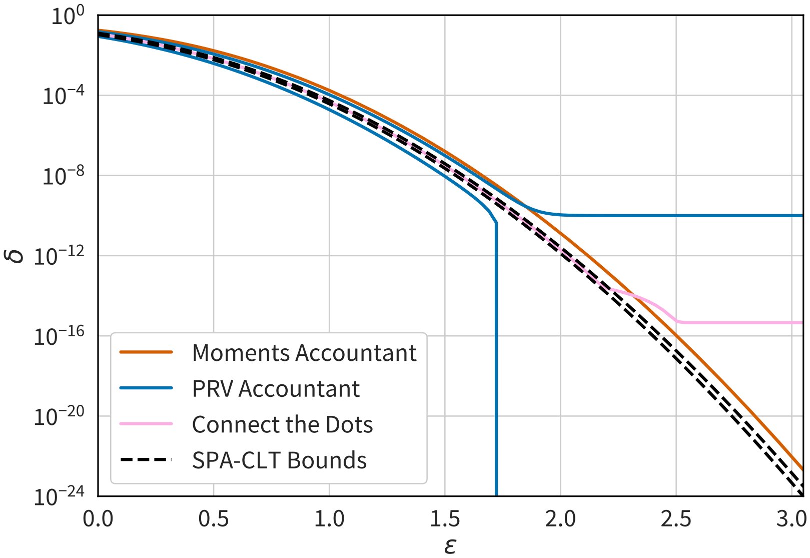


Figure 2.6: *Accounting for the composition of* 3000 *subsampled Gaussian mechanisms, with noise scale* $\sigma = 2$ *and subsampling rate* $\lambda = 0.01$*. The remaining FFT discretization parameters are set to the smallest that appear in their respective works, i.e.,* $\varepsilon_{\text{error}} = 0.1, \delta_{\text{error}} = 10^{-10}$ *for the PRV Accountant [GLW21], and discretization interval length of 0.005 for Connect the Dots [DGK+22].*

on its own, while maintaining the optimality of the runtime complexity.

The SPA works by estimating expectations of the privacy-loss random variable (PLRV), whose definition is recalled below.

Definition 2.11 (Privacy-Loss Random Variable (PLRV) [ZDW22])**.** A pair of probability measures (P, Q) is called a *dominating pair* for a mechanism (i.e., randomized algorithm) $\mathcal{M}$ if, for every $\varepsilon \geq 0$, event E, and neighboring datasets $D \simeq D'$, the following inequality holds:

$$\mathbb{P}\left[\mathcal{M}(D) \in E\right] - e^{\varepsilon}\, \mathbb{P}\left[\mathcal{M}(D') \in E\right] \leq P(E) - e^{\varepsilon} Q(E). \tag{2.122}$$

If (2.122) is tight, i.e., if

$$\sup_{D \simeq D'} \mathbb{P}\left[\mathcal{M}(D) \in E\right] - e^{\varepsilon}\, \mathbb{P}\left[\mathcal{M}(D') \in E\right] = P(E) - e^{\varepsilon} Q(E) \tag{2.123}$$

for each fixed $\varepsilon \geq 0$, then (P, Q) is said to be a *tightly dominating pair*. For any dominating pair (P, Q) consisting of equivalent measures, we associate a *privacy-loss random variable* (PLRV) that is defined as

$$L := \log \frac{dP}{dQ}(X), \qquad X \sim P. \tag{2.124}$$

It is not hard to see that a mechanism $\mathcal{M}$ having PLRV L will satisfy $(\varepsilon, \delta_L(\varepsilon))$-DP for every $\varepsilon \geq 0$, where we define the *privacy curve* (with $a^+ := \max(0, a)$)

$$\delta_L(\varepsilon) := \mathbb{E}\left[\left(1 - e^{\varepsilon - L}\right)^+\right]. \tag{2.125}$$

2.16.1 A Brief Overview of the Saddle-Point Accountant

Suppose that a DP mechanism has a privacy loss random variable whose cumulant-generating function $K(t)$ is finite for positive values of t. Note that $K(t)$ is a familiar quantity used in DP accounting; for instance, it can be verified that the mechanism satisfies exactly $(t+1, K(t)/t)$-Rényi-DP for each $t > 0$ [Mir17]. The SPA performs the following steps to estimate δ given ε:

1) Set $F(t) := K(t) - \varepsilon t - \log t - \log(t+1)$,

2) solve $F'(t) = 0$ over $t > 0$,

3) return $\delta(\varepsilon) \approx e^{F(t)} / \sqrt{2\pi F''(t)}$.

From this general workflow, it is clear that the SPA runs in *constant time* for n-fold self-composition; indeed, the cumulant-generating function for the composition is nK. Moreover, the root-finding in step 2 is similar to the one performed in the moments accountant [ACG⁺16], which solves $K'(t) - \varepsilon = 0$ instead.

We refer to the approximation returned by this simple procedure as SPA-MSD.[7] The reason SPA-MSD approximates the privacy curve well is the following three steps. First, we express the privacy curve as the following contour integral:

$$\delta(\varepsilon) = \frac{1}{2\pi i} \int_{t - i\infty}^{t + i\infty} e^{F(z)}\, dz, \tag{2.126}$$

which holds *independently* of the choice of $t > 0$. Second, we apply the method of steepest descent, which uses a judicious choice of the integration path in the complex plane: the line parallel to the imaginary axis with real part corresponding to the *saddle-point* of the integrand, i.e., the unique point $t > 0$ for which $F'(t) = 0$. This approach leads to a new series expansion for δ given a fixed ε (see (2.135)), where the first term of this series corresponding to the approximation in step 3 above.

Our experiments demonstrate that the SPA-MSD approximation is very accurate and can consistently achieve relative errors below 0.1% in ε for a fixed δ (see Figure 2.8). However, this approach *does not* provide a provable upper bound on the privacy curve—only an approximation. Consequently,

[7] MSD stands for "method of steepest descent."

we introduce another SPA, named SPA-CLT, where we first expand the K term in the integrand in (2.126) as an Edgeworth series [Hal13], then apply the Berry-Esseen theorem to prove upper and lower bounds on the privacy curve. This procedure is equivalent to applying CLT to a tilted version of the privacy loss random variable.

The SPA-CLT amounts to replacing step 3 above by a slightly different approximation given in Proposition 2.6. This second approximation also enjoys constant runtime, yields provable and accurate upper bounds for the privacy curve even for very small values of δ.

2.16.2 Subsampling and DP-SGD

In the context of differentially-private stochastic gradient descent (DP-SGD), one applies a DP mechanism on a subset of the dataset. The fraction of the batch size over the size of the dataset is called the *subsampling rate*, denoted by λ. Subsampling is known to amplify the privacy guarantees [BBG18]. In this setting, with $\mathcal{M}_\lambda$ denoting the subsampled mechanism, one should bound both orders $\mathsf{E}_{e^\varepsilon}(\mathcal{M}(D) \,\|\, \mathcal{M}_\lambda(D'))$ and $\mathsf{E}_{e^\varepsilon}(\mathcal{M}_\lambda(D) \,\|\, \mathcal{M}(D'))$ to obtain the value of δ. In the following lemma, we show that in fact one order dominates. See Appendix A.13 for the proof and further details on subsampling.

Lemma 2.4. *Fix a Borel probability measure P over $\mathbb{R}^n$ that is symmetric around the origin (i.e., $P(\mathcal{A}) = P(-\mathcal{A})$ for every Borel $\mathcal{A} \subset \mathbb{R}^n$), and fix constants $(s, \lambda, \gamma) \in \mathbb{R}^n \times [0,1] \times [1,\infty)$. Let T_sP be the probability measure given by $(T_sP)(\mathcal{A}) = P(\mathcal{A} - s)$, and let $Q = (1-\lambda)P + \lambda T_sP$. We have the inequality $\mathsf{E}_\gamma(P\|Q) \le \mathsf{E}_\gamma(Q\|P)$, with equality if and only if $(\gamma - 1)\,\lambda\,\|s\|\,\mathsf{E}_\gamma(Q\|P) = 0$.*

Proof. See Appendix A.13. □

2.16.3 The Method of Steepest Descent

We give a brief overview of the method of steepest descent (see Appendix A.14 for details). We need to compute

$$I_n = \frac{1}{2\pi i}\int_{t-i\infty}^{t+i\infty} e^{F_n(z)}\,dz \tag{2.127}$$

for a given F_n, provided that I_n is *independent* of the value of $t \in \mathbb{R}$. In a nutshell, the method of steepest descent is a powerful tool for choosing the best parameter t that renders the computation of I_n easiest. Namely, t is the *saddle-point* of F_n, defined as the unique solution to $F_n'(t_0) = 0$. Then, one would obtain the “asymptotic expansion”:

$$I_n \overset{\text{as. ex.}}{\sim} \frac{e^{F_n(t_0)}}{\sqrt{2\pi F_n''(t_0)}}\left(1+\sum_{m=2}^{\infty}\beta_{n,m}\right), \tag{2.128}$$

where we define the constants

$$\beta_{n,m} := \frac{(-1)^m B_{2m}(0,0,F_n^{(3)}(t_0),\ldots,F_n^{(2m)}(t_0))}{2^m m! F_n''(t_0)^m}. \tag{2.129}$$

Recall that this does not mean that the above equation holds for I_n with equality for any particular n. Rather, it is a heuristic indicating the potential for the truncated expansion to give close approximations for the intended integral I_n. In the remainder of this chapter, we will write $f_n \overset{\text{as. ex.}}{\sim} \sum_{k\in\mathbb{N}} a_{n,k}$ to indicate an asymptotic expansion, i.e., the series might not converge but the first few partial sums approximate f_n well.

In our application of the method of steepest descent to DP, we show in Theorem 2.15 that the privacy curve can be represented as the contour integral (2.127) for the choice of function

$$F_n(z) = K_{L^{(n)}}(z) - z\varepsilon - \log z - \log(1+z). \tag{2.130}$$

2.17 New Representations of the Privacy Curve

In Theorem 2.15, we derive two new formulas for the privacy curve. Then, we apply the method of steepest descent to the contour-integral formula (2.132). This yields the asymptotic expansion (2.135) of the privacy curve, which is the basis for the SPA-MSD as given by Definition 2.14. Later, in Section 2.19, we derive rigorous bounds on a CLT-based approximation that is inspired by the approximations in the present section.

We assume that we have access to a PLRV L for mechanism $\mathcal{M}$ (see Definition 2.11). In most cases, the relevant variable is $L^{(n)} = L_1 + \cdots + L_n$, such that $\delta_{L^{(n)}}$ is the composition curve for the adaptive composition $\mathcal{M}^{(n)} = \mathcal{M}_1 \circ \cdots \circ \mathcal{M}_n$ (and $L_1, \cdots, L_n$ are PLRVs for $\mathcal{M}_1, \cdots, \mathcal{M}_n$ that are independent). However, in this section we derive formulas for the privacy curve δ_L for any variable L. We note that for these formulas to be numerically computable, it suffices that the distribution of L be known to an extent that the derivatives of the MGF $M_L^{(k)}(t)$ can be computed.

2.17.1 The Privacy Curve as a Contour Integral

The privacy curve is defined in (2.125) as the expectation $\delta_L(\varepsilon) = \mathbb{E}[f(L)]$, where $f(x) = (1 - e^{\varepsilon - x})^+$. We want to transform this integral—via Parseval's identity—into the frequency domain. However, as $f \notin L^1(\mathbb{R})$, we cannot directly apply Parseval's identity. Nevertheless, exponentially tilting L, we may replace $f(x)$ by $e^{-tx}f(x)$, which decays fast. Recall that exponential tilting is defined as follows.

Definition 2.12. The *exponential tilting* with parameter $t \in \mathbb{R}$ of a random variable L having a finite MGF at t is the random variable $\widetilde{L}$ whose probability measure is given by $P_{\widetilde{L}}(B) := \frac{1}{M_L(t)} \int_B e^{tx}\, dP_L(x)$ for any Borel set B. If L has PDF p_L, then $\widetilde{L}$ is given by its PDF $p_{\widetilde{L}}(x) = e^{tx} p_L(x)/M_L(t)$.

We carry out the details of this idea in Appendix A.17 to obtain the following new formulas for δ_L.

Theorem 2.15. *If the PLRV L satisfies Assumption 2.5, then, for every $t > 0$, we may write the privacy curve as*

$$\delta_L(\varepsilon) = M_L(t)\, \mathbb{E}\left[e^{-t\widetilde{L}} \left(1 - e^{\varepsilon - \widetilde{L}}\right)^+ \right] \tag{2.131}$$

for all $\varepsilon \geq 0$, where $\widetilde{L}$ is the exponential tilting of L with parameter t (see Definition 2.12). If, in addition, L satisfies Assumption 2.6, then we also have the formula[8]

$$\delta_L(\varepsilon) = \frac{1}{2\pi i} \int_{t-i\infty}^{t+i\infty} e^{F_\varepsilon(z)}\, dz \tag{2.132}$$

for all $\varepsilon \geq 0$, where we define the exponent by[9]

$$F_\varepsilon(z) := K_L(z) - z\varepsilon - \log z - \log(1+z). \tag{2.133}$$

Proof. See Appendix A.17. □

The two formulas in (2.131)–(2.132) lead to two paths for approximating δ_L. The first is a direct application of the method of steepest descent, where F_ε is expanded around the saddle-point (see Section 2.16.3). The second simply approximates the expectation formula in (2.131) via the CLT, by replacing $\widetilde{L}$ with a Gaussian. The first path (described next) leads to better approximations numerically, but the second path is more amenable to an error analysis (see Section 2.19).

[8] The independence of formula (2.132) of t is not surprising, given Cauchy's integration theorem. More importantly, the theorem states that an integration path with real part t is actually equivalent to exponential tilting with parameter t.

[9] We use the principal branch of the complex logarithm, so F_ε is well defined and analytic over the half-plane $z \in (0, \infty) + i\mathbb{R}$.

2.17.2 The Privacy Curve in Terms of Bell Polynomials

As we have proved a formula in (2.132) for the privacy curve δ_L representing it as a contour integral like in (2.127), we can now apply the method of steepest descent to approximate it. Recall from Section 2.16.3 that the best choice for the real-axis intercept in (2.132) is the saddle-point.

Definition 2.13. The *saddle-point* associated with a PLRV L satisfying Assumption 2.5 and a privacy parameter ε satisfying $\varepsilon < \operatorname{ess\,sup} L$ is the unique $t_0 > 0$ such that $F_\varepsilon'(t_0) = 0$, or equivalently[10]

$$K_L'(t_0) = \varepsilon + \frac{1}{t_0} + \frac{1}{t_0+1}. \tag{2.134}$$

Remark 2.16. The original moments accountant aims to solve $K_L'(t) = \varepsilon$, indicating the connection between the moments accountant and the SPA, introduced formally in Section 2.17.3.

Applying the method of steepest descent to the contour integral in (2.132) with the choice of t being the saddle-point, we obtain the following asymptotic expansion for the privacy curve in terms of the derivatives of the MGF, connected via Bell polynomials (see Section 2.16.3).

Heuristic 2.1. *Let L be a PLRV satisfying Assumption 2.5. Then, for any $\varepsilon \in [0, \operatorname{ess\,sup} L)$, and with t_0 denoting the associated saddle-point, we have the asymptotic expansion*

$$\delta_L(\varepsilon) \overset{as.\ ex.}{\sim} \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F_\varepsilon''(t_0)}} \left(1 + \sum_{m=2}^{\infty} \beta_{\varepsilon,m}\right), \tag{2.135}$$

where, with $B_k(x_1, \ldots, x_k)$ denoting the k-th Bell polynomial and $F_\varepsilon^{(k)}$ the k-th derivative, we denote the constants

$$\beta_{\varepsilon,m} := \frac{(-1)^m B_{2m}(0, 0, F_\varepsilon^{(3)}(t_0), \ldots, F_\varepsilon^{(2m)}(t_0))}{2^m m! F_\varepsilon''(t_0)^m}. \tag{2.136}$$

Further, with $B_{k,j}(x_1, \cdots, x_k)$ denoting the (k, j)-th partial Bell polynomial, the derivatives of F_ε are[11] *(for $k \geq 2$)*

$$F_\varepsilon^{(k)}(t_0) = (-1)^{k-1}(k-1)! \left(\frac{1}{t_0^k} + \frac{1}{(t_0+1)^k}\right) + \sum_{j=1}^{k} \frac{(-1)^{j-1}(j-1)!}{M_L(t_0)^j} B_{k,j}(M'(t_0), \cdots, M^{(k)}(t_0)). \tag{2.137}$$

[10] For the well-definedness of the saddle-point, see Appendix A.16.

[11] The formula for $F_\varepsilon^{(k)}$ follows immediately by Faà di Bruno's formula for the derivatives of composition of functions.

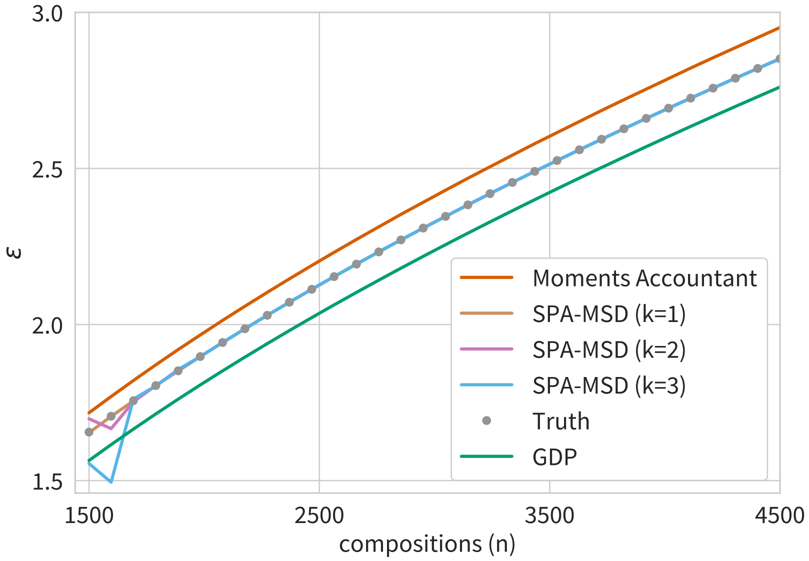


Figure 2.7: *Privacy budget ε of the subsampled Gaussian mechanism after $1500 \leq n \leq 4500$ compositions using the proposed SPA-MSD* (2.138) *and the other closed-form accountants. We use subsampling $\lambda = 0.01$, noise scale $\sigma = 2$, and $\delta = 10^{-15}$.*

2.17.3 Application: The Saddle-Point Accountant

Based on the asymptotic expansion in (2.135), we can derive various approximations of δ_L depending on how many terms we keep. This leads to the following versions of the *saddle-point accountant* (SPA).

Definition 2.14. The *order-k method-of-steepest-descent saddle-point accountant* (SPA-MSD) for the mechanism $\mathcal{M}$ with PLRV L satisfying Assumption 2.5 is defined by

$$\delta^{(k)}_{L,\text{SP-MSD}}(\varepsilon) := \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F''_\varepsilon(t_0)}} \left(1 + \sum_{m=2}^{k} \beta_{\varepsilon,m}\right) \tag{2.138}$$

when $\varepsilon < \operatorname{ess\,sup} L$, where $t_0 > 0$ is the saddle-point (i.e., $F'_\varepsilon(t_0) = 0$), and we set $\delta^{(k)}_{L,\text{SP-MSD}}(\varepsilon) = 0$ if $\varepsilon \geq \operatorname{ess\,sup} L$. Here, the $\beta_{\varepsilon,m}$ are as defined in (2.136).

The first SPA-MSD is $\delta^{(1)}_{L,\text{SP-MSD}}(\varepsilon) = e^{F_\varepsilon(t_0)} / \sqrt{2\pi F''_\varepsilon(t_0)}$, which can be expanded using the definition of F_ε as

$$\delta^{(1)}_{L,\text{SP-MSD}}(\varepsilon) = \frac{e^{K_L(t_0) - \varepsilon t_0}}{\sqrt{2\pi}\sqrt{t_0(t_0+1)K''_L(t_0) + t_0^2 + (1+t_0)^2}}. \tag{2.139}$$

The order-3 SPA-MSD is given by

$$\delta^{(3)}_{L,\text{SP-MSD}}(\varepsilon) = \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F''_\varepsilon(t_0)}} \cdot \left(1 + \frac{1}{8}\frac{F^{(4)}_\varepsilon(t_0)}{F''_\varepsilon(t_0)^2} - \frac{5}{24}\frac{F^{(3)}_\varepsilon(t_0)^2}{F''_\varepsilon(t_0)^3} - \frac{1}{48}\frac{F^{(6)}_\varepsilon(t_0)}{F''_\varepsilon(t_0)^3}\right) \tag{2.140}$$

and the order-2 SPA-MSD is obtained by keeping only the $1 + F^{(4)}_\varepsilon(t_0)/(8F''_\varepsilon(t_0)^2)$ term above.

Empirical Accuracy of SPA-MSD. The expressions for the SPA-MSD displayed in (2.139)–(2.140) can traverse privacy curves that are virtually indistinguishable from the ground-truth. We illustrate this in Figure 2.7 for the subsampled Gaussian, where we estimate ε (for fixed $\delta = 10^{-15}$) under a varying number of compositions. In this experiment, SPA-MSD improves on the other closed-form accountants (which run in constant time). Hence, SPA-MSD can be seen a correction to both the large deviation method and the CLT-based method found in the Moments Accountant and Gaussian-DP, respectively. See Appendix A.24 for the SPA-MSD pseudocode, Appendix A.25 for computing the ground-truth in Figure 2.7, and Appendix A.26 for more experiments.

2.18 Asymptotically Tight Composition Theorem

We show next that the lowest ε under composition cannot deviate from the mean of the PLRV by a large multiple of the standard deviation of the PLRV. This result is used afterwards to derive the asymptotic behavior of the saddle-point. The asymptotic behavior of the saddle-point, in turn, will be helpful in the next section to derive rigorous bounds on the SPA approximation error. We prove the following asymptotically tight DP composition theorem.

Theorem 2.16. *Let $\mathcal{M} = \mathcal{M}_1 \circ \cdots \circ \mathcal{M}_n$ have a PLRV $L = L_1 + \cdots + L_n$, where the L_j are PLRVs for the $\mathcal{M}_j$ that are independent. Assume that the L_j have finite absolute third moments, and $\mathrm{P}_0 = o(\sigma_L^3)$ as $n \to \infty$ (see (2.13)). Let $\delta \in (0,1/2)$ be such that $\limsup \delta < 1/2$ (so δ is allowed to vary with n). If $\sigma_L/(-\Phi^{-1}(\delta)) \to \infty$ as $n \to \infty$, then $\mathcal{M}$ is $(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L, \delta \cdot (1 + o(1)))$-DP. Conversely, this result is tight in the following sense. If $\delta_0 \in (0,1/2)$ is fixed, $\sigma_L \to \infty$, and $\mathcal{M}$ is $(\mathbb{E}[L] + b\sigma_L, \delta_0 + o(1))$-DP, then we must have $\liminf b \geq -\Phi^{-1}(\delta_0)$.*

Proof. See Appendix A.18. □

A more compact way to state the constant-δ claim in the theorem is that, for any fixed $\delta \in (0,1/2)$, we have

$$\delta_L(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L) \to \delta. \tag{2.141}$$

For example, Theorem 2.16 implies that $\delta_L(\varepsilon)$ is close to 10^{-10} if and only if ε is around $\mathbb{E}[L] + 6.4\sigma_L$ for all large n, since $-\Phi^{-1}(10^{-10}) \approx 6.4$. Thus, if one hopes to have a small value of δ, the only "interesting" values of ε, in the regime of high n, are those that are above $\mathbb{E}[L]$ by the derived multiple of σ_L.

2.18.1 Asymptotic Formula for the Saddle-Point

We re-parameterize $\varepsilon = \mathbb{E}[L] + b\sigma_L$, so b can be seen as the "Z-score" of ε, which is justified by Theorem 2.16. For this regime of values of ε, we prove the following asymptotic characterization of the saddle-point.

Theorem 2.17. *Let $L = L_1 + \cdots + L_n$ for independent L_j satisfying Assumption 2.5, and suppose that $(\mathbb{E}[L], \sigma_L^2) \sim n \cdot (\mathrm{KL}, \mathrm{V})$ for some constants $\mathrm{KL}, \mathrm{V} > 0$. Let $\varepsilon = \mathbb{E}[L] + b\sigma_L$, where $b > 0$ satisfies $b = o(n^{1/6})$, and assume that $\varepsilon < \operatorname{ess\,sup} L$. Then, the value of the saddle-point (as given by Definition 2.13) satisfies the asymptotic relation*

$$t_0 \sim \frac{b + \sqrt{b^2 + 4}}{2\sigma_L}. \tag{2.142}$$

Proof. See Appendix A.19. □

This asymptotic formula for the saddle-point will be useful in deriving the asymptotic rate of the approximation error of the SPA in the next section.

2.19 SPA Error Bound Analysis

While the approximations of Section 2.17 are often very precise (see Figure 2.7), they are merely *approximations*, and do not provide any hard guarantees on the (ε, δ)-DP of a given mechanism. In this section, we derive the alternative form of the SPA by applying the Berry-Esseen theorem to the saddle-point exponentially tilted PLRV, thereby obtaining upper and lower bounds on the achieved privacy parameters.

2.19.1 CLT Based Version of the SPA

We return to the expectation based formula for δ_L shown in Theorem 2.15, which can be rewritten as

$$\delta_L(\varepsilon) = e^{K_L(t)-\varepsilon t}\, \mathbb{E}\left[\bar{f}\left(\widetilde{L}-\varepsilon, t\right)\right], \tag{2.143}$$

where

$$\bar{f}(x,t) := e^{-xt}\left(1-e^{-x}\right)^{+}, \tag{2.144}$$

with $t > 0$ varying freely and $\widetilde{L}$ being the exponential tilting of L with parameter t. Here, $L = L_1 + \cdots + L_n$ for independent L_j satisfying Assumption 2.5. We will simply replace $\widetilde{L}$ by a Gaussian with the same first two moments,[12] and choose t to be the saddle-point of L as per Definition 2.13. Thus, we introduce the following version of the SPA.

Definition 2.15. Under Assumption 2.5, the *CLT version of the saddle-point accountant* (SPA-CLT) is defined by

$$\delta_{L,\text{SP-CLT}}(\varepsilon) := e^{K_L(t_0)-\varepsilon t_0}\, \mathbb{E}\left[\bar{f}(Z-\varepsilon, t_0)\right] \tag{2.145}$$

if $\varepsilon < \operatorname{ess\,sup} L$, where $Z \sim \mathcal{N}\left(K_L'(t_0), K_L''(t_0)\right)$, and t_0 is the saddle-point for L as given by Definition 2.13. We define $\delta_{L,\text{SP-CLT}}(\varepsilon) = 0$ for $\varepsilon \geq \operatorname{ess\,sup} L$.

Remark 2.17. The approach giving rise to $\delta_{L,\text{SP-CLT}}$ can be seen as a series expansion of the $e^{K_L(z)}$ part of the integrand in Theorem 2.15, or equivalently as an (order-0) Edgeworth expansion [Hal13] of the distribution of $\widetilde{L}$. However, the Edgeworth expansion approach delineated herein is different from what can be found in the DP literature [WGZ+22]. Specifically, we apply the Edgeworth expansion on the *tilted* random variable $\widetilde{L}$, whereas the approach of [WGZ+22] uses the Edgeworth expansion of the *non-tilted* version L. This distinction can yield very different approximations. We include a comparison between our approach and the standard CLT in Appendix A.20.

The following result expresses the CLT-based SPA in terms of easily computable functions. In what follows, we let $\delta_{L,\text{SP-CLT}}(\varepsilon; t)$ denote the same expression as in (2.145) but with t_0 replaced by a free $t > 0$, so in particular $\delta_{L,\text{SP-CLT}}(\varepsilon) = \delta_{L,\text{SP-CLT}}(\varepsilon; t_0)$.

Proposition 2.6. *Suppose Assumption 2.5 holds. Fix any $t > 0$ and $\varepsilon \in [0, \operatorname{ess\,sup} L)$, and denote*

$$\gamma := \frac{K_L'(t)-\varepsilon}{\sqrt{K_L''(t)}}, \quad \alpha := \sqrt{K_L''(t)}\; t - \gamma, \quad \beta := \sqrt{K_L''(t)}\; (t+1) - \gamma. \tag{2.146}$$

[12] It is not hard to see that the mean and variance of $\widetilde{L}$ are given by $\mathbb{E}[\widetilde{L}] = K_L'(t)$ and $\sigma_{\widetilde{L}}^2 = K_L''(t)$.

Then, we have that (with q as defined in (2.11))

$$\delta_{L,\text{SP-CLT}}(\varepsilon;t) = \frac{q(\alpha)-q(\beta)}{\sqrt{2\pi}} e^{K_L(t)-\varepsilon t-\gamma^2/2}. \tag{2.147}$$

Proof. See Appendix A.21.1. □

Remark 2.18. It holds that $0 < q(z) < \min(1/z, \sqrt{\pi/2})$ for all $z > 0$, and $q(z) \sim 1/z$ as $z \to \infty$ [NIS, Section 7.8].

While the two methods of approximation—the steepest descent as in Section 2.17.3, and the CLT approach in this section—lead to different approximations, these two approximations are closely related, as described by the following simple inequality.

Proposition 2.7. *Under Assumption 2.5, for any $t > 0$*

$$\delta_{L,\text{SP-CLT}}(\varepsilon;t) \le \frac{e^{F_\varepsilon(t)}}{\sqrt{2\pi K_L''(t)}}. \tag{2.148}$$

Proof. See Appendix A.21.2. □

Note that the only difference between the right-hand side of (2.148) and $\delta^{(1)}_{L,\text{SP-MSD}}(\varepsilon)$ is that the denominator involves K_L'' instead of F_ε''.

2.19.2 Finite-Composition Error Bound

Using the Berry-Esseen theorem, we prove the following theorem for the error bounds on the approximation $\delta_{L,\text{SP-CLT}}$ for arbitrary tilts.

Theorem 2.18. *Suppose Assumption 2.5 holds. For any $t > 0$ and $\varepsilon \ge 0$, there is a $\zeta \in [-1,1]$ such that*

$$\delta_L(\varepsilon) = e^{K_L(t)-\varepsilon t}\, \mathbb{E}\left[e^{-t(Z-\varepsilon)}\left(1-e^{-(Z-\varepsilon)}\right)^+\right] + \zeta\, \mathrm{err}_{\mathrm{SP}}(\varepsilon;t), \tag{2.149}$$

where $Z \sim \mathcal{N}(K_L'(t), K_L''(t))$ and the error is defined by

$$\mathrm{err}_{\mathrm{SP}}(\varepsilon;t) := e^{K_L(t)-\varepsilon t}\frac{t^t}{(1+t)^{1+t}} \cdot \frac{1.12\,\mathrm{P}_t}{K_L''(t)^{3/2}}. \tag{2.150}$$

Proof. See Appendix A.22. □

Note that omitting the ζ term in the right-hand side of (2.149) gives exactly $\delta_{L,\text{SP-CLT}}(\varepsilon;t)$ as per Definition 2.15. Thus, Theorem 2.18 can be equivalently restated as the following error bound for

SPA-CLT: for each $t > 0$ and $\varepsilon \geq 0$,

$$|\delta_L(\varepsilon) - \delta_{L,\text{SP-CLT}}(\varepsilon; t)| \leq \text{err}_{\text{SP}}(\varepsilon; t). \tag{2.151}$$

2.19.3 Asymptotic Error Rate

While Theorem 2.18 holds for any positive value of t around which the MGF is finite, a natural choice of t is the saddle-point t_0 itself, defined as the solution to (2.134). We analyze the ensuing error rate for this particular choice of tilt next.

Specifically, we show that the error rate in approximating δ_L by $\delta_{L,\text{SP-CLT}}$ decays roughly at least as fast as $1/(\sqrt{n}\, e^{b^2/2})$ for the choice $\varepsilon = \mathbb{E}[L] + b\sigma_L$, and we characterize the constant term too.

Theorem 2.19. *Let $L = L_1 + \cdots + L_n$ for independent PLRVs $L_1, \cdots, L_n$ that satisfy Assumption 2.5. Suppose that Assumption 2.7 holds too. Let $\varepsilon = \mathbb{E}[L] + b\sigma_L$ for $b > 0$ satisfying $b = o(n^{1/6})$, and let t_0 be the saddle-point of L (see Definition 2.13). Then, as $n \to \infty$, we have*

$$\text{err}_{\text{SP}}(\varepsilon; t_0) \sim \frac{1.12\sqrt{e}\,\mathrm{P}}{\mathrm{V}^{3/2} \cdot C(b)^{\tau} \cdot \sqrt{n}}, \tag{2.152}$$

where $\tau < 1$ satisfies $\tau \to 1$, and we define the term $C(b) := \exp\left((b^2 + b\sqrt{b^2+4})/4\right)$. Furthermore, writing $t_0 = \tau_0 \cdot \frac{b+\sqrt{b^2+4}}{2\sigma_L}$, we may take $\tau = (2-\tau_0)\tau_0$ in (2.152).

Proof. See Appendix A.23. □

Remark 2.19. In Appendix A.20, we illustrate the benefit of tilting the PLRV by comparing the error term in (2.152) with the corresponding standard CLT error (i.e., without tilting).

2.19.4 Relative-Error Comparisons

The SPA-CLT approximation (2.147) and its error bound (2.150) can approximate the privacy parameters accurately. In Figure 2.8, we plot the relative error[13] in estimating ε given $\delta = 10^{-15}$ incurred by SPA-CLT (both for the approximation in (2.147) and the approximation ± the error term (2.150)), SPA-MSD (for comparison), and the other closed-form accountants. The setting is for the subsampled Gaussian mechanism, with the same parameters as in Figure 2.7. Here, SPA improves on both the moments accountant and Gaussian-DP.

[13] We take the relative error of a privacy curve estimate $\hat{\varepsilon}(\delta)$, with a ground-truth of $\varepsilon(\delta)$, to be $|1 - \hat{\varepsilon}(\delta)/\varepsilon(\delta)|$.

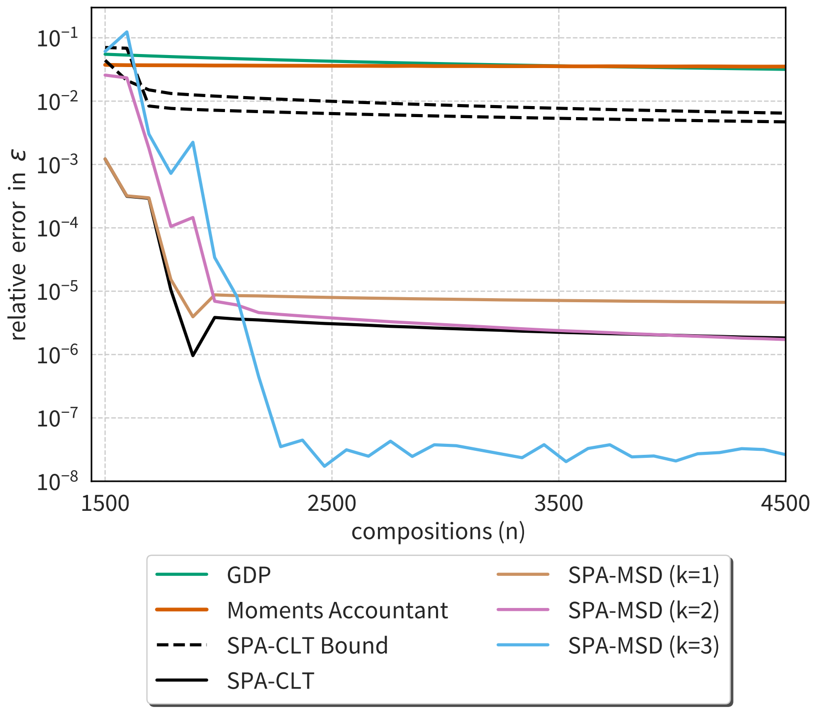


Figure 2.8: *Accounting for the privacy budget ε, given $\delta = 10^{-15}$, for the subsampled Gaussian mechanism, with subsampling rate $\lambda = 0.01$, and noise scale $\sigma = 2$. We plot the relative error in estimating ε (i.e., $|1 - \hat{\varepsilon}(\delta)/\varepsilon(\delta)|$ for an estimate $\hat{\varepsilon}$) versus the number of compositions, n. The SPA outperforms the other closed-form accountants for this experiment.*

2.20 Conclusion and Open Problems

We prove a large-composition theorem for DP, which reduces the ε DP parameter to a maximal KL-divergence term. We optimize the ensuing maximal KL-divergence, thereby obtaining optimized mechanisms for the large-composition regimes: the Cactus mechanism, the isotropic mechanism, and the Schrödinger mechanisms. We prove that these mechanisms perform arbitrarily close to optimal in their respective settings. We also show via numerical experiments that the proposed mechanisms outperform the subsampled Gaussian and Laplace mechanisms for finite-composition in terms of the DP parameters for the same cost constraint. It remains an interesting future line of work to refine our large-composition theorem, and to use these refinements in turn to optimize DP mechanisms for the large but fixed composition regime. Another intriguing line of work is comparing the accuracy-privacy trade-off curves of the proposed mechanisms and existing DP mechanisms in tasks such as DP stochastic gradient descent (DP-SGD).

We also introduce a novel application of the method of steepest descent in DP. First, using the exponentially-tilted version of the PLRV, we derive new formulas for the privacy curve (Theorem 2.15). Inspired by the method of steepest descent, we fix the exponential tilt to be the saddle-point of the integrand's exponent. This amounts to solving the 1-d equation $K_L'(t) = \varepsilon + 1/t + 1/(t+1)$. The ensuing closed-form formulas provide constant-runtime accurate approximations that can traverse the full privacy curve (e.g., for $\delta < 10^{-10}$). Our approach can be seen as a correction to both large-deviation methods (e.g., the moments accountant, via the additional $1/t + 1/(t+1)$ term) and CLT-based methods (e.g., Gaussian-DP, via preprocessing the PLRV with exponential tilting). This way, we retain the constant runtime of closed-form accountants without sacrificing accuracy, as demonstrated by our experiments. The saddle-point approach leaves a few questions open. The relative-error plot in Figure 2.8 indicates that, while the SPA-CLT bounds achieve reasonable relative error, the original approximation given by SPA-CLT and SPA-MSD seem to be several orders of magnitude more accurate than can be captured by the bounds we derive herein. Hence, it is an interesting future line of work to refine our bounds to further reveal the power of the saddle-point approximation. One promising path towards such a refinement might be through finding mechanism-specific bounds. Relatedly, such finer bounds would shed light on the question of "how large is large-enough n?" The additional experiments in Appendix A.23 show that n might only need to be of moderate size for the SPA to provide tight guarantees, yet a more complete answer requires additional techniques. Finally, it is interesting to use the SPA as a proxy for the privacy curve to optimize DP mechanisms.

Chapter 3

Optimal Multi-Class Fairness via Conditional Information Projection

Information projection [Che68, Csi75, CM03] is a fundamental formulation in several applications of information theory. Given a set of probability measures $\mathscr{C}$ and a reference measure P, a distribution $Q \in \mathscr{C}$ is said to be the *projection* of P onto $\mathscr{C}$ if it uniquely achieves the smallest KL-divergence $D_{\mathsf{kl}}(Q\|P)$ among all distributions in $\mathscr{C}$ [Csi75]. Both the minimizing distribution Q and the minimum divergence value are central quantities in large deviation theory [DZ96], universal source compression [YB17], hypothesis testing [Csi84], and beyond. Existence and uniqueness of the optimal distribution have been studied in [Csi75, CM03]. In particular, the optimal distribution has a simple closed-form given by an exponential tilting of the reference distribution P when the set $\mathscr{C}$ is determined by linear inequalities [Csi75]. Even though the information projection is most commonly defined with "distance" measured by the KL-divergence [Top79, Csi84, CM03, Bar00, Slo02, BC80], it has also been extended to Rényi divergences [AS16, KS15a, KS15b] and f-divergences [Csi95a, Csi95b].

We study a natural generalization of information projection: finding the "closest" *conditional* distribution (in a prescribed subset $\mathcal{F}$ of all possible conditional distributions) to a reference conditional distribution, where "distance" is measured by averaged (i.e., conditional) f-divergences. Motivated by applications in machine learning, we refer to this setting as *model projection*, since probabilistic classification models (e.g., logistic regression, neural networks with a softmax output layers) which map an input onto a probability distribution over predicted classes can be viewed as a conditional distribution. Analogous to the treatment of information projection, we start by proving the existence and uniqueness of the optimal conditional distribution. We then establish strong duality, which, in turn, leads to an equivalent formulation for obtaining the optimal conditional distribution. This dual formulation is easier to deal with since it converts an optimization with possibly infinitely many primal variables into a tractable, finite-dimensional optimization in Euclidean space. The optimal dual variables, in turn, allow the minimizing conditional distribution to be computed via a generalization of exponential "tilting": For a general f-divergence, one obtains the optimal conditional distribution by tilting the reference distribution by the inverse of the derivative of f. Naturally, this approach reduces to the usual exponential tilting when KL-divergence is the f-divergence of choice.

3.1 Application to Fair Machine Learning

Machine learning (ML) algorithms are increasingly used to automate decisions that have significant social consequences. This trend has led to a surge of research on designing and evaluating fairness interventions that prevent discrimination in ML models. When dealing with *group fairness*, fairness interventions aim to ensure that a ML model does not discriminate against different groups determined by, for example, race, sex, and/or nationality. Extensive comparisons between discrimination control methods can be found in [BDH$^+$18, FSV$^+$19, WRC21]. As these studies demonstrate, there is still no "best" fairness intervention for ML, and the majority of existing approaches are tailored to either binary classification tasks, binary population groups, or both.[1] Moreover, discrimination control methods are often tested on overused datasets of modest sizes collected in either the US or Europe (e.g., UCI Adult [Lic13] and COMPAS [ALMK16]).

Most fairness interventions in ML focus on binary outcomes. In this case, the classification output is either positive or negative, and group-fairness metrics are tailored to binary decisions [HPS$^+$16]. While binary classification covers a range of ML tasks of societal importance (e.g., whether to approve a loan, whether to admit a student), there are many cases where the predicted variable is not binary. For example, in education, grading algorithms assign one out of several grades to students. In healthcare, predicted outcomes are frequently not binary (e.g., severity of disease).

Using the model projection theory we put forth in this chapter, we introduce a theoretically-

[1] See Related Work and Table 3.1 for notable exceptions.

grounded discrimination control method called `FairProjection`. This method ensures group fairness in multi-class classification for several, potentially overlapping population groups. We consider group fairness metrics that are natural multi-class extensions of their binary classification counterparts, such as statistical parity [FFM+15], equalized odds [HPS+16], and error rate imbalance [PRW+17, Cho17]. When restricted to two predicted classes, `FairProjection` performs competitively against state-of-the-art fairness interventions tailored to binary classification tasks. `FairProjection` is model-agnostic (i.e., applicable to any model class) and scalable to datasets that are orders of magnitude larger than standard benchmarks found in the fair ML literature.

Prior work on information projection relies on a critical—and limiting—information-theoretic assumption: the underlying probability distributions are *known exactly*. This is infeasible in practical ML applications, where only a set of training examples sampled from the underlying data distribution is available. `FairProjection` fills this gap by using an efficient algorithm for computing the projected classifier with finite samples. We establish theoretical guarantees for this algorithm in terms of convergence and sample complexity.

Notably, our proposed fairness intervention is parallelizable (e.g., on a GPU). Hence, the algorithm `FairProjection` scales to datasets with the number of samples comparable to the population of many US states ($> 10^6$ samples). We provide a TensorFlow [AAB+15] implementation of `FairProjection` and apply it to post-process the outputs of probabilistic classifiers to ensure group fairness.

We benchmark our post-processing approach against several state-of-the-art fairness interventions selected based on the availability of reproducible code, and qualitatively compare it against many others. Our numerical results are among the most comprehensive comparisons of fairness interventions to date. We present performance results on the HSLS (High School Longitudinal Study, used in [JWC22]), Adult [Lic13], and COMPAS [ALMK16] datasets.

We also evaluate `FairProjection` on a dataset derived from open and anonymized data from Brazil's national high school exam—the *Exame Nacional do Ensino Médio* (ENEM)—with over 1 million samples. We made use of this dataset due to the need for large-scale benchmarks for evaluating fairness interventions in multi-class classification tasks. We also answer recent calls [BZZ+21, DHMS21] for moving away from overused datasets such as Adult [Lic13] and COMPAS [ALMK16]. We hope that the ENEM dataset encourages researchers in the field of fair ML to test their methods within broader contexts.[2]

In summary, our main contributions in the fairness intervention domain are: **(i)** We introduce a post-processing fairness intervention for multi-class classification problems that can account for multiple protected groups and is scalable to large datasets; **(ii)** We derive finite-sample guarantees and convergence-rate results for our post-processing method. Importantly, `FairProjection` makes information projection practical without requiring exact knowledge of probability distributions; **(iii)** We demonstrate the favourable performance of our approach through comprehensive benchmarks against state-of-the-art fairness interventions; **(iv)** We put forth a new large-scale dataset (ENEM) for benchmarking discrimination control methods in multi-class classification tasks; this dataset may encourage researchers in fair ML to evaluate their methods beyond Adult and COMPAS.

3.1.1 Related Fairness Intervention Methods

We summarize key differentiating factors from prior work in Table 3.1 and provide a more in-depth discussion in Appendix B.4.5. The fairness interventions that are the most similar to ours are the FairScoreTransformer [WRC20, WRC21, FST] and the pre-processing method in [JN20]. The FST and [JN20] can be viewed as instantiations of `FairProjection` when restricted to the binary classification setting and to cross-entropy (for FST) or KL-divergence (for [JN20]) as the f-divergence of choice. Thus, our approach is a generalization of both methods to multiple f-divergences. Importantly, unlike our method, [JN20] requires retraining a classifier multiple times.

A reductions approach for fair classification was introduced in [ABD$^+$18]. When restricted to binary classification, the benchmarks in Section 3.8 indicate that the reductions approach consistently achieves the most competitive fairness-accuracy trade-off compared to ours. `FairProjection` has two key differences from [ABD$^+$18]: it is not restricted to binary classification tasks and does not require refitting a classifier several times over the training dataset. These are also key differentiating points from [CHKV19], which presented a meta-algorithm for fair classification that accounts for multiple constraints and groups. The reductions approach was later significantly generalized in the GroupFair method by [YCK20] to account for overlapping groups and multiple predicted classes. Unlike [YCK20], we do not require retraining classifiers.

Several other recent fairness intervention methods consider optimizing accuracy under group-

[2] Since (to the best of our knowledge) the ENEM dataset has not been used in fair ML, we provide in Appendix B.5 a datasheet for the ENEM dataset. The data can be found at [INE20], and code for pre-processing the data and the implementation of `FairProjection` can be found at https://github.com/HsiangHsu/Fair-Projection.

Method			Feature				
	Multiclass	Multigroup	Scores	Curve	Parallel	Rate	Metric
Reductions [ABD^{+}18]	✗	✓	✓	✓	✗	✓	SP, (M)EO
Reject-option [KKZ12]	✗	✓	✗	✓	✗	✗	SP, (M)EO
EqOdds [HPS^{+}16]	✗	✓	✗	✗	✗	✓	EO
LevEqOpp [CDH^{+}19]	✗	✗	✗	✗	✗	✗	FNR
CalEqOdds [PRW^{+}17]	✗	✗	✓	✗	✗	✓	MEO
FACT [KCT20]	✗	✗	✗	✓	✗	✗	SP, (M)EO
Identifying[3] [JN20]	✓✗	✓	✓	✓	✗	✗	SP, (M)EO
FST [WRC20, WRC21]	✗	✓	✓	✓	✗	✓	SP, (M)EO
Overlapping [YCK20]	✓	✓	✓	✓	✗	✗	SP, (M)EO
Adversarial [ZLM18]	✓	✓	N/A[4]	✓	✓	✗	SP, (M)EO
`FairProjection` (ours)	✓	✓	✓	✓	✓	✓	SP, (M)EO

Table 3.1: *Comparison between benchmark methods.* ***Multiclass/multigroup****: implementation takes datasets with multiclass/multigroup labels;* ***Scores****: processes raw outputs of probabilistic classifiers;* ***Curve****: outputs fairness-accuracy tradeoff curves (instead of a single point);* ***Parallel****: parallel implementation (e.g., on GPU) is available;* ***Rate****: convergence rate or sample complexity guarantee is proved;* ***Metric****: applicable fairness metric, with SP↔Statistical Parity, EO↔Equalized Odds, MEO↔Mean EO. Since* `FairProjection` *is a post-processing method, we focus our comparison on post-processing fairness intervention methods, except for Reductions [ABD^{+}18], which is a representative in-processing method, and Adversarial [ZLM18], which we use to benchmark multi-class prediction. For comparing in-processing methods, see [LPB^{+}21, Table 1].*

fairness constraints. In [CJG^{+}19], a "proxy-Lagrangian" formulation was proposed for incorporating non-differentiable rate constraints, including group fairness constraints. We avoid non-differentiability issues by considering the probabilities (scores) at the output of the classifier instead of thresholded decisions. In [ZVRG17], a fairness-constrained optimization was introduced that is applicable to margin-based classifiers (our approach can be used on any probabilistic classifier). In [CDPF^{+}17] and [MW18], the fairness-accuracy trade-offs in binary classification tasks are characterized when the underlying distributions are known. A non-parity-based fairness notion was proposed in [KGZ19], called "multiaccuracy," which aims to ensure high accuracy for all subgroups even when the group information is not given in the data. We limit our analysis to parity notions of group fairness. To circumvent the non-differentiability of group-fairness constraints, approximate fairness constraints based on functionals found in information theory have been explored in [LPB^{+}21, Rényi mutual information], [BNBR19, Rényi maximal correlation], and [PQC^{+}19, maximum mean discrepancy]. We avoid such non-differentiability issues by casting group fairness constraints in the score domain.

3.2 Chapter Organization

We introduce model projection and compare it with information projection in Section 3.3. Explicit formulas for model projection are derived in Section 3.4. A finite-sample counterpart to model projection is introduced in Section 3.5. We introduce `FairProjection` in Section 3.6, which is an efficient procedure for computing model projection in practice. The connection between model projection and fair ML is explained in Section 3.7. Then, Numerical experiments are presented in Section 3.8.

3.2.1 Notation

Boldface Latin letters will always refer to vectors or matrices. The entries of a vector $\boldsymbol{z}$ are denoted by z_j, and those of a matrix $\boldsymbol{G}$ by $G_{i,j}$. The i-th row and j-th column of $\boldsymbol{G}$ are denoted by $\boldsymbol{G}_{i,:}$ and $\boldsymbol{G}_{:,j}$. The all-1 and all-0 vectors are denoted by $\mathbf{1}$ and $\mathbf{0}$. We set $[N] := \{1, \cdots, N\}$ and $\mathbb{R}_+ := [0, \infty)$. For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$, we write $\boldsymbol{a} \leq \boldsymbol{b}$ to indicate that $a_i \leq b_i$ for all $i \in [N]$. The probability simplex over $[N]$ is denoted by $\boldsymbol{\Delta}_N := \{\boldsymbol{p} \in \mathbb{R}_+^N \,;\, \mathbf{1}^T \boldsymbol{p} = 1\}$, and $\boldsymbol{\Delta}_N^+$ is its (relative) interior. The set of all probability measures definable on a general measurable space $(\mathcal{Y}, \Sigma)$ is denoted by $\boldsymbol{\Delta}_{\mathcal{Y}}$. If P is a Borel probability measure over $\mathbb{R}^N$, $Z \sim P$ is a random variable, and $f : \mathbb{R}^N \to \mathbb{R}^K$ is Borel, then the expectation of $f(Z)$ is denoted by $\mathbb{E}[f(Z)] = \mathbb{E}_P[f] = \mathbb{E}_P[f(Z)] = \mathbb{E}_{Z \sim P}[f(Z)]$. We use the standard asymptotic notations O, Θ, and Ω.

3.3 Model Projection Formulation

We recall the definition of information projection and some of its properties. Then we formally introduce model projection, which can be viewed as an extension of information projection. We prove the existence and uniqueness of the optimal model and establish strong duality in the next section.

[3] [JN20] mention that their method can be applied to multi-class classification, but their reported benchmarks are only for binary classification tasks.

[4] [ZLM18] is an in-processing method unlike other benchmarks in the table. It does not take a pre-trained classifier as an input.

3.3.1 Information Projection

For a given reference probability distribution and a set of distributions, information projection seeks to find the "closest" distribution within this set to the reference one. Fix a probability space (Ω, Σ, P). For any subset $\mathscr{C} \subset \mathbf{\Delta}_\Omega$, let

$$D_f(\mathscr{C} \| P) := \inf_{Q \in \mathscr{C}} D_f(Q \| P). \tag{3.1}$$

Here for a convex $f : (0, \infty) \to \mathbb{R}$ the f-divergence [AS66, Csi67] is given by

$$D_f(Q \| P) := \mathbb{E}_P \left[f\left(\frac{dQ}{dP} \right) \right] - f(1) \tag{3.2}$$

whenever Q is absolutely continuous with respect to (w.r.t.) P. We say that a $Q \in \mathscr{C}$ is the D_f-projection of P onto $\mathscr{C}$ if

$$D_f(Q \| P) = D_f(\mathscr{C} \| P) \tag{3.3}$$

and $D_f(R \| P) > D_f(\mathscr{C} \| P)$ whenever $Q \neq R \in \mathscr{C}$. The existence and uniqueness of the D_f-projection has been established under certain assumptions [Csi95a, Csi95b]. Furthermore, an explicit formula for the D_{KL}-projection (also termed I-projection) under linear constraints is proved [Csi67].

3.3.2 Model Projection: Problem Setup

We introduce next the definition of model projection.

Definition 3.1. Fix two measurable spaces $(\mathcal{X}, \Sigma)$ and $(\mathcal{Y}, \Gamma)$, a probability measure P_X on $(\mathcal{X}, \Sigma)$, a transition probability kernel $P_{Y|X}$ from $(\mathcal{X}, \Sigma)$ into $(\mathcal{Y}, \Gamma)$, and a set $\mathcal{F}$ of transition probability kernels from $(\mathcal{X}, \Sigma)$ into $(\mathcal{Y}, \Gamma)$. The *model projection* (MP) of $P_{Y|X}$ onto $\mathcal{F}$ is the unique solution (if it exists) to the minimization problem:

$$\inf_{Q_{Y|X} \in \mathcal{F}} D_f\left(Q_{Y|X} \,\|\, P_{Y|X} \mid P_X \right). \tag{3.4}$$

The model projection is the "closest" model to the prescribed model $P_{Y|X}$, where we use the f-divergence to measure the "closeness". The choice of the f-divergence is determined by the application at hand.

In what follows, let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = [C]$. In this setting, conditional distributions from $\mathcal{X}$ to $\mathcal{Y}$ become simply vector-valued functions. We denote the base model $P_{Y|X}$ by $\boldsymbol{h}^{\text{base}} : \mathcal{X} \to \mathbf{\Delta}_C$, i.e.,

$$\boldsymbol{h}^{\text{base}}(x) := (P_{Y|X}(1|x), \cdots, P_{Y|X}(C|x)), \qquad \text{for every } x \in \mathcal{X}. \tag{3.5}$$

An arbitrary conditional distribution from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by a vector-valued function $\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C$. Then, (3.4) becomes

$$\inf_{\boldsymbol{h} \in \mathcal{F}} \mathbb{E}_X \left[D_f \left(\boldsymbol{h}(X) \parallel \boldsymbol{h}^{\text{base}}(X) \right) \right]. \tag{3.6}$$

The choice of the constraint set $\mathcal{F}$ is usually application-dependent. Throughout this chapter, we consider a special case in which the constraint set is constructed via linear inequalities. In other words, for some given matrix-valued function $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{K \times C}$ the constraint set is in the form

$$\mathcal{F} = \{ \boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C \mid \mathbb{E}\left[\boldsymbol{G}(X)\boldsymbol{h}(X) \right] \le \boldsymbol{0} \}. \tag{3.7}$$

Note that the convex combination $\sum_{i=1}^{\ell} t_i \boldsymbol{h}^{(i)}$ is in $\mathcal{F}$ if each $\boldsymbol{h}^{(i)}$ is in $\mathcal{F}$ (and $\boldsymbol{t} \in \boldsymbol{\Delta}_\ell$). This linear form of the constraint set captures the following general setup, which encapsulates our fairness intervention application of model projection.

Lemma 3.1. *Let $X \sim P_X$, and let $\boldsymbol{g} : \mathcal{X} \times \mathcal{Y}^2 \to \mathbb{R}^K$ be such that $g(\,\cdot\,, c, c')$ is P_X-integrable for each fixed $(c, c') \in \mathcal{Y}^2$. Suppose we have a Markov chain Y–X–$\widehat{Y}$, where, for each $x \in \mathcal{X}$, we have $Y \mid X = x \sim \boldsymbol{h}^{\text{base}}(x)$ and $\widehat{Y} \mid X = x \sim \boldsymbol{h}(x)$. Then, the inequality $\mathbb{E}[\boldsymbol{g}(X, Y, \widehat{Y})] \le \boldsymbol{b}$ (for fixed $\boldsymbol{b} \in \mathbb{R}^K$) can be written in the linear form (3.7).*

Proof. Since Y and $\widehat{Y}$ are independent given X, we may write

$$\mathbb{E}[g_k(X, Y, \widehat{Y})] - b_k = \mathbb{E}\left[\mathbb{E}[g_k(X, Y, \widehat{Y}) \mid X] \right] - b_k = \mathbb{E}\left[\sum_{c, c' \in [C]} h_c^{\text{base}}(X) h_{c'}(X) g_k(X, c, c') \right] - b_k \tag{3.8}$$

$$= \mathbb{E}\left[\sum_{c' \in [C]} \left(-b_k + \sum_{c \in [C]} h_c^{\text{base}}(X) g_k(X, c, c') \right) h_{c'}(X) \right] = \mathbb{E}\left[\boldsymbol{G}_{k,:}(X) \boldsymbol{h}(X) \right], \tag{3.9}$$

where we define the matrix-valued function $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{K \times C}$ by

$$G_{k,c'}(x) = -b_k + \sum_{c \in [C]} h_c^{\text{base}}(x) g_k(x, c, c') \tag{3.10}$$

for each $(k, c', x) \in [K] \times [C] \times \mathcal{X}$. □

3.3.3 Connection between Information Projection and Model Projection

We connect model projection (3.4) with information projection (3.1) next. Keeping the notation before equation (3.1), suppose $\Omega = \mathcal{X} \times \mathcal{Y}$ and that $P_{X,Y} \in \boldsymbol{\Delta}_\Omega$ is a probability measure that disintegrates into P_X and $P_{Y|X}$. Let $\mathscr{P} \subset \boldsymbol{\Delta}_\Omega$ be the subset of all probability measures that marginalize to P_X on $\mathcal{X}$, i.e.,

$$\mathscr{P} := \{Q \in \mathbf{\Delta}_\Omega \mid Q(A \times \mathcal{Y}) = P_X(A) \text{ for all } A \times \mathcal{Y} \subset \Sigma\}.$$

Then the model projection (3.4) is information projection onto a subset of $\mathscr{P}$. In other words, for a set $\mathcal{F}$ of conditional distributions, the model projection of $P_{Y|X}$ onto $\mathcal{F}$ is exactly information projection of $P_{X,Y}$ onto

$$\mathscr{C} := \{P_X W_{Y|X} \mid W_{Y|X} \in \mathcal{F}\} \subset \mathscr{P}. \tag{3.11}$$

It is important to note that $\mathscr{P}$ cannot be described by finitely many linear constraints, precisely because a distribution may not be determined by finitely many of its moments. Hence, the results on information projection subject to finitely many linear constraints do not seem applicable to model projection.

On the other direction, observe that model projection subsumes information projection. This fact is rather trivial, since for a singleton $\mathcal{X} = \{x\}$ the set $\Omega = \mathcal{X} \times \mathcal{Y}$ can be identified with $\mathcal{Y}$ via $(x, y) \leftrightarrow y$. Then, P_X is a trivial atom $P_X = \delta_x$ (and $\mathscr{P} = \mathbf{\Delta}_\Omega$) so the averaging in (3.4) collapses into only one term, whose minimization is precisely the problem of information projection.

3.4 Model Projection Theory

In this section, we first prove the existence and uniqueness of the model projection onto a linear subset under the general f-divergence setting. For the information projection framework with f-divergence measuring "distance," this problem has been studied [Csi95a] under the condition $f'(0^+) = -\infty$ to ensure that the projection onto the linear set belongs to the interior of $\mathbf{\Delta}_C$. This condition also appears in our result. Then we compute the model projection by establishing strong duality for a functional optimization over the complete metric space $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_C)$ of continuous conditional distributions.[5]

3.4.1 Existence, Uniqueness, and Explicit Formulas

To start with, we introduce four assumptions, which will be the premises of our main theorems. These assumptions restrict the behavior of the f-divergence, the linear constraints (see (3.7)), the feasibility set, and the given conditional distribution $\boldsymbol{h}^{\text{base}}$, respectively. Our optimization is carried

[5] We endow $\mathcal{X} = \mathbb{R}^m$ with the standard topology, and $\mathbf{\Delta}_C \subset \mathbb{R}^C$ with the subspace topology, so continuity of $\boldsymbol{h} : \mathcal{X} \to \mathbf{\Delta}_C$ refers to the usual definition of continuous functions between Euclidean spaces. Then, endowing $\mathcal{C}(\mathcal{X}, \mathbf{\Delta}_c)$ with the sup-norm, $\|\boldsymbol{h}\|_\infty = \sup_{x \in \mathcal{X}} \|\boldsymbol{h}(x)\|$, turns it into a convex complete metric space.

over the "interior"

$$\mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C) := \left\{ \boldsymbol{h} \in \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C) \mid \inf_{(x,c)\in\mathcal{X}\times[C]} h_c(x) > 0 \right\}. \tag{3.12}$$

Assumption 3.1. *The functions $f : (0, \infty) \to \mathbb{R}$, $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{K\times C}$, and $\boldsymbol{h}^{\text{base}} : \mathcal{X} \to \boldsymbol{\Delta}_C$ satisfy the following:*

(a) f is twice continuously-differentiable, $f(1) = 0$, $f'(0^+) = -\infty$, and $f''(t) > 0$ for every $t > 0$,

(b) each function $G_{k,c} : \mathcal{X} \to \mathbb{R}$ is bounded and differentiable with bounded gradient,

(c) there exists at least one conditional distribution $\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$ satisfying $\mathbb{E}[\boldsymbol{G}\boldsymbol{h}] < \boldsymbol{0}$, and

(d) the function $\boldsymbol{h}^{\text{base}}$ belongs to $\mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$, and each h_c^{base} has bounded partial derivatives.

The following theorem guarantees the existence and uniqueness of the model projection.

Theorem 3.1. *Under Assumption 3.1, there exists a unique $\boldsymbol{h}^{\text{opt}} \in \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$ solving the model projection problem*

$$\begin{aligned} \min_{\boldsymbol{h}\in\mathcal{C}_+(\mathcal{X},\boldsymbol{\Delta}_C)} \quad & \mathbb{E}\left[D_f(\boldsymbol{h}(X) \parallel \boldsymbol{h}^{\text{base}}(X)) \right], \\ \text{s.t.} \quad & \mathbb{E}[\boldsymbol{G}\boldsymbol{h}] \leq \boldsymbol{0}. \end{aligned} \tag{3.13}$$

Proof. See Appendix B.1. □

In fact, the optimal model $\boldsymbol{h}^{\text{opt}}$ of Theorem 3.1 owns an explicit formula in terms of the convex conjugate of the f-divergence. Recall that the convex conjugate D_f^{conj} is defined as follows.

Definition 3.2. Fix a convex $f : (0, \infty) \to \mathbb{R}$ and $\boldsymbol{p} \in \boldsymbol{\Delta}_C$. The convex conjugate of the f-divergence $\boldsymbol{q} \mapsto D_f(\boldsymbol{q} \parallel \boldsymbol{p})$ is the function $\boldsymbol{v} \mapsto D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p})$ defined by the formula

$$D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}) := \sup_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} \boldsymbol{v}^T\boldsymbol{q} - D_f(\boldsymbol{q}\|\boldsymbol{p}) \tag{3.14}$$

at each $\boldsymbol{v} \in \mathbb{R}^C$.

The formula of the optimal model shows that the model projection onto a set constructed by linear constraints can be obtained by tilting the reference model, where the tilting is parametrized in terms of the function $\boldsymbol{v} : \mathcal{X} \times \mathbb{R}^K \to \mathbb{R}^C$ that we define by the matrix-vector multiplication

$$\boldsymbol{v}(x; \boldsymbol{\lambda}) := -\boldsymbol{G}(x)^T\boldsymbol{\lambda}. \tag{3.15}$$

Further, the tilting function is the inverse of f'; note that, under item (a) of Assumption 3.1, the derivative f' is strictly increasing, so one can define its inverse

$$\phi : (-\infty, M) \to (0, \infty) \tag{3.16}$$

by $\phi(u) := (f')^{-1}(u)$, where $M = \sup_{t>0} f'(t)$. We prove that model projection has the following formula.

Theorem 3.2. *Under Assumption 3.1, the model projection $\boldsymbol{h}^{\text{opt}}$ of the base model $\boldsymbol{h}^{\text{base}}$ (see Theorem 3.1) has the formula*

$$h_c^{\text{opt}}(x) = h_c^{\text{base}}(x)\, \phi(\gamma(x) + v_c(x; \boldsymbol{\lambda}^\star)), \qquad \textit{for every } (c, x) \in [C] \times \mathcal{X}, \tag{3.17}$$

where the function $\gamma : \mathcal{X} \to \mathbb{R}$ is uniquely defined by

$$\mathbb{E}_{c \sim \boldsymbol{h}^{\text{base}}(x)} \left[\phi(\gamma(x) + v_c(x; \boldsymbol{\lambda}^\star))\right] = 1, \qquad \textit{for every } x \in \mathcal{X}, \tag{3.18}$$

and $\boldsymbol{\lambda}^\star \in \mathbb{R}^K$ is any solution to the convex optimization problem

$$D^\star := \min_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \mathbb{E}\left[D_f^{\text{conj}}(\boldsymbol{v}(X; \boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X))\right]. \tag{3.19}$$

Proof. See Appendix B.1. □

Remark 3.1. If $\mathcal{X}$ is finite, then Theorems 3.1 and 3.2 hold without the differentiability assumptions on the $G_{k,c}$ and on the h_c^{base}.

The duality approach reduces the infinite-dimensional optimization (3.13) into a tractable finite-dimensional one (3.19). Note that in our setting, a simple application of duality is inaccessible. The primal optimization (3.13) is equivalent to

$$\inf_{\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)} \sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \mathbb{E}\left[D_f(\boldsymbol{h}(X) \parallel \boldsymbol{h}^{\text{base}}(X)) + \boldsymbol{\lambda}^T \boldsymbol{G}(X) \boldsymbol{h}(X)\right], \tag{3.20}$$

which is not necessarily equal to the dual optimization

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{\boldsymbol{h} \in \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)} \mathbb{E}\left[D_f(\boldsymbol{h}(X) \parallel \boldsymbol{h}^{\text{base}}(X)) + \boldsymbol{\lambda}^T \boldsymbol{G}(X) \boldsymbol{h}(X)\right]. \tag{3.21}$$

The difficulty here is that the space $\mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$ is not precompact when $\mathcal{X}$ is an infinite set. The minimax property does not hold in general if neither of the two optimization spaces is precompact. Our approach shows that, nevertheless, one may carve a precompact subset $\mathcal{Q} \subset \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$ that is guaranteed to contain the sought optimizer. Note that strict convexity of f implies that the unique solution of the inner minimization in the dual (3.21) at any outer maximizer $\boldsymbol{\lambda}^\star$ is in fact the unique solution to the primal problem (3.20) (i.e., it is the sought model projection of $\boldsymbol{h}^{\text{base}}$ onto $\mathcal{F} \cap \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$, see (3.7) and (3.12)). We take $\mathcal{Q}$ to be the collections of all such "potential" optimizers, where $\boldsymbol{\lambda}$ ranges over a bounded set; we find such bounded set that suffices, and we prove that $\mathcal{Q}$ is precompact, which allows us to use known minimax theorems to derive our formulas for model projection.

3.4.2 Comparison with the Information Projection Formula

Notably, for the KL-divergence, the model projection formula closely resembles that of the information projection. Analogous to the information projection formula under linear constraints, the model projection formula (3.17) for a fixed $x \in \mathcal{X}$ is an exponential tilt since for $f(t) = t \log t$ we have $\phi(u) = e^{u-1}$. The difference between the two projections is how the tilt is computed (i.e., in the value of the parameters $\boldsymbol{\lambda}^\star$) where its value under the model projection setting reflects the fact that we are penalizing the average distance. We need the following formula for the convex conjugate of the KL-divergence

$$D_{\mathsf{KL}}^{\mathsf{conj}}(\boldsymbol{v}, \boldsymbol{p}) = \log \sum_{c \in [C]} p_c e^{v_c}. \tag{3.22}$$

The optimal parameters $\boldsymbol{\lambda}^\star$ for the D_{KL}-projection is computed as follows. Consider a probability measure P defined on measurable space (Ω, Σ), a vector-valued function $\boldsymbol{g}$, and a linear set of probability measures

$$\mathscr{C} = \left\{ Q \in \boldsymbol{\Delta}_\Omega \mid \mathbb{E}_{Z \sim Q}\left[\boldsymbol{g}(Z)\right] \leq \mathbf{0} \right\}. \tag{3.23}$$

The probability measure $Q^\star$ that is the D_{kl}-projection of P onto $\mathscr{C}$ is given by the same formula (3.17), albeit with a slightly different $\boldsymbol{\lambda}^\star$. More precisely,

$$\frac{dQ^\star}{dP}(z) = e^{\gamma + \boldsymbol{g}(z)^T \boldsymbol{\lambda}^\star}, \tag{3.24}$$

where γ is a normalizing constant, and the optimal parameters $\boldsymbol{\lambda}^\star$ are exactly the minimizers of

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \log \mathbb{E}_{Z \sim P}\left[e^{v(Z;\boldsymbol{\lambda})}\right] \tag{3.25}$$

where $v(Z; \boldsymbol{\lambda}) := -\boldsymbol{g}(Z)^T \boldsymbol{\lambda}$. On the other hand, by plugging (3.22) into (3.19) the optimal parameters for the model projection problem are solutions to

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbb{E}\left[\log \mathbb{E}\left[e^{v_Y(X;\boldsymbol{\lambda})} \,\middle|\, X\right]\right]. \tag{3.26}$$

3.5 A Finite-sample Approximation of Model Projection

In practice, P_X is unknown and only data points $\mathbb{X} := \{X_i\}_{i \in [N]} \subset \mathcal{X}$, drawn from P_X, are available. Thus, we consider the following robust finite-sample optimization problem. We search for a (multi-class) classifier $\boldsymbol{h} : \mathbb{X} \to \boldsymbol{\Delta}_C$ that solves the following:

$$\begin{aligned} \underset{\substack{\boldsymbol{h}:\mathbb{X}\to\boldsymbol{\Delta}_C \\ \boldsymbol{a}:\mathbb{X}\to\mathbb{R}^C, \boldsymbol{b}\in\mathbb{R}^K}}{\text{minimize}} \quad & D_f\left(\boldsymbol{h} \,\|\, \boldsymbol{h}^{\text{base}} \mid \widehat{P}_X\right) + \tau_1 \cdot \left(\mathbb{E}_{X\sim\widehat{P}_X}\left[\|\boldsymbol{a}(X)\|_2^2\right] + \|\boldsymbol{b}\|_2^2\right) \\ \text{subject to} \quad & \mathbb{E}_{\widehat{P}_X}\left[\boldsymbol{G}\cdot(\boldsymbol{h}+\tau_2\boldsymbol{a})\right] \leq \tau_2 \boldsymbol{b}, \end{aligned} \tag{3.27}$$

with $\widehat{P}_X$ being the empirical measure (e.g., obtained from a dataset), and $\tau_1, \tau_2 > 0$ prescribed constants. The terms $\boldsymbol{a}$ and $\boldsymbol{b}$ are added to circumvent infeasibility issues and aid convergence of our numerical procedure. We show in the following theorem that there is a unique solution for (3.27), and that it is given by a tilt (i.e., multiplicative factor) of $\boldsymbol{h}^{\text{base}}$. The tilting parameter is the solution of a finite-dimensional strongly convex optimization problem.

Theorem 3.3. *Suppose Assumption 3.1 holds, and set $\zeta := \tau_2^2/(2\tau_1)$. There exists a unique solution $\boldsymbol{h}^{\text{opt},N}$ to (3.27), and it is given by the formula*

$$h_c^{\text{opt},N}(x) = h_c^{\text{base}}(x) \cdot \phi\left(v_c(x;\boldsymbol{\lambda}_{\zeta,N}^\star) + \gamma(x;\boldsymbol{\lambda}_{\zeta,N}^\star)\right), \quad (x,c) \in \mathbb{X} \times [C], \tag{3.28}$$

with $\boldsymbol{v}, \phi, \gamma$ as in Theorem 3.2, and $\boldsymbol{\lambda}_{\zeta,N}^\star \in \mathbb{R}^K$ is the unique solution to the strongly convex problem

$$D_{\zeta,N}^\star := \min_{\boldsymbol{\lambda}\in\mathbb{R}_+^K} \mathbb{E}_{\widehat{P}_X}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X)\right)\right] + \frac{\zeta}{2}\left\|\boldsymbol{\mathcal{G}}_N^T \boldsymbol{\lambda}\right\|_2^2 \tag{3.29}$$

where $\boldsymbol{\mathcal{G}}_N := \left(\boldsymbol{G}(X_1)/\sqrt{N}, \cdots, \boldsymbol{G}(X_N)/\sqrt{N}, \boldsymbol{I}_K\right) \in \mathbb{R}^{K\times(NC+K)}$.

Proof. See Appendix B.2. □

Theorem 3.3 shows that: strong duality holds between the primal (3.27) and (the negative of) the dual (3.29); there is a unique classifier $\boldsymbol{h}^{\text{opt},N}$ minimizing our optimization problem (3.27); there is a unique solution $\boldsymbol{\lambda}_{\zeta,N}^\star$ to the dual (3.19); and there is an explicit functional form of $\boldsymbol{h}^{\text{opt},N}$ in terms of $\boldsymbol{\lambda}_{\zeta,N}^\star$ in (3.28). Moreover, Theorem 3.3 yields a *practical* two-step procedure for solving the functional optimization in equation (3.27): (i) compute the dual variables $\boldsymbol{\lambda}$ by solving the strongly convex optimization in (3.29); (ii) tilt the base classifier by using the dual variables according to (3.28). This process is applied on real-world datasets using `FairProjection` (see Algorithm 1) in the following sections.

The key distinctions between the finite-sample formulation (3.27) and Theorem 3.2 are that we use the empirical measure $\widehat{P}_X$ (e.g., produced using a dataset with i.i.d. samples), we have a *strongly* convex dual problem in (3.29) (in contrast to the convex program in (3.19)), and we prove strong duality in Theorem 3.3 (whereas an analogous strong duality is absent in the proof of Theorem 3.2).

Remark 3.2. In practice, Assumption 3.1 is not a limiting factor for Theorem 3.3 and `FairProjection`. This is because: we are considering here a finite-set domain so continuity is automatic; we can perturb $\boldsymbol{h}^{\text{base}}$ by negligible noise to push it away from the simplex boundary; and the uniform classifier is strictly feasible. Nevertheless, Assumption 3.1 simplifies the derivation of our theoretical results.

3.6 Fair Projection: Numerical Model Projection

We introduce a parallelizable algorithm—which we call `FairProjection` since our application of interest is fair ML—that solves (3.27) using N i.i.d. data points. We prove that its utility converges to $D^\star$ (see (3.19)) in the population limit and establish both sample-complexity and convergence rate guarantees. Applying `FairProjection` to the group-fairness intervention problem in the following section yields the optimal parameters in (3.28) for post-processing (i.e., tilting) the output of a multi-class classifier in order to satisfy target fairness constraints.

The `FairProjection` algorithm uses ADMM [BPC+11] to solve the convex program in (3.29). Recall that it suffices to optimize (3.29) for computing (3.27) as proved in Theorem 3.3. Algorithm 1 presents the steps of `FairProjection`, and its detailed derivation is given in Appendix B.3.1. A salient feature of `FairProjection` is its *parallelizability*. Each step that is done for i varying over $[N]$ can be executed for each i separately and in parallel. In particular, this applies to the most computationally intensive step, the $\boldsymbol{v}_i$-update step. We discuss next how the $\boldsymbol{v}_i$-update step is carried out.

Algorithm 1 : `FairProjection` for solving (3.29).

```
1:  Input: divergence $f$, predictions $\{\boldsymbol{p}_i := \boldsymbol{h}^{\text{base}}(X_i)\}_{i\in[N]}$, constraints $\{\boldsymbol{G}_i := \boldsymbol{G}(X_i)\}_{i\in[N]}$, regularizer $\zeta$, ADMM penalty $\rho$, and initializers $\boldsymbol{\lambda}$ and $(\boldsymbol{w}_i)_{i\in[N]}$.
2:  Output: $h_c^{\text{opt},N}(x) := h_c^{\text{base}}(x) \cdot \phi(\gamma(x;\boldsymbol{\lambda}) + v_c(x;\boldsymbol{\lambda}))$.
3:  $\boldsymbol{Q} \leftarrow \frac{\zeta}{2}\boldsymbol{I} + \frac{\rho}{2N}\sum_{i\in[N]} \boldsymbol{G}_i\boldsymbol{G}_i^T$
4:  for $t = 1, 2, \cdots, t'$ do
5:      $\boldsymbol{a}_i \leftarrow \boldsymbol{w}_i + \rho\boldsymbol{G}_i^T\boldsymbol{\lambda}$                                          $i \in [N]$
6:      $\boldsymbol{v}_i \leftarrow \operatorname{argmin}_{\boldsymbol{v}\in\mathbb{R}^C} D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i) + \frac{\rho+\zeta}{2}\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}_i^T\boldsymbol{v}$           $i \in [N]$
7:      $\boldsymbol{q} \leftarrow \frac{1}{N}\sum_{i\in[N]} \boldsymbol{G}_i \cdot (\boldsymbol{w}_i + \boldsymbol{v}_i)$
8:      $\boldsymbol{\lambda} \leftarrow \operatorname{argmin}_{\boldsymbol{\ell}\in\mathbb{R}_+^K} \boldsymbol{\ell}^T\boldsymbol{Q}\boldsymbol{\ell} + \boldsymbol{q}^T\boldsymbol{\ell}$
9:      $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + \rho \cdot (\boldsymbol{v}_i + \boldsymbol{G}_i^T\boldsymbol{\lambda})$                                $i \in [N]$
10: end for
```

3.6.1 Inner Iterations

One approach to carry out the inner iteration in Algorithm 1 that updates $\boldsymbol{v}_i$ is to study the vanishing of the gradient of $\boldsymbol{v} \mapsto D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}_i^T\boldsymbol{v}$ (where $\xi = (\rho+\zeta)/2$ and $\boldsymbol{a}_i \in \mathbb{R}^C$ is some vector). In the KL-divergence case, $D_{\mathsf{KL}}^{\text{conj}}$ is given by a log-sum-exp function, so its gradient is given by a softmax function, and equating the gradient to zero becomes a fixed-point equation. We give an iterative routine to solve this fixed point equation in Appendix B.3.2, whose proof of convergence is discussed in the same section. It is worth noting that carrying out this inner step for the KL-divergence case can be guaranteed to converge faster if one can bound the ℓ_2-norm Lipschitz constant of the softmax function. As shown in [GP17, Prop. 4], the softmax function is known to be 1-Lipschitz. A result we prove—that is of independent interest—is that this Lipschitz constant can in fact be reduced to 1/2.

Proposition 3.1. *For any $n \in \mathbb{N}$, the softmax function $\sigma(\boldsymbol{z}) := \left(\frac{e^{z_j}}{\sum_{i=1}^n e^{z_i}}\right)_{j\in[n]}$ is $\frac{1}{2}$-Lipschitz with respect to the ℓ_2 norm.*

Proof. See Appendix B.3.3. □

Beyond the KL-divergence case, setting the gradient to zero does not seem to be an analytically tractable problem. Nevertheless, we may reduce the vector minimization in Line 6 of Algorithm 1 to a tractable 1-dimensional root-finding problem, as the following result aids in showing.

Lemma 3.2. *For $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$, $\boldsymbol{a} \in \mathbb{R}^C$, and $\xi > 0$, if f satisfies Assumption 3.1, we have that*

$$\min_{\boldsymbol{v} \in \mathbb{R}^C} D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}) + \xi \|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T \boldsymbol{v} = -\sup_{\theta \in \mathbb{R}} -\theta + \sum_{c \in [C]} \min_{q_c \geq 0} p_c f\left(\frac{q_c}{p_c}\right) + \frac{(a_c + q_c)^2}{4\xi} + \theta q_c. \tag{3.30}$$

Proof. See Appendix B.3.2. □

We note that the $\boldsymbol{v}_i$-update steps for both KL and CE (provided in detail in Appendix B.3.2) give, as a byproduct, the implicitly defined function $\gamma(x; \boldsymbol{\lambda})$ (see the statements of Theorems 3.2–3.3).

3.6.2 Convergence Guarantees

Our proposed algorithm, `FairProjection`, enjoys the following convergence guarantees. The output after the t-th iteration $\boldsymbol{\lambda}_{\zeta,N}^{(t)}$ converges exponentially fast to $\boldsymbol{\lambda}_{\zeta,N}^{\star}$ (see (3.29)).

Theorem 3.4. *Suppose Assumption 3.1 holds, and that the matrix $(\boldsymbol{G}(X_i))_{i \in N} \in \mathbb{R}^{K \times NC}$ has full row-rank. Let $\boldsymbol{\lambda}_{\zeta,N}^{(t)}$ and $\boldsymbol{h}^{(t)}$ be the t-th iteration outputs of `FairProjection` for the KL-divergence case. Then, we have the exponential decay of errors $\|\boldsymbol{\lambda}_{\zeta,N}^{(t)} - \boldsymbol{\lambda}_{\zeta,N}^{\star}\|_2 = e^{-\Omega(t)}$ and $\boldsymbol{h}^{(t)}(x) = \boldsymbol{h}^{\text{opt},N}(x) \cdot \left(1 \pm e^{-\Omega(t)}\right)$ uniformly in $x \in \mathbb{X}$ as $t \to \infty$.*

Proof. See Appendix B.3.4. □

Remark 3.3. The full-rank assumption on the matrix $(\boldsymbol{G}(X_i))_{i \in N} \in \mathbb{R}^{K \times NC}$ can be ensured by adding negligible noise to it. Further, although Theorem 3.4 is shown for the KL-divergence, the proof directly extends to general f-divergences satisfying Assumption 3.1 (see Appendix B.3.5 for further discussions).

Further, we show in the next theorem that carrying $t = \Omega(\log N)$ iterations of `FairProjection`, with regularizer $\zeta = \Theta(N^{-1/2})$, yields a parameter $\boldsymbol{\lambda}_{\zeta,N}^{(t)}$ that works well for the *population* problem for information projection (3.19); this makes `FairProjection` have a computational runtime of $O(N \log N)$.

Theorem 3.5. *Suppose Assumption 3.1 holds, let $\mathcal{X} = \mathbb{R}^d$, and consider the KL-divergence case. Then, choosing $\zeta = \Theta\left(N^{-1/2}\right)$ and $t = \Omega(\log N)$ we obtain for any $\delta \in (0,1)$ that (see (3.19))*

$$\Pr\left\{\mathbb{E}_X\left[D_{\mathsf{KL}}^{\text{conj}}\left(\boldsymbol{v}\left(X; \boldsymbol{\lambda}_{\zeta,N}^{(t)}\right), \boldsymbol{h}^{\text{base}}(X)\right)\right] > D^\star + O\left(\frac{1}{\sqrt{N}}\right)\right\} \leq \delta. \tag{3.31}$$

Proof. See Appendix B.3.6. □

3.6.3 Benefit of Parallelization

The parallelizability of `FairProjection` provides significant speedup. In Appendix B.4.2, we provide an ablation study comparing the speedup due to parallelization. For the ENEM dataset (discussed in Section 3.8), parallelization yields a 15-fold reduction in runtime in our experiments. In addition to the parallel advantage of `FairProjection`, its inherent mathematical approach is more advantageous than gradient-based solutions. When numerically solving the dual problem (3.29) (or any close variant) via gradient methods, the gradient of D_f^{conj} (the convex conjugate of an f-divergence) must be computed. However, this gradient is tractable in only a very limited number of relevant instances of f-divergences. `FairProjection` tackles this intractability through having its subroutines be informed by Lemma 3.2 and the discussion preceding it.

3.7 Application to Fair Machine Learning

In this section, we aim at designing a fairness-aware classifier using machinery of model projection. We formalize an optimization for this purpose which coincides with the model projection framework explored in the previous sections. Prior works attempt to design fair classifiers by implicitly solving a model projection problem, where accuracy is measured by, for example, KL-divergence [JN19] and cross-entropy [WRC20]. Here we provide a general framework in the setting of multiclass classification, and this approach allows the usage of any f-divergence. In what follows, we formally introduce our formulation.

3.7.1 Classification Tasks

The essential objects in (multi-class) classification are the input sample space $\mathcal{X}$, the predicted classes $\mathcal{Y}$, and the classifiers. We fix two random variables X and Y, taking values in sets $\mathcal{X}$ and $\mathcal{Y} := [C]$. Here, (X, Y) is a pair comprised of an input sample X (e.g., criminal history) and corresponding class label Y (e.g., criminal recidivism) randomly drawn from $\mathcal{X} \times \mathcal{Y}$ with distribution $P_{X,Y}$. A probabilistic classifier is a function $\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C$, where $h_c(x)$ represents the probability of sample $x \in \mathcal{X}$ falling in class $c \in \mathcal{Y}$. Thus, $\boldsymbol{h}$ gives rise to a $\mathcal{Y}$-valued random variable $\widehat{Y}$ via the distribution $P_{\widehat{Y}|X=x}(c) := h_c(x)$.

Let S be a group attribute (e.g., race and/or sex), with support $\mathcal{S} := [A]$. We work under the setting of a Markov chain $(Y, S) - X - \widehat{Y}$. We also assume that $P_{S|Y=c}(a) > 0$ for each $(a, c) \in$

$[A] \times [C]$, i.e., each pair comprised of a protected group and predicted class intersect nontrivially.[6]

We assume that we have in hand a well-calibrated classifier that approximates $P_{Y,S|X}$, i.e., that predicts both group membership S and the true label Y from input variables X. This classifier can be directly marginalized into the model $\boldsymbol{h}^{\text{base}}$ (i.e., the distribution $P_{Y|X}$) and the classifier $\boldsymbol{s} : \mathcal{X} \times \mathcal{Y} \to \boldsymbol{\Delta}_A$ defined by

$$s_a(x, c) := P_{S|X=x,Y=c}(a). \tag{3.32}$$

If the group attribute S is part of the input feature X, then $P_{S|X,Y}$ is simply replaced with an indicator function. Otherwise, we can approximate this conditional distribution by training a probabilistic classifier. Thus, we henceforth treat $\boldsymbol{s}$ as given.

3.7.2 Group-Fairness Constraints

We consider multi-class generalization of three commonly used group fairness criteria in Table 3.2. As observed by existing works (see, e.g., [ABD+18, MW18, CHKV19, WRC20, AAW+20]), each of these fairness constraints[7] can be written in the vector-inequality form $\mathbb{E}_{P_X}[\boldsymbol{G}\boldsymbol{h}] \leq \boldsymbol{0}$ for a closed-form matrix-valued function $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{K \times C}$. This linearity is not hard to see in view of Lemma 3.1. Consider for example the case of statistical parity. We may rewrite its defining condition as the collection of inequalities

$$(-1)^{\delta} P_{\widehat{Y}|S=a}(c') - ((-1)^{\delta} + \alpha) P_{\widehat{Y}}(c') \leq 0 \qquad \text{for every } (\delta, a, c') \in \{0,1\} \times [A] \times [C]. \tag{3.33}$$

Using the Markov chain condition $(Y, S) - X - \widehat{Y}$ and the law of total probability, the functions in the inequalities (3.33) can be transformed into the form

$$(-1)^{\delta} P_{\widehat{Y}|S=a}(c') - ((-1)^{\delta} + \alpha) P_{\widehat{Y}}(c') = \mathbb{E}\left[g_{\delta,a,c'}(X, Y, \widehat{Y})\right] \tag{3.34}$$

where

$$g_{\delta,a,c'}(x, y, y') = \left((-1)^{\delta} P_S(a)^{-1} s_a(x, y) - ((-1)^{\delta} + \alpha)\right) 1_{\{c'\}}(y'). \tag{3.35}$$

This means that statistical parity can be written in the form $\mathbb{E}[\boldsymbol{G}(X)\boldsymbol{h}(X)] \leq \boldsymbol{0}$, where the $\boldsymbol{G}$ matrix evaluated at a fixed individual $x \in \mathcal{X}$ has $K = 2AC$ rows indexed by $(\delta, a, c') \in \{0,1\} \times [A] \times [C]$, where the (δ, a, c')-th row is equal to $\left((-1)^{\delta} P_S(a)^{-1} \sum_{c \in [C]} s_a(x, c) h_c^{\text{base}}(x) - (\alpha + (-1)^{\delta})\right) \boldsymbol{e}_{c'}$, with

[6] This restriction is only necessary for the linearization of the equalized-odds fairness metric (see Table 3.2).

[7] We remark that our framework can be applied to other fairness constraints, e.g., the ones in [WRC20].

Fairness Criterion	Statistical parity	Equalized odds	Overall accuracy equality
Expression	$\left\lvert \frac{P_{\widehat{Y}\mid S=a}(c')}{P_{\widehat{Y}}(c')} - 1 \right\rvert \leq \alpha$	$\left\lvert \frac{P_{\widehat{Y}\mid Y=c, S=a}(c')}{P_{\widehat{Y}\mid Y=c}(c')} - 1 \right\rvert \leq \alpha$	$\left\lvert \frac{P(\widehat{Y} = Y \mid S = a)}{P(\widehat{Y} = Y)} - 1 \right\rvert \leq \alpha$

Table 3.2: *Standard multi-class group fairness criteria; one fixes $\alpha > 0$ and iterates over all $(a, c, c') \in [A] \times [C]^2$.*

$\boldsymbol{e}_1, \cdots, \boldsymbol{e}_C$ denoting the standard basis for $\mathbb{R}^C$. The expressions for the $\boldsymbol{G}$ matrix corresponding to the other fairness metrics are given in the following lemma.

Lemma 3.3. *Every fairness criterion listed in Table 3.2 can be written in the linear form $\mathbb{E}[\boldsymbol{G}(X)\boldsymbol{h}(X)] \leq \boldsymbol{0}$ for a matrix $\boldsymbol{G}$ that is completely determined by the classifiers $\boldsymbol{h}^{\text{base}}$ and $\boldsymbol{s}$ and the population distribution $P_{S,Y}$. Explicitly, with $\odot$ denoting element-wise multiplication, for statistical parity the matrix $\boldsymbol{G}$ has the $K = 2AC$ rows*

$$\boldsymbol{G}_{k,:}(x) = \left((-1)^{\delta} \frac{s_{a'}(x, \cdot) \odot \boldsymbol{h}^{\text{base}}(x)}{P_S(a')} - \left(\alpha + (-1)^{\delta} \right) \right) \boldsymbol{e}_{c'} \tag{3.36}$$

indexed by $k = (\delta, a', c') \in \{0,1\} \times [A] \times [C]$; for equalized odds, $\boldsymbol{G}$ has the $K = 2AC^2$ rows

$$\boldsymbol{G}_{k,:}(x) = \left((-1)^{\delta} \frac{s_{a'}(x, c) h_c^{\text{base}}(x)}{P_{S\mid Y=c}(a')} - \left(\alpha + (-1)^{\delta} \right) h_c^{\text{base}}(x) \right) \boldsymbol{e}_{c'} \tag{3.37}$$

indexed by $k = (\delta, a', c, c') \in \{0,1\} \times [A] \times [C]^2$; and for overall accuracy equality, $\boldsymbol{G}$ has the $K = 2A$ rows

$$\boldsymbol{G}_{k,:}(x) = (-1)^{\delta} \frac{s_{a'}(x, \cdot) \odot \boldsymbol{h}^{\text{base}}(x)}{P_S(a')} - \left(\alpha + (-1)^{\delta} \right) \boldsymbol{h}^{\text{base}}(x) \tag{3.38}$$

indexed by $k = (\delta, a') \in \{0,1\} \times [A]$.

3.7.3 Fairness Through Model Projection

Our goal is to design an efficient post-processing method that takes a pre-trained classifier $\boldsymbol{h}^{\text{base}}$ that may violate some target group-fairness criteria and finds a fair classifier that has the most similar outputs (i.e., closest utility performance) to that of $\boldsymbol{h}^{\text{base}}$. We formulate this fairness intervention problem mathematically using model projection, as follows. For a fixed search space $\mathcal{H} \subset \boldsymbol{\Delta}_C^{\mathcal{X}} := \{\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C\}$, a loss function $\mathsf{err} : \boldsymbol{\Delta}_C^{\mathcal{X}} \times \boldsymbol{\Delta}_C^{\mathcal{X}} \to \mathbb{R}$, and a base classifier $\boldsymbol{h}^{\text{base}} \in \boldsymbol{\Delta}_C^{\mathcal{X}}$, one seeks to solve:

$$\underset{\boldsymbol{h} \in \mathcal{H}}{\text{minimize}} \ \mathsf{err}\left(\boldsymbol{h}, \boldsymbol{h}^{\text{base}} \right) \quad \text{subject to } \mathbb{E}_{P_X}\left[\boldsymbol{G}\boldsymbol{h} \right] \leq \boldsymbol{0}. \tag{3.39}$$

The function err quantifies the "closeness" between the scores given by $\boldsymbol{h}$ and $\boldsymbol{h}^{\text{base}}$. The constraint on $\boldsymbol{h}$ can encode any arbitrary statistical information about the joint distribution induced on the Markov chain $(Y,S) - X - \widehat{Y}$ (see Lemma 3.1). Thus, solving the optimization (3.39) amounts to finding the minimal necessary perturbation to the base classifier $\boldsymbol{h}^{\text{base}}$ to make it satisfy a given on-average constraint. Since we consider raw output scores, we measure "closeness" via f-divergences:

$$\mathsf{err}\left(\boldsymbol{h}, \boldsymbol{h}^{\text{base}}\right) = D_f(\boldsymbol{h} \parallel \boldsymbol{h}^{\text{base}} \mid P_X) := \mathbb{E}_{P_X}\left[\sum_{c\in[C]} h_c^{\text{base}}(X) f\left(\frac{h_c(X)}{h_c^{\text{base}}(X)}\right)\right] - f(1), \tag{3.40}$$

where f is a convex function over $(0,\infty)$. By varying different choices of f, we can obtain e.g., cross-entropy (CE, $f(t) = -\log t$) and KL-divergence ($f(t) = t\log t$). For a chosen f-divergence, the optimization problem (3.39) becomes a generalization of *information projection* [Csi75]; specifically, it becomes thee *model projection* problem we introduce in the previous sections.

Thus, the pipeline for applying model projection to fair ML is summarized in the following steps:

1. Assume access to a potentially unfair base multi-class predictor $\boldsymbol{h}^{\text{base}} : \mathcal{X} \to \boldsymbol{\Delta}_C$, and also a set of N training samples from the dataset.

2. Fix a collection of desired linearizable group-fairness constraints (e.g., from among the ones in Table 3.2). Let $\boldsymbol{G}$ be the ensuing matrix-valued function (see Lemma 3.3).

3. Pick an f-divergence, and a regularizer $\zeta \propto 1/\sqrt{N}$ (see Theorem 3.5), and consider the finite-sample formulation in (3.27).

4. Use `FairProjection` (Algorithm 1) and Theorem 3.3 to compute the optimal parameters $\boldsymbol{\lambda}^{\star}_{\zeta,N}$ and project $\boldsymbol{h}^{\text{base}}$ to obtain the optimal fair model $\boldsymbol{h}^{\text{opt},N}$.

The way we design the fair classifier falls into the post-processing category. This is because the optimal fair classifier is a tilting of the label classifier (see Theorems 3.2 and 3.3). Notably, the formulations (3.13) and (3.27) do not *a priori* assume a post-processing design procedure. Nevertheless, the projected classifier turns out to own an optimality guarantee among all classifiers.

We point out that the formulation in [WRC20] presents a special case of the model projection theory using cross-entropy as the f-divergence of choice and assuming Y and S are binary. While computationally lightweight, the experiments in [WRC20, Section 6] demonstrate that the model projection formulation may perform favorably compared to state-of-the-art fairness intervention mechanisms. Here, we provide a general theoretical work that allows usage of a wide class of

f-divergences. In the next section, we provide more comprehensive numerical experiments including for multiple f-divergences and for multi-class prediction.

3.8 Numerical benchmarks

We present empirical results and show that `FairProjection` has competitive performance both in terms of runtime and fairness-accuracy trade-off curves compared to benchmarks—most notably the reductions approach in [ABD+18], which requires retraining. Extensive additional benchmarks and experiment details are reported in Appendix B.4.

3.8.1 Setup

We consider three base classifiers (`Base`): gradient boosting (GBM), logistic regression (LR), and random forest (RF), implemented by Scikit-learn [PVG+11]. For `FairProjection` (the constrained optimization in (3.27)), we use cross-entropy (`FairProjection-CE`) and KL-divergence (`FairProjection-KL`) as the loss function[8]. We consider two fairness constraints: mean equalized odds (MEO) and statistical parity (SP) (cf. Table 3.2). Particularly, to measure multi-class performance, we extend the definition of MEO as

$$\mathsf{MEO} = \max_{i \in \mathcal{Y}} \max_{s_1, s_2 \in \mathcal{S}} \left(|\mathsf{TPR}_i(s_1) - \mathsf{TPR}_i(s_2)| + |\mathsf{FPR}_i(s_1) - \mathsf{FPR}_i(s_2)| \right)/2 \tag{3.41}$$

where $\mathsf{TPR}_i(s) = P(\widehat{Y} = i | Y = i, S = s)$, and $\mathsf{FPR}_i(s) = P(\widehat{Y} = i | Y \neq i, S = s)$. The definition of multi-class statistical parity is provided in Appendix B.4.4. All values reported in this section are from the test set with 70/30 train-test split. When benchmarking against methods tailored to binary classification, we restrict our results to both binary Y and S since, unlike `FairProjection`, competing methods cannot necessarily handle multi-class predictions and multiple groups.

3.8.2 Datasets

We evaluate `FairProjection` and all benchmarks on four datasets. We use two datasets in the education domain: the high-school longitudinal study (HSLS) dataset [IPH+11, JWC22] and a novel dataset ENEM [INE20] (details in Appendix B.4.1). The ENEM dataset contains Brazilian college

[8] We focus on `FairProjection-CE` and random forest here; results for `FairProjection-KL` and other models are in Appendix B.4.

entrance exam scores along with student demographic information and socio-economic questionnaire answers (e.g., if they own a computer). After pre-processing, the dataset contains ~1.4 million samples with 139 features. Race was used as the group attribute S, and Humanities exam score is used as the label Y. The score can be quantized into an arbitrary number of classes. For binary experiments, we quantize Y into two classes, and for multi-class, we quantize it to 5 classes. The race feature S has 5 categories, but we binarize it into White and Asian ($S = 1$) and others ($S = 0$). We call the entire ENEM dataset ENEM-1.4M. We also created smaller versions of the dataset with 50k samples: ENEM-50k-2C (binary classes) and ENEM-50k-5C (5 classes).[9] For completeness, we report results on UCI Adult [Lic13] and COMPAS [ALMK16].

3.8.3 Benchmarks

For the binary classification experiments, we compare our method with five existing fair learning algorithms: `Reduction` [ABD+18], reject-option classifier [KKZ12, `Rejection`], equalized-odds [HPS+16, `EqOdds`], calibrated equalized-odds [PRW+17, `CalEqOdds`], and leveraging equal opportunity [CDH+19, `LevEqOpp`].[10] The choice of benchmarks is based on the availability of reproducible codes. For the first four baselines, we use IBM AIF360 library [BDH+18]. For `Reduction` and `Rejection`, we vary the tolerance to achieve different operation points on the fairness-accuracy trade-off curves. As `EqOdds`, `CalEqOdds` and `LevEqOpp` only allow hard equality constraint on equalized odds, they each produce a single point on the plot (see Fig. 3.1). We include the group attribute as a feature in the training set following the same benchmark procedure described in [ABD+18, WRC21] for a consistent comparison. For multi-class classification experiments, we did not find methods that can be easily compared against `FairProjection` and use the multi-class extensions of mean equalized odds and statistical parity. For the sake of completeness, we modified the codes of adversarial debiasing [ZLM18, `Adversarial`], and compare our method against it. Note that `Reduction` [ABD+18] and `Adversarial` [ZLM18] are in-processing methods, and the rest of the benchmark algorithms are post-processing methods like `FairProjection`. Additional comparisons to [KCT20] are given in Appendix B.4.4.

There are four methods in Table 3.1 we did not include the experiments: FACT [KCT20], Identifying [JN20], FST [WRC21], and Overlapping [YCK20], as explained in Appendix B.4.1.

[9] A datasheet (see [GMV+21]) for ENEM is given in Appendix B.5.

[10] https://github.com/lucaoneto/NIPS2019_Fairness.

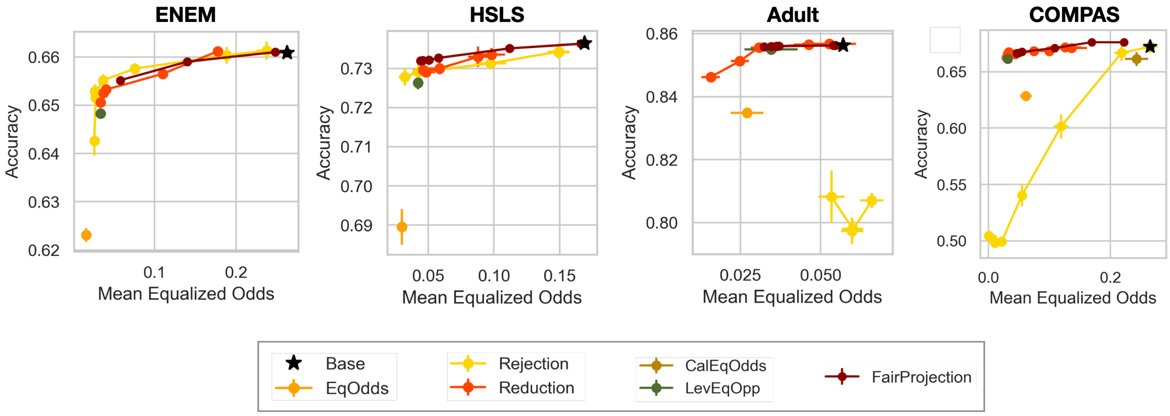
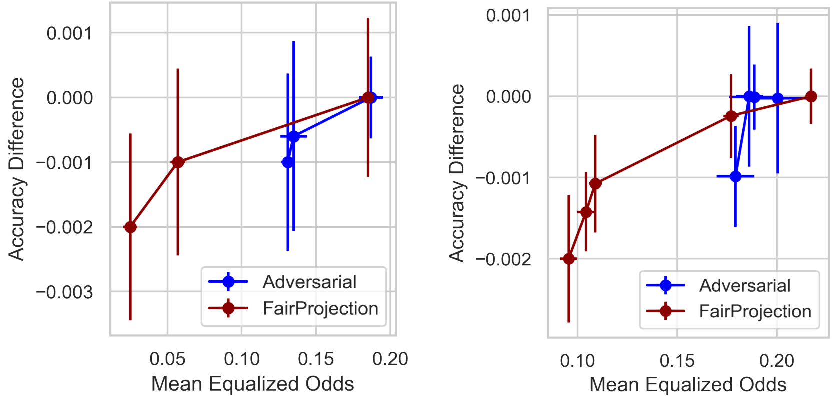


Figure 3.1: *Fairness-accuracy trade-off comparisons between* `FairProjection` *and five baselines on ENEM-50k-2C, HSLS, Adult and COMPAS datasets. For all methods, we used random forest as a base classifier. Note that* `EqOdds`*,* `CalEqOdds`*, and* `LevEqOpp` *only produce a single accuracy-fairness trade-off point, whereas the rest of the methods are capable of producing the accuracy-fairness trade-off curves by varying the fairness budget α for the group fairness criteria listed in Table 3.2—a smaller α corresponds to a lefter point on the accuracy-fairness trade-off curve.*

3.8.4 Binary Classification Results

We compare `FairProjection` with benchmarks tailored to binary classification in terms of the MEO-accuracy trade-off on the ENEM-50k-2C, HSLS, Adult, and COMPAS datasets in Fig. 3.1. Each point is obtained by averaging 10 runs with different train-test splits. `FairProjection-CE` curves were obtained by varying α values (cf. Table 3.2). When $\alpha = 1.0$, the outputs of `FairProjection-CE` are equivalent to the base classifier RF.

We observe that `FairProjection-CE` and `Reduction` have the overall best and most consistent performances. On ENEM-50k-2C and HSLS datasets, although `EqOdds` achieves the best fairness, that fairness comes at the cost of 4% accuracy drop (additively). The other four methods, on the other hand, produce comparatively good fairness with an accuracy loss of $< 1\%$. In particular, `FairProjection-CE` has the smallest accuracy drop whilst improving MEO from 0.17 to 0.04 on HSLS. `CalEqOdds` requires strict calibration requirements and yields inconsistent performance when these requirements are not met. On ENEM-50k-2C and HSLS, `LevEqOpp` achieves comparable MEO with a slight accuracy drop, and on COMPAS, `LevEqOpp` performs equally well as `FairProjection-CE` and `Reduction`. Note that with high fairness constraints (i.e., small tolerance), the accuracy of `Rejection` deteriorates.



(a) *HSLS.* **(b)** *ENEM-50k-5C.*

Figure 3.2: *Fairness-accuracy trade-off for multi-class prediction on HSLS and ENEM-50k-5C. FairProjection is* `FairProjection-CE` *with LR base classifier.*

3.8.5 Multi-Class Results

We illustrate how `FairProjection` performs on multi-class prediction using HSLS and ENEM-50k-5C. For HSLS, we divided student math performance into quartiles and generated four classes. In Figure 3.2, we plot fairness-accuracy trade-off of `FairProjection-CE` with logistic regression and adversarial debiasing [ZLM18, `Adversarial`]. As their base classifiers are different (`Adversarial` is a GAN-based method), we plot accuracy difference compared to the base classifier instead of plotting the absolute value of accuracy[11]. `FairProjection` reduces MEO significantly with very small loss in accuracy. While `Adversarial` is also able to reduce MEO with negligible accuracy drop, it does not reduce the MEO as much as `FairProjection`. We show more extensive results with multi-group and multi-class ($|\mathcal{Y}| = 5, = |\mathcal{S}| = 5$) in Appendix B.4.4.

3.8.6 Runtime Comparisons

To demonstrate the scalability of `FairProjection`, we record in Table 3.3 the runtime of the instantiations `FairProjection-CE` and `-KL` with the five benchmarks on ENEM-1.4M-2C, which is the biggest dataset we have. These experiments were run on a machine with AMD Ryzen 2990WX 64-thread 32-Core CPU and NVIDIA TITAN Xp 12-GB GPU. For consistency, we used the same fairness metric (MEO, $\alpha = 0.01$), base classifier (GBM), and train/test split, and each number is the average of 2

[11] Base accuracy for `FairProjection` $= 0.336$, `Adversarial` $= 0.307$. Random guessing accuracy $= 0.2$.

Method	***Reduction*** [ABD+18]	**Rejection** [KKZ12]	EqOdds [HPS+16]	LevEqOpp [CDH+19]	CalEqOdds [PRW+17]	***FairProjection (ours)*** CE	 KL
Runtime	223.6	16.9	5.9	7.9	5.3	11.3	11.6

Table 3.3: *Execution time of* `FairProjection` *on the ENEM-1.4M-2C compared with five baseline methods (time shown in minutes). Methods in* ***bold*** *are capable of producing a fairness-accuracy trade-off curve. Methods that are italicized have a uniformly superior performance. The time reported here for* `FairProjection` *includes the time to fit the base classifiers. If base classifiers are given, the runtime of e.g.* `FairProjection-KL` *is 1.63 mins. The runtimes are consistent with small standard deviations across repeated experiments.*

repeated experiments. `EqOdds`, `LevEqOpp`, and `CalEqOdds` are faster than `FairProjection` since they are optimized to produce one trade-off point (cf. Fig. 3.1). Compared to baselines that produce full fairness-accuracy trade-off curves (i.e., `Reduction` and `Rejection`), `FairProjection` has the fastest runtime. Also, the non-parallel implementation of `FairProjection-KL` takes 25.3 mins—parallelization attains 15× speedup (detailed results in Appendix B.4.2). We further compare the runtime results for the binary HSLS, which is the second biggest dataset, with the baselines that produce full fairness-accuracy trade-off curves. The runtimes for `Reduction`, `Rejection` and `FairProjection-CE` are 81.1 sec, 9.73 sec and 4.50 sec respectively—again, `FairProjection` has the fastest runtime. For a theoretical comparison between the runtime of `FairProjection` and `Reduction`, see Appendix B.4.3.

3.9 Conclusion and Open Problems

We introduce model projection, a generalization of information projection to conditional distributions. We prove existence, uniqueness, and explicit formulas for the model projection. Instantiating our model projection theory to the domain of group-fairness, we introduce a theoretically-grounded and versatile fairness intervention method, `FairProjection`, and showcase its favorable performance in extensive experiments. We encourage the reader to peruse our theoretical guarantees in Appendix B.3 and extensive additional numerical benchmarks in Appendix B.4. `FairProjection` is able to correct bias for multigroup/multiclass datasets, and it enjoys a fast runtime thanks to its parallelizability. We also evaluate our method on the ENEM dataset (see Appendix B.5 for a detailed description of the dataset). Our benchmarks are a step forward in moving away from the overused COMPAS and UCI Adult datasets.

We only consider group-fairness, and it would be interesting to try to incorporate other fairness notions (e.g., individual fairness [DHP+12]) into our formulation. We assume that $\boldsymbol{h}^{\text{base}}$ is a pre-trained accurate (and potentially unfair) classifier; one future research direction is understanding how the accuracy of $\boldsymbol{h}^{\text{base}}$ influences the performance of the projected classifier. Finally, the performance of `FairProjection` is inherently constrained by data availability. Performance may degrade with intersectional increases of the number of groups, the number of labels, and the number of fairness constraints.

Chapter 4

Measuring Information from Moments

A fundamental formula in information theory is the I-MMSE relation [GSV05], which shows that in Gaussian channels the mutual information is the integral of the minimum mean-square error (MMSE):

$$I(X; \sqrt{\gamma}X + N) = \frac{1}{2} \int_0^{\gamma} \text{mmse}\left(X \mid \sqrt{t}X + N\right) \, dt. \tag{4.1}$$

Here, X has finite variance and N is a standard normal random variable independent of X. In this chapter, we build on this relation to express information measures of two random variables X and Y as functions of their moments. For example, whenever X and Y are continuous with finite moment-generating functions around the origin, there is a sequence of rational functions $\{\rho_n\}_{n\in\mathbb{N}}$—each completely determined by finitely many moments of X and Y—such that the mutual information is

$$I(X;Y) = \lim_{n\to\infty} \int_0^{\infty} \rho_n(t) \, dt. \tag{4.2}$$

We derive the new expression (4.2) and a similar formula for differential entropy in three steps. First, we produce polynomial approximations of conditional expectations. Second, we apply these approximations to bound the mean-square error of reconstructing a hidden variable X from an observation Y using an estimator that is a polynomial in Y. We call this approximation the PMMSE, in short for Polynomial MMSE. Finally, we use the PMMSE in the I-MMSE relation (4.1) to approximate mutual information (as in (4.2)) and differential entropy.

4.1 Overview of the Main Results

The crux of our work is the study of polynomial approximations of conditional expectations. A surprising result that motivates this study is a negative answer to the question: If X and $N \sim \mathcal{N}(0,1)$ are independent random variables, can $y \mapsto \mathbb{E}[X \mid X+N=y]$ be a nonlinear polynomial? Proposition 4.1, stated below, shows that if X is integrable (i.e., $\mathbb{E}[|X|] < \infty$), the only way that $\mathbb{E}[X \mid X+N]$ can be a polynomial is if X is Gaussian or constant. In other words, if $y \mapsto \mathbb{E}[X \mid X+N=y]$ is a polynomial, then it is of degree at most 1.

Proposition 4.1 ([AC21c, Theorem 1])**.** *For $Y = X+N$ where X is an integrable random variable and $N \sim \mathcal{N}(0,1)$ independent of X, the conditional expectation $\mathbb{E}[X \mid Y]$ cannot be a polynomial in Y with degree greater than* 1*. Therefore, the MMSE estimator in a Gaussian channel with finite-variance input is a polynomial if and only if the input is Gaussian or constant.*

Despite the negative result in Proposition 4.1, we produce a sequence of polynomials converging to the conditional expectation $\mathbb{E}[X \mid Y]$, provided that X has finite variance and Y is light-tailed. For each $n \in \mathbb{N}$, we consider the orthogonal projection of X onto the subspace[1] $\mathscr{P}_n(Y) \subset L^2(P_Y)$ of polynomials in Y with real coefficients and of degree at most n, where it is assumed that $\mathbb{E}[X^2], \mathbb{E}[Y^{2n}] < \infty$. The standard theory of orthogonal projections in Hilbert spaces yields that the orthogonal projection of X onto $\mathscr{P}_n(Y)$, which we denote by $E_n[X \mid Y]$, exists and is unique; indeed, being finite-dimensional, the subspace $\mathscr{P}_n(Y)$ is closed. Further, it is well-known that the orthogonal projection $E_n[X \mid Y]$ is the unique best polynomial approximation of both X and $\mathbb{E}[X \mid Y]$ in the $L^2(P_Y)$ norm (see, e.g., [SS19, Section 4.4]). From an estimation-theoretic point of view, the operators E_n are natural generalizations of the linear minimum mean-square error (LMMSE) estimate. Hence, we call this process polynomial minimum mean-square (PMMSE) estimation. We collect these observations in the following definition, in which we denote the random vector $\boldsymbol{Y}^{(n)} := (1, Y, \cdots, Y^n)^T$.

Definition 4.1 (Polynomial MMSE)**.** Fix $n \in \mathbb{N}$ and two random variables X and Y satisfying $\mathbb{E}\left[X^2\right] < \infty$ and $\mathbb{E}\left[Y^{2n}\right] < \infty$. We define the n-th order *polynomial minimum mean-square error*

[1] Throughout, we fix a probability space $(\Omega, \mathcal{F}, P)$, over which random variables are defined. For $q \geq 1$, the Banach space $L^q(P)$ consists of all q-integrable random variables Z, i.e., $\|Z\|_q := \left(\int_\Omega |Z|^q \, dP\right)^{1/q} < \infty$. The inner product of the Hilbert space $L^2(P)$ is denoted by $\langle \cdot , \cdot \rangle$. The Borel probability measure on $\mathbb{R}$ induced by Y is denoted by P_Y. The Banach subspace $L^q(P_Y) \subset L^q(P)$ consists of $\sigma(Y)$-measurable and q-integrable random variables.

(PMMSE) for estimating X given Y by

$$\mathrm{pmmse}_n(X \mid Y) := \min_{\boldsymbol{c}\in\mathbb{R}^{n+1}} \mathbb{E}\left[\left(X - \boldsymbol{c}^T\boldsymbol{Y}^{(n)}\right)^2\right]. \tag{4.3}$$

We define the n-th order PMMSE *estimate* of X given Y by $E_n[X \mid Y] := \boldsymbol{c}^T\boldsymbol{Y}^{(n)} \in \mathscr{P}_n(Y)$ for any minimizer $\boldsymbol{c} \in \mathbb{R}^{n+1}$ in (4.3).

The PMMSE estimate is the unique minimizer (in $L^2(P_Y)$) of the following two minimization problems

$$E_n[X \mid Y] = \operatorname*{argmin}_{q(Y)\in\mathscr{P}_n(Y)} \mathbb{E}\left[(q(Y) - \mathbb{E}[X \mid Y])^2\right] \tag{4.4}$$

$$= \operatorname*{argmin}_{q(Y)\in\mathscr{P}_n(Y)} \mathbb{E}\left[(q(Y) - X)^2\right]. \tag{4.5}$$

Furthermore, we have that the PMMSE satisfies the equality $\mathrm{pmmse}_n(X \mid Y) = \mathbb{E}[(X - E_n[X \mid Y])^2]$. We show in the following result that the PMMSE indeed converges to the MMSE, provided that Y is light-tailed, and we also give an explicit formula for the PMMSE. Recall that Y is said to satisfy Carleman's condition if $\sum_{n=1}^{\infty} \mathbb{E}\left[Y^{2n}\right]^{-1/(2n)} = \infty$, which holds if, e.g., Y has a moment-generating function (MGF) [Sch17, Sec. 4.2]. For $n \in \mathbb{N}$, we denote the n-th order Hankel matrix[2] of moments of Y by $\boldsymbol{M}_{Y,n} := \left(\mathbb{E}\left[Y^{i+j}\right]\right)_{0\le i,j\le n}$.

Theorem 4.1. *If X has finite variance and Y satisfies Carleman's condition, then, as $n \to \infty$, we have the convergences $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$ in $L^2(P_Y)$-norm and* $\mathrm{pmmse}_n(X \mid Y) \searrow \mathrm{mmse}(X \mid Y)$. *Further, for each $n \in \mathbb{N}$, if $|\mathrm{supp}(Y)| > n$ then $E_n[X \mid Y] = \mathbb{E}\left[(X, XY, \cdots, XY^n)\right] \boldsymbol{M}_{Y,n}^{-1} \left(1, Y, \cdots, Y^n\right)^T$.*

Proof. We may assume Y has infinite support, for otherwise we would have $\mathbb{E}[X \mid Y] \in \mathscr{P}_{|\mathrm{supp}(Y)|-1}(Y)$ and $\mathbb{E}[X \mid Y] = E_n[X \mid Y]$ for every $n \ge |\mathrm{supp}(Y)| - 1$. Since Y satisfies Carleman's condition, polynomials are dense in $L^2(P_Y)$ [Sch17, Sec. 4.2]. Let $\{p_j(Y) \in \mathscr{P}_j(Y)\}_{j\in\mathbb{N}}$ be the complete orthonormal set in $L^2(P_Y)$ that results from applying Gram-Schmidt orthonormalization to the monomials $\{Y^j\}_{j\in\mathbb{N}}$. By definition of $E_n[X \mid Y]$ as the orthogonal projection of $\mathbb{E}[X \mid Y]$ onto $\mathscr{P}_n(Y)$, we have that $E_n[X \mid Y] = \sum_{j=0}^{n} \langle \mathbb{E}[X \mid Y], p_j(Y)\rangle p_j(Y)$. The $L^2(P_Y)$-norm convergence $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$ follows. Furthermore, by the orthogonality principle of $\mathbb{E}[X \mid Y]$, we have that

$$\mathrm{pmmse}_n(X, t) - \mathrm{mmse}(X, t) = \mathbb{E}\left[(E_n[X \mid Y] - \mathbb{E}[X \mid Y])^2\right]. \tag{4.6}$$

[2] Hankel matrices are square matrices with constant skew diagonals.

Since $\mathscr{P}_0(Y) \subset \mathscr{P}_1(Y) \subset \cdots$, we deduce the monotone convergence $\text{pmmse}_n(X \mid Y) \searrow \text{mmse}(X \mid Y)$ from the $L^2(P_Y)$ convergence $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$. Finally, the formula for $E_n[X \mid Y]$ is shown in Lemma 4.1. □

Remark 4.1. The convergences in Theorem 4.1 are stated for Y that is not necessarily a Gaussian perturbation of X. In general, when stating the results of this chapter we do not make an implicit assumption on the relationship between X and Y.

We investigate the PMMSE in more detail in the case when Y is the output of a Gaussian channel whose input is X, i.e., $Y = \sqrt{t}X + N$ where $N \sim \mathcal{N}(0,1)$ is independent of X and $t \geq 0$ is constant. In this case, we show the following rationality of the PMMSE in signal-to-noise ratio (SNR), t. We use the shorthand

$$\text{pmmse}_n(X,t) := \text{pmmse}_n(X \mid \sqrt{t}X + N). \tag{4.7}$$

Theorem 4.2. *Fix $n \in \mathbb{N}_{>0}$ and a random variable X satisfying $\mathbb{E}\left[X^{2n}\right] < \infty$. The mapping $t \mapsto \text{pmmse}_n(X,t)$ over $[0,\infty)$ is a rational function, with leading coefficients given by*

$$\text{pmmse}_n(X,t) = \frac{\sigma_X^2 G(n+2) + \cdots + (\det \boldsymbol{M}_{X,n})t^{d_n-1}}{G(n+2) + \left(\sigma_X^2 G(n+2) d_n\right) t + \cdots + (\det \boldsymbol{M}_{X,n})t^{d_n}}, \tag{4.8}$$

where $d_n := \binom{n+1}{2}$ and $G(k) := \prod_{j=1}^{k-2} j!$ (for integers $k \geq 1$) is the Barnes G-function [Ada01]. Further, each coefficient in the numerator or denominator of $\text{pmmse}_n(X,t)$ is a multivariate polynomial in $(\mathbb{E}[X], \cdots, \mathbb{E}[X^{2n}])$.

Proof. See Section 4.3.1 and Appendix C.2. □

Remark 4.2. The PMMSE definition naturally generalizes to random vectors, where orthogonal projection is then done over spaces of multivariate polynomials. In this case, if $\boldsymbol{X}$ is an m-dimensional random vector that is independent of $\boldsymbol{N} \sim \mathcal{N}(0, \boldsymbol{I}_m)$, the leading terms in the PMMSE formula become

$$\text{pmmse}_n(\boldsymbol{X},t) = \frac{(\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{X}}) \det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + (\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{N}})\left(\det \boldsymbol{M}_{\boldsymbol{X},n}\right)\ t^{d_{n,m}-1}}{\det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + \left(\det \boldsymbol{M}_{\boldsymbol{X},n}\right)\ t^{d_{n,m}}}, \tag{4.9}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{N}}$ are the covariance matrices and $d_{n,m} = m\binom{n+m}{m+1}$; the matrices $\boldsymbol{M}_{\boldsymbol{X},n}$ and $\boldsymbol{M}_{\boldsymbol{N},n}$ are also natural generalizations of the real-valued case, see Appendix C.4 for the details.

The intermediate terms in the rational function $\text{pmmse}_n(X,t)$ can also be given explicitly via Theorem 4.1. For example, if X is zero-mean and unit-variance, denoting $\mathcal{X}_k = \mathbb{E}[X^k]$, we have the

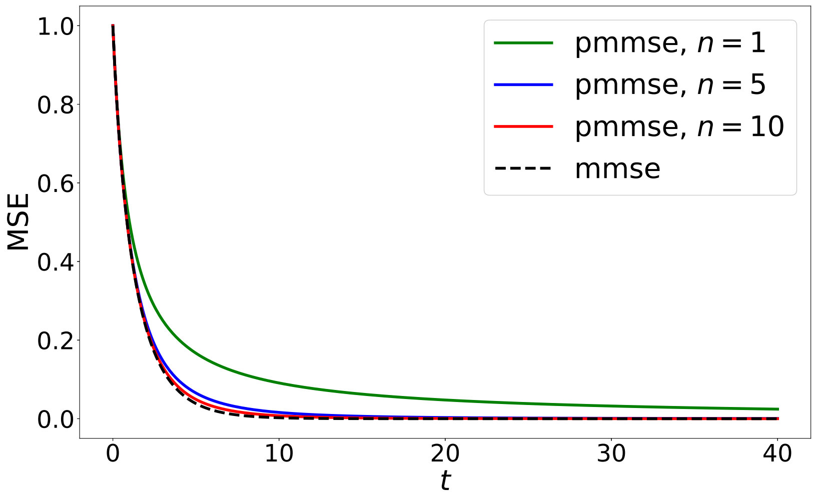


Figure 4.1: *Comparison of the graphs of the functions* $t \mapsto \mathrm{pmmse}_n(X,t)$ *(solid lines) against the function* $t \mapsto \mathrm{mmse}(X,t)$ *(dashed black line) for* $n \in \{1,5,10\}$ *and* $X \sim \mathrm{Unif}(\{\pm 1\})$*.*

formula

$$\mathrm{pmmse}_2(X,t) = \frac{2+4t+(\mathcal{X}_4-\mathcal{X}_3^2-1)t^2}{2+6t+(\mathcal{X}_4+3)t^2+(\mathcal{X}_4-\mathcal{X}_3^2-1)t^3}. \tag{4.10}$$

For a general $n \in \mathbb{N}$, the coefficients in both the numerator and denominator of the PMMSE in (4.8) are “homogeneous" polynomials in the moments of X (i.e., for a single coefficient $c(X)$ there is a $k_c \in \mathbb{N}$ such that $c(\alpha X) = \alpha^{k_c} c(X)$).

The expression (4.8) of the PMMSE in terms of moments gives a simple yet powerful method for approximating the MMSE. Figure 4.1 shows an example of how the PMMSE approximates the MMSE for a random variable X that takes the values 1 and -1 equiprobably, where we are also using the shorthand $\mathrm{mmse}(X,t) := \mathrm{mmse}(X \mid \sqrt{t}X + N)$ for $N \sim \mathcal{N}(0,1)$ independent of X. In this case, the MMSE is given by

$$\mathrm{mmse}(X,t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \tanh(z\sqrt{t})^2 e^{-(z+\sqrt{t})^2/2}\, dz, \tag{4.11}$$

whereas the functions $\mathrm{pmmse}_n(X,t)$ are rational in t, e.g., for $n = 1$ we have the LMMSE $\mathrm{pmmse}_1(X,t) = 1/(1+t)$, and for $n = 5$ we have the 5-th degree PMMSE[3]

$$\mathrm{pmmse}_5(X,t) = \frac{45+360t+675t^2+300t^3}{45+405t+1035t^2+1005t^3+450t^4+96t^5+8t^6}. \tag{4.12}$$

[3] In general, $\mathrm{pmmse}_5(Z,t)$ is a ratio of a degree-14 polynomial by a degree-15 polynomial as in equation (4.8). In the special case of a Rademacher random variable, significant cancellations occur and we obtain equation (4.12).

Comparing the curves in Figure 4.1 hints that the convergence $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ is uniform in the SNR t; note that the corresponding pointwise convergence is an immediate corollary of Theorem 4.1. We show in the following result that the PMMSE indeed converges uniformly to the MMSE provided that X has a MGF (i.e., its MGF is finite over a neighborhood of the origin). We also show, under additional assumptions on the distribution of X, that for fixed t the pointwise-convergence rate of $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ is faster than any polynomial in n.

Theorem 4.3. *If the MGF of a random variable X exists,*[4] *then we have the uniform and monotone convergence*

$$\sup_{t\geq 0} \ \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) \searrow 0 \tag{4.13}$$

as $n \to \infty$. If, in addition, X has a probability density function or a probability mass function p_X that is compactly-supported, even, and decreasing over $[0,\infty) \cap \mathrm{supp}(p_X)$, then for all $k, t \geq 0$ we have that

$$\lim_{n\to\infty} n^k \cdot (\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)) = 0. \tag{4.14}$$

Proof. See Section 4.3.2 and Appendix C.3. □

Remark 4.3. By the orthogonality property of the conditional expectation, we have the equality of approximation errors

$$\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) = \mathbb{E}\left[\left(E_n[X \mid \sqrt{t}X + N] - \mathbb{E}[X \mid \sqrt{t}X + N]\right)^2\right], \tag{4.15}$$

where $N \sim \mathcal{N}(0,1)$ is independent of X. Thus, the convergence rate (4.14) is equivalent to

$$\lim_{n\to\infty} n^k \mathbb{E}\left[\left(E_n[X \mid \sqrt{t}X + N] - \mathbb{E}[X \mid \sqrt{t}X + N]\right)^2\right] = 0. \tag{4.16}$$

Equipped with the PMMSE functional, we are able to derive new formulas for differential entropy and mutual information in terms of moments. A corollary of the I-MMSE relation states that the differential entropy of a finite-variance continuous random variable X can be expressed in terms of the MMSE as [GSV05]

$$h(X) = \frac{1}{2}\int_0^\infty \mathrm{mmse}(X,t) - \frac{1}{2\pi e + t}\, dt. \tag{4.17}$$

Naturally, we consider the functionals obtained by replacing the MMSE with the PMMSE, which we

[4] The assumption that the MGF of X exists is imposed so that $\sqrt{t}X + N$ satisfies Carleman's condition (for $N \sim \mathcal{N}(0,1)$ independent of X and $t \geq 0$ fixed), which holds because $\sqrt{t}X + N$ will then have a MGF. It is not true in general that Carleman's condition is satisfied by the sum of two independent random variables each satisfying Carleman's condition, see [Ber85, Proposition 3.1].

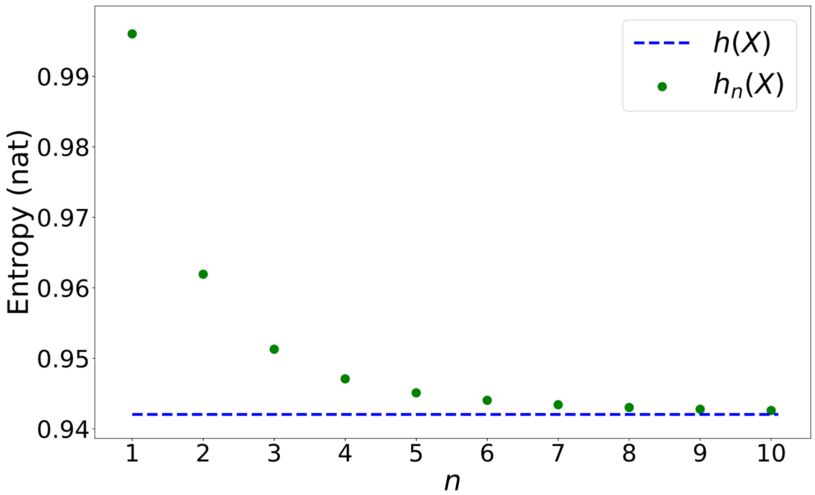


Figure 4.2: *Comparison of the values of $h_n(X)$ (green dots) against the true value $h(X)$ (dashed blue line) for $n \in \{1, \cdots, 10\}$ and $X \sim \chi_2$. We have that $h(X) < h_{10}(X) < h(X) + 6 \cdot 10^{-4}$.*

show converge to the differential entropy monotonically from above.

Theorem 4.4. *Let $\boldsymbol{X}$ be a continuous m-dimensional random vector whose MGF exists. Consider the functionals*

$$h_n(\boldsymbol{X}) := \frac{1}{2}\int_0^\infty \mathrm{pmmse}_n(\boldsymbol{X}, t) - \frac{m}{2\pi e + t}\, dt \tag{4.18}$$

for each $n \in \mathbb{N}_{>0}$. Then, we have a decreasing sequence

$$h(\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{X}})) = \frac{1}{2}\log\left((2\pi e)^m \det \boldsymbol{\Sigma}_{\boldsymbol{X}}\right) \tag{4.19}$$

$$= h_1(\boldsymbol{X}) \geq h_2(\boldsymbol{X}) \geq \cdots \geq h(\boldsymbol{X}) \tag{4.20}$$

converging to the differential entropy, $h_n(\boldsymbol{X}) \searrow h(\boldsymbol{X})$.

Proof. See Section 4.4 and Appendix C.5.2. □

Figure 4.2 illustrates how $h_n(X)$ approximates $h(X)$, where X has a chi distribution with two degrees of freedom (commonly denoted by χ_2). It is evident from the figure that $h_n(X)$ approximates the differential entropy of X monotonically more accurately as n grows; indeed, this is true in general in view of the monotonicity of the convergence $\mathrm{pmmse}_n(X \mid Y) \searrow \mathrm{mmse}(X \mid Y)$ as in Theorem 4.1.

A noteworthy implication of Theorem 4.4 is that it gives a formula for the differential entropy $h(X)$ that, in view of Theorem 4.2, is entirely in terms of the moments of X. Furthermore, closure properties of polynomial subspaces under affine transformations imply that the PMMSE behaves under affine transformations exactly as the MMSE does: if $\mathbb{E}[X^2], \mathbb{E}[Y^{2n}] < \infty$ then

$$\text{pmmse}_n(aX+b \mid cY+d) = a^2 \,\text{pmmse}_n(X \mid Y) \tag{4.21}$$

for all constants a, b, c, and d such that $c \neq 0$ (Lemma 4.2). Thus, the distribution functionals h_n behave under affine transformations exactly as differential entropy does, namely, if $\mathbb{E}[X^{2n}] < \infty$ then

$$h_n(aX+b) = h_n(X) + \log|a| \tag{4.22}$$

for $a \neq 0$ (Corollary 4.2).

The moment-based differential entropy formula in Theorem 4.4 gives rise to formulas of mutual information primarily in terms of moments.

Theorem 4.5. *If the mutual information $I(X;Y)$ exists (but possibly infinite), then it can be written in terms of the underlying moments in the following two cases:*

1. *Suppose X is discrete with finite support, and Y is continuous whose MGF exists and that satisfies $h(Y) > -\infty$. Then, letting $Y^{(x)}$ denote the random variable obtained from Y by conditioning on $\{X = x\}$, we have*

$$I(X;Y) = \frac{1}{2} \lim_{n\to\infty} \int_0^\infty \text{pmmse}_n(Y,t) - \mathbb{E}_X\left[\text{pmmse}_n(Y^{(X)},t)\right] \, dt. \tag{4.23}$$

2. *Suppose that X and Y are continuous whose MGFs exist and that satisfy $h(X), h(Y) > -\infty$. Suppose also that $I(X;Y) < \infty$ or else (X,Y) is not continuous. Then,*

$$I(X;Y) = \frac{1}{2} \lim_{n\to\infty} \int_0^\infty \text{pmmse}_n(X,t) + \text{pmmse}_n(Y,t) - \text{pmmse}_n((X,Y),t) \, dt. \tag{4.24}$$

Proof. See Section 4.4 and Appendix C.5.3. □

One result that helps in the proof of Theorem 4.5 in the second scenario is the following generalization of the MMSE dimension to random vectors.

Theorem 4.6. *Fix two square-integrable continuous m-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{N}$ that are independent. Suppose that $p_{\boldsymbol{N}}$ is bounded and*[5] *$p_{\boldsymbol{N}}(\boldsymbol{z}) = O\left(\|\boldsymbol{z}\|^{-(m+2)}\right)$ as $\|\boldsymbol{z}\| \to \infty$. Then, we have that*

$$\lim_{t\to\infty} t \cdot \text{mmse}\left(\boldsymbol{X} \mid \sqrt{t}\boldsymbol{X} + \boldsymbol{N}\right) = \text{tr}\, \boldsymbol{\Sigma}_{\boldsymbol{N}}. \tag{4.25}$$

[5] The exponent $m+2$ in the decay rate may be replaced with $m+1+\varepsilon$ for any $\varepsilon > 0$, see [SS19, Section 3.2]

Proof. This result follows by a straightforward extension of the proof for the one-dimensional case given in [WV11]; see [AC21a, Appendix I] for the full details. □

We introduce new estimators of information measures by approximating the PMMSE in (4.8) via plugging in sample moments in place of moments. If $\{X_j\}_{j=1}^m$ are i.i.d. samples taken from the distribution of X, then a uniform random variable over the samples $U \sim \text{Unif}(\{X_j\}_{j=1}^m)$ provides an estimate $\text{pmmse}_n(U,t)$ for $\text{pmmse}_n(X,t)$. The moments of U converge to the moments of X by the law of large numbers. Further, using $\text{pmmse}_n(U,t)$ to estimate $\text{pmmse}_n(X,t)$ is a consistent estimator by the continuous mapping theorem, as the PMMSE is a continuous function of the moments. The same can be said of $h_n(U)$ as an estimate of $h_n(X)$, or of $I_n(U;V)$ as an estimate of $I(X;Y)$ when $(U,V) \sim \text{Unif}(\{(X_j,Y_j)\}_{j=1}^m)$ where $\{(X_j,Y_j)\}_{j=1}^m$ are i.i.d. samples drawn according to the distribution of (X,Y) (where I_n is the functional given by the expressions inside the limits in Theorem 4.5). These estimators also satisfy some desirable properties. For example, the behavior of the PMMSE under affine transformations (4.21) implies that the estimate of the PMMSE from data is robust to (injective) affine transformations, the functionals h_n behave under affine transformations exactly as differential entropy does, and the same is true for I_n and I.

The rest of the chapter is organized as follows. We introduce the PMMSE, provide an explicit formula for it, prove its convergence to the MMSE (Theorem 4.1), and exhibit some of its properties in Section 4.2. A more detailed treatment of the Gaussian-channel case occupies Section 4.3. Specifically, we show rationality of the PMMSE (Theorem 4.2) in Section 4.3.1, then prove the uniform convergence of the PMMSE to the MMSE and bound the pointwise-convergence rate (Theorem 4.3) in Section 4.3.2. Building on the derived results about the PMMSE, we prove new moments-based formulas for differential entropy and mutual information in Section 4.4. Our formulas then give rise to a new estimator that we introduce in Section 4.5, where simulations also illustrate the estimator's performance.

4.1.1 Related Literature

The mutual information between the input and output of the Gaussian channel is known to have an integral relation with the MMSE, referred to in the literature as the I-MMSE relation. This connection was made in the work of Guo, Shamai, and Verdú in [GSV05]. Extensions of the I-MMSE relation were investigated in [Zak05, Guo09, Ver10, GWSV11, WV12, AVW14, HJW15, DBPS17, DV20], and applications have been established, e.g., in optimal power allocation [LTV06] and monotonicity of non-Gaussianness [TV06]. Our work is inscribed within this literature.

We introduce the PMMSE approximation of the MMSE, derive new representations of distribution functionals in terms of moments, and introduce estimators based on these new representations. We note that utilizing higher-order polynomials as proxies of the MMSE has appeared, e.g., in approaches to denoising [CM18]. Works such as [DSGW03] and [Don88] show some impossibility results for estimating the MMSE in the general case. Recent work by Diaz et al. [DKS21] gives lower bounds for the MMSE via estimating by neural networks. Also, studying smoothed distributions, e.g., via convolutions with Gaussians, has generated recent interest in the context of information theory [CPW18, PW16] and learning theory [GGWP19, GGNWP20].

At the heart of our work is the Bernstein approximation problem, on which a vast literature exists within approximation theory. The original Bernstein approximation problem extends Weierstrass approximation to the whole real line by investigating whether polynomials are dense in $L^\infty(\mu)$ for a measure μ that is absolutely continuous with respect to the Lebesgue measure. Works such as those by Carleson [Car51] and Freud [Fre77], and eventually the more comprehensive solution given by Ditzian and Totik [DT87]—which introduces moduli of smoothness, a natural extension of the modulus of continuity—show that tools used to solve the Bernstein approximation problem can be useful for the more general question of denseness of polynomials in $L^p(\mu)$ for all $p \geq 1$ (see [Lub07] for a comprehensive survey). In particular, the case $p = 2$ has a close relationship with the Hamburger moment problem, described next.

The Hamburger moment problem asks whether a countably-infinite sequence of real numbers corresponds uniquely to the moments of a positive Borel measure on $\mathbb{R}$. A connection between this problem and the Bernstein approximation problem is that if the Hamburger moment problem has a positive answer for the sequence of moments of μ then polynomials are dense in $L^2(\mu)$, see [BC81]. In the context of information theory, the application of the Bernstein approximation problem and the Hamburger moment problem has appeared in [MZ17].

The denominator of the PMMSE in Gaussian channels, which is given by $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$, as well as the leading coefficient of both the numerator and the denominator, $\det \boldsymbol{M}_{X,n}$, can be seen as generalizations of the Selberg integral. Denote

$$\mathcal{I}_n(\varphi) = \int_{\mathbb{R}^{n+1}} \prod_{0\leq i<j\leq n} (y_i - y_j)^2 \prod_{i=0}^{n} \varphi(y_i) \, dy_0 \cdots dy_n. \tag{4.26}$$

If φ is the PDF of a Beta distribution or a standard normal distribution, then $\mathcal{I}_n(\varphi)$ is the Selberg integral or the Mehta integral, respectively (both with parameter $\gamma = 1$) [FW08]. For a continuous random variable Y whose PDF is p_Y,

$$\det \boldsymbol{M}_{Y,n} = \frac{1}{(n+1)!}\mathcal{I}_n(p_Y). \tag{4.27}$$

The Vandermonde-determinant power $\prod_{i<j}(y_i - y_j)^2$ in the integrand in (4.26) bears a close connection with the quantum Hall effect [STW94, KTW04]. The connection arises via expanding powers of the Vandermonde determinant and investigating which of the ensuing monomials have nonzero coefficients.

We quantify the rate of convergence of the PMMSE to the MMSE in Theorem 4.3, for which the key ingredient is the bound in Lemma 4.9 on the derivatives of the conditional expectation. The first-order derivative of the conditional expectation in Gaussian channels has been treated in [DPS20]. We note that in parallel to this work the authors were made aware that the higher-order derivative expressions in Proposition 4.3 were also derived in [DPS21]. We also extend the proofs for the MMSE dimension in the continuous case as given in [WV11] to higher dimensions.

Distribution functionals, such as mutual information, are popular metrics for quantifying associations between data (e.g., [GNO$^+$12, CLA$^+$10, Fle04]), yet reliably estimating distributional functions directly from samples is a non-trivial task. The naive route of first estimating the underlying distribution is generally impractical and imprecise. To address this challenge, a growing number of distribution functionals' estimators have recently been proposed within the information theory and computer science communities (see, e.g., [KSG04, VV11, JVHW15, WY16, GKOV17]). The estimators we propose satisfy desirable properties, such as shift invariance and scale resiliency, without the need to estimate the underlying distributions.

4.1.2 Notation

Throughout, we fix a probability space $(\Omega, \mathcal{F}, P)$. For $q \geq 1$, the Banach space $L^q(P)$ consists of all q-times integrable real-valued random variables with norm denoted by $\| \cdot \|_q$. The Borel probability measure induced by a random variable Y is denoted by P_Y. The subspace $L^q(P_Y) \subset L^q(P)$ consists of q-times integrable and $\sigma(Y)$-measurable random variables. The inner product of $L^2(P)$ is denoted by $\langle \cdot , \cdot \rangle$. The Banach space $L^q(\mathbb{R})$ consists of all q-times Lebesgue integrable functions from $\mathbb{R}$ to itself, with norm denoted by $\| \cdot \|_{L^q(\mathbb{R})}$. We say that Y has a moment-generating function (MGF) if $\mathbb{E}[e^{tY}] < \infty$ over some nonempty interval $t \in (-\delta, \delta)$. We let $\mathrm{supp}(Y)$ denote the support of Y. We denote the cardinality of a set S by $|S|$, and say that Y has *infinite support* if $|\mathrm{supp}(Y)| = \infty$. If $\mathbb{E}\left[Y^{2n}\right] < \infty$, we denote the Hankel matrix of moments by $\boldsymbol{M}_{Y,n} := \left(\mathbb{E}\left[Y^{i+j}\right]\right)_{0 \le i,j \le n}$. We denote the random vector $\boldsymbol{Y}^{(n)} := (1, Y, \cdots, Y^n)^T$. Note that $\boldsymbol{M}_{Y,n}$ is the expectation of the outer product of $\boldsymbol{Y}^{(n)}$, i.e., $\boldsymbol{M}_{Y,n} = \mathbb{E}\left[\boldsymbol{Y}^{(n)}\left(\boldsymbol{Y}^{(n)}\right)^T\right]$. Therefore, $\boldsymbol{M}_{Y,n}$ is a rank-1 perturbation of the covariance matrix of $\boldsymbol{Y}^{(n)}$, denoted $\boldsymbol{\Sigma}_{\boldsymbol{Y}^{(n)}}$. We let $\mathscr{P}_n$ denote the collection of all polynomials of degree at most n with real coefficients, and we set $\mathscr{P}_n(Y) := \{q(Y) \ ; \ q \in \mathscr{P}_n\}$. For $n \in \mathbb{N}$, we set $[n] := \{0, 1, \cdots, n\}$. Vectors are denoted by boldface letters, in which case subscripted regular letters refer to the entries. The $n \times n$ identity matrix is denoted by $\boldsymbol{I}_n$. The closure of a set S will be denoted by $\overline{S}$. We use the shorthand $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$, and the notation $\mathcal{Y}_k$ is defined analogously.

4.2 Polynomial MMSE

We give in this section a brief overview of the Polynomial MMSE (PMMSE). The PMMSE, introduced in Definition 4.1, can be characterized in two equivalent ways: it is the orthogonal projection onto subspaces of polynomials of bounded degree, and it is also a natural generalization of the Linear MMSE (LMMSE) to higher-degree polynomials. Recall that standard results on orthogonal projections in Hilbert spaces (see, e.g., [SS19, Section 4.4]) yield that the minimum in (4.3) is always attained, and that the polynomials $\boldsymbol{c}^T \boldsymbol{Y}^{(n)}$ represent the same element of $\mathscr{P}_n(Y)$ for all minimizers $\boldsymbol{c}$ of (4.3). In other words, the PMMSE estimate $E_n[X \mid Y]$ as given by Definition 4.1 is a well-defined element in $\mathscr{P}_n(Y) \subset L^2(P_Y)$.[6]

Unlike the case of the MMSE, working with the PMMSE is tractable and allows for explicit formulas. For instance, the PMMSE in Gaussian channels is a rational function in the SNR; more precisely, the formula for $t \mapsto \mathrm{pmmse}_n(X \mid \sqrt{t}X + N)$ stated in Theorem 4.2 reveals that this mapping is a rational function of t (where $N \sim \mathcal{N}(0,1)$ is independent of X). In addition, as shown in Theorem 4.1, we have the strong convergence (i.e., in the strong operator topology) of orthogonal projection operators $E_n[\,\cdot \mid Y] \to \mathbb{E}[\,\cdot \mid Y]$ provided that polynomials in Y are dense in $L^2(P_Y)$.

[6] Uniqueness of the minimizing polynomial $E_n[X \mid Y]$ should not be confused with the possible non-uniqueness of the vector $\boldsymbol{c} \in \mathbb{R}^{n+1}$ in the relation $E_n[X \mid Y] = \boldsymbol{c}^T \boldsymbol{Y}^{(n)}$. For example, if Y is binary and $n = 2$, then $Y^2 = Y$, so for any $c_0, c_1, c_2 \in \mathbb{R}$ for which $E_2[X \mid Y] = c_0 + c_1 Y + c_2 Y^2$ we also have $E_2[X \mid Y] = c_0 + (c_1 - 1)Y + (c_2 + 1)Y^2$. In particular, there is no unique quadratic $p \in \mathscr{P}_2$ for which $E_2[X \mid Y] = p(Y)$. Nevertheless, in the problems of interest to us, uniqueness of $\boldsymbol{c}$ is also attained (e.g., if Y is continuous); in fact, $\boldsymbol{c}$ is unique if and only if $|\mathrm{supp}(Y)| > n$ holds.

4.2.1 PMMSE Formula

We show next explicit PMMSE formulas. We build on these formulas in the next section to prove the rationality of $t \mapsto \text{pmmse}_n(X,t)$ stated in Theorem 4.2, which in turn will simplify the proof of consistency of the estimators for information measures introduced in Section 4.5.

Lemma 4.1. *Fix $n \in \mathbb{N}$, and let X and Y be random variables such that $\mathbb{E}\left[X^2\right], \mathbb{E}\left[Y^{2n}\right] < \infty$. We have that $\boldsymbol{M}_{Y,n}$ is invertible if and only if $|\text{supp}(Y)| > n$. Further, if it is the case that $|\text{supp}(Y)| > n$, then the PMMSE estimator is given by*

$$E_n[X \mid Y] = \mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\boldsymbol{Y}^{(n)}, \tag{4.28}$$

and the PMMSE by

$$\text{pmmse}_n(X \mid Y) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right], \tag{4.29}$$

which then satisfy the relation

$$\text{pmmse}_n(X \mid Y) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[XE_n[X \mid Y]\right]. \tag{4.30}$$

Proof. See Appendix C.1.1. □

To expound on the formulas given by Lemma 4.1, we instantiate them next for the cases $n \in \{1,2\}$. By definition of the PMMSE, these expressions recover the LMMSE and "quadratic" MMSE. Polynomial regression is also shown below to be an instantiation of the PMMSE.

Example 1. For $n = 1$, if $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$ and $|\text{supp}(Y)| > 1$, we have from (4.28) that

$$E_1[X \mid Y] = (\mathbb{E}[X], \mathbb{E}[XY]) \begin{pmatrix} 1 & \mathbb{E}[Y] \\ \mathbb{E}[Y] & \mathbb{E}\left[Y^2\right] \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ Y \end{pmatrix}. \tag{4.31}$$

Computing the matrix inverse and multiplying out, we obtain

$$E_1[X \mid Y] = \mathbb{E}[X] + \frac{\text{cov}(X,Y)}{\sigma_Y^2}\left(Y - \mathbb{E}[Y]\right), \tag{4.32}$$

where $\text{cov}(X,Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is the covariance between X and Y. Formula (4.32) indeed gives the LMMSE estimate. Via the relation in (4.30), we recover

$$\text{pmmse}_1(X \mid Y) = \sigma_X^2 - \frac{\text{cov}(X,Y)^2}{\sigma_Y^2} = \sigma_X^2 \cdot (1 - \rho_{X,Y}^2), \tag{4.33}$$

with $\rho_{X,Y} := \mathrm{cov}(X,Y)/(\sigma_X \sigma_Y)$ the Pearson correlation coefficient between X and Y (when $\sigma_X \neq 0$). Formula (4.33) verifies that $\mathrm{pmmse}_1(X \mid Y)$ is the LMMSE. □

Example 2. We will use the notation $\mathcal{Y}_k := \mathbb{E}\left[Y^k\right]$ for short. For $n = 2$, and assuming $\mathbb{E}[X^2], \mathbb{E}[Y^4] < \infty$ and $|\mathrm{supp}(Y)| > 2$, Lemma 4.1 gives the quadratic $E_2[X \mid Y] = \frac{\alpha_0}{\delta} + \frac{\alpha_1}{\delta}Y + \frac{\alpha_2}{\delta}Y^2$ where

$$\alpha_0 = (\mathcal{Y}_2\mathcal{Y}_4 - \mathcal{Y}_3^2)\mathbb{E}[X] + (\mathcal{Y}_2\mathcal{Y}_3 - \mathcal{Y}_1\mathcal{Y}_4)\mathbb{E}[XY] + (\mathcal{Y}_1\mathcal{Y}_3 - \mathcal{Y}_2^2)\mathbb{E}[XY^2] \tag{4.34}$$

$$\alpha_1 = (\mathcal{Y}_2\mathcal{Y}_3 - \mathcal{Y}_1\mathcal{Y}_4)\mathbb{E}[X] + (\mathcal{Y}_4 - \mathcal{Y}_2^2)\mathbb{E}[XY] + (\mathcal{Y}_1\mathcal{Y}_2 - \mathcal{Y}_3)\mathbb{E}[XY^2] \tag{4.35}$$

$$\alpha_2 = (\mathcal{Y}_1\mathcal{Y}_3 - \mathcal{Y}_2^2)\mathbb{E}[X] + (\mathcal{Y}_1\mathcal{Y}_2 - \mathcal{Y}_3)\mathbb{E}[XY] + (\mathcal{Y}_2 - \mathcal{Y}_1^2)\mathbb{E}[XY^2] \tag{4.36}$$

and

$$\delta = \mathcal{Y}_2\mathcal{Y}_4 - \mathcal{Y}_1^2\mathcal{Y}_4 - \mathcal{Y}_2^3 - \mathcal{Y}_3^2 + 2\mathcal{Y}_1\mathcal{Y}_2\mathcal{Y}_3. \tag{4.37}$$

Note that $\delta = \det \boldsymbol{M}_{Y,2} \neq 0$ by Lemma 4.1. Relation (4.30) then yields the formula

$$\mathrm{pmmse}_2(X \mid Y) = \mathbb{E}\left[X^2\right] - \delta^{-1}\sum_{k=0}^{2} \alpha_k \mathbb{E}\left[XY^k\right]. \tag{4.38}$$

□

Example 3. Finding the PMMSE estimate can be seen as a generalization of modeling via polynomial regression. The goal of single-variable polynomial regression is to model a random variable X as a polynomial in a random variable Y, i.e., $X = \beta_0 + \beta_1 Y + \cdots + \beta_n Y^n + \varepsilon$ for a modeling-error random variable ε and constants β_j to be determined from data. Given access to samples $\{(x_i, y_i)\}_{i=1}^m$, this model leads to the equation $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X} = (x_1, \cdots, x_m)^T$, $\boldsymbol{Y} = (y_i^j)_{i\in\{1,\cdots,m\},j\in[n]}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_m)^T$ where the ε_j are samples from ε, and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_n)^T$. It is assumed that the number of distinct y_i is strictly larger than n, so $\boldsymbol{Y}$ has full column-rank. A value of $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{\varepsilon}\|$ is known from polynomial regression to be $\boldsymbol{\beta}^T = \boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}$. This formula follows from the PMMSE estimate formula in Lemma 4.1. Indeed, minimizing $\|\boldsymbol{\varepsilon}\|$ in polynomial regression amounts to finding the PMMSE estimate $E_n[U \mid V]$, where $(U,V) \sim \mathrm{Unif}(\{(x_i,y_i)\}_{i=1}^m)$. By the PMMSE formula in Lemma 4.1, we have that

$$\boldsymbol{\beta}^T = \mathbb{E}\left[U\boldsymbol{V}^{(n)}\right]^T \boldsymbol{M}_{V,n}^{-1} \tag{4.39}$$

By definition of (U,V), we also have that $\boldsymbol{X}^T\boldsymbol{Y} = m\mathbb{E}\left[U\boldsymbol{V}^{(n)}\right]^T$ and $(\boldsymbol{Y}^T\boldsymbol{Y})^{-1} = \frac{1}{m}\boldsymbol{M}_{V,n}^{-1}$. Multiplying the latter two equations together, we obtain $\boldsymbol{\beta}^T = \boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}$ in view of (4.39). To sum up, the polynomial regression approach solves the restricted problem of finding the PMMSE $E_n[X' \mid Y']$ when both X' and Y' are discrete with PMFs that evaluate to rational numbers, i.e., when the distribution of (X', Y') is uniform over a finite dataset $\{(x_i', y_i')\}_{i=1}^m$. □

Remark 4.4. We note that $E_n[\,\cdot \mid Y]$ is not in general a conditional expectation operator, i.e., there are some $n \in \mathbb{N}$ and $Y \in L^{2n}(P)$ such that for every sub-σ-algebra $\Sigma \subset \mathcal{F}$ we have $E_n[\,\cdot \mid Y] \neq \mathbb{E}[\,\cdot \mid \Sigma]$. One way to see this is that $E_n[\,\cdot \mid Y]$ might not preserve positivity. For example, if $X \sim \mathrm{Unif}(0,1)$ and $Y = X + N$ for $N \sim \mathcal{N}(0,1)$ independent of X, we have that $E_1[X \mid Y] = (Y+6)/13$ (see (4.32)). Therefore, the probability that $E_1[X \mid Y] < 0$ is $P_Y((-\infty, -6)) > 0$. In other words, although X is non-negative, $E_1[X \mid Y]$ is not; in contrast, $\mathbb{E}[X \mid \Sigma]$ is non-negative for every sub-σ-algebra $\Sigma \subset \mathcal{F}$.

Remark 4.5. We may define the pointwise PMMSE estimate $E_n[X \mid Y = y]$ for $y \in \mathrm{supp}(Y)$ by the equation $E_n[X \mid Y = y] := \sum_{j \in [n]} c_j y^j$ where $\boldsymbol{c} = (c_0, \cdots, c_n)^T$ is any minimizer in (4.3), and a direct verification shows that this makes $E_n[X \mid Y = y]$ well-defined.

Remark 4.6. The PMMSE, as we introduce it in Definition 4.1, can be equivalently written in vector LMMSE notation as $\mathrm{pmmse}_n(X \mid Y) = \mathrm{lmmse}(X \mid \boldsymbol{Y}^{(n)})$. However, even when the channel producing Y from X is additive, the same might not be true of that producing $\boldsymbol{Y}^{(n)}$ from X. For example, if $Y = X + N$, then Y^2 contains the cross term XN. For this reason, we use the introduced PMMSE notation in place of the vector LMMSE notation.

4.2.2 PMMSE Properties

We investigate next the behavior of the PMMSE under affine transformations, and exhibit a few additional properties of the PMMSE that parallel those of the MMSE. The behavior of the PMMSE under affine transformations, shown in Lemma 4.2 below, has desirable implications on the moments-based approximations of differential entropy and mutual information that we introduce in Section 4.4. For example, recall that differential entropy satisfies $h(aY + b) = h(Y) + \log|a|$ for any $a, b \in \mathbb{R}$ with $a \neq 0$. Because of Lemma 4.2, the same property holds for the approximations h_n (as given by (4.18)), i.e., $h_n(aY + b) = h_n(Y) + \log|a|$.

For random variables X and Y such that $\mathbb{E}[X^2] < \infty$ and constants $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ such that $\gamma \neq 0$, one has $\mathrm{mmse}(\alpha X + \beta \mid \gamma Y + \delta) = \alpha^2 \mathrm{mmse}(X \mid Y)$ (see, e.g., [GSV05]). This property of the MMSE holds because $\mathrm{mmse}(\,\cdot \mid Y)$ measures the distance to $L^2(P_Y)$, which is a space that is invariant under (injective) affine transformations of Y. A similar reasoning yields an analogous property for the PMMSE.

Lemma 4.2. *Let X and Y be two random variables and $n \in \mathbb{N}$, and assume that both $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[Y^{2n}\right]$ are finite. For any $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ such that $\gamma \neq 0$, we have that $\operatorname{pmmse}_n(\alpha X + \beta \mid \gamma Y + \delta) = \alpha^2 \operatorname{pmmse}_n(X \mid Y)$.*

Proof. This property follows from the fact that $\mathscr{P}_n(aY + b) - c = \mathscr{P}_n(Y)$ for $a, b, c \in \mathbb{R}$ with $a \neq 0$. □

We show next that the operator $E_n[\,\cdot \mid Y]$ satisfies several properties analogously to the conditional expectation $\mathbb{E}[\,\cdot \mid Y]$. Note that the properties we derive for the PMMSE cannot be straightforwardly deduced from analogous properties that the conditional expectation satisfies, since $E_n[\,\cdot \mid Y]$ is not in general a conditional expectation operator (see Remark 4.4). Nevertheless, we have the following PMMSE operator properties.

Lemma 4.3. *For $n \in \mathbb{N}$ and random variables X, Y, and Z such that $\sigma_X, \sigma_Y, \mathbb{E}[Z^{2n}] < \infty$, the following hold:*

(i) Linearity: $E_n[aX + bY \mid Z] = aE_n[X \mid Z] + bE_n[Y \mid Z]$ for any $a, b \in \mathbb{R}$.

(ii) Invariance: $E_n[p(Z) \mid Z] = p(Z)$ for any $p \in \mathscr{P}_n$.

(iii) Idempotence: $E_n\left[E_n\left[X \mid Z\right] \mid Z\right] = E_n[X \mid Z]$.

(iv) Contractivity: $\|E_n[X \mid Z]\|_2 \leq \|X\|_2$.

(v) Self-Adjointness: $\mathbb{E}\left[E_n[X \mid Z]Y\right] = \mathbb{E}\left[XE_n[Y \mid Z]\right]$, i.e., $E_n[\,\cdot \mid Z]$ is self-adjoint.

(vi) Orthogonality: $\mathbb{E}[(X - E_n[X \mid Z])p(Z)] = 0$ for $p \in \mathscr{P}_n$, and $E_n[Y \mid Z] = 0$ if and only if $Y \in \mathscr{P}_n(Z)^\perp$.

(vii) Total expectation: $\mathbb{E}[E_n[X \mid Z]] = \mathbb{E}[X]$.

(viii) Independence: If X and Z are independent, then $E_n[X \mid Z] = \mathbb{E}[X]$.

(ix) Markov Chain: If X—Y—Z forms a Markov chain, then $E_n\left[\mathbb{E}[X \mid Y] \mid Z\right] = E_n[X \mid Z]$.

Proof. Properties (i)–(vi) follow immediately from the characterization of $E_n[\,\cdot \mid Z]$ as an orthogonal projection from $L^2(P)$ onto $\mathscr{P}_n(Z)$. Property (vii) follows from the first part of (vi) via linearity of expectation by choosing the constant polynomial $p \equiv 1$. If X and Z are independent, then $X - \mathbb{E}[X] \in \mathscr{P}_n(Z)^\perp$, so we deduce (viii) from the second part of (vi) by choosing $Y = X - \mathbb{E}[X]$. Finally, (ix) is a restatement of $X - \mathbb{E}[X \mid Y] \in \mathscr{P}_n(Z)^\perp$, which can be easily seen to hold when X—Y—Z forms a Markov chain. □

Remark 4.7. In view of properties (vii)–(viii), one may define the unconditional version of E_n as $E_n[X] := \mathbb{E}[X]$ for $X \in L^2(P)$. With this definition, the total expectation property (vii) becomes $E_n[E_n[X \mid Z]] = E_n[X]$, and the independence property (viii) becomes $E_n[X \mid Z] = E_n[X]$ for independent X and Z. This definition of $E_n[X]$ is also consistent with defining it as $E_n[X \mid 1]$, because $\mathbb{E}[X]$ is the closest constant to X in $L^2(P)$.

We also show that the PMMSE estimate satisfies the "tower property" similarly to the conditional expectation. This property is relegated Proposition C.1 in Appendix C.4.2, where we extend our results on the PMMSE to multiple dimensions.

Next we show that, if X and Z are symmetric random variables[7] that are independent, then the polynomial in $X+Z$ closest to X is always of odd degree or is a constant.

Lemma 4.4. *For $k \in \mathbb{N}_{\geq 1}$ and symmetric and independent random variables X and Z satisfying $\mathbb{E}\left[Z^2\right], \mathbb{E}\left[X^{4k}\right] < \infty$ and $|\mathrm{supp}(X+Z)| > 2k$, we have that $E_{2k}[X \mid X+Z] = E_{2k-1}[X \mid X+Z]$.*

Proof. See Appendix C.1.2. □

Finally, we show that the pointwise PMMSE estimate $E_n[X \mid Y = y]$ (see Remark 4.5) satisfies the following convergence theorems.

Lemma 4.5 (Convergence Theorems)**.** *Fix a sequence of square-integrable random variables $\{X_k\}_{k\in\mathbb{N}}$, and let $n \in \mathbb{N}$ and the random variable Y be such that $\mathbb{E}\left[Y^{2n}\right] < \infty$ and $|\mathrm{supp}(Y)| > n$. For every $y \in \mathbb{R}$, the following hold:*

(i) Monotone Convergence: If $\{X_k\}_{k\in\mathbb{N}}$ is monotone with square-integrable pointwise limit $X = \lim_{k\to\infty} X_k$, and either $Y \geq 0$ or $Y \leq 0$ holds almost surely, then

$$E_n[X \mid Y = y] = \lim_{k\to\infty} E_n\left[X_k \mid Y = y\right]. \tag{4.40}$$

(ii) Dominated Convergence: If there is a square-integrable random variable M such that $\sup_{k\in\mathbb{N}} |X_k| \leq M$, and if the pointwise limit $X := \lim_{k\to\infty} X_k$ exists, then

$$E_n[X \mid Y = y] = \lim_{k\to\infty} E_n\left[X_k \mid Y = y\right]. \tag{4.41}$$

Proof. See Appendix C.1.3. □

[7] A random variable Y is symmetric if $P_{Y-a} = P_{-(Y-a)}$ for some $a \in \mathbb{R}$.

4.3 PMMSE in Gaussian Channels

We focus in this section on the case $Y = \sqrt{t}X + N$ for $t \geq 0$ and $N \sim \mathcal{N}(0,1)$ independent of X. We prove rationality of the PMMSE (Theorem 4.2), uniform convergence of the PMMSE to the MMSE, and a pointwise-convergence rate bound (Theorem 4.3). Investigating the PMMSE in Gaussian channels allows us to extrapolate—via the I-MMSE relation—new formulas for differential entropy and mutual information primarily in terms of moments in the next section, which then pave the way for new estimators for these information measures in Section 4.5. We write

$$\mathrm{pmmse}_n(X,t) := \mathrm{pmmse}_n(X \mid \sqrt{t}X + N), \tag{4.42}$$

$$\mathrm{mmse}(X,t) := \mathrm{mmse}(X \mid \sqrt{t}X + N), \tag{4.43}$$

$$\mathrm{lmmse}(X,t) := \mathrm{lmmse}(X \mid \sqrt{t}X + N). \tag{4.44}$$

Omitted proofs of results stated in this section can be found in Appendix C.2 (for Section 4.3.1) and Appendix C.3 (for Section 4.3.2).

4.3.1 Rationality of the PMMSE: Proof of Theorem 4.2

Fix an integer $n > 0$, and let X be a random variable such that $\mathbb{E}[X^{2n}] < \infty$. We denote the moments of X by $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$, where $\mathcal{X}_0 := 1$. We begin by rewriting the PMMSE as

$$\mathrm{pmmse}_n(X,t) = \frac{\mathrm{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n}}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}}, \tag{4.45}$$

where $N \sim \mathcal{N}(0,1)$ is independent of X. With some algebra, one can show that the above expresses the PMMSE as a rational function.

Lemma 4.6. *Fix $n \in \mathbb{N}$, $N \sim \mathcal{N}(0,1)$, and a random variable X that is independent of N and which satisfies $\mathbb{E}\left[X^{2n}\right] < \infty$. Over $t \in [0,\infty)$, the function $t \mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most $d_n := \binom{n+1}{2}$, whose coefficient of t^{d_n} is $\det \boldsymbol{M}_{X,n}$, coefficient of t is $\sigma_X^2 G(n+2) d_n$, and constant term is $G(n+2)$, where $G(n+2) := \prod_{k=1}^n k!$ is the Barnes G-function. In addition, over $t \in [0,\infty)$, the function $t \mapsto \mathrm{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most $d_n - 1$, whose constant term is $\sigma_X^2 G(n+2)$. Furthermore, each coefficient in either of these two polynomials stays unchanged if X is shifted by a constant.*

Proof. See Appendix C.2.1. □

According to Lemma 4.6, we may define constants $a_X^{n,j}$ and $b_X^{n,j}$ by the polynomial identities

$$\mathrm{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n-1]} a_X^{n,j}\ t^j, \tag{4.46}$$

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j}\ t^j, \tag{4.47}$$

and taking the ratio of these two polynomials yields the following rational expression for the PMMSE

$$\mathrm{pmmse}_n(X,t) = \frac{\sum_{j\in[d_n-1]} a_X^{n,j}\ t^j}{\sum_{j\in[d_n]} b_X^{n,j}\ t^j}. \tag{4.48}$$

Lemma 4.6 also derives a subset of the desired coefficients values[8]

$$\left(a_X^{n,0}, b_X^{n,0}, b_X^{n,1}, b_X^{n,d_n}\right) = \left(\sigma_X^2 G(n+2), G(n+2), \sigma_X^2 G(n+2) d_n, \det \boldsymbol{M}_{X,n}\right), \tag{4.49}$$

so it only remains to derive the value of a_X^{n,d_n-1}.

Remark 4.8. We give fully-expanded formulas for each of the $a_X^{n,j}$ and $b_X^{n,j}$ in Appendix C.2.2, expressing them as integer-coefficient multivariate polynomials in the first $2n$ moments of X. Examining these expressions gives a strengthening of Theorem 4.2 in which the specific moments that could appear in any of the $a_X^{n,j}$ or $b_X^{n,j}$ are further restricted.

To complete the proof, we show that the value of the leading term in the numerator in (4.48) is given by

$$a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}. \tag{4.50}$$

We prove (4.50) next for continuous X, then generalize for every random variable X.

Assume for now that X is continuous. In particular, $|\mathrm{supp}(X)| = \infty$, so $\det \boldsymbol{M}_{X,n} \neq 0$ according to Lemma 4.1. In view of $b_X^{n,d_n} = \det \boldsymbol{M}_{X,n}$ (see (4.49)), showing $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ becomes equivalent to showing $\mathrm{pmmse}_n(X,t) \sim 1/t$ as $t \to \infty$ (see (4.48)). In addition, the PMMSE is bounded by the LMMSE and the MMSE,

$$\mathrm{mmse}(X,t) \leq \mathrm{pmmse}_n(X,t) \leq \mathrm{lmmse}(X,t). \tag{4.51}$$

We have that $\mathrm{lmmse}(X,t) \sim 1/t$ as $t \to \infty$. Further, the assumption of continuity of X implies that $\mathrm{mmse}(X,t) \sim 1/t$ too [WV11]. Thus, by (4.51), we obtain $\mathrm{pmmse}_n(X,t) \sim 1/t$ as $t \to \infty$. We have

[8] Note that Lemma 4.6 also shows that $a_{X+s}^{n,j} = a_X^{n,j}$ and $b_{X+s}^{n,\ell} = b_X^{n,\ell}$ for each $(j,\ell,s) \in [d_n-1] \times [d_n] \times \mathbb{R}$, which is a stronger result than shift-invariance of the PMMSE (see Lemma 4.2); however, we do not utilize this fact in the remainder of the proof.

thus shown the desired equation (4.50) when X is continuous.

We now return to the general case (i.e., not necessarily continuous X). Note that the quantity $a_X^{n,d_n-1} - \det \boldsymbol{M}_{X,n}$ is a multivariate polynomial in the first $2n$ moments of X. By showing $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ in the previous paragraph for every continuous X, we have established the vanishing of a multivariate polynomial in the first $2n$ moments of every continuous $2n$-times integrable random variable X. We show in Proposition 4.2 below that such a set of zeros is in fact too large to be contained in the zero-locus of any nonzero polynomial, i.e., that such a polynomial must vanish identically (equivalently, $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ must hold even when X is not continuous). For the proof of the latter claim, we first derive a moment-approximation intermediate result.

Lemma 4.7. *Fix $m \in \mathbb{N}_{>0}$, set $\ell = \lfloor m/2 \rfloor$ and $\mu_0 = 1$, and let $(\mu_1, \cdots, \mu_m) \in \mathbb{R}^m$ be such that $\left(\mu_{i+j}\right)_{(i,j)\in[\ell]^2}$ is positive definite. For every $\varepsilon > 0$, there exists a continuous random variable Z such that $\left|\mathbb{E}\left[Z^k\right] - \mu_k\right| < \varepsilon$ for every $k \in [m]$.*

Proof. Since $\left(\mu_{i+j}\right)_{(i,j)\in[\ell]^2}$ is assumed to be positive definite, the solution to the truncated Hamburger moment problem implies that there is a finitely-supported discrete random variable W such that $\mathbb{E}\left[W^k\right] = \mu_k$ for each $k \in [2\ell+1]$ (see [CF91, Theorem 3.1, items (iii) and (v)]). Let $U \sim \text{Unif}(0,1)$ be independent of W, and consider the continuous random variables $Z_\eta = W + \eta U$ for $\eta > 0$. For each $k \in [m]$, $Z_\eta^k \to W^k$ in distribution as $\eta \to 0^+$. Further, the set $\{Z_\eta^k\}_{0<\eta\le 1}$ is uniformly integrable since $|Z_\eta^k| \le (|W|+1)^k \in L^1(P)$. By the Lebesgue-Vitali theorem [Bog07, Theorem 4.5.4], we get $\mathbb{E}[Z_\eta^k] \to \mathbb{E}[W^k] = \mu_k$ for each $k \in [m]$ as $\eta \to 0^+$. Hence, for each $\varepsilon > 0$, we may choose $\eta > 0$ small enough so that $|\mathbb{E}[Z_\eta^k] - \mu_k| < \varepsilon$ for every $k \in [m]$, completing the proof. □

In the other direction, if $\mu_0 = 1$ and $(\mu_1, \cdots, \mu_{2\ell}) \in \mathbb{R}^{2\ell}$ come from a continuous random variable Z, i.e., $\mathbb{E}\left[Z^k\right] = \mu_k$ for each $k \in [2\ell]$, then it must be that the Hankel matrix $\boldsymbol{H} = \left(\mu_{i+j}\right)_{(i,j)\in[\ell]^2}$ is positive definite. Indeed, since $|\text{supp}(Z)| = \infty$, we have that $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} = \left\|\sum_{k\in[\ell]} v_k Z^k\right\|_2^2 > 0$ for every nonzero real vector $\boldsymbol{v} = (v_0, \cdots, v_\ell)^T$.

For each integer $m \ge 2$, let $\mathcal{R}_m \subset L^m(P)$ be the set of all continuous random variables X such that $\mathbb{E}[|X|^m] < \infty$. Consider the set $\mathcal{C}_m \subset \mathbb{R}^m$ defined by $\mathcal{C}_m = \{(\mathbb{E}[X], \cdots, \mathbb{E}[X^m])\ ;\ X \in \mathcal{R}_m\}$. We have the following result.

Proposition 4.2. *Let p be a polynomial in m variables with real coefficients. If $p\left(\mathbb{E}[X], \cdots, \mathbb{E}[X^m]\right) = 0$ for every continuous random variable X satisfying $\mathbb{E}[|X|^m] < \infty$, then p is the zero polynomial.*

Proof. See Appendix C.2.3. □

Proposition 4.2 completes the proof of Theorem 4.2. Indeed, since $a_X^{n,d_n-1} - \det \boldsymbol{M}_{X,n} = p\left(\mathbb{E}[X], \cdots, \mathbb{E}[X^{2n}]\right)$ for some multivariate polynomial p, and since we have shown above that p vanishes over $\mathcal{C}_m$, we conclude from Proposition 4.2 that p vanishes identically. In other words, the equation $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ holds for any random variable X satisfying $\mathbb{E}[X^{2n}] < \infty$ (regardless of whether X is continuous).[9] This completes the proof of Theorem 4.2. □

We note the following corollary of Theorem 4.2.

Corollary 4.1. *For a random variable X satisfying $\mathbb{E}\left[X^{2n}\right] < \infty$, we have that $\text{pmmse}_n(X,0) = \sigma_X^2$, for every $t > 0$ we have the inequalities*

$$\text{pmmse}_n(X,t) \le \frac{\sigma_X^2}{1+\sigma_X^2 t} < \frac{1}{t}, \tag{4.52}$$

and the function $t \mapsto \text{pmmse}_n(X,t)$ is real-analytic at each $t \in [0,\infty)$. If X also satisfies $|\text{supp}(X)| > n$, then as $t \to \infty$ we have the asymptotic

$$\text{pmmse}_n(X,t) = \frac{1}{t} + O(t^{-2}). \tag{4.53}$$

Proof. That $\text{pmmse}_n(X,0) = \sigma_X^2$ follows by setting $t = 0$ in (4.8) or in the definition of the PMMSE. The inequalities in (4.52) follow since $\text{pmmse}_n(X,t) \le \text{lmmse}(X,t) = \sigma_X^2/(1+\sigma_X^2 t)$. In addition, a rational function is analytic at each point in its domain. For each $t \ge 0$, $|\text{supp}(\sqrt{t}X + N)| = \infty$ where $N \sim \mathcal{N}(0,1)$ independent of X. Therefore, $\boldsymbol{M}_{\sqrt{t}X+N}$ is invertible for every $t \ge 0$, i.e., the denominator in (4.8) is never zero for $t \ge 0$, so we infer analyticity of $\text{pmmse}_n(X,t)$. Finally, if $|\text{supp}(X)| > n$ then $\det \boldsymbol{M}_{X,n} \ne 0$, so (4.53) follows from (4.8). □

4.3.2 Convergence of PMMSE to MMSE: Proof of Theorem 4.3

In Appendix C.3.1 we give the proof of the uniform convergence in (4.13), namely, that as $n \to \infty$ we have

$$\sup_{t \ge 0}\ \text{pmmse}_n(X,t) - \text{mmse}(X,t) \searrow 0 \tag{4.54}$$

for X having a MGF. In a nutshell, the proof follows from Cantor's intersection theorem in view of continuity of the PMMSE and the MMSE in the SNR, t, and monotonicity of the PMMSE in the polynomial degree, n.

In this subsection, we prove the asymptotic convergence rate stated in (4.14). Specifically, let

[9] In [AC21a, Appendix L], an alternative proof of $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ is given via a self-contained algebraic argument.

$\mathscr{D}$ denote the set of all PDFs or PMFs p that are compactly-supported, even, and decreasing over $[0,\infty)\cap \mathrm{supp}(p)$. Suppose that X is continuous or discrete, with PDF or PMF $p_X \in \mathscr{D}$. We prove next that for any fixed $k,t \geq 0$ we have

$$\lim_{n\to\infty} n^k \cdot (\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)) = 0. \tag{4.55}$$

Let $N \sim \mathcal{N}(0,1)$ be independent of X, and set $Y = X + N$.

The proof of the convergence rate in (4.14) relies on results on the Bernstein approximation problem in weighted L^p spaces. In particular, we consider the Freud case [Lub07, Definition 3.3], where the weight is of the form e^{-Q} for Q of polynomial growth, e.g., a Gaussian weight.

Definition 4.2 (Freud Weight, [Lub07, Definition 3.3])**.** A function $W : \mathbb{R} \to (0,\infty)$ is called a *Freud Weight*, and we write $W \in \mathscr{F}$, if it is of the form $W = e^{-Q}$ for $Q : \mathbb{R} \to \mathbb{R}$ satisfying:

(1) Q is even,

(2) Q is differentiable, and $Q'(y) > 0$ for $y > 0$,

(3) $y \mapsto yQ'(y)$ is strictly increasing over $(0,\infty)$,

(4) $yQ'(y) \to 0$ as $y \to 0^+$, and

(5) there exist $\lambda, a, b, c > 1$ such that for every $y > c$ we have $a \leq \frac{Q'(\lambda y)}{Q'(y)} \leq b$.

One may associate to each Freud weight $W = e^{-Q}$ its Mhaskar–Rakhmanov–Saff numbers $a_n(Q)$, defined next.

Definition 4.3. If $Q : \mathbb{R} \to \mathbb{R}$ satisfies conditions (2)–(4) in Definition 4.2, and if $yQ'(y) \to \infty$ as $y \to \infty$, then the *n-th Mhaskar–Rakhmanov–Saff (MRS) number $a_n(Q)$ of Q* is defined as the unique positive root a_n of the equation

$$n = \frac{2}{\pi}\int_0^1 \frac{a_n t Q'(a_n t)}{\sqrt{1-t^2}}\, dt. \tag{4.56}$$

Remark 4.9. The condition $yQ'(y) \to \infty$ as $y \to \infty$ in Definition 4.3 is satisfied if e^{-Q} is a Freud weight. Indeed, in view of properties (2)–(3) in Definition 4.2, the quantity $\ell := \lim_{y\to\infty} yQ'(y)$ is well-defined and it belongs to $(0,\infty]$. If $\ell \neq \infty$, then because $\lim_{y\to\infty} \lambda y Q'(\lambda y) = \ell$ too, property (5) would imply that $a \leq 1/\lambda \leq b$ contradicting that $\lambda, a > 1$. Therefore, $\ell = \infty$.

For example, the Gaussian weight $W(y) = e^{-y^2}$ is a Freud weight for which $Q(y) = y^2$, and it has the MRS numbers $a_n(Q) = \sqrt{n}$ since $\int_0^1 t^2/\sqrt{1-t^2}\, dt = \frac{\pi}{4}$.

We apply the following Jackson-type theorem.

Theorem 4.7 ([Lub07, Corollary 3.6])**.** *Fix $W \in \mathscr{F}$, and let u be an r-times continuously differentiable function such that $u^{(r)}$ is absolutely continuous. Let $a_n = a_n(Q)$ where $W = e^{-Q}$, and fix $1 \leq s \leq \infty$. Then, for some constant $D(W, r, s)$ and every $n \geq \max(r-1, 1)$*

$$\inf_{q \in \mathscr{P}_n} \|(q-u)W\|_{L^s(\mathbb{R})} \leq D(W, r, s) \left(\frac{a_n}{n}\right)^r \|u^{(r)}W\|_{L^s(\mathbb{R})}. \tag{4.57}$$

We will apply the polynomial approximation result stated in Theorem 4.7 for the $L^2(P_Y)$ norm, i.e., we set $s = 2$, $W = \sqrt{p_Y}$, and $u(y) = \mathbb{E}[X \mid Y = y]$ in Theorem 4.7. To this end, we will establish the following three facts:

(i) $\sqrt{p_Y} \in \mathscr{F}$,

(ii) $a_n(-\frac{1}{2}\log p_Y) = O_{p_X}(\sqrt{n})$, and

(iii) $\|(d^r/dy^r)\mathbb{E}[X \mid Y = y]\|_2 = O_r(1)$.

The former two facts are established in the following lemma.

Lemma 4.8. *If $X \sim p$ for some $p \in \mathscr{D}$, and $N \sim \mathcal{N}(0,1)$ is independent of X, then p_{X+N}^s is a Freud weight for any fixed constant $s > 0$. Further, suppose $M > 0$ is such that $\mathrm{supp}(p) \subset [-M, M]$, and denote $Q = -\log p_{X+N}$. Then, for each integer $n \geq 1$ and real $s > 0$, we have the bound*

$$a_n(sQ) \leq \left(2M + \sqrt{2}\right)\sqrt{n/s}. \tag{4.58}$$

Proof. See Appendix C.3.2. □

Next, we derive a bound on $\|(d^r/dy^r)\mathbb{E}[X \mid Y = y]\|_2$ that depends only on r. We will need the following result showing that the higher-order derivatives of the conditional expectation are given by the conditional cumulants.

Proposition 4.3 ([AC21c, Proposition 1], [DPS21, Proposition 7])**.** *Fix an integrable random variable X and an independent $N \sim \mathcal{N}(0,1)$, and let $Y = X + N$. For each integer $r \geq 1$ and real y, we have the formula*

$$\frac{d^{r-1}}{dy^{r-1}}\,\mathbb{E}\left[X \mid Y = y\right] = \kappa_r(X \mid Y = y), \tag{4.59}$$

where $\kappa_r(X \mid Y = y) := \frac{\partial^r}{\partial \tau^r} \log \mathbb{E}\left[e^{\tau X} \mid Y = y\right]\big|_{\tau=0}$ is the r-th conditional cumulant of X given $\{Y = y\}$.

Using Proposition 4.3, we obtain the following bound on the second moment of the derivatives of the conditional expectation via Hölder's inequality.

Lemma 4.9. *Fix an integrable random variable X and an independent $N \sim \mathcal{N}(0,1)$, let $Y = X + N$, and fix an integer $r \geq 2$. Denote the constants $q_r := \lfloor (\sqrt{8r+9} - 3)/2 \rfloor$, $\gamma_r := (2rq_r)!^{1/(4q_r)}$, and*

$$C_r = \sum_{k=1}^{r} (k-1)! \sum_{j=0}^{k} (-1)^j \binom{r}{j} \left\{ {r-j \atop k-j} \right\}, \tag{4.60}$$

where $\left\{ {r \atop k} \right\}$ denotes Stirling's number of the second kind.[10] *We have the bound*

$$\left\| \frac{d^{r-1}}{dy^{r-1}} \mathbb{E}[X \mid Y = y] \right\|_2 \leq 2^r C_r \min\left(\gamma_r, \|X\|_{2rq_r}^r \right). \tag{4.61}$$

Proof. See Appendix C.3.3. □

Remark 4.10. For $2 \leq r \leq 7$, we obtain the first few values of q_r as $1,1,1,2,2,2$, and we have $q_r \sim \sqrt{2r}$ as $r \to \infty$ (see Remark C.3 at the end of the proof in Appendix C.3.3 for a way to reduce q_r). The first few values of C_r (for $2 \leq r \leq 7$) are given by $1,1,4,11,56,267$, and as $r \to \infty$ we have the asymptotic $C_r \sim (r-1)!/\alpha^r$ for some constant $\alpha \approx 1.146$ (see [OEI]). The crude bound $C_r < r^r$ can also be seen by a combinatorial argument.

We now apply the results of Lemmas 4.8–4.9 in Theorem 4.7 to complete the proof of the convergence rate in (4.14). Fix a real $k \geq 0$, set $r = \lceil k+1 \rceil$, and let $n \geq \max(r-1,1)$ be an integer. We apply Theorem 4.7 for the conditional expectation function $u(y) = \mathbb{E}[X \mid Y = y]$, the weight $W = \sqrt{p_Y}$, and the exponent $s = 2$. By our choice of weight, $\|vW\|_{L^2(\mathbb{R})} = \|v(Y)\|_2$ for any Borel function $v : \mathbb{R} \to \mathbb{R}$; in particular, this holds for the choice $v(y) = q(y) - \mathbb{E}[X \mid Y = y]$ for any $q \in \mathscr{P}_n$, and also for $v(y) = \frac{d^r}{dy^r}\mathbb{E}[X \mid Y = y]$. Recall from (4.4) that $E_n[X \mid Y]$ minimizes $\|q(Y) - \mathbb{E}[X \mid Y\|_2$ over $q(Y) \in \mathscr{P}_n(Y)$. Hence, with our choice of W and u, we have

$$\|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2 = \inf_{q \in \mathscr{P}_n} \|(q-u)W\|_{L^2(\mathbb{R})}. \tag{4.62}$$

By Lemma 4.8, $W = \sqrt{p_Y}$ is a Freud weight, and we have a bound $a_n(Q) = O_{p_X}(\sqrt{n})$ where $W = e^{-Q}$. In addition, by Lemma 4.9, we have a bound $\|\frac{d^r}{dy^r}\mathbb{E}[X \mid Y = y]\|_2 = O_r(1)$. Therefore, by

[10] The integer $\left\{ {r \atop k} \right\}$ equals the number of unordered set-partitions of an r-element set into k nonempty subsets. The integer C_r equals the number of cyclically-invariant ordered set-partitions of an r-element set into subsets of sizes at least 2, see sequence A032181 at [OEI].

Theorem 4.7, we obtain a constant $D'(p_X,k)$ (depending on $D(\sqrt{p_Y},r,2)$, see (4.57)) such that

$$\|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2 \le \frac{D'(p_X,k)}{n^{\lceil k+1 \rceil/2}}. \tag{4.63}$$

From (4.63), we conclude

$$n^k \; \|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2^2 \le \frac{D'(p_X,k)^2}{n}. \tag{4.64}$$

Further, by the orthogonality principle of $\mathbb{E}[X \mid Y]$, we have that (see (4.6))

$$\text{pmmse}_n(X,1) - \text{mmse}(X,1) = \|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2^2. \tag{4.65}$$

Hence, we conclude from (4.64) that

$$\lim_{n\to\infty} n^k \; (\text{pmmse}_n(X,1) - \text{mmse}(X,1)) = 0. \tag{4.66}$$

Finally, note that the premises of the theorem are also satisfied by $\sqrt{t}X$ for any $t>0$, so we have

$$\lim_{n\to\infty} n^k \; \left(\text{pmmse}_n(\sqrt{t}X,1) - \text{mmse}(\sqrt{t}X,1)\right) = 0. \tag{4.67}$$

Also, one straightforwardly obtains from Lemma 4.2 that

$$\text{pmmse}_n(X,t) - \text{mmse}(X,t) = \frac{1}{t}\left(\text{pmmse}_n(\sqrt{t}X,1) - \text{mmse}(\sqrt{t}X,1)\right). \tag{4.68}$$

Thus, we conclude from (4.67) the desired asymptotic result that $n^k\;(\text{pmmse}_n(X,t) - \text{mmse}(X,t)) \to 0$ as $n\to\infty$ for any fixed reals $k,t\ge 0$ (note that the limit trivially holds for $t=0$ since then both the PMMSE and the MMSE are equal to σ_X^2). □

Remark 4.11. The convergence rate proved in Theorem 4.3 is an asymptotic one, and obtaining a finitary version hinges on having explicit characterization of the constants $D(W,r,s)$ in Theorem 4.7. However, no explicit formula for $D(W,r,s)$ exists in the literature, to the best of our knowledge. To give more details, note that we show in (4.63) a bound for finite n. Namely, for $k\ge 0$, $r=\lceil k+1\rceil$, and $n\ge \max(r-1,1)$ we have the bound

$$\|E_n[X \mid X+N] - \mathbb{E}[X \mid X+N]\|_2 \le \frac{D'(p_X,k)}{n^{r/2}}, \tag{4.69}$$

where the constant $D'(p_X,k)$ can be chosen as, e.g., with $\text{supp}(p_X)\subset[-M,M]$,

$$D'(p_X,k) = D(\sqrt{p_{X+N}},r,2)\cdot\left(2\left(\sqrt{2}M+1\right)\right)^r\cdot 2^r C_r \min(\gamma_r,M^r). \tag{4.70}$$

Thus, to make explicit the constant of interest to us, $D'(p_X, k)$, it suffices to have an explicit bound on $D(\sqrt{p_{X+N}}, r, 2)$. However, this latter result, to the best of our knowledge, does not exist in the literature; further, distilling an explicit form for $D(W, r, s)$ from existing proofs is a nontrivial matter. The constants $D(W, r, s)$ carry over from [Lub07, Corollary 3.6], a result that was first proved in [DL97] (specifically, it is the combination of Theorem 1.2 and Corollary 1.8 in [DL97]). The constant $D(W, r, s)$ is a universal constant in the sense that Theorem 4.7 is a Jackson-type theorem, i.e., it gives a polynomial-approximation bound that holds uniformly for all admissible functions u that are to be approximated (although the weight W is fixed). Thus, making $D(W, r, s)$ explicit is in fact a significant improvement on the general approximation-theoretic problem. Note that we do not need to utilize this universality for our PMMSE convergence-rate analysis, since we only need to apply the bound in Theorem 4.7 for the specific choice of u being the conditional expectation function. This in particular implies the potential of the constant $D(\sqrt{p_{X+N}}, r, 2)$ being improved for our purposes. Yet, we note that the closely related Jackson-type theorem shown in [Mha96, Theorem 4.1.1] can potentially lead to explicit constants more easily; this result derives inequality (4.57) in Theorem 4.7, but with the MRS number a_n replaced with the Freud number q_n (the positive solution to $q_n Q'(q_n) = n$), and it is also premised on a few assumptions on Q''. Finally, since we are interested in guaranteeing convergence in n, the derivation in Theorem 4.3 is sufficient for our PMMSE analysis. See Remark 4.13 for further discussion.

Remark 4.12. Examining the proof of the asymptotic convergence rate in Theorem 4.3 reveals that it is possible to show that the same convergence rate holds beyond Gaussian channels. Specifically, the following is a blueprint for showing that

$$\lim_{n\to\infty} n^k \left(\mathrm{pmmse}_n(X \mid \sqrt{t}X + Z) - \mathrm{mmse}(X \mid \sqrt{t}X + Z) \right) = 0 \tag{4.71}$$

for every $k, t \geq 0$, where Z a (non-necessarily Gaussian) continuous noise that is independent of X:

1. Suppose that the random variable $Y = \sqrt{t}X + Z$ is such that the conditional PDFs $p_{Y|X=x}$ form an exponential family. From [DC21, Proposition 3], the higher-derivative formulas $\frac{d^{r-1}}{dy^{r-1}}\mathbb{E}[X \mid Y = y] = \kappa_r(X \mid Y = y)$ (as in Proposition 4.3) carries over to this case.

2. The proof of Lemma 4.9 carries over verbatim to obtain a bound $\left\| \frac{d^{r-1}}{dy^{r-1}}\mathbb{E}[X \mid Y = y] \right\|_2 \leq 2^r C_r \|X\|_{2rq_r}^r$.

3. Assume that p_Z is a Freud weight, say $p_Z = e^{-Q}$ for $Q(z) \sim z^\ell$ as $z \to \infty$ for some fixed

$\ell > 1$. Then, the proof of Lemma 4.8 can be adapted to show that (if, e.g., $p_X \in \mathscr{D}$, where $\mathscr{D}$ is as defined in the beginning of this subsection) the PDF p_Y is also a Freud weight with MRS number of order $n^{1/\ell}$.

4. Applying the Bernstein approximation result stated in Theorem 4.7, we obtain an upper bound on the approximation error $\text{pmmse}_n(X \mid Y) - \text{mmse}(X \mid Y)$ of order $n^{-k(1-1/\ell)}$ as $n \to \infty$. As this is true for every $k \geq 0$, we conclude the asymptotic rate of convergence $n^k \cdot (\text{pmmse}_n(X \mid Y) - \text{mmse}(X \mid Y)) \to 0$ for every $k \geq 0$ and every $t \geq 0$.

4.4 New Formulas for Information Measures in Terms of Moments

We apply the derived PMMSE results in the I-MMSE relation to express the differential entropy and mutual information in terms of moments. For example, combining Theorems 4.2 and 4.4 shows that for any continuous random variable X that has a MGF, we may express differential entropy as (see (4.8))

$$h(X) = \frac{1}{2} \lim_{n \to \infty} \int_0^\infty -\frac{1}{2\pi e + t} + \frac{\sigma_X^2 G(n+2) + \cdots + (\det \boldsymbol{M}_{X,n}) t^{d_n - 1}}{G(n+2) + \left(\sigma_X^2 G(n+2) d_n\right) t + \cdots + (\det \boldsymbol{M}_{X,n}) t^{d_n}} \, dt, \tag{4.72}$$

where the coefficients of the integrand are all multivariate polynomials in the moments of X. The starting point in deriving this formula is the I-MMSE relation, which we briefly review first.

Theorem 4.8 (I-MMSE relation, [GSV05])**.** *For any square-integrable random variable X, an independent $N \sim \mathcal{N}(0,1)$, and $\gamma > 0$, we have that*

$$I(X; \sqrt{\gamma} X + N) = \frac{1}{2} \int_0^\gamma \text{mmse}(X, t) \, dt. \tag{4.73}$$

The I-MMSE relation directly yields the following formula for differential entropy: for a square-integrable continuous random variable X we have that [GSV05]

$$h(X) = \frac{1}{2} \log\left(2\pi e \sigma_X^2\right) - \frac{1}{2} \int_0^\infty \frac{\sigma_X^2}{1 + \sigma_X^2 t} - \text{mmse}(X, t) \, dt. \tag{4.74}$$

Since $\int_0^\infty \frac{a}{1+at} - \frac{b}{1+bt} \, dt = \log \frac{a}{b}$ for any $a, b > 0$, we may simplify (4.74) to become

$$h(X) = \frac{1}{2} \int_0^\infty \text{mmse}(X, t) - \frac{1}{2\pi e + t} \, dt. \tag{4.75}$$

We further extend the representation in (4.75) to higher dimensions.

Lemma 4.10. *If the m-dimensional continuous random vector* $\boldsymbol{X}$ *has a finite covariance matrix, then*

$$h(\boldsymbol{X}) = \frac{1}{2}\int_0^\infty \operatorname{mmse}(\boldsymbol{X},t) - \frac{m}{2\pi e + t}\,dt. \tag{4.76}$$

Proof. See Appendix C.5.1. □

The MMSE term in the expression for $h(\boldsymbol{X})$ given in Lemma 4.10 can be approximated by the PMMSE, resulting in an expression for differential entropy as a function of moments of X. From (4.74) and (4.75), and since $\operatorname{mmse}(X,t) \le \operatorname{lmmse}(X,t)$, replacing the MMSE with the LMMSE gives the upper bound on differential entropy $h(X)$

$$\begin{aligned} h(X) \le h_1(X) &:= \frac{1}{2}\int_0^\infty \operatorname{lmmse}(X,t) - \frac{1}{2\pi e + t}\,dt & (4.77)\\ &= \frac{1}{2}\log\left(2\pi e \sigma_X^2\right) = h(\mathcal{N}(0,\sigma_X^2)), & (4.78)\end{aligned}$$

which is the maximum possible differential entropy for a continuous random variable with a prescribed variance of σ_X^2. We take this a step further and introduce for each integer $n \ge 1$ (assuming only $\mathbb{E}[X^{2n}] < \infty$) the functional

$$h_n(X) := \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(X,t) - \frac{1}{2\pi e + t}\,dt. \tag{4.79}$$

By the monotonicity $\operatorname{pmmse}_1(X,t) \ge \operatorname{pmmse}_2(X,t) \ge \cdots \ge \operatorname{mmse}(X,t)$, we also have a monotone sequence $h_1(X) \ge h_2(X) \ge \cdots \ge h(X)$ for a random variable X having moments of all orders. As stated in Theorem 4.4, which we prove next in the 1-dimensional case, if X also has a MGF then $h_n(X) \searrow h(X)$. The proof for arbitrary dimensions requires extending our PMMSE results to higher dimensions (which we give in Appendix C.4), hence we relegate it to Appendix C.5.2.

Proof of Theorem 4.4 (for the 1-dimensional case). The functions $g_n(t) := \operatorname{lmmse}(X,t) - \operatorname{pmmse}_n(X,t)$ are nonnegative and nondecreasing. By Theorem 4.3, $g_n \nearrow g$ pointwise, where $g(t) := \operatorname{lmmse}(X,t) - \operatorname{mmse}(t)$. Therefore, by the monotone convergence theorem, $\int_0^\infty g_n(t)\,dt \nearrow \int_0^\infty g(t)\,dt$. Adding and subtracting $1/(2\pi e + t)$ to each integrand, and noting that $t \mapsto \operatorname{lmmse}(X,t) - 1/(2\pi e+t)$ is absolutely integrable, we conclude that $h_n(X) \searrow h(X)$. □

Remark 4.13. It remains a topic of ongoing investigation to derive the convergence rate of the limit $h_n(X) \searrow h(X)$ shown in Theorem 4.4. Note that we may write the convergence error as

$$h_n(X) - h(X) = \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(X,t) - \operatorname{mmse}(X,t)\,dt. \tag{4.80}$$

Hence, the convergence rate of $h_n(X) \searrow h(X)$ can be shown if one has the convergence rate of $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ as a function of t. However, the asymptotic convergence rate bound we show in Theorem 4.3 does not depend on the parameter t. As discussed in Remark 4.11, finer characterization of the PMMSE convergence rate hinges on having explicit bounds on the constant $D(W,r,s)$ (see the statement of Theorem 4.7). This constant is only given implicitly in [DL97], which is likely due to the universality it enjoys, i.e., the approximation error in Theorem 4.7 is controlled by $D(W,r,s)$ for a fixed W and *every* function u that is to be approximated by polynomials. In our case, however, we need another type of universality. Precisely, we need to control the best-polynomial error when approximating the class of functions $u_t(y) := \mathbb{E}[X \mid \sqrt{t}X + N = y]$ in their respective weighted Hilbert spaces with weights $W_t := \sqrt{p_{\sqrt{t}X+N}}$ for *every* $t \geq 0$. To the best of our knowledge, no such universality result where the weight can vary parametrically exists in the literature.

The behavior of the PMMSE under affine transformations shown in Lemma 4.2 implies that each approximation h_n behaves under (injective) affine transformations exactly as differential entropy does.

Corollary 4.2. *If X is a random variable satisfying $\mathbb{E}[X^{2n}] < \infty$, and $(\alpha, \beta) \in \mathbb{R}^2$ with $\alpha \neq 0$, then we have*

$$h_n(\alpha X + \beta) = h_n(X) + \log|\alpha|. \tag{4.81}$$

In addition, if X and Y are independent with finite $2n$-th moments, then $h_n(X,Y) = h_n(X) + h_n(Y)$.

The moments-based formula for differential entropy shown in Theorem 4.4 yields moments-based formulas for mutual information in view of the expansions $I(X;Y) = h(Y) - h(Y \mid X)$ in the discrete-continuous case and $h(X,Y) = h(X) + h(Y) - h(X,Y)$ in the purely continuous case. The proof of these formulas, stated in Theorem 4.5, is given in Appendix C.5.3. We discuss here a few implications. If X is discrete and Y is continuous, and if they satisfy the assumptions in the first case of Theorem 4.5, then we denote the functionals

$$I_n(X;Y) := \frac{1}{2}\int_0^\infty \mathrm{pmmse}_n(Y,t) - \mathbb{E}_X\left[\mathrm{pmmse}_n(Y^{(X)},t)\right] dt. \tag{4.82}$$

Recall that we denote by $Y^{(x)}$ the random variable obtained from Y by conditioning on $\{X = x\}$. If X and Y are continuous satisfying the premises of the second case of Theorem 4.5, then we denote the functional

$$I_n(X;Y) := \frac{1}{2}\int_0^\infty \mathrm{pmmse}_n(X,t) + \mathrm{pmmse}_n(Y,t) - \mathrm{pmmse}_n((X,Y),t)\, dt. \tag{4.83}$$

The statement of Theorem 4.5 is that $I_n(X;Y) \to I(X;Y)$ as $n \to \infty$.

The functionals I_n enjoy properties that resemble those for the mutual information. First, the behavior of the PMMSE under affine transformations exhibited in Lemma 4.2 implies that $I_n(X;Y)$ is invariant under injective affine transformations of Y. Indeed, this can be seen immediately from the behavior of h_n in Corollary 4.2. Also, the approximations $I_n(X;Y)$ detect independence exactly.

Corollary 4.3. *Suppose X and Y are random variables satisfying the premises of Theorem 4.5 (in either case 1 or case 2). For any constants $(\alpha,\beta) \in \mathbb{R}^2$ with $\alpha \neq 0$, and for any $n \in \mathbb{N}$, we have*

$$I_n(X;\alpha Y+\beta) = I_n(X;Y). \tag{4.84}$$

In addition, if X and Y are independent, then $I_n(X;Y)=0$ for every n.

We give full expressions for the first two approximants of mutual information that are generated by the LMMSE and quadratic MMSE, in the discrete-continuous case.

Example 4. When $n=1$, we obtain

$$I_1(X;Y) = \log \sigma_Y - \mathbb{E}_X\left[\log \sigma_{Y^{(X)}}\right], \tag{4.85}$$

which is the exact formula for $I(X;Y)$ when both Y is Gaussian and each $Y^{(x)}$ (for $x \in \mathrm{supp}(X)$) is Gaussian; indeed, in such a case, the MMSE is just the LMMSE. □

Example 5. For $n=2$, we obtain the formula

$$\begin{aligned} I_2(X;Y) = \frac{1}{6}\log \frac{b_Y^{2,3}}{\prod_{x\in\mathrm{supp}(X)}\left(b_{Y^{(x)}}^{2,3}\right)^{P_X(x)}} \\ + \frac{1}{2}\int_0^\infty \frac{a_Y^{2,1}\, t}{2+b_Y^{2,1}\, t+b_Y^{2,2}\, t^2+b_Y^{2,3}\, t^3} - \mathbb{E}_X\left[\frac{a_{Y^{(X)}}^{2,1}\, t}{2+b_{Y^{(X)}}^{2,1}\, t+b_{Y^{(X)}}^{2,2}\, t^2+b_{Y^{(X)}}^{2,3}\, t^3}\right] dt \end{aligned} \tag{4.86}$$

where we may compute for any $R \in L^4(P)$

$$b_R^{2,3} := \begin{vmatrix} 1 & \mathbb{E}[R] & \mathbb{E}[R^2] \\ \mathbb{E}[R] & \mathbb{E}[R^2] & \mathbb{E}[R^3] \\ \mathbb{E}[R^2] & \mathbb{E}[R^3] & \mathbb{E}[R^4] \end{vmatrix} \tag{4.87}$$

$$= \sigma_R^2\mathbb{E}[R^4] + 2\mathbb{E}[R]\mathbb{E}[R^2]\mathbb{E}[R^3] - \mathbb{E}[R^2]^3 - \mathbb{E}[R^3]^2, \tag{4.88}$$

which is strictly positive when $|\mathrm{supp}(R)| > 2$, and

$$b_R^{2,2} = -4\mathbb{E}[R]\mathbb{E}[R^3] + 3\mathbb{E}[R^2]^2 + \mathbb{E}[R^4] \tag{4.89}$$

$$b_R^{2,1} = 6\sigma_R^2 \tag{4.90}$$

$$a_R^{2,1} = 4\mathbb{E}[R]^4 - 8\mathbb{E}[R]^2\mathbb{E}[R^2] + \frac{8}{3}\mathbb{E}[R]\mathbb{E}[R^3] + 2\mathbb{E}[R^2]^2 - \frac{2}{3}\mathbb{E}[R^4]. \tag{4.91}$$

□

4.5 Application: Estimation of Information Measures from Data

The approximations introduced in the previous sections naturally motivate estimators for information measures. These estimators are based on (i) approximating moments with sample moments, then (ii) plugging the sample moments into the formulas we have developed for information measures. Since the formulas for information measures depend continuously on the underlying moments, the resulting estimators are asymptotically consistent. Moreover, the estimators also behave as the target information measure under affine transformations, being inherently robust to, for example, rescaling of the samples.

We estimate $h(X)$ from i.i.d. samples $X_1, \cdots, X_m$ as $h_n(U)$ for $U \sim \mathrm{Unif}(\{X_1, \cdots, X_m\})$. More precisely, we introduce the following estimator of differential entropy.

Definition 4.4. Let $X, X_1, \cdots, X_m$ be i.i.d. continuous random variables, and denote $\mathcal{S} = \{X_j\}_{j=1}^m$. We define the n-th estimate $\widehat{h}_n(\mathcal{S})$ of the differential entropy $h(X)$ as the functional that takes the value $h_n(X)$ if the first $2n$ moments of X are replaced by their respective sample moments. In other words, with $U \sim \mathrm{Unif}(\mathcal{S})$, we set $\widehat{h}_n(\mathcal{S}) := h_n(U)$.

The estimator of mutual information $I(X;Y)$ between a discrete X and a continuous Y is defined next. We utilize Theorem 4.5. We will need to invert the Hankel matrices of moments $(\mathbb{E}[V^{i+j} \mid U = u])_{i,j\in[n]}$ for each $u \in \mathrm{supp}(U)$, where (U,V) is uniformly distributed over the samples $\mathcal{S} = \{(X_j, Y_j)\}_{j=1}^m$. These Hankel matrices are invertible if and only if for each $u \in \{X_j\}_{j=1}^m$ there are more than n distinct samples (X_j, Y_j) for which $X_j = u$; equivalently, the size of the support set of the random variable V conditioned on $U = u$ exceeds n. Thus, we remove all values u that appear at most n times in the samples $\mathcal{S}$. In other words, we replace $\mathcal{S}$ with the subset

$$\mathcal{S}^{(n)} := \left\{(X',Y') \in \mathcal{S} \; ; \; |\{1 \le i \le m \; ; \; X_i = X'\}| > n\right\}. \tag{4.92}$$

Definition 4.5. Let $(X, Y), (X_1, Y_1), \cdots, (X_m, Y_m)$ be i.i.d. 2-dimensional random vectors such that X is discrete with finite support and Y is continuous, and denote $\mathcal{S} = \{(X_j, Y_j)\}_{j=1}^m$. Define $\mathcal{S}^{(1)} \supseteq \mathcal{S}^{(2)} \supseteq \cdots$ by

$$\mathcal{S}^{(n)} := \left\{ (X', Y') \in \mathcal{S} \, ; \, |\{1 \le i \le m \, ; \, X_i = X'\}| > n \right\}. \tag{4.93}$$

For each $n \geq 1$ such that $\mathcal{S}^{(n)}$ is nonempty, let $(U^{(n)}, V^{(n)}) \sim \text{Unif}(\mathcal{S}^{(n)})$. We define the n-th estimate $\widehat{I}_n(\mathcal{S})$ of the mutual information $I(X; Y)$ by $\widehat{I}_n(\mathcal{S}) := I_n(U^{(n)}; V^{(n)})$.

We show in this Appendix C.6 how to implement the proposed estimators numerically. In this section, we prove that the estimators are consistent, and discuss their sample complexity. We end the section by empirically comparing the estimators' performance with other estimators from the literature.

4.5.1 Consistency

As sample moments converge almost surely to the moments, and as our expressions for differential entropy and mutual information depend continuously on the moments, the continuous mapping theorem yields that the estimators of differential entropy and mutual information introduced in the beginning of this section are consistent.

Theorem 4.9. *Let X be a continuous random variable that has a MGF. Let $\{X_j\}_{j=1}^\infty$ be i.i.d. samples drawn according to P_X. Then, for every $n \in \mathbb{N}$, we have the almost-sure convergence*

$$\lim_{m \to \infty} \widehat{h}_n \left(\{X_j\}_{j=1}^m \right) = h_n(X). \tag{4.94}$$

Furthermore, we have that

$$h(X) = \lim_{n \to \infty} \lim_{m \to \infty} \widehat{h}_n \left(\{X_j\}_{j=1}^m \right) \tag{4.95}$$

where the convergence in m is almost-sure convergence.

Proof. See Appendix C.7.1. □

Corollary 4.4. *Let X be discrete random variable with finite support, and Y be a continuous random variable with a MGF and satisfying $h(Y) > -\infty$. Let $\{(X_j, Y_j)\}_{j=1}^\infty$ be i.i.d. samples drawn according to $P_{X,Y}$. For every $n \in \mathbb{N}$, we have the almost-sure convergence*

$$\lim_{m \to \infty} \widehat{I}_n \left(\{(X_j, Y_j)\}_{j=1}^m \right) = I_n(X; Y). \tag{4.96}$$

Furthermore,

$$I(X;Y) = \lim_{n\to\infty} \lim_{m\to\infty} \widehat{I}_n \left(\{(X_j, Y_j)\}_{j=1}^m \right) \tag{4.97}$$

where the convergence in m is almost-sure convergence.

Proof. See Appendix C.7.2. □

4.5.2 Sample Complexity

When X is a continuous random variable of bounded support, we may derive the following sample complexity of the estimator of differential entropy in Definition 4.4 from Hoeffding's inequality.

Proposition 4.4. *Fix a bounded-support continuous random variable $X \in L^{2n}(P)$. There is a constant $C = C(X, n)$ such that, for all small enough $\varepsilon, \delta > 0$, any collection $\mathcal{S}$ of i.i.d. samples drawn according to P_X of size*

$$|\mathcal{S}| > \frac{C}{\varepsilon^2} \log \frac{1}{\delta} \tag{4.98}$$

must satisfy

$$\Pr \left\{ \left| \widehat{h}_n(\mathcal{S}) - h_n(X) \right| < \varepsilon \right\} \geq 1 - \delta. \tag{4.99}$$

Proof. See Appendix C.8. □

Remark 4.14. The sample complexity bound may be rearranged as follows. With $m = |\mathcal{S}|$ denoting the sample size, we have that

$$\Pr \left\{ \left| \widehat{h}_n(\mathcal{S}) - h_n(X) \right| \geq \frac{C_1 \sqrt{\log(1/\delta)}}{\sqrt{m}} \right\} \leq \delta, \tag{4.100}$$

where C_1 is a constant depending only on p_X and n. There are existing results on the sample complexity rates for estimators that are minimax optimal (see the analysis on the modified Kernel Density Estimator, KDE, in [HJWW20]) or near-optimal (see the analysis of the fixed k-nearest neighbor, k-NN, estimator in [JGH18]). These analyses show an upper bound on the root mean-square error $\mathbb{E}\left[\left(\widehat{h}(\mathcal{S}) - h(\boldsymbol{X}) \right)^2 \right]^{1/2}$ that is roughly of the order $(m \log m)^{-s/(s+d)} + m^{-1/2}$ or $m^{-s/(s+d)} \log m + m^{-1/2}$; here, $\boldsymbol{X}$ is a d-dimensional random vector satisfying certain regularity assumptions that are controlled by the smoothness parameter $s \in (0, 2]$, $\mathcal{S}$ is a set of m i.i.d. samples drawn according to $P_{\boldsymbol{X}}$, and $\widehat{h}$ is the modified KDE or k-NN estimator. When $d = 1$ and $s < 1$ (roughly, X is compactly supported and either does not vanish, or does not vanish smoothly, at the boundary), then the first terms in either of these bounds dominates the $m^{-1/2}$ term. Our bound in (4.100) contains the relevant asymptotic term $m^{-1/2}$, but it is given instead in terms of probability. Nevertheless, it may be converted to a root mean-square bound of order $\sqrt{(\log m)/m}$ (by choosing $\delta = 1/m$) under the assumption that the probability that the samples $\mathcal{S}$ are well-spaced is not too small, since then one may bound $\widehat{h}_n(\mathcal{S})$ almost surely and apply the reverse Markov inequality. It is worth noting that the sample complexity bound we give in Proposition 4.4 and (4.100) holds for all (compactly-supported) PDFs without any regularity assumptions of any kind. However, we also note that the constant in this bound is PDF-dependent.

From Proposition 4.4, we may also obtain a sample complexity result for the estimate $\widehat{I}_n$ in Definition 4.5.

Proposition 4.5. *Fix a finitely-supported discrete random variable X and a bounded-support continuous random variable $Y \in L^{2n}(P)$. There is a constant $C = C(X, Y, n)$ such that, for all small enough $\varepsilon, \delta > 0$, any collection $\mathcal{S}$ of i.i.d. samples drawn according to $P_{X,Y}$ of size*

$$|\mathcal{S}| > \frac{C}{\varepsilon^2} \log \frac{1}{\delta} \tag{4.101}$$

must satisfy

$$\Pr\left\{ \left| \widehat{I}_n(\mathcal{S}) - I_n(X;Y) \right| < \varepsilon \right\} \geq 1 - \delta. \tag{4.102}$$

Proof. See Appendix C.8.4. □

4.5.3 Numerical Results

We compare via synthetic experiments the performance of our estimators[11] against some of the estimators in the literature.

Our proposed estimator for differential entropy is $\widehat{h}_{10}$, i.e., given samples $\mathcal{S}$ of X we estimate $h(X)$ by $\widehat{h}_{10}(\mathcal{S})$ as given by Definition 4.4, for a large sample size (e.g., $|\mathcal{S}| > 600$), and it is $\widehat{h}_5$ for a smaller sample size (e.g., $|\mathcal{S}| \leq 600$). We compare this estimator with two estimation methods: k-Nearest-Neighbors (k-NN), and Kernel Density Estimation (KDE). The k-NN-based method we compare against is as provided by the Python package 'entropy_estimators' [Ste14], which we will refer to in this section as KSG. The kernel used for the KDE method is Gaussian, and it is obtained by computing from a set of samples $\{X_j\}_{j=1}^m$ a kernel Φ via the Python function 'scipy.stats.gaussian_kde' [VGO+20]; then, the estimate for differential entropy will be $\frac{-1}{m} \sum_{j=1}^m \log \Phi(X_j)$. The parameters for the KSG

[11] A Python code can be found at [AC21b].

and the KDE estimators are the default parameters, namely, $k = 3$ for the KSG estimator, and the bandwidth for the KDE estimator is chosen according to Scott's rule (i.e., $m^{-1/(d+4)}$ for a set of m samples of a d-dimensional random vector). We note that a more recent iteration of KDE has been proposed by Han *et al.* in [HJWW20], which improves the estimation for the non-smooth part of a PDF.

The mutual information is estimated using $\widehat{I}_5$, i.e., given samples $\mathcal{S}$ of (X, Y) our estimate for $I(X;Y)$ will be $\widehat{I}_5(\mathcal{S})$ as given by Definition 4.5. This estimator is compared against the partitioning estimator and the Mixed KSG estimator [GKOV17] (which is a k-NN-based estimator); we utilize the implementation in [GKOV17] for both estimators. In particular, the parameters are fixed throughout, namely, we utilize the parameters used in [GKOV17] ($k = 5$ for the Mixed KSG, and 8 bins per dimension for the partitioning estimator).

We perform 250 independent trials for each experiment and each fixed sample size, then plot the absolute error as a percentage of the true value (except for the last experiment, where the ground truth is 0, so we plot the absolute error) against the sample size. The sample sizes chosen for our experiments parallel those in [GKOV17], namely, $\{800, 1600, 2400, 3200, 4000\}$. To illustrate the smaller sample size regime, we repeat our Experiment 1 (estimating the differential entropy of Wigner's semicircle law) for sample sizes among $\{100, 200, 400, 600\}$. Since the PMMSE theory we developed in this chapter applies only to light-tailed distributions (e.g., those with MGFs), we restrict our experiments to such distributions.

We note that we also performed the mutual information experiments for the Noisy KSG estimator based on the estimator in [KSG04] (with noise strength $\sigma = 0.01$ as in [GKOV17]), but its performance was much worse than the other estimators, so we do not include it in the plots.

Remark 4.15. There is a trade-off between the approximation error $h_n(X) - h(X)$ and the estimation error $|\widehat{h}_n(\mathcal{S}) - h_n(X)|$ as the choice of the polynomial degree n varies. Indeed, as n increases, the approximation error vanishes, since we know that $h_n(X) \searrow h(X)$ by Theorem 4.4. On the other hand, the estimation error is expected to increase for large n, since the quality of estimating moments via sample moments deteriorates for higher moments and a fixed sample size. Evidently, similar trade-offs can be observed for other estimators in the literature, e.g., for the k-NN estimator one has bias-variance trade-off as k varies. Proposition 4.4 gives a characterization of the estimation error. To fully understand the best choice of n, one would need both a finer characterization of the constant $C(X, n)$ in Proposition 4.4 (namely, its dependence on n), and also a convergence rate refinement for $h_n(X) \searrow h(X)$ in Theorem 4.4 (see Remark 4.13). Note that the approximation error can be efficiently numerically computed for a given X and n (see Figure 4.2), and we report this value for the experiments we perform in this section. These experiments show that $n = 5$ gives a favorable estimation error compared to state-of-the-art estimators for moderate sample sizes ($m \leq 600$) and similarly $n = 10$ for larger support sizes ($m > 600$). We note that the compute time it takes to estimate $h(X)$ by $\widehat{h}_5(\mathcal{S})$ is comparable to that of both the k-NN and KDE estimators (in the order of seconds on a commercial laptop), and the compute time for $\widehat{h}_{10}(\mathcal{S})$ is in the order two minutes.

Experiment 1. We estimate the differential entropy of a random variable X distributed according to Wigner's semicircle distribution, i.e.,

$$p_X(x) := \frac{2}{\pi}\sqrt{1-x^2} \cdot 1_{[-1,1]}(x). \tag{4.103}$$

The ground truth is $h(X) \approx 0.64473$ nats. We generate a set $\mathcal{S}$ of i.i.d. samples distributed according to P_X. The size of $\mathcal{S}$ ranges from 800 to 4000 in increments of 800, and for each fixed sample size we independently generate 250 such sets $\mathcal{S}$ (so we generate a total of 1250 sets of samples). The differential entropy $h(X)$ is estimated by three methods: the moments-based estimator that we propose $\widehat{h}_{10}$, the k-NN-based estimator implemented in [Ste14] (which we refer to as the KSG estimator), and the Gaussian KDE estimator. For the proposed estimator, we use $\widehat{h}_{10}(\mathcal{S})$ as an estimate for $h(X)$. For the KSG estimator, we use the default setting, for which $k = 3$. We also use the default setting for the Gaussian KDE estimator; in particular, the bandwidth is chosen according to Scott's Rule as $m^{-1/(d+4)}$ where $m = |\mathcal{S}|$ and $d = 1$ is the dimensionality of X. The percentage relative absolute error in the estimation (e.g., $100 \cdot |\widehat{h}_{10}(\mathcal{S})/h(X) - 1|$, in %) is plotted against the sample size for the three estimators in Figure 4.3. The solid lines in Figure 4.3 are the means of the errors, i.e., the mean in the 250 independent trials of the percentage relative absolute error for each fixed sample size in $\{800, 1600, 2400, 3200, 4000\}$. Via bootstrapping, we infer confidence intervals, which are indicated by the shaded areas around the solid lines in Figure 4.3. We see that the proposed estimator outperforms the KSG estimator and the KDE estimator for this experiment. We note that we have the true value of the functional $h_{10}(X) \approx 0.64632$ nats (i.e., this is the value if we use the true first 20 moments of X instead of the corresponding sample moments obtained from i.i.d. samples). Hence, the approximation error is $h_{10}(X) - h(X) \approx 0.00159$ nats, i.e., $h_{10}(X)$ is approximately 99.75% accurate when approximating the ground truth $h(X)$ (so $100 - 100 \cdot (h_{10}(X) - h(X))/h(X) \approx 99.75$). For the sake of illustrating the case of smaller sample

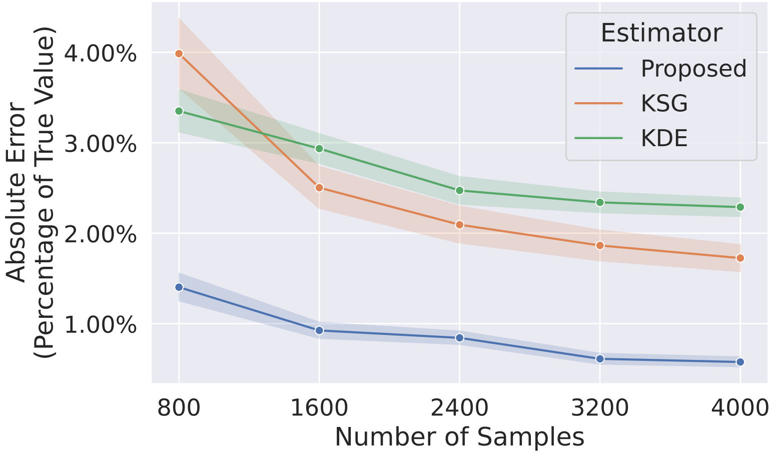


Figure 4.3: *Estimation of differential entropy for a semicircle distribution as in Experiment 1. The vertical axis shows the percentage relative absolute error in the estimation, e.g., for the proposed estimator it is* $100 \cdot |\widehat{h}_{10}(\mathcal{S})/h(X) - 1|$ *(%) where* $\mathcal{S}$ *is the set of samples and* $h(X) \approx 0.64473$ *nats is the ground truth. The horizontal axis shows* $|\mathcal{S}|$*, the sample size. The proposed estimator* $\widehat{h}_{10}$ *outperforms the k-NN-based estimator (denoted KSG) and the Gaussian KDE estimator for this experiment.*

sizes, we further carry out this experiment with sample sizes in the set $\{100, 200, 400, 600\}$. In this regime, we choose $n = 5$, i.e., our estimator is $\widehat{h}_5$. The results are illustrated in Figure 4.4. We also notice that the proposed estimator outperforms both the KSG and KDE estimators in this regime. In this case, $h_5(X) \approx 0.6509$ nats, so $h_5(X) - h(X) \approx 0.00617$ nats, giving $h_5(X)$ a 99.04% accuracy as an approximation for $h(X)$.

Experiment 2. We estimate the differential entropy $h(\boldsymbol{X})$ of a random vector $\boldsymbol{X} = (X_1, X_2)^T$ where X_1 and X_2 are i.i.d. distributed according to Wigner's semicircle distribution, namely, $\boldsymbol{X}$ has the PDF

$$p_{\boldsymbol{X}}(x, y) = \frac{4}{\pi^2}\sqrt{(1-x^2)(1-y^2)} \cdot 1_{[-1,1]\times[-1,1]}(x, y). \tag{4.104}$$

The ground truth is $h(\boldsymbol{X}) \approx 1.28946$ nats. The same numerical setup as in Experiment 1 is performed here. The results are plotted in Figure 4.5, where we see a similar behavior to the comparison in the 1-dimensional case; in particular, the proposed estimator outperforms the KSG estimator and the KDE estimator for this experiment. By independence of X_1 and X_2, we know that $h(\boldsymbol{X}) = 2h(X_1)$ and $h_{10}(\boldsymbol{X}) = 2h_{10}(X_1)$. Thus, we get the same relative approximation errors as in Experiment 1, namely, $h_{10}(\boldsymbol{X}) - h(\boldsymbol{X}) \approx 0.00318$ nats so $h_{10}(\boldsymbol{X})$ is approximately 99.75% accurate in approximating $h(\boldsymbol{X})$.

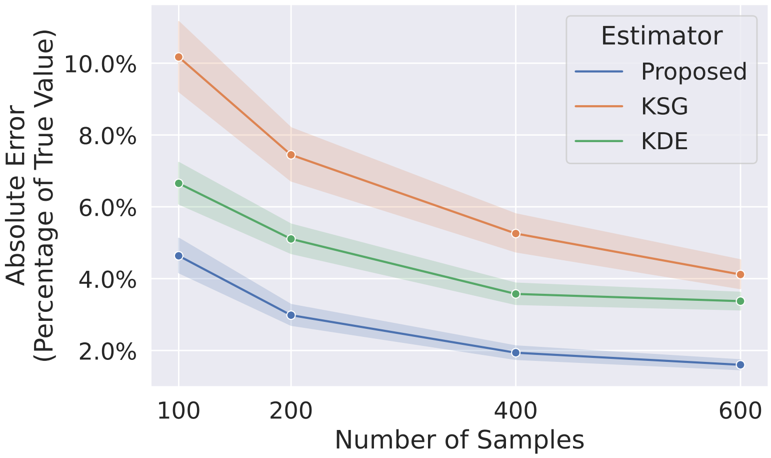


Figure 4.4: *Estimation of differential entropy for a semicircle distribution as in Experiment 1 for the small sample size regime* ($100 \leq m \leq 600$)*. In this regime, the plotted proposed estimator curve refers to the estimation of differential entropy using* $\widehat{h}_5$*, i.e.,* $n = 5$*. The proposed estimator outperforms both the KSG and the KDE estimators for this experiment in the small sample size regime too.*

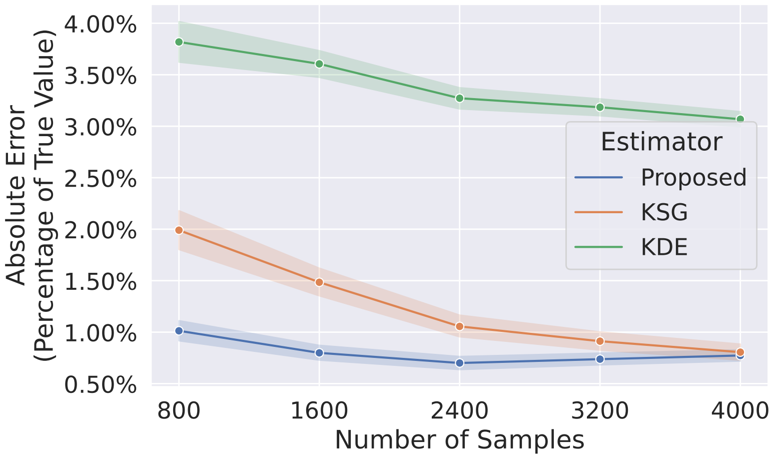


Figure 4.5: *Estimation of differential entropy for a 2-dimensional semicircle distribution as in Experiment 2. The proposed estimator* $\widehat{h}_{10}$ *outperforms both the KSG and the KDE estimators for this experiment.*

Experiment 3. We estimate the differential entropy $h(X)$ of a Gaussian mixture X whose PDF is given by

$$p_X(x) = \sum_{i=1}^{4} \frac{p_i}{\sqrt{2\pi\sigma_i^2}} e^{-(x-\mu_i)^2/(2\sigma_i^2)}, \tag{4.105}$$

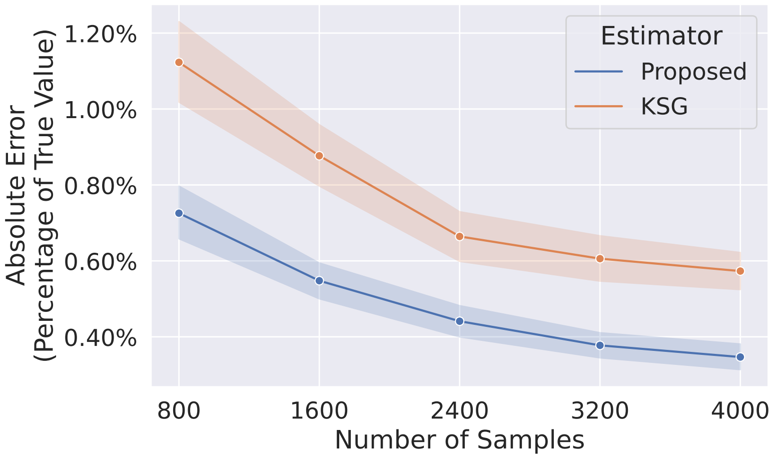


Figure 4.6: *Estimation of differential entropy for a Gaussian mixture as in Experiment 3. The proposed estimator $\widehat{h}_{10}$ outperforms both the KSG and KDE estimators for this experiment. The plot of the KDE estimator's performance is omitted to avoid cluttering, as it lies just above the line for the proposed estimator but overlaps significantly with its uncertainty region.*

where

$$\boldsymbol{p} = (0.1, 0.2, 0.3, 0.4) \tag{4.106}$$

$$\boldsymbol{\mu} = (-2, 0, 1, 5) \tag{4.107}$$

$$\boldsymbol{\sigma} = (1.5, 1, 2, 1). \tag{4.108}$$

The ground truth is $h(X) \approx 2.34249$ nats. The same numerical setup in Experiments 1 and 2 is used here. The results are plotted in Figure 4.6. For this experiment, the proposed estimator outperforms the KSG estimator, and it is essentially indistinguishable from the KDE estimator. Note that it is expected that the KDE estimator performs well in this Gaussian mixture experiment, since it is designed specifically to approximate densities by Gaussian mixtures. We have the true value $h_{10}(X) \approx 2.34817$ nats, so the approximation error is $h_{10}(X) - h(X) \approx 0.00568$ nats, making $h_{10}(X)$ approximately 99.76% accurate in approximating the true differential entropy $h(X)$.

Experiment 4. We estimate the differential entropy $h(\boldsymbol{X})$ of a random vector $\boldsymbol{X}$ that is a mixture of two Gaussians, namely, $\boldsymbol{X}$ has the PDF

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{4\pi\sqrt{\det(\boldsymbol{A})}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2} + \frac{1}{4\pi\sqrt{\det(\boldsymbol{B})}} e^{-(\boldsymbol{x}-\boldsymbol{\nu})^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{\nu})/2}, \tag{4.109}$$

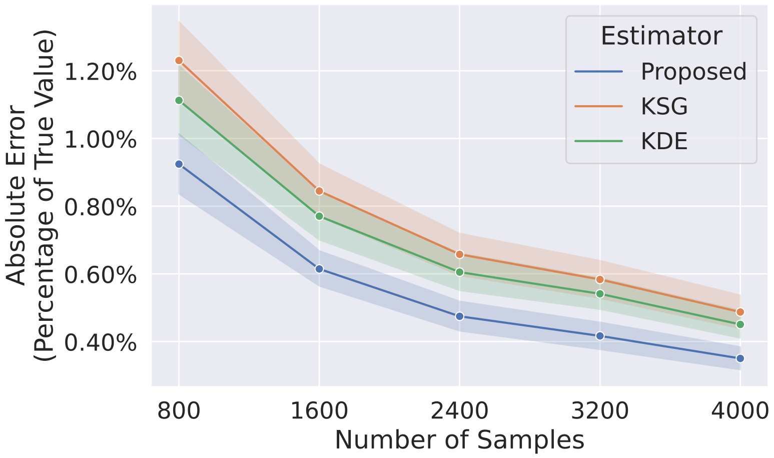


Figure 4.7: *Estimation of differential entropy for a vector Gaussian mixture as in Experiment 4. The proposed estimator* $\widehat{h}_{10}$ *outperforms both the KSG and KDE estimators for this experiment.*

where we have the means $\boldsymbol{\mu} = (-1,-1)^T$ and $\boldsymbol{\nu} = (1,1)^T$, and the covariance matrices

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \tag{4.110}$$

and $\boldsymbol{B} = \boldsymbol{I}_2$. The ground truth is $h(\boldsymbol{X}) \approx 3.22406$ nats. The same numerical setup as in Experiments 1–3 is performed here. The results are plotted in Figure 4.7. As in the 1-dimensional case in Experiment 3, the proposed estimator outperforms the KSG estimator for this experiment. Further, the proposed estimator also outperforms the KDE estimator in this 2-dimensional setting. We have the true value $h_5(\boldsymbol{X}) \approx 3.22846$ nats, so the approximation error is $h_5(\boldsymbol{X}) - h(\boldsymbol{X}) \approx 0.0044$ nats, making $h_5(\boldsymbol{X})$ approximately 99.86% accurate in approximating the true differential entropy $h(\boldsymbol{X})$.

Experiment 5. We replicate the mixture-distribution part of the zero-inflated Poissonization experiment of [GKOV17]. In detail, we let $Y \sim \mathrm{Exp}(1)$, and let $X = 0$ with probability 0.15 and $X \sim \mathrm{Pois}(y)$ given that $Y = y$ with probability 0.85. The quantity to be estimated is the mutual information $I(X;Y)$, and the ground truth is $I(X;Y) \approx 0.25606$ nats. We generate a set of i.i.d. samples $\mathcal{S}$ according to the distribution $P_{X,Y}$, where $\mathcal{S}$ has size in $\{800, 1600, 2400, 3200\}$. We estimate $I(X;Y)$ via the proposed estimator by $\widehat{I}_5(\mathcal{S})$, and we also consider the estimates given by the Mixed KSG estimator and the partitioning estimator, both as implemented in [GKOV17] (including the parameters used therein). This estimation process is repeated independently 250 times. The comparison of estimators' performance is plotted in Figure 4.8. The solid lines indicate the mean percentage relative absolute error, and the shaded areas indicate confidence intervals obtained via bootstrapping. We see in Figure 4.8 that the proposed estimator outperforms the other considered estimators for this experiment. We note that we have the true value $I_5(X;Y) \approx 0.24677$ nats, which gives an approximation error $|I_5(X;Y) - I(X;Y)| \approx 0.00929$ nats, i.e., $I_5(X;Y)$ is approximately 96.37% accurate in approximating $I(X;Y)$. We also test the affine-transformation invariance property of the proposed estimator. In particular, we consider estimating the mutual information from the scaled samples $\mathcal{S}'$ obtained from $\mathcal{S}$ via scaling the Y samples by 10^4, i.e.,

$$\mathcal{S}' := \{(A, 10^4 B) \ ; \ (A,B) \in \mathcal{S}\}. \tag{4.111}$$

Plotted in Figure 4.9 is a comparison of the same estimators using the same samples as those used to generate Figure 4.8, but where Y is processed through this affine transformation. The ground truth stays unchanged, and so do our estimator and the partitioning estimator, but the Mixed KSG estimates change. This experiment illustrates the resiliency of the proposed estimator to affine transformations. In fact, the computed numerical values in the modified setting by the proposed estimator differ by no more than 10^{-15} nats from those numerically computed in the original setting for each of the 1000 different sets of samples $\mathcal{S}$; in theory, these pairs of values are identical, and the less than 10^{-15} discrepancy is an artifact of the computer implementation. Finally, we note that although the setup is more general than the assumptions we prove our results under in this chapter (as X here is not finitely supported), the proposed estimator outperformed the other estimators.

Experiment 6. We test for independence under the following settings. We consider independent $X \sim \text{Bernoulli}(0.5)$ and $Y \sim \text{Unif}([0,2])$. We estimate $I(X;Y)$, whose true value is $I(X;Y) = 0$. We employ the same estimation procedure as in Experiment 5. The results are plotted in Figure 4.10, which shows that the proposed estimator predicted independence more accurately than the other estimators for the same sample size. Note that in this case the plot shows the absolute error (in nats) rather than the relative absolute error, as the ground truth is zero. In this case, the true value of $I_5(X;Y)$ is exactly equal to $I(X;Y)$, i.e., $I_5(X;Y)$ is 100% accurate in approximating $I(X;Y)$.

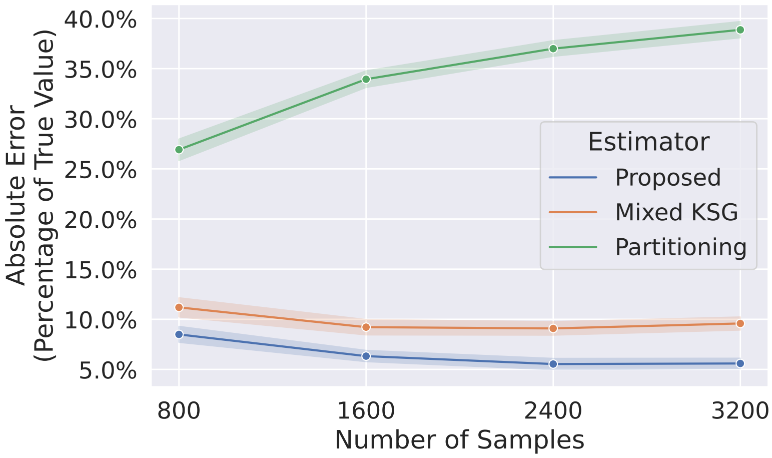


Figure 4.8: *Percentage relative absolute error vs. sample size for unscaled zero-inflated poissonization in Experiment 5. The proposed estimator $\widehat{I}_5$ outperforms both the k-NN-based estimator (denoted Mixed KSG) and the partitioning estimator.*

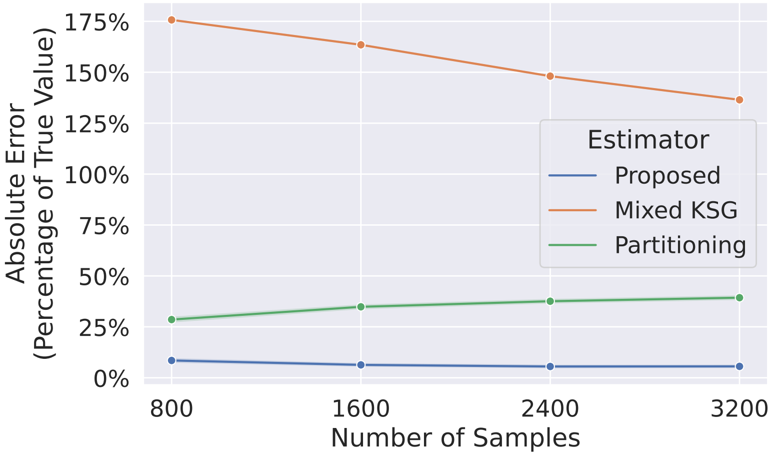


Figure 4.9: *Percentage relative absolute error vs. sample size for the scaled zero-inflated poissonization in Experiment 5. To generate these plots, we use the same samples that yield the plots in Figure 4.8, but we process them through an affine transformation. Specifically, each sample (A, B) is replaced with $(A, 10^4 B)$. Then the samples are passed to the three estimators. We see that the proposed estimator $\widehat{I}_5$ is resilient to scaling, i.e., the same performance line in Figure 4.8 is observed here too. This is in contrast to the performance of the Mixed KSG estimator. The partitioning estimator is resilient to scaling, but its performance is not favorable in this experiment (with above 25% relative absolute error).*

4.6 Conclusion

We investigate in this work the interplay between information measures and moments. Via developing the PMMSE, we give polynomial approximations of the conditional expectation. The PMMSE in turn yields new formulas for the differential entropy and mutual information in terms of the underlying

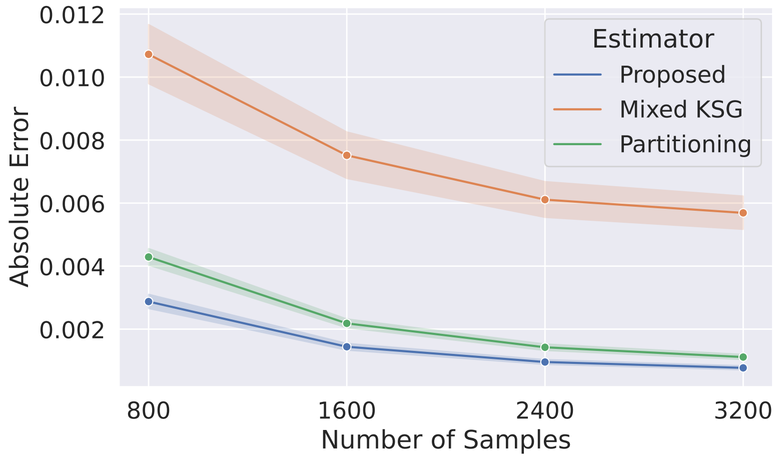


Figure 4.10: *Absolute error (in nats) vs. sample size for the independence testing in Experiment 6. The proposed estimator $\widehat{I}_5$ outperforms the Mixed KSG and the partitioning estimators in this experiment.*

moments. These formulas gave rise to a new estimator from data, where simply the moments are estimated from sample moments. The estimator is illustrated in several experiments that indicate a favorable performance as compared to the Gaussian KDE and k-NN estimators. For future work, it is worth investigating the finitary version of the convergence rate of the PMMSE to the MMSE, which would naturally yield convergence rates for the functionals h_n and I_n to the differential entropy and mutual information, and these in turn would tighten the sample complexity analysis. The proposed estimator's performance could also be compared with more recently developed estimators. It is interesting also to apply the PMMSE to the problem of estimating Fisher information, which is tightly related to the MMSE via Brown's identity [CDFP21]. Finally, the I-MMSE relation has been extended beyond Gaussian channels (e.g., Poisson channels [GSV08]), and it remains to be seen how the framework we develop in this chapter can shed light on those channels.

References

[AAB$^+$15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AAC$^+$a] Wael Alghamdi, Shahab Asoodeh, Flavio P. Calmon, Juan F. Gomez, Oliver Kosut, and Lalitha Sankar. Optimal differential privacy mechanisms for small sensitivity. Accepted in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023.

[AAC$^+$b] Wael Alghamdi, Shahab Asoodeh, Flavio P. Calmon, Juan F. Gomez, Oliver Kosut, and Lalitha Sankar. Optimal multidimensional differentially private mechanisms in the large-composition regime. Accepted in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023.

[AAC$^+$c] Wael Alghamdi, Shahab Asoodeh, Flavio P. Calmon, Juan F. Gomez, Oliver Kosut, and Lalitha Sankar. The saddle-point method in differential privacy. *Under review.*

[AAC$^+$22] Wael Alghamdi, Shahab Asoodeh, Flavio P. Calmon, Oliver Kosut, Lalitha Sankar, and Fei Wei. Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1838–1843, 2022.

[AAV19] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

[AAW$^+$20] Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P. Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2711–2716, 2020.

[ABD$^+$18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[AC19] W. Alghamdi and F. P. Calmon. Mutual information as a function of moments. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 3122–3126, 2019.

[AC21a] Wael Alghamdi and Flavio P. Calmon. Measuring information from moments. https://arxiv.org/abs/2109.00649v1, 2021.

[AC21b] Wael Alghamdi and Flavio P. Calmon. Measuring Information from Moments. https://github.com/WaelAlghamdi/MIE, 2021.

[AC21c] Wael Alghamdi and Flavio P. Calmon. Polynomial approximations of conditional expectations in scalar gaussian channels. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 420–425, 2021.

[AC22] Wael Alghamdi and Flavio P. Calmon. Measuring information from moments. *IEEE Transactions on Information Theory*, pages 1–1, 2022.

[ACG$^+$16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[Ada01] V. S. Adamchik. On the Barnes function. In *Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation*, ISSAC '01, page 15–20, New York, NY, USA, 2001. Association for Computing Machinery.

[AHJ$^+$22] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Winston Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and COMPAS: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*, 2022.

[ALC$^+$21] Shahab Asoodeh, Jiachun Liao, Flavio P Calmon, Oliver Kosut, and Lalitha Sankar. Three variants of differential privacy: Lossless conversion and applications. *IEEE Journal on Selected Areas in Information Theory*, 2(1):208–222, 2021.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[AS66] Sami M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of Royal Statistics*, 28:131–142, 1966.

[AS16] M. Ashok Kumar and I. Sason. Projection theorems for the Rényi divergence on α -convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935, Sep. 2016.

[AVW14] Himanshu Asnani, Kartik Venkat, and Tsachy Weissman. Relations Between Information and Estimation in the Presence of Feedback. *Lecture Notes in Control and Information Sciences*, 450 LNCIS:157–175, 2014.

[Bar00] A. R. Barron. Limits of information, markov chains, and projection. In *2000 IEEE International Symposium on Information Theory (Cat. No.00CH37060)*, pages 25–, June 2000.

[BBG18] B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NeurIPS*, pages 6280–6290, 2018.

[BC80] Robert M. Bell and Thomas M. Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, 5(2):161–166, 1980.

[BC81] Christian Berg and J. P. Reus Christensen. Density questions in the classical theory of moments. *Annales de l'institut Fourier*, 1981.

[BDH$^+$18] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[BDLS20] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie Su. Deep learning with Gaussian differential privacy. *Harvard Data Science Review*, 2(3), sep 30 2020. https://hdsr.mitpress.mit.edu/pub/u24wj42y.

[BDNP19] Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[Ber85] Christian Berg. On the preservation of determinacy under convolution. *Proceedings of the American Mathematical Society*, 93(2):351–357, 1985.

[BFG87] Jonathan M Borwein, SP Fitzpatrick, and JR Giles. The differentiability of real functions on normed linear space using generalized subgradients. *Journal of mathematical analysis and applications*, 128(2):512–534, 1987.

[BNBR19] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, page 97–110, 2013.

[BO13] Gilles Barthe and Federico Olmedo. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *International Colloquium on Automata, Languages, and Programming*, pages 49–60. Springer, 2013.

[Bog07] V. I. Bogachev. *Measure theory*. Springer, Berlin, 2007.

[BPC$^+$11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, jan 2011.

[BR19] Beáta Bényi and José L Ramírez. Some applications of S-restricted set partitions. *Periodica mathematica Hungarica*, 78(1):110–127, 2019.

[BS91] F. A. Berezin and M. Shubin. *The Schrödinger Equation*. Springer, Dordrecht, 1991.

[BV09] J. F. Bercher and C. Vignat. On minimum fisher information distributions with restricted support and fixed variance. *Inf. Sci.*, 179(22):3832–3842, 2009.

[BZZ$^+$21] Michelle Bao, Angela Zhou, Samantha A Zottola, Brian Brubach, Sarah Desmarais, Aaron Seth Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[Car51] Lennart Carleson. On Bernstein's approximation problem. *Proceedings of the American Mathematical Society*, 1951.

[Ç11] Erhan Çınlar. *Probability and Stochastics*. Springer, New York, NY, 2011.

[CDFP21] Wei Cao, Alex Dytso, Michael Fauß, and H. Vincent Poor. Finite-sample bounds on the accuracy of plug-in estimators of Fisher information. *Entropy*, 23(5), 2021.

[CDH$^+$19] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[CDPF$^+$17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[CF91] Raul E. Curto and Lawrence A. Fialkow. Recursiveness, positivity, and truncated moment problems. *Houston J. Math.*, 17(4):603–635, 1991.

[Che68] N. N. Chentsov. Nonsymmetrical distance between probability distributions, entropy and the theorem of pythagoras. *Mathematical notes of the Academy of Sciences of the USSR*, 4(3):686–691, Sep 1968.

[CHKV19] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CHS20] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems*, 33:15088–15099, 2020.

[CJG$^+$19] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20(172):1–59, 2019.

[CKV20] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, pages 1349–1359. PMLR, 2020.

[CLA$^+$10] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J. Bollo, Xudong Zhao, Evan Y. Snyder, Erik P. Sulman, Sandrine L. Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano, and Antonio Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 2010.

[CM03] I. Csiszar and F. Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, June 2003.

[CM18] S. Cha and T. Moon. Neural adaptive image denoiser. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2981–2985, 2018.

[Com] Creative Commons. Creative commons attribution-noderivs 3.0 unported license. https://creativecommons.org/licenses/by-nd/3.0/deed.en. 05/25/2022.

[CPW18] Flavio du Pin Calmon, Yury Polyanskiy, and Yihong Wu. Strong data processing inequalities for input constrained additive noise channels. *IEEE Transactions on Information Theory*, 64(3):1879–1892, 2018.

[Csi67] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[Csi75] Imre Csiszar. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 02 1975.

[Csi84] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 08 1984.

[Csi95a] Imre Csiszár. Generalized projections for non-negative functions. In *Proceedings of 1995 IEEE International Symposium on Information Theory*, pages 6–, Sep. 1995.

[Csi95b] Imre Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68(1):161–186, Mar 1995.

[CST17] Liang Chen, Defeng Sun, and Kim-Chuan Toh. A note on the convergence of admm for linearly constrained convex optimization problems. *Comput. Optim. Appl.*, 66(2):327–343, mar 2017.

[DBH$^+$22] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[DBPS17] Alex Dytso, Ronit Bustin, H. Vincent Poor, and Shlomo Shamai. A view of information-estimation relations in Gaussian networks. *Entropy*, 19(8):1–51, 2017.

[DC21] Alex Dytso and Martina Cardone. A general derivative identity for the conditional expectation with focus on the exponential family. In *2021 IEEE Information Theory Workshop (ITW)*, pages 1–6, 2021.

[DEHH21] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

[DGK$^+$22] Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. In *Privacy Enhancing Technologies Symposium (PETS)*, 2022.

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

[DHP$^+$12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Dif17] Differential privacy team Apple. Learning with privacy at scale, 2017.

[DKM$^+$06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *EUROCRYPT*, pages 486–503, 2006.

[DKS21] Mario Diaz, Peter Kairouz, and Lalitha Sankar. Lower bounds for the minimum mean-square error via neural network-based estimation. https://arxiv.org/abs/2108.12851, 2021.

[DL97] Z Ditzian and D. S Lubinsky. Jackson and smoothness theorems for Freud weights in L_p $(0 < p \leq \infty)$. *Constructive approximation*, 13(1):99–152, 1997.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006.

[DOBD$^+$18] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[Don88] David L. Donoho. One-sided inference about functionals of a density. *The Annals of Statistics*, 16(4):1390 – 1420, 1988.

[DPS20] A. Dytso, H. V. Poor, and S. Shamai Shitz. A general derivative identity for the conditional mean estimator in gaussian noise and some applications. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1183–1188, 2020.

[DPS21] Alex Dytso, H. Vincent Poor, and S. Shamai (Shitz). A general derivative identity for the conditional mean estimator in Gaussian noise and some applications. https://arxiv.org/abs/2104.01883, 2021.

[DRS22] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 02 2022.

[DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[DSGW03] Luc Devroye, Dominik Schäfer, László Györfi, and Harro Walk. The estimation problem of minimum mean squared error. *Statistics & Decisions*, 21(1):15–28, 2003.

[DT87] Z. Ditzian and V. Totik. *Moduli of Smoothness*. Springer New York, 1987.

[DV75] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[DV20] Alex Dytso and H. Vincent Poor. Estimation in Poisson noise: Properties of the conditional mean estimator. *IEEE Transactions on Information Theory*, 66(7):4304–4323, 2020.

[DY16] Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66:889–916, 2016.

[DZ96] Amir Dembo and Ofer Zeitouni. Refinements of the gibbs conditioning principle. *Probability theory and related fields*, 104(1):1–14, 1996.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[Ern17] Philip A. Ernst. Minimizing fisher information with absolute moment constraints. *Statistics & Probability Letters*, 129:167–170, 2017.

[ET99] Ivar Ekeland and Roger Témam. *Convex analysis and variational problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1999.

[FFM$^+$15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[Fle04] Francois Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 2004.

[Fre77] Geza Freud. On Markov-Bernstein-type inequalities and their applications. *Journal of Approximation Theory*, 1977.

[FS18] Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2018.

[FS19] Farhad Farokhi and Henrik Sandberg. Ensuring privacy with constrained additive noise by minimizing fisher information. *Automatica*, 99:275–288, 2019.

[FSV$^+$19] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[FW08] Peter Forrester and SVEN Warnaar. The importance of the selberg integral. *Bulletin of the American Mathematical Society*, 45(4):489–534, 2008.

[GDGK19] Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Optimal noise-adding mechanism in additive differential privacy. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 11–20. PMLR, 16–18 Apr 2019.

[GDGK20] Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 89–99, 2020.

[GGNWP20] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

[GGWP19] Ziv Goldfeld, Kristjan Greenewald, Jonathan Weed, and Yury Polyanskiy. Optimality of the plug-in estimator for differential entropy estimation under Gaussian convolutions. In *IEEE International Symposium on Information Theory - Proceedings*, 2019.

[GKKM22] Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Faster privacy accounting via evolving discretization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7470–7483. PMLR, 17–23 Jul 2022.

[GKOV15] Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

[GKOV17] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. *Advances in Neural Information Processing Systems*, 2017-Decem:5987–5998, 2017.

[GLW21] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[GMV$^+$21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[GNO$^+$12] Hani Goodarzi, Hamed S. Najafabadi, Panos Oikonomou, Todd M. Greco, Lisa Fish, Reza Salavati, Ileana M. Cristea, and Saeed Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 2012.

[GP17] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[GRS12] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

[GS10] Mangesh Gupte and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, page 135–146, 2010.

[GSV05] Dongning Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[GSV08] Dongning Guo, Shlomo Shamai, and Sergio Verdu. Mutual information and conditional mean estimation in Poisson channels. *IEEE Transactions on Information Theory*, 54(5):1837–1849, 2008.

[Guo09] Dongning Guo. Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation. *IEEE International Symposium on Information Theory - Proceedings*, pages 814–818, 2009.

[GV15] Quan Geng and Pramod Viswanath. The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2):925–951, 2015.

[GV16] Quan Geng and Pramod Viswanath. Optimal noise adding mechanisms for approximate differential privacy. *IEEE Transactions on Information Theory*, 62(2):952–969, 2016.

[GWSV11] Dongning Guo, Yihong Wu, Shlomo Shamai, and Sergio Verdú. Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4):2371–2385, 2011.

[Hal13] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[HJW15] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

[HJWW20] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228 – 3250, 2020.

[HPS$^+$16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[HR09] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics, Second Edition*. Wiley, 2009.

[HR19] Bruce Hajek and Maxim Raginsky. Statistical learning theory. *Lecture Notes*, 387, 2019.

[INE20] INEP. Instituto nacional de estudos e pesquisas educaionais anísio teixeira, microdados do ENEM. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, 2020. Accessed: 2022-05-23.

[IPH$^+$11] Steven J Ingels, Daniel J Pratt, Deborah R Herget, Laura J Burns, Jill A Dever, Randolph Ottem, James E Rogers, Ying Jin, and Steve Leinwand. High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*, 2011.

[JGH18] Jiantao Jiao, Weihao Gao, and Yanjun Han. The nearest neighbor information estimator is adaptively near minimax rate-optimal. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[JJ99] Harold Jeffreys and Bertha Jeffreys. *Methods of Mathematical Physics*. Cambridge University Press, 3rd edition, 1999.

[JN19] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*, 2019.

[JN20] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[JSW22] Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[JVHW15] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[JWC22] Haewon Jeong, Hao Wang, and Flavio Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[Kag86] A. M. Kagan. Information property of exponential families. *Theory of Probability & Its Applications,*, 30(4):831–835, 1986.

[KCT20] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. FACT: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR, 2020.

[KGZ19] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[KH21] Antti Koskela and Antti Honkela. Computing differential privacy guarantees for heterogeneous compositions using fft. *CoRR*, abs/2102.12412, 2021.

[KJH20] Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR, 2020.

[KJPH21] Antti Koskela, Joonas Jälkö, Lukas Prediger, and Antti Honkela. Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using fft. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3358–3366. PMLR, 13–15 Apr 2021.

[KJW$^+$21] Anilesh Krishnaswamy, Zhihao Jiang, Kangning Wang, Yu Cheng, and Kamesh Munagala. Fair for all: Best-effort fairness guarantees for classification. In *International Conference on Artificial Intelligence and Statistics*, pages 3259–3267. PMLR, 2021.

[KKZ12] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, Dec 2012.

[KLN$^+$11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, jun 2011.

[KMR$^+$20] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. Guidelines for implementing and auditing differentially private systems. *ArXiv*, abs/2002.04049, 2020.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1376–1385, 2015.

[KS15a] M. A. Kumar and R. Sundaresan. Minimization problems based on relative α-entropy i: Forward projection. *IEEE Transactions on Information Theory*, 61(9):5063–5080, Sep. 2015.

[KS15b] M. A. Kumar and R. Sundaresan. Minimization problems based on relative α-entropy ii: Reverse projection. *IEEE Transactions on Information Theory*, 61(9):5081–5095, Sep. 2015.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(6):16, 2004.

[KTW04] R. C. King, F. Toumazet, and B. G. Wybourne. The square of the Vandermonde determinant and its q-generalization. *Journal of Physics A: Mathematical and General*, 37(3):735–767, 2004.

[Kul59] Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959.

[KZ06] Andrew J Kurdila and Michael Zabarankin. *Convex functional analysis*. Springer Science & Business Media, 2006.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[LPB$^+$21] Andrew Lowy, Rakesh Pavan, Sina Baharlouei, Meisam Razaviyayn, and Ahmad Beirami. Fermi: Fair empirical risk minimization via exponential Rényi mutual information. *arXiv preprint arXiv:2102.12586*, 2021.

[LTV06] Angel Lozano, Antonia M. Tulino, and Sergio Verdú. Optimum power allocation for parallel Gaussian channels with arbitrary input distributions. *IEEE Transactions on Information Theory*, 52(7):3033–3051, 2006.

[Lub07] D.s Lubinsky. A survey of weighted polynomial approximation with exponential weights. *Surveys in Approximation Theory*, 3:1–105, 2007.

[Mha96] Hrushikesh Narhar Mhaskar. Introduction to the theory of weighted polynomial approximation. In *Series in approximations and decompositions*, 1996.

[Mir17] Ilya Mironov. Rényi differential privacy. In *Proc. IEEE Comp. Security Foundations Symp. (CSF)*, pages 263–275, 2017.

[MM18] Sebastian Meiser and Esfandiar Mohammadi. Tight on budget? tight bounds for r-fold approximate differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 247–264, 2018.

[MTZ19] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[MV16] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Proc. Int. Conf. Theory of Cryptography*, pages 157–175, 2016.

[MW18] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[MZ17] Anuran Makur and Lizhong Zheng. Polynomial singular value decompositions of a family of source-channel models. *IEEE Transactions on Information Theory*, 63(12):7716–7728, 2017.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, Boston, MA, 2004.

[NIS] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.4 of 2022-01-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[OEI] OEIS Foundation Inc. (2021), The On-Line Encyclopedia of Integer Sequences.

[Pet82] L. C. Petersen. On the relation between the multidimensional moment problem and the one-dimensional moment problem. *Mathematica Scandinavica*, 51:361–366, Jun. 1982.

[Pol19] Yury Polyanskiy. Lecture notes on information theory, 2019.

[Pos75] E. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

[PPV10] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

[PQC$^+$19] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

[PRW$^+$17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[PW16] Y. Polyanskiy and Y. Wu. Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55, 2016.

[Roc09] R. Tyrrell Rockafellar. *Variational analysis*. Grundlehren der mathematischen Wissenschaften ; 317. Springer, Berlin ; Heidelberg, 1st ed. 1998. edition, 2009.

[SCDF13] Jordi Soria-Comas and Josep Domingo-Ferrer. Optimal data-independent noise for differential privacy. *Information Sciences*, 250(Complete):200–214, 2013.

[Sch17] Konrad Schmüdgen. *The Moment Problem*. Springer Cham, 2017.

[Sim98] Barry Simon. The classical moment problem as a self-adjoint finite difference operator. *Advances in Mathematics*, 137(1):82–203, 1998.

[Slo02] Noam Slonim. The information bottleneck: Theory and applications. *Ph.D Thesis*, 01 2002.

[SMM19] David M Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Proceedings on Privacy Enhancing Technologies*, 2019(2):245–269, 2019.

[SS19] Elias M. Stein and Rami Shakarchi. *Real Analysis*. Princeton University Press, 2019.

[Ste14] Greg Ver Steeg. NPEET. https://github.com/gregversteeg/NPEET, 2014.

[STW94] T. Scharf, J. Thibon, and B. G. Wybourne. Powers of the Vandermonde determinant and the quantum Hall effect. *Journal of Physics A: General Physics*, 27(12):4211–4219, 1994.

[Top79] Flemming Topsøe. Information-theoretical optimization techniques. *Kybernetika*, 15(1):(8)–27, 1979.

[TV06] Antonio M. Tulino and Sergio Verdú. Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof. *IEEE Transactions on Information Theory*, 52(9):4295–4297, 2006.

[UK95] Elke Uhrmann-Klingen. Minimal fisher information distributions with compact-supports. *Sankhyā: The Indian Journal of Statistics*, 57(3):360–374, 1995.

[Ver10] Sergio Verdú. Mismatched estimation and relative entropy. *IEEE Transactions on Information Theory*, 56(8):3712–3720, 2010.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. *Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 685–694, 2011.

[WGZ$^+$22] Hua Wang, Sheng Gao, Huanyu Zhang, Milan Shen, and Weijie J. Su. Analytical composition of differential privacy via the edgeworth accountant. *arXiv preprint arXiv:2206.04236*, 2022.

[WRC20] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. Optimized score transformation for fair classification. In *23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[WRC21] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. Optimized score transformation for consistent fair classification. *Journal of Machine Learning Research*, 22(258):1–78, 2021.

[WV11] Yihong Wu and Sergio Verdú. MMSE dimension. *IEEE Transactions on Information Theory*, 57(8):4857–4879, 2011.

[WV12] Y. Wu and S. Verdu. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, 2012.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[YB17] Xiao Yang and Andrew R Barron. Minimax compression and large alphabet approximation through poissonization and tilting. *IEEE Transactions on Information Theory*, 63(5):2866–2884, 2017.

[YCK20] Forest Yang, Mouhamadou Cisse, and Oluwasanmi O Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

[YX20] Qing Ye and Weijun Xie. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*, 2020.

[Zak05] Moshe Zakai. On mutual information, likelihood ratios, and estimation error for the additive Gaussian channel. *IEEE Transactions on Information Theory*, 51(9):3017–3024, 2005.

[ZDW21] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. *arXiv preprint arXiv:2106.08567*, 2021.

[ZDW22] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[ZVRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[ZW19] Yuqing Zhu and Yu-Xiang Wang. Poission subsampled rényi differential privacy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97, pages 7634–7642, 09–15 Jun 2019.

Appendix A

Appendix to Chapter 2

A.1 Further Comparisons with the Literature

We contrast in this appendix our Fisher information minimization contribution with the relevant literature.

- [HR09, Example 5.1]: Although there is no general statement (e.g., a theorem) in [HR09] showing a result similar to our result in Theorem 2.13, one can distill from Section 4.5 in [HR09] a claim that roughly translates as follows. For a PDF p to uniquely minimize the Fisher information over all PDFs satisfying $\mathbb{E}_p[c] \le C$, it suffices to satisfy the following: **(i)** p is strictly positive, absolutely continuous, and twice differentiable, **(ii)** the following integration by parts holds[1] for the ratio $\psi = p'/p$

  $$\int_{\mathbb{R}} \psi(x)(q'(x) - p'(x))\, dx = -\int_{\mathbb{R}} \psi'(x)(q(x) - p(x))\, dx \tag{A.1}$$

  for *every* PDF q with $I(q) < \infty$ and $\mathbb{E}_q[c] \le C$, and **(iii)** there is a $\theta > 0$ such that $y = \sqrt{p}$ uniquely solves the Schrödinger equation $y'' = (\theta c - E)y$ with E being the smallest possible constant. Example 5.1 of [HR09] gives full details for the special case when $c(x) = -a \cdot 1_{|x| \le 1} + b \cdot 1_{|x|>1}$ (and notes the well-known case $c(x) = x^2$). In contrast, our results on Fisher information minimization assumes none of the assumptions made in [HR09]; rather, we *derive* similar results that are required for our proof technique to follow through (e.g., via proving

[1] We note that the integration by parts in equation (A.1) should not be expected to hold for arbitrary cost c.

Proposition 2.2).

- [Ern17]: The derivations therein assume without proof some of the above mentioned properties regarding [HR09], such as positivity, smoothness, and the validity of the integration by parts in (A.1); there are no worked examples in [Ern17]. Similarly to our comparison with [HR09], we derive rather than assume the required properties.

- The use of Fisher information for optimizing privacy has appeared in [FS18, FS19]. However, in these papers, rather than connecting DP to Fisher information, the authors set up the privacy problem as one where Fisher information is to be minimized. Then, the problem of minimizing Fisher information is connected to the Schrödinger equation. However, we note that the mathematical setup for the Fisher-information minimization problems in [FS18, FS19] is different from, and less general than, what we consider herein. Recall that we derive the unique minimizers of the Fisher information $I(p)$ for $p \in \mathcal{P}(\mathbb{R})$, i.e., over *all* possible PDFs, subject to the constraint $\mathbb{E}_p[c] \le C$ where c satisfies Assumption 2.2. In contrast, [FS18] considers only bounded-support PDFs that are also twice continuously differentiable. The work in [FS18] is extended in [FS19] to consider unbounded-support PDFs, but subject to two restrictions: the PDF must be twice continuously differentiable, and the cost constraint is the variance cost constraint. Again, we do not assume these properties *a priori*, but derive whatever properties are necessary for our approach.

A.2 Proof of Theorem 2.1

First, if $\mathrm{KL}_{\max} = 0$, then we have both $\varepsilon_{P_{Y|X}^{\circ k}}(\delta) = 0$ and $\mathrm{V}_{\max} = 0$, and there is nothing to prove. So, assume $\mathrm{KL}_{\max} > 0$. It is not hard to see that this implies $\mathrm{V}_{\max} > 0$ too. The proof is divided into the following steps.

• *Step 1: applying the CLT.*

By the definition of $\delta_{P_{Y|X}^{\circ k}}$ given in (2.27), we may write

$$\delta_{P_{Y|X}^{\circ k}}(\varepsilon) = \sup_{\|u_j - v_j\| \le s,\, j \in [k]} \mathsf{E}_{e^{\varepsilon}}\left(\prod_{j\in[k]} P_{Y|X=u_j} \middle\| \prod_{j\in[k]} P_{Y|X=v_j} \right). \tag{A.2}$$

Define the functions $f_k : (\mathbb{R}^m)^k \times (\mathbb{R}^m)^k \times [0,\infty) \to [0,1]$ by

$$f_k(u^{(k)}, v^{(k)}, \varepsilon) := \mathsf{E}_{e^{\varepsilon}}\left(\prod_{j\in[k]} P_{Y|X=u_j} \Big\| \prod_{j\in[k]} P_{Y|X=v_j}\right), \tag{A.3}$$

so we have

$$\delta_{P_{Y|X}^{\circ k}}(\varepsilon) = \sup_{\|u_j - v_j\| \le s,\, j\in[k]} f_k(u^{(k)}, v^{(k)}, \varepsilon). \tag{A.4}$$

For each pair $u^{(k)}, v^{(k)} \in (\mathbb{R}^m)^k$ with $\|u_j - v_j\| \le s$ for $1 \le j \le k$, let $L_{u^{(k)},v^{(k)}} := \sum_{j\in[k]} L_{u_j,v_j}$ where the L_{u_j,v_j} are independent PLRVs defined as in (2.2). By assumption of finiteness of the KL-divergence, we have the equivalence of measures, so we may write

$$f_k(u^{(k)}, v^{(k)}, \varepsilon) = \mathbb{E}\left[\left(1 - e^{\varepsilon - L_{u^{(k)},v^{(k)}}}\right)^+\right]. \tag{A.5}$$

We apply the CLT to this expectation.

First, we note that we may through away pairs $(u^{(k)}, v^{(k)}) \in \mathbb{R}^{2mk}$ for which $\mathrm{V}_{u^{(k)},v^{(k)}} = 0$. Indeed, if $\mathrm{V}_{u^{(k)},v^{(k)}} = 0$, then $L_{u^{(k)},v^{(k)}} = \mathrm{KL}_{u^{(k)},v^{(k)}}$ almost surely. But then we would have $1 = \mathbb{E}[e^{-L_{u^{(k)},v^{(k)}}}] = \prod_{j\in[k]} e^{-\mathrm{KL}_{u_j,v_j}}$, which would imply in view of nonnegativity of the KL-divergence that $L_{u^{(k)},v^{(k)}} = 0$. In this case, we have $f_k(u^{(k)}, v^{(k)}, \varepsilon) = 0$ for every $\varepsilon \ge 0$. Therefore, by nonnegativity of the f_k, for the purpose of maximizing $(u^{(k)}, v^{(k)}) \mapsto f_k(u^{(k)}, v^{(k)}, \varepsilon)$ we may exclude pairs $(u^{(k)}, v^{(k)})$ for which $\mathrm{V}_{u^{(k)},v^{(k)}} = 0$. Denote the restricted sets

$$\mathcal{V}_k := \left\{(u^{(k)}, v^{(k)}) \in \mathbb{R}^{2mk} \;:\; \|u_j - v_j\| \le s \text{ for } 1 \le j \le k, \text{ and } \mathrm{V}_{u^{(k)},v^{(k)}} > 0\right\}. \tag{A.6}$$

Then, for each $k \in \mathbb{N}$ and $\varepsilon \ge 0$,

$$\delta_{P_{Y|X}^{\circ k}}(\varepsilon) = \sup_{\|u_j - v_j\| \le s, j\in[k]} f_k(u^{(k)}, v^{(k)}, \varepsilon) = \sup_{(u^{(k)},v^{(k)})\in\mathcal{V}_k} f_k(u^{(k)}, v^{(k)}, \varepsilon). \tag{A.7}$$

Now, fix $(u^{(k)}, v^{(k)}) \in \mathcal{V}_k$, i.e., $\|u_j - v_j\| \le s$ for $j \in [k]$ and $\mathrm{V}_{u^{(k)},v^{(k)}} > 0$, and we will derive bounds on $f_k(u^{(k)}, v^{(k)}, \varepsilon)$. Consider $W_{u^{(k)},v^{(k)}} \sim \mathcal{N}(\mathrm{KL}_{u^{(k)},v^{(k)}}, \mathrm{V}_{u^{(k)},v^{(k)}})$. By the CLT there is a function $r(k) = o(1)$ (uniformly in $(u^{(k)}, v^{(k)}, \varepsilon)$ by assumption of uniformly bounded variances) such that

$$\begin{aligned} f_k(u^{(k)}, v^{(k)}, \varepsilon) &= \mathbb{E}\left[\left(1 - e^{\varepsilon - L_{u^{(k)},v^{(k)}}}\right)^+\right] & \text{(A.8)}\\ &= \int_0^1 \mathbb{P}\left[L_{u^{(k)},v^{(k)}} > \varepsilon - \log(1-u)\right] du & \text{(A.9)}\\ &= \int_0^1 \mathbb{P}\left[W_{u^{(k)},v^{(k)}} > \varepsilon - \log(1-u)\right] du + r(k) & \text{(A.10)} \end{aligned}$$

$$= \mathbb{E}\left[\left(1 - e^{\varepsilon - \mathrm{W}_{u^{(k)},v^{(k)}}}\right)^{+}\right] + r(k) \tag{A.11}$$

$$= Q\left(\frac{\varepsilon - \mathrm{KL}_{u^{(k)},v^{(k)}}}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}\right) - e^{\varepsilon - \mathrm{KL}_{u^{(k)},v^{(k)}} + \mathrm{V}_{u^{(k)},v^{(k)}}/2}\, Q\left(\frac{\varepsilon - \mathrm{KL}_{u^{(k)},v^{(k)}} + \mathrm{V}_{u^{(k)},v^{(k)}}}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}\right) + r(k) \tag{A.12}$$

where Q denotes the Gaussian Q-function. Next, we use (A.12) to investigate the limits of $f_k(u^{(k)}, v^{(k)}, \underline{\varepsilon}_k(\delta))$ and $f_k(u^{(k)}, v^{(k)}, \overline{\varepsilon}_k(\delta))$ for specific values of $\underline{\varepsilon}_k(\delta)$ and $\overline{\varepsilon}_k(\delta)$.

• *Step 2: an upper bound on ε.*

Let $k_0 \in \mathbb{N}$ be such that $r(k) \in (\delta - 1/2, \delta)$ whenever $k \geq k_0$. For each $k \geq k_0$, define the constant

$$\overline{\varepsilon}_k(\delta) := k \cdot \mathrm{KL}_{\max} - \Phi^{-1}(\delta - r(k)) \cdot \sqrt{k \cdot \mathrm{V}_{\max}}, \tag{A.13}$$

which we will show is an upper bound on $\varepsilon_{P_{Y|X}^{\circ k}}(\delta)$. We do this by showing the bound $\delta_{P_{Y|X}^{\circ k}}(\overline{\varepsilon}_k(\delta)) \leq \delta$ using (A.12), then inverting it. Specifically, an upper bound on f_k may be given by (note that $Q(z) = \Phi(-z)$ for $z \geq 0$)

$$f_k(u^{(k)}, v^{(k)}, \overline{\varepsilon}_k(\delta)) \leq \Phi\left(\sup_{(u^{(k)},v^{(k)}) \in \mathcal{V}_k} \frac{\mathrm{KL}_{u^{(k)},v^{(k)}} - \overline{\varepsilon}_k(\delta)}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}\right) + r(k) \tag{A.14}$$

for every $(u^{(k)}, v^{(k)}) \in \mathcal{V}_k$. Since $\mathrm{KL}_{u^{(k)},v^{(k)}} \leq k \cdot \mathrm{KL}_{\max} < \overline{\varepsilon}_k(\delta)$, we conclude from the definition of $\overline{\varepsilon}_k(\delta)$ that

$$f_k(u^{(k)}, v^{(k)}, \overline{\varepsilon}_k(\delta)) \leq \Phi\left(\Phi^{-1}(\delta - r(k))\right) + r(k) \leq \delta. \tag{A.15}$$

Therefore, maximizing over $(u^{(k)}, v^{(k)}) \in \mathcal{V}_k$, we conclude that (see (A.7))

$$\delta_{P_{Y|X}^{\circ k}}(\overline{\varepsilon}_k(\delta)) \leq \delta. \tag{A.16}$$

Inverting this inequality, we get the upper bound

$$\varepsilon_{P_{Y|X}^{\circ k}}(\delta) \leq \overline{\varepsilon}_k(\delta). \tag{A.17}$$

• *Step 3: a general lower bound on ε.*

We similarly lower bound f_k, but we now require a more delicate argument. We first show a general asymptotic lower bound $k \cdot \mathrm{KL}_{\max}$ on $\varepsilon_{P_{Y|X}^{\circ k}}(\delta)$, then we refine it in the case $\underline{\mathrm{V}} > 0$.

For the general case, fix any $\tau \in (0, \mathrm{KL}_{\max})$, and we will show that $\varepsilon_{P_{Y|X}^{\circ k}}(\delta) \geq k\tau$ for all large k by showing that $\delta_{P_{Y|X}^{\circ k}}(k\tau) > \delta$. Let $\{(u_j, v_j)\}_{j \in \mathbb{N}} \subset \mathbb{R}^{2m}$ be a sequence with $\|u_j - v_j\| \leq s$ and $\mathrm{KL}_{u_j,v_j} \to$

$\mathrm{KL}_{\max}$ as $j \to \infty$. We may assume that $\mathrm{KL}_{u_j,v_j} > 0$ for each j. Let $\mathrm{V}' := \liminf_{j\to\infty} \mathrm{V}_{u_j,v_j}$. If necessary, we replace the (u_j, v_j) by a subsequence so that $\mathrm{V}_{u_j,v_j} \to \mathrm{V}'$. Note that $\mathrm{V}_{(u_j,v_j)} > 0$ for each j since $\mathrm{KL}_{u_j,v_j} > 0$. Denote the constant k-tuples $u_j^{(k)} := (u_j, \cdots, u_j) \in \mathbb{R}^{mk}$ and $v_j^{(k)} := (v_j, \cdots, v_j) \in \mathbb{R}^{mk}$. For all large j, we have that $\mathrm{KL}_{u_j,v_j} > \tau$. Hence, for all large j,

$$Q\left(\frac{k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}}}{\sqrt{\mathrm{V}_{u_j^{(k)},v_j^{(k)}}}}\right) \geq Q(0) = \frac{1}{2}. \tag{A.18}$$

Next, we show that the second term in (A.12) can be made arbitrarily small. We have the limit

$$\lim_{j\to\infty} k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \frac{1}{2}\cdot \mathrm{V}_{u_j^{(k)},v_j^{(k)}} = k \cdot \left(\tau - \mathrm{KL}_{\max} + \frac{1}{2}\cdot \mathrm{V}'\right). \tag{A.19}$$

We consider two cases according to whether $\mathrm{V}' < 2(\mathrm{KL}_{\max} - \tau)$. Assume for now that $\mathrm{V}' < 2(\mathrm{KL}_{\max} - \tau)$ holds. Then, the limit in (A.19) is negative and the second term in (A.12) can be made arbitrarily small by choosing $(u^{(k)}, v^{(k)}) = (u_j^{(k)}, v_j^{(k)})$ for large j and k. Indeed, let $j_0 \in \mathbb{N}$ be such that $j \geq j_0$ implies

$$\tau - \mathrm{KL}_{u_j,v_j} + \frac{1}{2}\cdot \mathrm{V}_{u_j,v_j} < \frac{1}{2}\cdot\left(\tau - \mathrm{KL}_{\max} + \frac{1}{2}\cdot \mathrm{V}'\right) =: \theta_0 < 0. \tag{A.20}$$

Hence, in this case, bounding the Q-function from above by 1, the second term in (A.12) is bounded by

$$\exp\left(k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \frac{1}{2}\mathrm{V}_{u_j^{(k)},v_j^{(k)}}\right)\ Q\left(\frac{k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \mathrm{V}_{u_j^{(k)},v_j^{(k)}}}{\sqrt{\mathrm{V}_{u_j^{(k)},v_j^{(k)}}}}\right) \leq e^{\theta_0 k} \tag{A.21}$$

for $j \geq j_0$. Hence, in this case, we obtain the bound (see (A.18))

$$\sup_{(u^{(k)},v^{(k)})\in\mathcal{V}_k} f_k(u^{(k)}, v^{(k)}, k\tau) \geq \frac{1}{2} - o(1) \tag{A.22}$$

for a function $o(1)$ that goes to zero as $k \to \infty$ (for instance, it may be taken as $e^{\theta_0 k} - r(k)$). Now, assume instead that $\mathrm{V}' \geq 2(\mathrm{KL}_{\max} - \tau)$. Let $j_1 \in \mathbb{N}$ be such that $j \geq j_1$ implies

$$\tau - \mathrm{KL}_{u_j,v_j} + \cdot \mathrm{V}_{u_j,v_j} \geq \frac{1}{2}\cdot(\mathrm{KL}_{\max} - \tau) =: \theta_1 > 0. \tag{A.23}$$

In this case, using the bound $Q(z) \leq \frac{1}{\sqrt{2\pi}z}e^{-z^2/2}$ for $z > 0$, we may bound the second term in (A.12) by

$$\begin{aligned}
&\exp\left(k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \frac{1}{2}\mathrm{V}_{u_j^{(k)},v_j^{(k)}}\right)\ Q\left(\frac{k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \mathrm{V}_{u_j^{(k)},v_j^{(k)}}}{\sqrt{\mathrm{V}_{u_j^{(k)},v_j^{(k)}}}}\right)\\
&\quad\le \frac{\sqrt{\mathrm{V}_{u_j^{(k)},v_j^{(k)}}}}{\sqrt{2\pi}\cdot\left(k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}} + \mathrm{V}_{u_j^{(k)},v_j^{(k)}}\right)}\exp\left(-\frac{1}{2}\left(\frac{k\tau - \mathrm{KL}_{u_j^{(k)},v_j^{(k)}}}{\sqrt{\mathrm{V}_{u_j^{(k)},v_j^{(k)}}}}\right)^2\right) \qquad \text{(A.24)}\\
&\quad\le \frac{\sqrt{\mathrm{V}_{\max}}}{\sqrt{2\pi}\cdot\theta_1}\cdot\frac{1}{\sqrt{k}}
\end{aligned}$$

for all large $j \ge j_1$. In particular, in this case too we obtain the bound (A.22), namely,

$$\sup_{(u^{(k)},v^{(k)})\in\mathcal{V}_k} f_k(u^{(k)},v^{(k)},k\tau) \ge \frac{1}{2} - o(1) \qquad \text{(A.25)}$$

for a function $o(1)$ that goes to zero as $k\to\infty$. Thus, we always have that

$$\delta_{P_{Y|X}^{\circ k}}(k\tau) = \sup_{(u^{(k)},v^{(k)})\in\mathcal{V}_k} f_k(u^{(k)},v^{(k)},k\tau) \ge \frac{1}{2} - o(1) > \delta \qquad \text{(A.26)}$$

as $k\to\infty$. Inverting this inequality, we obtain that

$$k\tau \le \varepsilon_{P_{Y|X}^{\circ k}}(\delta) \qquad \text{(A.27)}$$

for all large k. In particular,

$$\tau \le \liminf_{k\to\infty} \frac{\varepsilon_{P_{Y|X}^{\circ k}}(\delta)}{k} \qquad \text{(A.28)}$$

for every $\tau \in (0, \mathrm{KL}_{\max})$. Taking $\tau \nearrow \mathrm{KL}_{\max}$, we conclude that

$$\mathrm{KL}_{\max} \le \liminf_{k\to\infty} \frac{\varepsilon_{P_{Y|X}^{\circ k}}(\delta)}{k}. \qquad \text{(A.29)}$$

Combining (A.17) and (A.29), we infer the inequalities (2.35) in the theorem statement, i.e.,

$$k\cdot(\mathrm{KL}_{\max} - o(1)) \le \varepsilon_{P_{Y|X}^{\circ k}}(\delta) \le k\cdot\mathrm{KL}_{\max} + \left(-\Phi^{-1}(\delta) + o(1)\right)\sqrt{k\cdot\mathrm{V}_{\max}}, \qquad \text{(A.30)}$$

Next, we derive the expansion (2.36) via deriving a refined lower bound on $\varepsilon_{P_{Y|X}^{\circ k}}(\delta)$ under the assumption that $\underline{\mathrm{V}} > 0$.

- *Step 4: a lower bound on ε when $\underline{\mathrm{V}} > 0$.*

Recall that $\underline{\mathrm{V}}$ is defined by (2.37) in the theorem statement as the minimal value $\underline{\mathrm{V}}$ that $\mathrm{V}_{x,x'}$ can take if $\mathrm{KL}_{x,x'}$ is arbitrarily close to maximal:

$$\underline{\mathrm{V}} := \inf\left\{\liminf_{\ell\to\infty}\ \mathrm{V}_{x_\ell,x'_\ell}\ :\ x_\ell, x'_\ell \in \mathbb{R}^m,\ \sup_{\ell\in\mathbb{N}}\|x_\ell - x'_\ell\| \leq s,\ \lim_{\ell\to\infty}\ \mathrm{KL}_{x_\ell,x'_\ell} = \mathrm{KL}_{\max}\right\}. \tag{A.31}$$

From Assumption 2.3, it can be inferred that $\underline{\mathrm{V}} > 0$. Indeed, if it were the case that $\underline{\mathrm{V}} = 0$, then we may take far enough elements of a sequence of sequences of pairs $\{(x_{m,\ell}, x'_{m,\ell})\}_{m,\ell}$ (so that $\lim_{m\to\infty}\ \liminf_{\ell\to\infty}\ \mathrm{V}_{x_{m,\ell},x'_{m,\ell}} = 0$, and $\lim_{\ell\to\infty}\ \mathrm{KL}_{x_{m,\ell},x'_{m,\ell}} = \mathrm{KL}_{\max}$ for each m) to produce a new sequence $\{(\xi_m, \xi'_m) := (x_{m,\ell_m}, x'_{m,\ell_m})\}_m$ (for large enough ℓ_m) for which $\mathrm{KL}_{\xi_m,\xi'_m} \to \mathrm{KL}_{\max}$ but $\mathrm{V}_{\xi_m,\xi'_m} \to 0$ (possibly after passing to a subsequence), thereby violating Assumption 2.3. Consider also the constant

$$\alpha := \inf_{\|x-x'\|\leq s}\ \mathrm{KL}_{\max} - \mathrm{KL}_{x,x'} + \mathrm{V}_{x,x'}. \tag{A.32}$$

As $\underline{\mathrm{V}} > 0$, we have that $\alpha > 0$.

Denote the constant $\gamma := \frac{1}{\alpha}\sqrt{\frac{\mathrm{V}_{\max}}{2\pi}}$, and fix an arbitrary $\eta \in (0, 1/2)$. For all large k, define the two functions

$$\underline{\varepsilon}_k(\delta) := k\cdot\mathrm{KL}_{\max} - \Phi^{-1}\left(\delta + \eta + r(k) + \frac{\gamma}{\sqrt{k}}\right)\cdot\sqrt{k\cdot\underline{\mathrm{V}}}. \tag{A.33}$$

We assume here, and for the remainder of the proof, that k is large enough that the argument of Φ^{-1} above falls inside the interval $(0,1)$. We will show the bound $\delta < \delta_{P^{\circ k}_{Y|X}}(\underline{\varepsilon}_k(\delta))$. This bound may be inverted to obtain $\underline{\varepsilon}_k(\delta) \leq \varepsilon_{P^{\circ k}_{Y|X}}(\delta)$, from which the desired asymptotic result follows readily.

From (A.12), we have that, for any $(u^{(k)}, v^{(k)}) \in \mathcal{V}_k$,

$$f_k(u^{(k)}, v^{(k)}, \underline{\varepsilon}_k(\delta)) \geq g_k(u^{(k)}, v^{(k)}) - h_k(u^{(k)}, v^{(k)}) + r(k), \tag{A.34}$$

where we define the functions

$$g_k(u^{(k)}, v^{(k)}) := \Phi\left(\frac{\mathrm{KL}_{u^{(k)},v^{(k)}} - \underline{\varepsilon}_k(\delta)}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}\right), \tag{A.35}$$

$$h_k(u^{(k)}, v^{(k)}) := e^{\underline{\varepsilon}_k(\delta) - \mathrm{KL}_{u^{(k)},v^{(k)}} + \mathrm{V}_{u^{(k)},v^{(k)}}/2}\ \Phi\left(\frac{\mathrm{KL}_{u^{(k)},v^{(k)}} - \mathrm{V}_{u^{(k)},v^{(k)}} - \underline{\varepsilon}_k(\delta)}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}\right). \tag{A.36}$$

We upper bound h_k. Note that

$$\Phi(-z) < \frac{1}{z\sqrt{2\pi}}e^{-z^2/2} \tag{A.37}$$

for $z > 0$. Therefore,

$$h_k(u^{(k)}, v^{(k)}) < \frac{1/\sqrt{2\pi}}{\exp\left(w_{k,u^{(k)},v^{(k)}}^2/2\right) \cdot \left(w_{k,u^{(k)},v^{(k)}} + \sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}\right)} < \frac{1/\sqrt{2\pi}}{w_{k,u^{(k)},v^{(k)}} + \sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}}. \tag{A.38}$$

where $w_{k,u^{(k)},v^{(k)}} := (\underline{\varepsilon}_k(\delta) - \mathrm{KL}_{u^{(k)},v^{(k)}})/\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}$. Therefore, we have that

$$\begin{aligned} w_{k,u^{(k)},v^{(k)}} + \sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}} &= \frac{1}{\sqrt{\mathrm{V}_{u^{(k)},v^{(k)}}}} \cdot \left(\underline{\varepsilon}_k(\delta) - \mathrm{KL}_{u^{(k)},v^{(k)}} + \mathrm{V}_{u^{(k)},v^{(k)}}\right) && \text{(A.39)} \\ &\geq \frac{1}{\sqrt{k \cdot \mathrm{V}_{\max}}} \cdot \left(k \cdot \mathrm{KL}_{\max} - \mathrm{KL}_{u^{(k)},v^{(k)}} + \mathrm{V}_{u^{(k)},v^{(k)}}\right) && \text{(A.40)} \\ &\geq \sqrt{k} \cdot \frac{\alpha}{\sqrt{\mathrm{V}_{\max}}}, && \text{(A.41)} \end{aligned}$$

where the last line follows by definition of α (see (A.32)). Therefore, we have that

$$\sup_{(u^{(k)},v^{(k)}) \in \mathcal{V}_k} h_k(u^{(k)}, v^{(k)}) \leq \frac{\sqrt{\mathrm{V}_{\max}}}{\alpha\sqrt{2\pi k}}. \tag{A.42}$$

Next, we lower bound the supremum of g_k. Let $\{(x_\ell, x'_\ell)\}_{\ell \in \mathbb{N}}$ be a sequence with $\|x_\ell - x'_\ell\| \leq s$ such that $\mathrm{KL}_{x_\ell, x'_\ell} \to \mathrm{KL}_{\max}$ and $\mathrm{V}_{x_\ell, x'_\ell} \to \underline{\mathrm{V}}$. Consider the length-$k$ sequences of repeated vectors $u_\ell^{(k)} = (x_\ell, \cdots, x_\ell)$ and $v_\ell^{(k)} = (x'_\ell, \cdots, x'_\ell)$, and note that we have $(u_\ell^{(k)}, v_\ell^{(k)}) \in \mathcal{V}_k$ for all large ℓ. We have the limit

$$\lim_{\ell \to \infty} \frac{\mathrm{KL}_{u_\ell^{(k)}, v_\ell^{(k)}} - \underline{\varepsilon}_k(\delta)}{\sqrt{\mathrm{V}_{u_\ell^{(k)}, v_\ell^{(k)}}}} = \Phi^{-1}\left(\delta + \eta + r(k) + \frac{\gamma}{\sqrt{k}}\right). \tag{A.43}$$

Therefore, we have the lower bound

$$\sup_{(u^{(k)},v^{(k)}) \in \mathcal{V}_k} g_k(u^{(k)}, v^{(k)}) \geq \delta + \eta + r(k) + \frac{\gamma}{\sqrt{k}}. \tag{A.44}$$

Putting (A.42) and (A.44) together, we conclude the lower bound

$$\delta_{P_{Y|X}^{\circ k}}(\underline{\varepsilon}_k(\delta)) \geq \delta + \eta + r(k) > \delta \tag{A.45}$$

for all large k. From (A.17) and (A.45), and by definition of $\varepsilon_{P_{Y|X}^{\circ k}}$ as an inverse of $\delta_{P_{Y|X}^{\circ k}}$, we arrive at the bounds

$$\underline{\varepsilon}_k(\delta) \leq \varepsilon_{P_{Y|X}^{\circ k}}(\delta) \leq \overline{\varepsilon}_k(\delta). \tag{A.46}$$

Plugging in the definitions of $\underline{\varepsilon}_k(\delta)$ and $\overline{\varepsilon}_k(\delta)$, then rearranging and taking $k \to \infty$ then $\eta \to 0^+$, we obtain the bounds

$$0 < \sqrt{\underline{\mathrm{V}}} \leq \liminf_{k\to\infty} \frac{\varepsilon_{P_{Y|X}^{\circ k}}(\delta) - k\cdot \mathrm{KL}_{\max}}{-\Phi^{-1}(\delta)\cdot\sqrt{k}} \leq \limsup_{k\to\infty} \frac{\varepsilon_{P_{Y|X}^{\circ k}}(\delta) - k\cdot \mathrm{KL}_{\max}}{-\Phi^{-1}(\delta)\cdot\sqrt{k}} \leq \sqrt{\mathrm{V}_{\max}} < \infty. \tag{A.47}$$

This completes the proof of the theorem.

A.3 Proof of Theorem 2.4: Additive Mechanisms are Optimal

Let $F : \mathscr{R} \to [0,\infty]$ denote the objective function in (2.8), i.e.,

$$F(P_{Y|X}) := \sup_{\|u-v\|\leq s} D(P_{Y|X=u} \parallel P_{Y|X=v}). \tag{A.48}$$

Thus,

$$\mathrm{KL}^{\star} = \inf_{P_{Y|X}\in\mathscr{P}} F(P_{Y|X}). \tag{A.49}$$

Fix a sequence of conditional distributions

$$\left\{P_{Y|X}^{(k)}\right\}_{k\in\mathbb{N}} \subset \mathscr{P} \tag{A.50}$$

satisfying

$$\mathrm{KL}^{\star} = \lim_{k\to\infty} F\left(P_{Y|X}^{(k)}\right). \tag{A.51}$$

Recall that by assumption, the version of each conditional distribution $P_{Y|X}^{(k)}$ we choose is regular, i.e., $x \mapsto P_{Y|X=x}^{(k)}(B)$ is a Borel function for each Borel set $B \subset \mathbb{R}^m$. Note that $\mathrm{KL}^{\star} < \infty$. Throwing away the first few elements in the sequence, we assume that $F\left(P_{Y|X}^{(k)}\right) < \infty$ for each $k \in \mathbb{N}$. Let λ denote the Lebesgue measure on $\mathbb{R}^m$, and $B_r(x) \subset \mathbb{R}^m$ the open ball around x of radius r.

We break the proof down into several steps:

1. Introduce Markov kernels $\overline{P}_{Y|X}^{(k)}$ as "continuous" convex combinations of the $P_{Y|X}^{(k)}$.

2. The $\overline{P}_{Y|X}^{(k)}$ also satisfy the cost constraint.

3. The $\overline{P}_{Y|X}^{(k)}$ asymptotically achieve $\mathrm{KL}^{\star}$.

4. The $\overline{P}_{Y|X=x}^{(k)}$ are asymptotically shifted versions $T_x P^{\star}$ of a fixed $P^{\star} \in \mathscr{B}$.

5. $P^{\star}$ achieves $\mathrm{KL}^{\star}$.

- *Step 1*: Averaging the $P_{Y|X}^{(k)}$.

For $k \in \mathbb{N}$, we will define the Markov kernel $\overline{P}^{(k)}_{Y|X} \in \mathscr{R}$ by

$$\overline{P}^{(k)}_{Y|X=x}(A) := \frac{1}{\lambda(B_k(0))} \int_{B_k(0)} P^{(k)}_{Y|X=x+z}(A+z)\, dz. \tag{A.52}$$

Of course, we need to check that (A.52) indeed yields a Markov kernel $\overline{P}^{(k)}_{Y|X}$. In view of Fubini's theorem, it suffices to check that the map $(x,z) \mapsto P^{(k)}_{Y|X=x+z}(A+z)$ is jointly Borel (for every fixed Borel set $A \subset \mathbb{R}^m$). This joint measurability is not self-evident, so we check next that it indeed holds.

Let the transition probability kernel $L^{(k)} : \mathbb{R}^{2m} \times \mathcal{B}(\mathbb{R}^m) \to [0,1]$ be defined by

$$L^{(k)}((x,z),A) := P^{(k)}_{Y|X=x+z}(A). \tag{A.53}$$

Let $N^{(k)} : \mathbb{R}^{2m} \times \mathcal{B}(\mathbb{R}^m) \to [0,1]$ denote the map

$$N^{(k)}((x,z),A) := P^{(k)}_{Y|X=x+z}(A+z). \tag{A.54}$$

For each $(x,z) \in \mathbb{R}^{2m}$ and Borel set $A \subset \mathbb{R}^m$, we may write $N^{(k)}((x,z),A)$ as the integral of a nonnegative Borel function against $L((x,z),dy)$, namely,

$$N^{(k)}((x,z),A) = \int_{\mathbb{R}^m} 1_A(y-z)\, L((x,z),dy), \tag{A.55}$$

where $((x,z),y) \mapsto 1_A(y-z)$ is Borel. Hence (see, e.g., [Ç11, Chapter 1, Proposition 6.9]) $(x,z) \mapsto N^{(k)}((x,z),A)$ is a Borel function. Hence, $\overline{P}^{(k)}_{Y|X}$ as given by (A.52) is indeed a well-defined Markov kernel on $\mathbb{R}^m$.

For the next steps, we will use the following notation

$$R^{(k,x)}_{Y|X=z}(A) := P^{(k)}_{Y|X=x+z}(A+z), \tag{A.56}$$

$$P^{(k,x)}(A) := \overline{P}^{(k)}_{Y|X=x}(A), \tag{A.57}$$

$$U^{(k)}(A) := \frac{\lambda(A \cap B_k(0))}{\lambda(B_k(0))}. \tag{A.58}$$

Note that $R^{(k,x)}_{Y|X} \in \mathscr{R}$ and $P^{(k,x)} \in \mathscr{B}$ for each fixed $(k,x) \in \mathbb{N} \times \mathbb{R}^m$, and (A.52) may be rewritten as

$$P^{(k,x)} = R^{(k,x)}_{Y|X} \circ U^{(k)}. \tag{A.59}$$

- *Step 2*: The $\overline{P}^{(k)}_{Y|X}$ satisfy the cost constraint.

Fix $k \in \mathbb{N}$, and we will show next that $\overline{P}^{(k)}_{Y|X} \in \mathscr{P}$, i.e., that $\overline{P}^{(k)}_{Y|X}$ satisfies the cost constraint.

Recall that a Markov kernel $P_{Y|X} \in \mathscr{R}$ belongs to $\mathscr{P}$ if and only if it satisfies

$$\sup_{x \in \mathbb{R}^m} \mathbb{E}_{P_{Y|X=x}} [T_x c] \leq C. \tag{A.60}$$

By the assumption that $P^{(k)}_{Y|X} \in \mathscr{P}$, we have that

$$\mathbb{E}_{P^{(k)}_{Y|X=x}} [T_x c] \leq C \tag{A.61}$$

for every $x \in \mathbb{R}^m$. Shifting the variable of integration in (A.61) by a fixed constant $-z$, we obtain that

$$\mathbb{E}_{T_{-z} P^{(k)}_{Y|X=x}} [T_{x-z} c] \leq C \tag{A.62}$$

for every $(x, z) \in \mathbb{R}^{2m}$. Replacing x by $x + z$ in (A.62), we conclude that (see (A.56))

$$\mathbb{E}_{R^{(k,x)}_{Y|X=z}} [T_x c] \leq C \tag{A.63}$$

for every $(x, z) \in \mathbb{R}^{2m}$. We proceed via the following standard approximation by simple functions argument.

Fix $x \in \mathbb{R}^m$, and let $\sum_j a_j 1_{A_j}(y)$ be a nonnegative simple function upper bounded by $(T_x c)(y)$. Integrating against $R^{(k,x)}_{Y|X=z}(dy)$ we deduce from (A.63) that

$$\sum_j a_j R^{(k,x)}_{Y|X=z}(A_j) \leq C \tag{A.64}$$

for every $z \in \mathbb{R}^m$. Integrating (A.64) against $U^{(k)}(dz)$, and noting that $P^{(k,x)} = R^{(k,x)}_{Y|X} \circ U^{(k)}$ (see (A.59)), we deduce that

$$\sum_j a_j P^{(k,x)}(A_j) \leq C. \tag{A.65}$$

Now, as (A.65) holds for all nonnegative simple functions below $T_x c$, taking an increasing sequence of nonnegative simple function converging pointwise to $T_x c$ we conclude that

$$\mathbb{E}_{P^{(k,x)}} [T_x c] \leq C. \tag{A.66}$$

In other words (see (A.57)),

$$\mathbb{E}_{\overline{P}^{(k)}_{Y|X=x}} [T_x c] \leq C. \tag{A.67}$$

As (A.67) holds for all $x \in \mathbb{R}^m$, we have shown that $\overline{P}^{(k)}_{Y|X} \in \mathscr{P}$.

- *Step 3*: The $\overline{P}^{(k)}_{Y|X}$ are asymptotically optimal.

Next, we use monotonicity of the KL-divergence under conditioning (see Lemma A.1) to show the limit

$$\mathrm{KL}^{\star} = \lim_{k\to\infty} F\left(\overline{P}^{(k)}_{Y|X}\right). \tag{A.68}$$

Shift-invariance of the KL-divergence implies that, for each $x, x', z \in \mathbb{R}^m$,

$$D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) = D\left(P^{(k)}_{Y|X=x+z} \| P^{(k)}_{Y|X=x'+z}\right). \tag{A.69}$$

Thus, as $(x+z)-(x'+z) = x-x'$, we conclude that

$$\sup_{\substack{\|x-x'\|\le s \\ z\in\mathbb{R}^m}} D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) = F\left(P^{(k)}_{Y|X}\right) \tag{A.70}$$

By assumption of optimality of the $P^{(k)}_{Y|X}$ (see (A.51)), for each $\delta > 0$, there exists a k_0 such that for all $k \ge k_0$,

$$\sup_{\substack{\|x-x'\|\le s \\ z\in\mathbb{R}^m}} D\left(R^{(k,x)}_{Y|X=z} \| R^{(k,x')}_{Y|X=z}\right) \le \mathrm{KL}^{\star} + \delta. \tag{A.71}$$

By definition of KL-divergence, we infer $R^{(k,x)}_{Y|X=z} \ll R^{(k,x')}_{Y|X=z}$ for all $z \in \mathbb{R}^m$ and $\|x - x'\| \le s$. Also, (A.71) shows in particular that

$$\sup_{\|x-x'\|\le s} \mathbb{E}_{\xi\sim U^{(k)}}\left[D\left(R^{(k,x)}_{Y|X=\xi} \| R^{(k,x')}_{Y|X=\xi}\right)\right] \le \mathrm{KL}^{\star} + \delta. \tag{A.72}$$

Using (A.59), Lemma A.1 yields that

$$\sup_{\|x-x'\|\le s} D\left(P^{(k,x)} \| P^{(k,x')}\right) \le \mathrm{KL}^{\star} + \delta. \tag{A.73}$$

Taking $\delta \to 0^+$, we see that (A.68) holds.

- *Step 4*: $P^{(k,x)}$ is asymptotically $T_x P^{\star}$ for a fixed $P^{\star}$.

Next, we show that there is a measure $P^{\star} \in \mathscr{B}$ such that, for every $x \in \mathbb{R}^m$, we have the weak convergence

$$P^{(k,x)} \to T_x P^{\star} \tag{A.74}$$

as $k \to \infty$.

First, for each fixed $x \in \mathbb{R}^m$, we establish the total-variation distance convergence

$$\lim_{k\to\infty} \left\| P^{(k,x)} - T_x P^{(k,0)} \right\|_{\mathrm{TV}} = 0. \tag{A.75}$$

We may write

$$\left(T_x P^{(k,0)}\right)(A) = \frac{1}{\lambda(B_k(0))} \int_{B_k(-x)} R^{(k,x)}_{Y|X=z}(A)\, dz. \tag{A.76}$$

Therefore, for any Borel set $A \subset \mathbb{R}^m$ we have that

$$P^{(k,x)}(A) - T_x P^{(k,0)}(A) \leq \frac{1}{\lambda(B_k(0))} \int_{B_k(0)\setminus B_k(-x)} R^{(k,x)}_{Y|X=z}(A)\, dz \leq \frac{\lambda(B_k(0) \setminus B_k(-x))}{\lambda(B_k(0))}. \tag{A.77}$$

Now, applying a rotation, we note that

$$\lambda(B_k(0) \setminus B_k(-x)) = \lambda(B_k(0) \setminus B_k(\|x\| e_1)) \leq \lambda(B_k(0) \setminus B_k(se_1)) \tag{A.78}$$

where $e_1 = (1, 0 \cdots, 0) \in \mathbb{R}^m$. Furthermore, the triangle inequality yields that $B_{k-s/2}((s/2)e_1) \subset B_k(se_1) \cap B_k(0)$; indeed, if $\|z - (s/2)e_1\| < k - s/2$ then $\|z\|, \|z - se_1\| \leq \|z - (s/2)e_1\| + \|(s/2)e_1\| < k$. Therefore, we have that

$$B_k(0) \setminus B_k(se_1) \subset B_k(0) \setminus B_{k-s/2}((s/2)e_1), \tag{A.79}$$

and

$$\lambda\left(B_k(0) \setminus B_{k-s/2}((s/2)e_1)\right) = \lambda\left(B_k(0)\right) - \lambda\left(B_{k-s/2}((s/2)e_1)\right). \tag{A.80}$$

Therefore, combining (A.78)–(A.80), we obtain

$$\frac{\lambda(B_k(0) \setminus B_k(-x))}{\lambda(B_k(0))} \leq \frac{k^m - (k - s/2)^m}{k^m}. \tag{A.81}$$

Further, by Bernoulli's inequality, for every $k > s/2$,

$$\frac{k^m - (k - s/2)^m}{k^m} = 1 - \left(1 - \frac{s}{2k}\right)^m \leq 1 - \left(1 - \frac{ms}{2k}\right) = \frac{ms}{2k}. \tag{A.82}$$

From (A.77), (A.81), and (A.82), we conclude that

$$\left\| P^{(k,x)} - T_x P^{(k,0)} \right\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R}^m)} P^{(k,x)}(A) - T_x P^{(k,0)}(A) \leq \frac{ms}{2k}. \tag{A.83}$$

Hence, the total-variation distance convergence in (A.75) follows.

The next ingredient we need is that the set $\{P^{(k,0)}\}_{k \in \mathbb{N}} \subset \mathscr{B}$ is tight, i.e., that

$$\lim_{n \to \infty} \sup_{k \in \mathbb{N}} P^{(k,0)}(\mathbb{R}^m \setminus B_n(0)) = 0. \tag{A.84}$$

By Step 2 of this proof, we have that $P^{(k,0)}$ satisfies the cost constraint, i.e., $\mathbb{E}_{P^{(k,0)}}[c] \leq C$. By the assumption of isotropy of c, there is a function $\widetilde{c}$ such that $c(y) = \widetilde{c}(\|y\|)$. Then, by the assumption of monotonicity of $\widetilde{c}$,

$$P^{(k,0)}(\mathbb{R}^m \setminus B_n(0)) \cdot \widetilde{c}(n) \leq \int_{\mathbb{R}^m \setminus B_n(0)} \widetilde{c}(\|y\|)\, dP^{(k,0)}(y) \leq C. \tag{A.85}$$

Since $\widetilde{c}(n) \to \infty$ as $n \to \infty$ by assumption, we conclude that

$$\limsup_{n\to\infty} \sup_{k\in\mathbb{N}} P^{(k,0)}(\mathbb{R}^m \setminus B_n(0)) \leq \limsup_{n\to\infty} \frac{C}{\widetilde{c}(n)} = 0. \tag{A.86}$$

Hence (A.84) follows, i.e., $\{P^{(k,0)}\}_{k\in\mathbb{N}}$ is tight.

By tightness of $\{P^{(k,0)}\}_{k\in\mathbb{N}}$, we conclude via Prokhorov's theorem [Ç11, Chapter 3, Theorem 5.13] after passing to a subsequence that there is a $P^\star \in \mathscr{B}$ such that $P^{(k,0)} \to P^\star$ weakly as $k \to \infty$, i.e., for every continuous and bounded function $f : \mathbb{R}^m \to \mathbb{R}$ we have

$$\lim_{k\to\infty} \mathbb{E}_{P^{(k,0)}}[f] = \mathbb{E}_{P^\star}[f]. \tag{A.87}$$

This immediately implies that, for each $x \in \mathbb{R}^m$, we also have

$$T_x P^{(k,0)} \to T_x P^\star \tag{A.88}$$

weakly as $k \to \infty$. As convergence in total variation is stronger than weak convergence, we conclude from (A.75) and (A.88) that for every $x \in \mathbb{R}^m$

$$P^{(k,x)} \to T_x P^\star \tag{A.89}$$

weakly as $k \to \infty$.

- *Step 5*: The additive mechanism $P^\star$ is optimal.

The final step is showing that $P^\star$ attains $\mathrm{KL}^\star$ and satisfies the cost constraint. By joint lower-semicontinuity of the KL-divergence [Pos75, Theorem 1], we deduce from (A.89) that for each $x \in \mathbb{R}^m$

$$D(P^\star \| T_x P^\star) \leq \liminf_{k\to\infty} D\left(P^{(k,0)} \| P^{(k,x)}\right). \tag{A.90}$$

But we also have

$$\sup_{\|x\|\leq s} D\left(P^{(k,0)} \| P^{(k,x)}\right) \leq F\left(\overline{P}^{(k)}_{Y|X}\right). \tag{A.91}$$

Therefore, taking the supremum over $\|x\| \leq s$ in (A.90), we infer from (A.68) that

$$\sup_{\|x\|\leq s} D(P^\star \| T_x P^\star) \leq \mathrm{KL}^\star. \tag{A.92}$$

Hence, it only remains to check that $P^\star \in \mathscr{P}_{\mathrm{add}}$ for us to conclude that equality holds in (A.92).

For every $r > 0$ and $x \in \mathbb{R}^m$, the function $1_{B_r(0)} \cdot T_x c$ is lower semicontinuous and bounded. Hence, the weak convergence $P^{(k,x)} \to T_x P^\star$ yields

$$\mathbb{E}_{T_x P^\star}\left[1_{B_r(0)} \cdot T_x c\right] \leq \liminf_{k\to\infty} \mathbb{E}_{P^{(k,x)}}\left[1_{B_r(0)} \cdot T_x c\right]. \tag{A.93}$$

As $\overline{P}^{(k)}_{Y|X} \in \mathscr{P}$, nonnegativity of c implies in view of (A.93) that

$$\mathbb{E}_{T_x P^\star}\left[1_{B_r(0)} \cdot T_x c\right] \leq C. \tag{A.94}$$

By the monotone convergence theorem, taking $r \to \infty$ yields

$$\mathbb{E}_{T_x P^\star}\left[T_x c\right] \leq C, \tag{A.95}$$

In other words, $P^\star \in \mathscr{P}_{\mathrm{add}}$. Therefore, we must have

$$\mathrm{KL}^\star \leq \mathrm{KL}^\star_{\mathrm{add}} \leq \sup_{\|x\|\leq s} D(P^\star \| T_x P^\star). \tag{A.96}$$

Combining this inequality with (A.92), we conclude that

$$\mathrm{KL}^\star = \mathrm{KL}^\star_{\mathrm{add}} = \sup_{\|x\|\leq s} D(P^\star \| T_x P^\star). \tag{A.97}$$

This completes the proof of the theorem.

Remark A.1. The lemma stated below, showing that conditioning increases divergence, is a well-known fact. It is shown in the literature under various assumptions on the underlying distributions (see, e.g., [Pol19, Theorem 2.2 and Section 2.6]). We use it in the proof of Theorem 2.3 in the specific situation where one of the conditional distributions is absolutely continuous with respect to the other for each individual input. As in [Pol19, Remark 2.4], Doob's version of the Radon-Nikodym theorem can be used to derive that conditioning increases divergence in our case. For completeness, we add a proof of this lemma here.

Lemma A.1 (Conditioning increases divergence)**.** *Let $P_{Y|X}, P'_{Y|X}$ be Markov kernels on $\mathbb{R}^m$ such that $P_{Y|X=x} \ll P'_{Y|X=x}$ for every $x \in \mathbb{R}^m$. Fix a Borel probability measure P_X on $\mathbb{R}^m$. Denote the marginalizations* *of $P_{X,Y} := P_{Y|X} \otimes P_X, P'_{X,Y} := P'_{Y|X} \otimes P_X$ in the second coordinate by P_Y, P'_Y. Then, we have the inequality*

$$D\left(P_Y \parallel P'_Y\right) \leq \mathbb{E}_{\xi \sim P_X}\left[D\left(P_{Y|X=\xi} \parallel P'_{Y|X=\xi}\right)\right]. \tag{A.98}$$

Proof. Since by assumption $P_{Y|X=x} \ll P'_{Y|X=x}$ for every $x \in \mathbb{R}^m$, a generalization of the Radon-Nikodym theorem by Doob (see [Ç11, Chapter 5, Theorem 4.44]) yields the existence of a version of the Radon-Nikodym derivatives $dP_{Y|X=x}/dP'_{Y|X=x}$ such that the function

$$(x,y) \mapsto \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y) \tag{A.99}$$

is jointly measurable. We show that this function is a version of $dP_{X,Y}/dP'_{X,Y}$. First, note that $P_{X,Y} \ll P'_{X,Y}$. Indeed, for any Borel set $E \subset \mathbb{R}^m$, denoting the sections by $E_x := \{y \in \mathbb{R}^m \,;\, (x,y) \in E\}$, we have that $P_{X,Y}(E) = 0$ if and only if $P_{Y|X=x}(E_x) = 0$ for P_X-a.e. x, and a similar statement holds for $P'_{X,Y}$. By assumption, $P_{Y|X=x} \ll P'_{Y|X=x}$ for each x, so we obtain $P_{X,Y} \ll P'_{X,Y}$. By joint measurability and nonnegativity, using the disintegration theorem (see, e.g., [Ç11, Chapter 1, Theorem 6.11]) we obtain that for any Borel $E \subset \mathbb{R}^{2m}$

$$\int_E \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\, dP'_{X,Y}(x,y) = \int_{\mathbb{R}^m}\int_{E_x} \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\, dP'_{Y|X=x}(y)\, dP_X(x) \tag{A.100}$$

$$= \int_{\mathbb{R}^m}\int_{E_x} dP_{Y|X=x}(y)\, dP_X(x) \tag{A.101}$$

$$= P_{X,Y}(E). \tag{A.102}$$

Thus, we have the equality

$$\frac{dP_{X,Y}}{dP'_{X,Y}}(x,y) = \frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y) \tag{A.103}$$

for $P'_{X,Y}$-a.e. (x,y).

Define $f : [0,\infty) \to [-1/e,\infty)$ by $f(0) = 0$ and $f(t) = t\log t$ for $t > 0$. By the disintegration theorem and (A.103), we have the equality

$$D\left(P_{X,Y} \parallel P'_{X,Y}\right) = \int_{\mathbb{R}^{2m}} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}\right) dP'_{X,Y} \tag{A.104}$$

$$= \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} f\left(\frac{dP_{Y|X=x}}{dP'_{Y|X=x}}(y)\right) dP'_{Y|X=x}\, dP_X(x) \tag{A.105}$$

$$= \mathbb{E}_{\xi \sim P_X}\left[D\left(P_{Y|X=\xi} \parallel P'_{Y|X=\xi}\right)\right]. \tag{A.106}$$

On the other hand, disintegration with respect to Y yields the following bound. Denote by

$P_{X|Y}, P'_{X|Y}$ the disintegrations of $P_{X,Y}, P'_{X,Y}$ with respect to P_Y, P'_Y. In particular, $P_{X|Y}$ and $P'_{X|Y}$ are Markov kernels on $\mathbb{R}^m$. By the disintegration theorem and Jensen's inequality,

$$D\left(P_{X,Y} \parallel P'_{X,Y}\right) = \int_{\mathbb{R}^{2m}} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}\right) dP'_{X,Y} \tag{A.107}$$

$$= \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} f\left(\frac{dP_{X,Y}}{dP'_{X,Y}}(x,y)\right) dP'_{X|Y=x}(x)\, dP'_Y(y) \tag{A.108}$$

$$\geq \int_{\mathbb{R}^m} f\left(g(y)\right) dP'_Y(y) \tag{A.109}$$

where

$$g(y) := \int_{\mathbb{R}^m} \frac{dP_{X,Y}}{dP'_{X,Y}}(x,y)\, dP'_{X|Y=x}(x). \tag{A.110}$$

For this application of Jensen's inequality, we use the fact, shown next, that g is finite P'_Y-a.e. In fact, we show that g is a version of dP_Y/dP'_Y. Note that $P_{X,Y} \ll P'_{X,Y}$ implies that $P_Y \ll P'_Y$. Now, for any Borel $B \subset \mathbb{R}^m$, the disintegration theorem yields that

$$\int_B g\, dP'_Y = \int_B \int_{\mathbb{R}^m} \frac{dP_{X,Y}}{dP'_{X,Y}}(x,y)\, dP'_{X|Y=x}(x)\, dP'_Y(y) \tag{A.111}$$

$$= \int_{\mathbb{R}^m \times B} \frac{dP_{X,Y}}{dP'_{X,Y}}\, dP'_{X,Y} \tag{A.112}$$

$$= P_{X,Y}(\mathbb{R}^m \times B) = P_Y(B). \tag{A.113}$$

Thus, we have that

$$g(y) = \frac{dP_Y}{dP'_Y}(y). \tag{A.114}$$

for P'_Y-a.e. y. Hence, we obtain from inequality (A.109) that

$$D\left(P_{X,Y} \parallel P'_{X,Y}\right) \geq D\left(P_Y \parallel P'_Y\right). \tag{A.115}$$

Combining inequality (A.115) and equation (A.106) we obtain the desired inequality (A.98). □

A.4 Proof of Theorem 2.7: Finite-Dimensionality

Note that the vector $\boldsymbol{p}$ only includes p_i for $0 \leq i \leq N$. We will simplify our analysis by defining p_i for all integers i. Specifically, for $i \in \mathbb{Z} \setminus \{0, \cdots, N\}$, we denote

$$p_i := \begin{cases} p_{|i|}, & \text{if } -N \leq i \leq -1, \\ p_N r^{|i|-N}, & \text{if } |i| > N. \end{cases} \tag{A.116}$$

Thus we may rewrite the formula for $f_{n,r,\boldsymbol{p}}$ in (2.65) as

$$f_{n,r,\boldsymbol{p}}(x) = np_i \quad \text{if } x \in \mathcal{J}_{n,i}. \tag{A.117}$$

We show first that

$$\sup_{a\in\mathbb{R}:|a|\le 1} D(P_{n,r,\boldsymbol{p}} \| T_a P_{n,r,\boldsymbol{p}}) = \max_{k\in\mathbb{Z}:|k|\le n} \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}}, \tag{A.118}$$

then we show that this formula is equal to the objective function in (2.82). For convenience, we drop the subscripts on $f_{n,r,\boldsymbol{p}}$ and $P_{n,r,\boldsymbol{p}}$ throughout this proof. We may assume $\boldsymbol{p} > \boldsymbol{0}$, since any vector $\boldsymbol{p}$ with some zero coordinate will be infeasible in both optimization problems (2.80) and (2.82).

Fix $a \in [-1,1]$. For each $i \in \mathbb{Z}$, let $\mathcal{J}^{\circ}_{n,i} = \left(\frac{i-1/2}{n}, \frac{i+1/2}{n}\right)$ denote the interior of $\mathcal{J}_{n,i}$. We start by showing that the function

$$F_a := f \log \frac{f}{T_{-a} f} \tag{A.119}$$

is integrable, which would allow us to use countable additivity of the Lebesgue integral to split $D(P \| T_{-a} P)$ into a sum of integrals over the $\mathcal{J}^{\circ}_{n,i}$. Let $k \in \mathbb{Z}$ be the unique integer such that $a + \frac{1}{2n} \in \mathcal{J}_{n,k}$, and denote $\Delta := k - an$. From

$$\frac{k-1/2}{n} \le a + \frac{1}{2n} \le \frac{k+1/2}{n}, \tag{A.120}$$

we conclude that $0 \le \Delta \le 1$. Consider an integer i and a real $x \in \mathcal{J}^{\circ}_{n,i}$. If $x < (i-1/2+\Delta)/n$, then

$$x + a = x + \frac{k-\Delta}{n} < \frac{i+k-1/2}{n} = \frac{(i+k-1)+1/2}{n} \tag{A.121}$$

and, since $\Delta \le 1$,

$$x + a = x + \frac{k-\Delta}{n} > \frac{i-1/2}{n} + \frac{k-1}{n} = \frac{(i+k-1)-1/2}{n}. \tag{A.122}$$

Inequalities (A.121) and (A.122) together imply that $x + a \in \mathcal{J}^{\circ}_{n,i+k-1}$. Similarly, if $x > (i - 1/2 + \Delta)/n$ then $x + a \in \mathcal{J}^{\circ}_{n,i+k}$. We may ignore the countably many cases $x = (i-1/2+\Delta)/n$ (as i varies over $\mathbb{Z}$) for the sake of integrating F_a. We conclude that for every $x \in \mathbb{R}$ such that $nx - \Delta + \frac{1}{2}$ is not an integer,

$$F_a(x) = \begin{cases} np_i \log \frac{p_i}{p_{i+k-1}}, & \text{if } x \in \mathcal{J}_{n,i}, \quad x < \frac{i-1/2+\Delta}{n}, \\ np_i \log \frac{p_i}{p_{i+k}}, & \text{if } x \in \mathcal{J}_{n,i}, \quad x > \frac{i-1/2+\Delta}{n}. \end{cases} \tag{A.123}$$

Since $\int_{\mathbb{R}} |F_a| = \sum_{i\in\mathbb{Z}} \int_{\mathcal{J}^\circ_{n,i}} |F_a|$, we obtain

$$\int_{\mathbb{R}} |F_a| = \sum_{i\in\mathbb{Z}} p_i \left(\Delta \left| \log \frac{p_i}{p_{i+k-1}} \right| + (1-\Delta) \left| \log \frac{p_i}{p_{i+k}} \right| \right). \tag{A.124}$$

Now, we may conclude that $F_a \in L^1(\mathbb{R})$ by comparison with a geometric series. Indeed, we show the convergence of the series

$$S_\ell := \sum_{i\in\mathbb{Z}} p_i \left| \log \frac{p_i}{p_{i+\ell}} \right| \tag{A.125}$$

for each fixed $\ell \in \mathbb{Z}$. Consider the set of indices

$$I = \mathbb{Z} \setminus \{-N - |\ell|, \cdots, N + |\ell|\}, \tag{A.126}$$

and note that for each $i \in I$ we have $p_{i+j} = p_N r^{|i+j|-N}$ for both values $j \in \{0, \ell\}$. In particular, for $i \in I$ we have that

$$\left| \log \frac{p_i}{p_{i+\ell}} \right| = ||i| - |i+\ell|| \cdot \log \frac{1}{r} \leq |\ell| \cdot \log \frac{1}{r}. \tag{A.127}$$

Therefore, we obtain the bound

$$S_\ell \leq \frac{|\ell| p_N \log \frac{1}{r}}{r^N} \cdot \frac{1+r}{1-r} + \sum_{|i| \leq N + |\ell|} p_i \left| \log \frac{p_i}{p_{i+\ell}} \right| < \infty. \tag{A.128}$$

As S_k and S_{k-1} are both finite, we conclude from (A.124) that $F_a \in L^1(\mathbb{R})$. Therefore, by countable additivity,

$$D(P \| T_{-a} P) = \sum_{i\in\mathbb{Z}} \int_{\mathcal{J}^\circ_{n,i}} F_a, \tag{A.129}$$

i.e.,

$$D(P \| T_{-a} P) = \sum_{i\in\mathbb{Z}} p_i \left(\Delta \log \frac{p_i}{p_{i+k-1}} + (1-\Delta) \log \frac{p_i}{p_{i+k}} \right). \tag{A.130}$$

Let B_ℓ denote the same sum as S_ℓ but without the absolute value sign,

$$B_\ell := \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+\ell}}. \tag{A.131}$$

Finiteness of the S_ℓ yields from (A.130) that

$$D(P \| T_{-a} P) = \Delta B_{k-1} + (1-\Delta) B_k. \tag{A.132}$$

Also, the relation we are aiming to prove (A.118) can be restated as

$$\sup_{|d| \leq 1} D(P \| T_d P) = \max_{|\ell| \leq n} B_\ell. \tag{A.133}$$

We deduce from $k = an + \Delta$, $|a| \leq 1$, and $0 \leq \Delta \leq 1$ that we must have $-n \leq k \leq n+1$. If it holds that $-n+1 \leq k \leq n$, then what we have shown in (A.132) implies, in view of $0 \leq \Delta \leq 1$, that

$$D(P\|T_{-a}P) \leq \max_{|\ell|\leq n} B_\ell. \tag{A.134}$$

We treat the remaining two extreme cases $k \in \{-n, n+1\}$ separately. First, if $k = -n$ then $\Delta = 0$, in which case

$$D(P\|T_{-a}P) = B_{-n} \leq \max_{|\ell|\leq n} B_\ell. \tag{A.135}$$

Second, if $k = n+1$ then $\Delta = 1$, in which case

$$D(P\|T_{-a}P) = B_n \leq \max_{|\ell|\leq n} B_\ell. \tag{A.136}$$

Combining all cases, we conclude that

$$\sup_{|d|\leq 1} D(P\|T_dP) \leq \max_{|\ell|\leq n} B_\ell. \tag{A.137}$$

We establish now that the reverse inequality in (A.137) also holds. Let $\ell \in \{0, \cdots, n\}$. The shift $a_\ell := \ell/n$ satisfies $|a_\ell| \leq 1$ and $a_\ell + \frac{1}{2n} \in \mathcal{J}_{n,\ell}$. Also, $\Delta_\ell := \ell - a_\ell n = 0$. Therefore, we conclude from (A.132) that

$$D(P\|T_{-a_\ell}P) = B_\ell. \tag{A.138}$$

This shows that

$$\sup_{|d|\leq 1} D(P\|T_dP) \geq \max_{0\leq \ell\leq n} B_\ell. \tag{A.139}$$

In addition, consider $\ell \in \{-n, \cdots, -1\}$ and the shift $a'_\ell := \ell/n$. Then, in this case $a'_\ell + \frac{1}{2n} \in \mathcal{J}_{n,\ell+1}$. Also, $\Delta'_\ell := (\ell+1) - a'_\ell n = 1$. Thus, by (A.132), we have that

$$D(P\|T_{-a'_\ell}P) = B_{(\ell+1)-1} = B_\ell. \tag{A.140}$$

Therefore,

$$\sup_{|d|\leq 1} D(P\|T_{-d}P) \geq \max_{-n\leq \ell\leq -1} B_\ell. \tag{A.141}$$

Combining (A.139) and (A.141), we conclude that

$$\sup_{|d|\leq 1} D(P\|T_{-d}P) \geq \max_{|\ell|\leq n} B_\ell. \tag{A.142}$$

Inequality (A.142) together with the reverse inequality (A.137) yield that the desired equation (A.118) holds, i.e.,

$$\sup_{|a|\le 1} D(P\|T_aP) = \max_{|k|\le n} \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}}. \tag{A.143}$$

Next, we show that the expression

$$\max_{|k|\le n} \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}} \tag{A.144}$$

reduces to the form given in the statement of the theorem. By construction, $p_i = p_{-i}$ for each $i \in \mathbb{Z}$. Thus, we have for each $k \in \mathbb{Z}$

$$B_k = \sum_{i\in\mathbb{Z}} p_i \log \frac{p_i}{p_{i+k}} = \sum_{j\in\mathbb{Z}} p_{-j} \log \frac{p_{-j}}{p_{-j+k}} = \sum_{j\in\mathbb{Z}} p_j \log \frac{p_j}{p_{j-k}} = B_{-k}. \tag{A.145}$$

Therefore, $B_k = (B_k + B_{-k})/2$ for every $k \in \mathbb{Z}$. Note that this is a symmetric expression in k. As $B_0 = 0$, the KL-divergence is nonnegative, and $B_k \ge 0$ for every $|k| \le n$ (see (A.138) and (A.140)), we conclude that

$$\sup_{|a|\le 1} D(P\|T_aP) = \max_{1\le k\le n} \frac{1}{2}(B_k + B_{-k}). \tag{A.146}$$

We now rewrite (A.146) in terms of p_i for only $0 \le i \le N$, by taking advantage of (A.116). Fix $k \in \{1, \cdots, n\}$. We may write

$$B_{-k} = \sum_{j\in\mathbb{Z}} p_j \log \frac{p_j}{p_{j-k}} = \sum_{i\in\mathbb{Z}} p_{i+k} \log \frac{p_{i+k}}{p_i}, \tag{A.147}$$

so

$$B_k + B_{-k} = \sum_{i\in\mathbb{Z}} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \tag{A.148}$$

We split this sum at the points $-N, N-k$, and N. For any $k \in \{1, \ldots, n\}$, using the assumption that $n < N$, we may write

$$\begin{aligned}\sum_{i\in\mathbb{Z}} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} &= \sum_{i=-N+1}^{N-k-1} (p_{|i|} - p_{|i+k|}) \log \frac{p_{|i|}}{p_{|i+k|}} \\ &\quad + \sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} + \sum_{i=-\infty}^{-N} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \end{aligned} \tag{A.149}$$

In fact, the third term in (A.149) is identical to the second. This is proved by

$$\sum_{i=-\infty}^{-N} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}} = \sum_{i=N}^{\infty} (p_{-i} - p_{-i+k}) \log \frac{p_{-i}}{p_{-i+k}} \tag{A.150}$$

$$= \sum_{i=N}^{\infty} (p_i - p_{i-k}) \log \frac{p_i}{p_{i-k}} \tag{A.151}$$

$$= \sum_{i=N-k}^{\infty} (p_{i+k} - p_i) \log \frac{p_{i+k}}{p_i} \tag{A.152}$$

$$= \sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}. \tag{A.153}$$

Moreover, we may rewrite this expression as

$$\sum_{i=N-k}^{\infty} (p_i - p_{i+k}) \log \frac{p_i}{p_{i+k}}$$

$$= \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} + \sum_{i=N}^{\infty} (p_N r^{i-N} - p_N r^{i+k-N}) \log \frac{p_N r^{i-N}}{p_N r^{i+k-N}} \tag{A.154}$$

$$= \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} + p_N \sum_{i=N}^{\infty} r^{i-N}(1 - r^k) \log r^{-k} \tag{A.155}$$

$$= \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} + p_N \frac{1-r^k}{1-r} k \log r^{-1}. \tag{A.156}$$

Putting all of the above together shows that (A.146) is exactly equal to the objective function in (2.82).

Finally, we show that the cost constraint

$$\mathbb{E}_P[c] \leq C \tag{A.157}$$

is equivalent to the one given in (2.82). By nonnegativity of c, we have that

$$\mathbb{E}_P[c] = \int_{\mathbb{R}} fc = \sum_{i \in \mathbb{Z}} \int_{\mathcal{J}_{n,i}} np_i c = \sum_{i \in \mathbb{Z}} p_i c_{n,i} = p_0 c_{n,0} + 2 \sum_{i=1}^{N-1} p_i c_{n,i} + 2p_N \sum_{i=N}^{\infty} c_{n,i} r^{i-N}, \tag{A.158}$$

and the proof is complete.

A.5 Proof of Theorem 2.8: Optimality of Cactus

We will use the integration shorthand

$$\int_A f := \int_A f(x)\, dx. \tag{A.159}$$

Define

$$\gamma := \begin{cases} 1/2 & \text{if } \alpha > 1, \\ \alpha/2 & \text{otherwise.} \end{cases} \tag{A.160}$$

Note that $\gamma \in (0, 1/2]$ and $\gamma < \alpha$. Define the PDF

$$\psi(x) := \exp\left(-|x|^{\gamma}\right) \cdot \chi^{-1}, \tag{A.161}$$

where

$$\chi := \int_{\mathbb{R}} \exp\left(-|x|^{\gamma}\right) \, dx \tag{A.162}$$

is the normalization constant. As $\gamma \in (0,1]$, the function $z \mapsto |z|^{\gamma}$ is subadditive. Hence, for any $x, y \in \mathbb{R}$ we have the inequality

$$\frac{\psi(x+y)}{\psi(x)} \leq \exp\left(|y|^{\gamma}\right). \tag{A.163}$$

For each $\sigma > 0$, denote the dilated PDF

$$\psi^{\sigma}(x) := \frac{1}{\sigma}\psi\left(\frac{x}{\sigma}\right). \tag{A.164}$$

We denote the result of convolving a PDF q with ψ^{σ} by q_{σ},

$$q_{\sigma} := q * \psi^{\sigma}. \tag{A.165}$$

For any $a \in \mathbb{R}$, it is easy to see that

$$T_a(q_{\sigma}) = (T_a q)_{\sigma}, \tag{A.166}$$

so we denote this common quantity by $T_a q_{\sigma}$.

Due to the length of the proof, we break down some of the initial steps into the following five auxiliary lemmas. The proof resumes in the subsequent subsection.

A.5.1 Auxiliary Lemmas

The first lemma helps reduce the problem to considering only continuous PDFs. Specifically, it shows that a convolution q_{σ} can perform arbitrarily close to how the original PDF q does.

Lemma A.2. *For any PDF q and constant $\eta > 0$, there is a constant $\sigma_0 \in (0,1)$ such that $\sigma \in (0, \sigma_0]$ implies the inequalities*

$$D(q_{\sigma} \| T_a q_{\sigma}) \leq D(q \| T_a q), \quad \textit{for all } a \in \mathbb{R}, \tag{A.167}$$

$$\mathbb{E}_{q_{\sigma}}[c] \leq \mathbb{E}_q[c] + \eta. \tag{A.168}$$

Proof. First, by the data-processing inequality, for any $a \in \mathbb{R}$ and $\sigma > 0$,

$$D(q_{\sigma} \| T_a q_{\sigma}) \leq D(q \| T_a q). \tag{A.169}$$

Thus, (A.167) always holds. We may assume that $\mathbb{E}_q[c] < \infty$, for otherwise (A.168) trivially holds.

Now, we will establish (A.168) for all small σ by proving the limit

$$\lim_{\sigma \to 0^+} \mathbb{E}_{q_\sigma}[c] = \mathbb{E}_q[c]. \tag{A.170}$$

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $Z, V : \Omega \to \mathbb{R}$ be independent random variables with PDFs q and ψ, respectively, with respect to λ, i.e., with $P_Z(B) := P(Z^{-1}(B))$ and $P_V(B) := P(V^{-1}(B))$ we have

$$\frac{dP_Z}{d\lambda} = q, \qquad \frac{dP_V}{d\lambda} = \psi. \tag{A.171}$$

Then, for any $\sigma > 0$, the random variable $Z_\sigma := Z + \sigma V$ has PDF q_σ (see equations (A.160)–(A.165)). Denote integration against P by $\mathbb{E}$; in particular,

$$\mathbb{E}[f(Z, V)] := \int_\Omega f(Z(\omega), V(\omega))\, dP(\omega) \tag{A.172}$$

for any Borel function $f : \mathbb{R}^2 \to \mathbb{R}$.

By Slutsky's theorem, we have that $Z_\sigma \to Z$ in distribution. By the continuous mapping theorem, we also have that $c(Z_\sigma) \to c(Z)$ in distribution. Thus, by the Lebesgue-Vitali theorem [Bog07, Theorem 4.5.4], to conclude that (A.170) holds, it suffices to show uniform integrability of $\{c(Z_\sigma)\}_{0<\sigma\leq 1}$, i.e., it suffices to show that

$$\lim_{K\to\infty} \sup_{0<\sigma\leq 1} \mathbb{E}\left[c(Z_\sigma) \cdot 1_{(K,\infty)}(c(Z_\sigma))\right] = 0. \tag{A.173}$$

To establish (A.173), it suffices to uniformly upper bound the $c(Z_\sigma)$ (for $\sigma \in (0, 1]$) by an integrable random variable. To see this, note that if

$$\sup_{0<\sigma\leq 1} c(Z_\sigma) \leq U \tag{A.174}$$

for some random variable $U : \Omega \to \mathbb{R}$ with $\mathbb{E}[U] < \infty$, then we have the inequality

$$\sup_{0<\sigma\leq 1} \mathbb{E}\left[c(Z_\sigma) \cdot 1_{(K,\infty)}(c(Z_\sigma))\right] \leq \mathbb{E}\left[U \cdot 1_{(K,\infty)}(U)\right], \tag{A.175}$$

and the limit

$$\lim_{K\to\infty} \mathbb{E}\left[U \cdot 1_{(K,\infty)}(U)\right] = 0 \tag{A.176}$$

follows by absolute continuity of the Lebesgue integral in view of $\mathbb{E}[U] < \infty$.

Now, we show that a uniform bound as in (A.174) holds. Recall that for any $(u, v) \in \mathbb{R}^2$ and

$0 < s < t$, denoting $\|(u,v)\|_s := (|u|^s + |v|^s)^{1/s}$, one has from Hölder's inequality that

$$\|(u,v)\|_t \le \|(u,v)\|_s \le 2^{\frac{1}{s}-\frac{1}{t}} \|(u,v)\|_t. \tag{A.177}$$

In particular, for any $r > 0$, denoting $\ell_r := \max(1, 2^{r-1})$, one has that

$$(|u| + |v|)^r \le \ell_r (|u|^r + |v|^r). \tag{A.178}$$

In addition, by the tail-regularity assumption on c, there is a constant $\beta_1 > 0$ such that

$$c(x) \le \beta_1 \left(1 + |x|^\alpha\right) \tag{A.179}$$

for every $x \in \mathbb{R}$. Then, for any $u, v \in \mathbb{R}$, we have that

$$c(u+v) \le \beta_1 \left(1 + \ell_\alpha \left(|u|^\alpha + |v|^\alpha\right)\right). \tag{A.180}$$

In particular, for every $\sigma \in (0,1]$,

$$c(Z_\sigma) \le \beta_1 \left(1 + \ell_\alpha \left(|Z|^\alpha + |V|^\alpha\right)\right) =: U. \tag{A.181}$$

Now, we have that $\mathbb{E}[|V|^\alpha] < \infty$ by definition of ψ. Further, by assumption on c, there are $A, \beta_2 > 0$ such that $|x| > A$ implies

$$\beta_2 |x|^\alpha \le c(x). \tag{A.182}$$

Then, as

$$|Z|^\alpha \le A^\alpha + |Z|^\alpha \cdot 1_{\mathbb{R}\setminus[-A,A]}(Z) \le A^\alpha + c(Z)/\beta_2 \tag{A.183}$$

and $\mathbb{E}[c(Z)] = \mathbb{E}_q[c] < \infty$ by assumption, we also have that $\mathbb{E}[|Z|^\alpha] < \infty$. Thus, $\mathbb{E}[U] < \infty$. Hence, by absolute continuity of the Lebesgue integral, the uniform bound in (A.181) implies the uniform integrability of the set $\{c(Z_\sigma)\}_{0<\sigma\le 1}$, so (A.170) follows by the Lebesgue-Vitali theorem, and the proof is complete. □

The following lemma shows that the integrands when computing $D(q_\sigma \| T_a q_\sigma)$ have equi-small tails as a varies over $[-1,1]$. This will allow us to focus on approximating q_σ by a cactus distribution only in a bounded interval.

Lemma A.3. *If the PDF q satisfies*

$$\sup_{|a|\le 1} D(q \| T_a q) < \infty \tag{A.184}$$

then for any $\sigma > 0$

$$\lim_{z\to\infty} \sup_{|a|\le 1} \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \left|\log \frac{q_\sigma}{T_a q_\sigma}\right| = 0. \tag{A.185}$$

Proof. Assume that q satisfies (A.184). By the data processing inequality, we also have

$$\sup_{|a|\le 1} D(q_\sigma \| T_a q_\sigma) < \infty. \tag{A.186}$$

Suppose, for the sake of contradiction, that (A.185) does not hold. That is, suppose there exists $\varepsilon > 0$ where

$$\limsup_{z\to\infty} \sup_{|a|\le 1} \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \left|\log \frac{q_\sigma}{T_a q_\sigma}\right| = \varepsilon. \tag{A.187}$$

This implies that there exists a sequence $\{(z_n, a_n)\}_{n\in\mathbb{N}}$, where $z_n \nearrow \infty$ and $\sup_{n\in\mathbb{N}} |a_n| \le 1$, such that for all n

$$\int_{\mathbb{R}\setminus[-z_n,z_n]} q_\sigma \left|\log \frac{q_\sigma}{T_{a_n} q_\sigma}\right| \ge \varepsilon/2. \tag{A.188}$$

Since $[-1,1]$ is a compact set, there exists a convergent subsequence $\{a_{n_k}\}_{k\in\mathbb{N}}$, say $a_{n_k} \to a$ where $a \in [-1,1]$. Moreover, for any $z > 0$, for sufficiently large k we have $z_{n_k} \ge z$, which implies

$$\limsup_{k\to\infty} \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \left|\log \frac{q_\sigma}{T_{a_{n_k}} q_\sigma}\right| \ge \varepsilon/2. \tag{A.189}$$

Recall that ψ is as defined in (A.161) and that, as shown in (A.163), it satisfies the inequality

$$\frac{\psi(x+y)}{\psi(x)} \le \exp\left(|y|^\gamma\right) \tag{A.190}$$

for every $x, y \in \mathbb{R}$. Thus, for any $a, b, z \in \mathbb{R}$,

$$\begin{aligned} (T_a q_\sigma)(z) &= q_\sigma(z-a) && \text{(A.191)}\\ &= \int_{\mathbb{R}} q(x) \frac{1}{\sigma} \psi\left(\frac{z-a-x}{\sigma}\right) dx && \text{(A.192)}\\ &\le e^{|a-b|^\gamma/\sigma^\gamma} \int_{\mathbb{R}} q(x) \frac{1}{\sigma} \psi\left(\frac{z-b-x}{\sigma}\right) dx && \text{(A.193)}\\ &= e^{|a-b|^\gamma/\sigma^\gamma} (T_b q_\sigma)(z). && \text{(A.194)} \end{aligned}$$

Thus, for any $a, b \in \mathbb{R}$, we have the uniform bound

$$\left\|\log \frac{T_a q_\sigma}{T_b q_\sigma}\right\|_{L^\infty(\mathbb{R})} \le \left(\frac{|a-b|}{\sigma}\right)^\gamma. \tag{A.195}$$

Applying this bound to the integral in (A.189) gives

$$\int_{\mathbb{R}\setminus[-z,z]} q_\sigma \cdot \left| \log \frac{q_\sigma}{T_{a_{n_k}} q_\sigma} \right| = \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \cdot \left| \log \frac{q_\sigma}{T_a q_\sigma} + \log \frac{T_a q_\sigma}{T_{a_{n_k}} q_\sigma} \right| \tag{A.196}$$

$$\leq \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \cdot \left(\left| \log \frac{q_\sigma}{T_a q_\sigma} \right| + \left(\frac{|a_{n_k} - a|}{\sigma} \right)^\gamma \right) \tag{A.197}$$

$$\leq \left(\frac{|a_{n_k} - a|}{\sigma} \right)^\gamma + \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \cdot \left| \log \frac{q_\sigma}{T_a q_\sigma} \right|. \tag{A.198}$$

Recalling inequality (A.189) and that $a_{n_k} \to a$ as $k \to \infty$, we have, for any $z > 0$,

$$\int_{\mathbb{R}\setminus[-z,z]} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| \geq \varepsilon/2. \tag{A.199}$$

Finally, note that by finiteness of the KL-divergence $D(q_\sigma \| T_a q_\sigma)$ (see (A.186)), we also have that

$$\int_{\mathbb{R}} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| < \infty. \tag{A.200}$$

Indeed, the function $f(t) := t \log t$ over $(0, \infty)$ is lower bounded by $-1/e$, so dividing the integration region over the two regions where f is positive or negative we obtain

$$\int_{\mathbb{R}} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| = \mathbb{E}_{T_a q_\sigma} \left[\left| f \circ \frac{q_\sigma}{T_a q_\sigma} \right| \right] \leq D\left(q_\sigma \| T_a q_\sigma \right) + \frac{2}{e} < \infty. \tag{A.201}$$

Thus, by the monotone convergence theorem, we must have

$$\lim_{z \to \infty} \int_{\mathbb{R}\setminus[-z,z]} q_\sigma \left| \log \frac{q_\sigma}{T_a q_\sigma} \right| = 0. \tag{A.202}$$

As this contradicts (A.199), the lemma is proved. □

The following lemma gives an $\exp(-O(w^\gamma))$ lower bound on the minimum value of q_σ over $[-w, w]$ and on the probability that $Z_\sigma \sim q_\sigma$ exceeds w, both as $w \to \infty$.

Lemma A.4. *For a PDF q and a constant $\sigma > 0$, we have that*

$$\int_{[w,\infty)} q_\sigma = \exp\left(-O(w^\gamma)\right) \tag{A.203}$$

and

$$\min_{|x| \leq w} q_\sigma(x) = \exp\left(-O(w^\gamma)\right), \tag{A.204}$$

both as $w \to \infty$.

Proof. First, we show that there is a bounded Borel set B with $\lambda(B) > 0$ such that

$$\mu := \inf_{x \in B} q(x) > 0. \tag{A.205}$$

Note that we may remove the boundedness condition on B. Indeed, if the Borel set B satisfies $\lambda(B) > 0$ and $\inf_{x \in B} q(x) > 0$, then the bounded Borel sets $A_m := B \cap [-m, m]$ also satisfy $\lambda(A_m) > 0$ and $\inf_{x \in A_m} q(x) > 0$ for all large m by continuity of λ and the definition of the infimum. Now, to see that such a B exists, consider the Borel sets $B_n := q^{-1}([1/n, \infty))$ for integers $n \geq 1$. For each $n \geq 1$, we have that $\inf_{x \in B_n} q(x) \geq 1/n$. Suppose, for the sake of contradiction, that $\lambda(B_n) = 0$ for each n. Then we would have

$$\lambda(q^{-1}((0,\infty))) = \lambda\left(q^{-1}\left(\bigcup_{n \geq 1} [1/n, \infty)\right)\right) = \lambda\left(\bigcup_{n \geq 1} B_n\right) = 0. \tag{A.206}$$

Hence, $q = 0$ a.e. However, this would contradict that q is a PDF. Thus, we conclude that $\lambda(B_n) > 0$ for some n. In short, there must exist a bounded Borel set B with $\lambda(B) > 0$ and $\inf_{x \in B} q(x) > 0$. Fix such a B, and let $x_0 > 0$ be such that $B \subset [-x_0, x_0]$.

Recall that we define $q_\sigma = q * \psi^\sigma$ (see equations (A.160)–(A.165)). For each $w \in \mathbb{R}$, Tonelli's theorem implies that

$$\int_{[w,\infty)} q_\sigma = \int_{\mathbb{R}} q(x) \int_w^\infty \psi\left(\frac{y - x}{\sigma}\right) \frac{1}{\sigma}\, dy\, dx. \tag{A.207}$$

Performing a change of variable, we have for every $x, w \in \mathbb{R}$

$$\int_w^\infty \psi\left(\frac{y - x}{\sigma}\right) \frac{1}{\sigma}\, dy = \int_{[(w-x)/\sigma, \infty)} \psi. \tag{A.208}$$

Further, for any $z \geq 0$, by definition of ψ, we have the bound

$$\int_{[z,\infty)} \psi \geq \int_{[z,z+1]} \psi \geq \exp\left(-(z+1)^\gamma\right) \cdot \chi^{-1}, \tag{A.209}$$

where $\chi = \int_{\mathbb{R}} \exp(-|u|^\gamma)\, du$ is the normalization constant for ψ. Therefore, whenever $w \geq x$ we have

$$\int_{[(w-x)/\sigma,\infty)} \psi \geq \exp\left(-\left(\frac{w - x + \sigma}{\sigma}\right)^\gamma\right) \cdot \chi^{-1}. \tag{A.210}$$

Now, combining (A.207) and (A.208), nonnegativity of the PDFs q and ψ implies the bound

$$\int_{[w,\infty)} q_\sigma \geq \int_B q(x) \int_{[(w-x)/\sigma,\infty)} \psi(u)\, du\, dx. \tag{A.211}$$

Since $B \subset [-x_0, x_0]$, we conclude from (A.210) that for every $w \geq x_0$

$$\int_{[w,\infty)} q_\sigma \geq \int_B \mu \cdot \exp\left(-\left(\frac{w-x+\sigma}{\sigma}\right)^\gamma\right) \cdot \chi^{-1}\, dx \tag{A.212}$$

$$\geq \lambda(B)\mu\chi^{-1} \cdot \exp\left(-\left(\frac{w+x_0+\sigma}{\sigma}\right)^\gamma\right). \tag{A.213}$$

The estimate in (A.203) follows by taking $w \to \infty$.

Finally, we show that (A.204) holds. Let $w_0 > 0$ be such that $\int_{[-w,w]} q \geq 1/2$ for every $w \geq w_0$. Then, for any $w \geq w_0$ and $x \in [-w, w]$,

$$q_\sigma(x) = \int_{\mathbb{R}} q(u)\psi^\sigma(x-u)\, du \tag{A.214}$$

$$= (\sigma\chi)^{-1} \int_{\mathbb{R}} q(u) \exp\left(-|x-u|^\gamma/\sigma^\gamma\right)\, du \tag{A.215}$$

$$\geq (\sigma\chi)^{-1} \int_{-w}^{w} q(u) \exp\left(-|x-u|^\gamma/\sigma^\gamma\right)\, du \tag{A.216}$$

$$\geq (\sigma\chi)^{-1} \exp\left(-(2/\sigma^\gamma)w^\gamma\right) \int_{[-w,w]} q \tag{A.217}$$

$$\geq (2\sigma\chi)^{-1} \exp\left(-(2/\sigma^\gamma)w^\gamma\right). \tag{A.218}$$

The estimate (A.204) follows by taking $w \to \infty$. □

Conversely, the following lemma gives an upper bound on the tail of any distribution that satisfies the cost constraint.

Lemma A.5. *For any $P \in \mathscr{B}$, if $\mathbb{E}_P[c] < \infty$ then*

$$P(\mathbb{R} \setminus [-x, x]) = o\left(c(x)^{-1}\right) \tag{A.219}$$

as $x \to \infty$.

Proof. By evenness of c, it suffices to show that $P((x,\infty)) = o(c(x)^{-1})$. By monotonicity of c,

$$c(x)P((x,\infty)) = c(x)\int_{(x,\infty)} dP \leq \int_{(x,\infty)} c(t)\, dP(t) \to 0, \tag{A.220}$$

as desired. □

The final auxiliary lemma gives an upper bound on the tail of the cost constraint incurred by a cactus distribution.

Lemma A.6. *Fix $r \in (0,1)$ and integers $N > n \geq 1$, and set $w = (N - 1/2)/n$. Assume that $c(x) \leq \beta_1 x^\alpha$*

for $x \geq w$. Then, we have the bound

$$\sum_{i \geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{2\left(\frac{\alpha}{e}\right)^\alpha \log \frac{1}{r} + \Gamma(\alpha+1)}{rn^\alpha \left(\log \frac{1}{r}\right)^{\alpha+1}} \right), \tag{A.221}$$

where $\ell_\alpha := \max(1, 2^{\alpha-1})$.

Proof. By monotonicity of c,

$$\begin{aligned}\sum_{i \geq N} c_{n,i} r^{i-N} &= \sum_{i \geq N} \int_{(i-1/2)/n}^{(i+1/2)/n} ncr^{i-N} && \text{(A.222)}\\ &\leq \sum_{i \geq N} \beta_1 \left(\frac{i+1/2}{n}\right)^\alpha r^{i-N} && \text{(A.223)}\\ &= \beta_1 \sum_{i \geq 0} \left(w + \frac{i+1}{n}\right)^\alpha r^i && \text{(A.224)}\\ &\leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{\mathrm{Li}_{-\alpha}(r)}{rn^\alpha}\right), && \text{(A.225)}\end{aligned}$$

where

$$\mathrm{Li}_{-\alpha}(r) := \sum_{k \geq 1} k^\alpha r^k \tag{A.226}$$

is the polylogarithm function. To finish the proof of the lemma, we show next that

$$\mathrm{Li}_{-\alpha}(r) \leq 2\left(\frac{\alpha}{e \log \frac{1}{r}}\right)^\alpha + \frac{\Gamma(\alpha+1)}{\left(\log \frac{1}{r}\right)^{\alpha+1}}. \tag{A.227}$$

Now, consider the function $g : (0, \infty) \to (0, \infty)$ defined by

$$g(x) := x^\alpha r^x. \tag{A.228}$$

We have that

$$g'(x) = (\alpha + x \log r)\, x^{\alpha-1} r^x. \tag{A.229}$$

Thus, g increases until it reaches a maximum at $x_0 = \alpha / \log \frac{1}{r}$ then it decreases. Thus,

$$\mathrm{Li}_{-\alpha}(r) \leq g(\lfloor x_0 \rfloor) + g(\lceil x_0 \rceil) + \int_{(0,\infty)} g. \tag{A.230}$$

We have

$$g(\lfloor x_0 \rfloor) + g(\lceil x_0 \rceil) \leq 2g(x_0) = 2\left(\frac{\alpha}{e \log \frac{1}{r}}\right)^\alpha, \tag{A.231}$$

and

$$\int_{(0,\infty)} g = \frac{\Gamma(\alpha+1)}{\left(\log \frac{1}{r}\right)^{\alpha+1}}. \tag{A.232}$$

The proof is thus complete. □

A.5.2 Proof of Theorem 2.8

By Theorem 2.3, there is an even PDF $q^\star$ that satisfies both

$$\sup_{|a|\le 1} D(q^\star \| T_a q^\star) = \mathrm{KL}^\star, \tag{A.233}$$

$$\mathbb{E}_{q^\star}[c] \le C. \tag{A.234}$$

Fix arbitrary constants $\delta, \eta > 0$, and we will find a cactus distribution that attains the KL-divergence (A.233) to within δ and the cost (A.234) to within η.

By Lemma A.2, there is a $\sigma > 0$ such that the PDF $q_\sigma^\star$ satisfies the bounds

$$\sup_{|a|\le 1} D(q_\sigma^\star \| T_a q_\sigma^\star) \le \mathrm{KL}^\star, \tag{A.235}$$

$$\mathbb{E}_{q_\sigma^\star}[c] \le C + \frac{\eta}{2}. \tag{A.236}$$

Throughout the proof, we will denote

$$q := q_\sigma^\star \tag{A.237}$$

for short. Let

$$Q(B) := \int_B q \tag{A.238}$$

be the probability measure induced by q. We will construct a cactus distribution that approximates q.

We first note a few properties of q. Note that q is an even PDF. Further, it is uniformly continuous, and strictly positive over $\mathbb{R}$. Thus, q is locally bounded away from zero. For each $z \ge 0$, denote the minimum

$$\mu_z := \min_{|x|\le z} q(x), \tag{A.239}$$

so $\mu_z > 0$ for every z. In addition, q is upper bounded: by Young's inequality, we have that

$$\|q\|_{L^\infty(\mathbb{R})} = \|q^\star * \psi^\sigma\|_{L^\infty(\mathbb{R})} \le \|q^\star\|_{L^1(\mathbb{R})} \cdot \|\psi^\sigma\|_{L^\infty(\mathbb{R})} = (\sigma\chi)^{-1} =: M. \tag{A.240}$$

In fact, q satisfies a property resembling local γ-Hölder continuity. Specifically, as in the proof of

Lemma A.3 (see (A.190)–(A.194)), we have that

$$q(x) \leq e^{|x-y|^\gamma/\sigma^\gamma} q(y) \tag{A.241}$$

for every $x, y \in \mathbb{R}$. Therefore, for some $|t_{x,y}| \leq 1$ we have

$$|q(x) - q(y)| = q(y) \left| e^{t_{x,y}|x-y|^\gamma/\sigma^\gamma} - 1 \right| \leq \frac{2M}{\sigma^\gamma} |x-y|^\gamma, \tag{A.242}$$

where the latter inequality follows whenever $|x - y| \leq \sigma$. In particular, for all $\varepsilon \in (0, 2M)$, we have that

$$|q(x) - q(y)| \leq \varepsilon \quad \text{whenever} \quad |x-y| \leq \sigma \cdot \left(\frac{\varepsilon}{2M}\right)^{1/\gamma}. \tag{A.243}$$

Before constructing the parameters (n, N, r) of the cactus distribution, we note a fundamental lower bound on n. For the cost constraint to be satisfied, we need $c_{n,0} < C$ to hold. Nevertheless, by continuity of c, every real number is a Lebesgue point of c. In particular, as 0 is a Lebesgue point of c, we obtain

$$c_{n,0} = \frac{\int_{[-1/(2n),1/(2n)]} c}{1/n} \to c(0) = 0 \tag{A.244}$$

as $n \to \infty$. Let $n_{\min}$ be the least positive integer such that

$$c_{n,0} < C \tag{A.245}$$

for every $n \geq n_{\min}$. Note that $n_{\min}$ depends only on c and C.

Now, we choose the integers n and N. Denote the constants

$$\theta_\alpha := 4\left(\frac{\alpha}{e}\right)^\alpha + 2\Gamma(\alpha+1) \tag{A.246}$$

$$\theta'_\alpha := (2\theta_\alpha)^{1/\alpha} \tag{A.247}$$

$$\gamma' := \frac{\gamma+\alpha}{2} \in (\gamma, \alpha) \tag{A.248}$$

$$\varepsilon_{\min} := 2M \cdot \min\left(\frac{2}{\sigma n_{\min}}, \frac{1}{\theta'_\alpha \sigma}\right)^\gamma \tag{A.249}$$

$$z_{\min,0} := \left(\log\left(\frac{4}{\sigma} \cdot \left(\frac{2M}{\varepsilon_{\min}}\right)^{1/\gamma}\right)\right)^{1/\gamma'} \tag{A.250}$$

$$z_{\min,1} := \left(\frac{\eta}{\delta} \cdot \frac{2e\theta'_\alpha}{\beta_1 \ell_\alpha}\right)^{1/(\alpha+1)} \tag{A.251}$$

$$z_{\min,2} := \left(\frac{2^{\alpha+1}}{\beta_1 \ell_\alpha}\right)^{1/(\alpha-\gamma')} \tag{A.252}$$

$$z_{\min} := \max\left(z_{\min,0}, z_{\min,1}, z_{\min,2}, \frac{\delta}{12M}\right). \tag{A.253}$$

Since $q = q_\sigma^\star$ (see (A.237)), Lemma A.3 yields the existence of a constant $z_0 > 0$ such that $z \geq z_0$ implies the uniform bound

$$\sup_{|a|\leq 1} \int_{\mathbb{R}\setminus[-z,z]} q \left| \log \frac{q}{T_a q} \right| \leq \frac{\delta}{3}. \tag{A.254}$$

In addition, Lemma A.4 yields the existence of constants $\tau, z_1 > 0$ such that $z \geq z_1$ implies (see (A.238) and (A.239))

$$\min\left(\mu_z, Q([z,\infty))\right) \geq \exp\left(-\tau z^\gamma\right). \tag{A.255}$$

By the tail-regularity assumption on c, there are constants $\beta_1, \beta_2, z_2 > 0$ such that

$$\beta_2 z^\alpha \leq c(z) \leq \beta_1 z^\alpha \tag{A.256}$$

for every $z \geq z_2$. By Lemma A.5, we have that (see (A.238))

$$\lim_{z\to\infty} Q(\mathbb{R} \setminus [-z,z]) c(z) = 0. \tag{A.257}$$

Let $z_3 > 0$ be large enough that $z \geq z_3$ implies

$$Q\left(\mathbb{R} \setminus [-z,z]\right) c(z) \leq \frac{\beta_2}{\beta_1 \ell_\alpha} \cdot \frac{\eta}{6}. \tag{A.258}$$

If $z \geq \max(z_2, z_3)$, then by (A.256) and (A.258) we may bound the tail of Q also by

$$Q\left(\mathbb{R} \setminus [-z,z]\right) \leq \frac{1}{\beta_1 \ell_\alpha z^\alpha} \cdot \frac{\eta}{6}. \tag{A.259}$$

Let $z_4 > 0$ be the smallest number such that both inequalities

$$e^{\tau z^\gamma} \geq \frac{\delta \beta_1 \ell_\alpha}{2M\eta} \cdot z^\alpha \tag{A.260}$$

$$e^{\gamma z^{\gamma'}} \geq \left(\frac{4}{\sigma}\right)^\gamma \cdot \frac{48M^2}{\delta} \cdot z e^{\tau z^\gamma} \tag{A.261}$$

hold for all $z \geq z_4$. Fix a rational number

$$z > \max(z_{\min}, z_0, z_1, z_2, z_3, z_4, 2\theta'_\alpha) \tag{A.262}$$

that is a ratio of an odd integer by an even integer, and set

$$w := z + 1. \tag{A.263}$$

We choose z (hence also w) here to belong in $\mathbb{N} + \frac{1}{2}$ for simplicity, but we note that any other choice (of denominator) is also valid provided that w is increased so that the subsequent choices in (A.268) below can be made. Set

$$\varepsilon := \frac{2^{2\gamma+1}M}{\sigma^{\gamma}} \cdot e^{-\gamma w^{\gamma'}}. \tag{A.264}$$

Denote

$$n_0 := \frac{2}{\sigma} \cdot \left(\frac{2M}{\varepsilon}\right)^{1/\gamma}, \tag{A.265}$$

By the uniform continuity of q shown in (A.243), we have that

$$|q(x) - q(y)| \leq \varepsilon \quad \text{whenever} \quad |x - y| \leq \frac{2}{n_0}. \tag{A.266}$$

Note that $n_{\min} < n_0$ since $\varepsilon < \varepsilon_{\min}$, which in turn follows because $w > z_{\min,0}$. We note also that $\varepsilon < \varepsilon_{\min}$ implies $2\theta'_\alpha < n_0$. Set

$$n_1 := e^{w^{\gamma'}}. \tag{A.267}$$

By construction, we have that $n_1 = 2n_0$. Thus, we may choose integers $n \in [n_0, n_1]$ and $N > n$ such that

$$w = \frac{2N-1}{2n} \tag{A.268}$$

Next, we choose the parameter r, thereby completing the cactus distribution construction. Define, for $i \in \{0, \cdots, N-1\}$,

$$p_i := \inf_{x \in \mathcal{J}_{n,i}} \frac{q(x)}{n}. \tag{A.269}$$

By evenness, continuity, and strict positivity of q, we have that

$$p_0 + \sum_{i=1}^{N-1} 2p_i = \int_{[-w,w]} \sum_{|i| \leq N-1} np_{|i|} \cdot 1_{\mathcal{J}_{n,i}} \leq \int_{[-w,w]} q < 1. \tag{A.270}$$

Thus, for any $r \in (0,1)$, setting

$$p_N := \frac{1-r}{2}\left(1 - \left(p_0 + \sum_{i=1}^{N-1} 2p_i\right)\right), \tag{A.271}$$

we infer from (A.270) that the vector $\boldsymbol{p} = (p_0, \cdots, p_N)$ belongs to $(0,1]^{N+1}$, and by construction it satisfies $S_{r,\boldsymbol{p}} = 1$. We will choose r as

$$r := 1 - \frac{\theta'_\alpha}{wn}, \tag{A.272}$$

and define p_N as in (A.271) for this choice of r.

Therefore, $f_{n,r,\boldsymbol{p}}$ is a valid cactus distribution. By uniform continuity of q (see (A.266)) and by definition of the p_i (see (A.269)), we have that $f_{n,r,\boldsymbol{p}}$ uniformly approximates q from below over

$[-w,w]$: for every $x \in [-w,w]$ we have that

$$0 \leq q(x) - f_{n,r,\boldsymbol{p}}(x) \leq \varepsilon. \tag{A.273}$$

We will deduce from the uniform bound (A.273) that $f_{n,r,\boldsymbol{p}}$ approximates q in the two senses:

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq \mathbb{E}_q[c] + \frac{\eta}{2} \tag{A.274}$$

and

$$\sup_{|a|\leq 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \leq \sup_{|a|\leq 1} D(q \| T_a q) + \delta. \tag{A.275}$$

Combined with (A.235)–(A.236), we would conclude from (A.274)–(A.275) that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq C + \eta \tag{A.276}$$

and

$$\sup_{|a|\leq 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \leq \mathrm{KL}^\star + \delta. \tag{A.277}$$

Now, we show that $f_{n,r,\boldsymbol{p}}$ satisfies the cost constraint (A.276). Since $f_{n,r,\boldsymbol{p}}|_{[-w,w]} \leq q|_{[-w,w]}$, we have that

$$\mathbb{E}_{f_{n,r,p}}[c \cdot 1_{[-w,w]}] \leq \mathbb{E}_q[c] \leq C + \frac{\eta}{2}. \tag{A.278}$$

We show next that

$$\mathbb{E}_{f_{n,r,p}}[c \cdot 1_{\mathbb{R}\setminus[-w,w]}] \leq \frac{\eta}{2}. \tag{A.279}$$

By construction of $f_{n,r,\boldsymbol{p}}$, and since $w = (N-1/2)/n$ (see (A.268)), we have the expression

$$\mathbb{E}_{f_{n,r,p}}[c \cdot 1_{\mathbb{R}\setminus[-w,w]}] = 2p_N \sum_{i\geq N} c_{n,i} r^{i-N}. \tag{A.280}$$

We bound the terms $2p_N$ and $\sum_{i\geq N} c_{n,i} r^{i-N}$ separately. By Lemma A.6, we have the bound

$$\sum_{i\geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{2\left(\frac{\alpha}{e}\right)^\alpha \log\frac{1}{r} + \Gamma(\alpha+1)}{r n^\alpha \left(\log\frac{1}{r}\right)^{\alpha+1}} \right). \tag{A.281}$$

By definition of r (see (A.272)), and since $w \geq 1$ and $n \geq 2\theta_\alpha'$, we have that $r \geq 1/2 > 1/e$. Thus, we deduce from (A.281) that

$$\sum_{i\geq N} c_{n,i} r^{i-N} \leq \beta_1 \ell_\alpha \left(\frac{w^\alpha}{1-r} + \frac{\theta_\alpha}{n^\alpha \left(\log\frac{1}{r}\right)^{\alpha+1}} \right), \tag{A.282}$$

where θ_α is as defined in (A.246). In addition, we have that (recall that we denote by $P_{n,r,\boldsymbol{p}}$ the probability measure associated with $f_{n,r,\boldsymbol{p}}$)

$$\frac{2p_N}{1-r} = P_{n,r,\boldsymbol{p}}\left(\mathbb{R} \setminus [-w,w]\right) = 1 - P_{n,r,\boldsymbol{p}}\left([-w,w]\right). \tag{A.283}$$

As $f_{n,r,\boldsymbol{p}}$ uniformly approximates q from below over $[-w,w]$ to within ε (see (A.273)), we have that

$$P_{n,r,\boldsymbol{p}}\left([-w,w]\right) \ge Q\left([-w,w]\right) - 2\varepsilon w. \tag{A.284}$$

Thus, by the bound on the tail of Q in (A.259)

$$\frac{2p_N}{1-r} \le Q\left(\mathbb{R} \setminus [-w,w]\right) + 2\varepsilon w \le \frac{1}{\beta_1 \ell_\alpha w^\alpha} \cdot \frac{\eta}{6} + 2\varepsilon w. \tag{A.285}$$

Further, combining inequalities (A.260)–(A.261) and using the definition of ε in (A.264), we obtain

$$\varepsilon \le \frac{\eta}{12\beta_1 \ell_\alpha w^{\alpha+1}}. \tag{A.286}$$

Thus, we deduce

$$2p_N \le \frac{\eta \cdot (1-r)}{3\beta_1 \ell_\alpha w^\alpha}. \tag{A.287}$$

From the expression in (A.280), multiplying inequalities (A.282) and (A.287) and noting that $1-r \le \log \frac{1}{r}$, we obtain

$$\mathbb{E}_{f_{n,r,p}}[c \cdot 1_{\mathbb{R}\setminus[-w,w]}] \le \frac{\eta}{3}\left(1 + \frac{\theta_\alpha}{\left(wn \log \frac{1}{r}\right)^\alpha}\right). \tag{A.288}$$

By definition of r, we have that

$$\log \frac{1}{r} \ge 1 - r = \frac{\theta'_\alpha}{wn}. \tag{A.289}$$

Using inequality (A.289) in (A.288), we obtain

$$\mathbb{E}_{f_{n,r,p}}[c \cdot 1_{\mathbb{R}\setminus[-w,w]}] \le \frac{\eta}{3} \cdot \frac{3}{2} = \frac{\eta}{2}, \tag{A.290}$$

which is inequality (A.279). Combining (A.278)–(A.279), we deduce (A.276), i.e.,

$$\mathbb{E}_{f_{n,r,p}}[c] \le C + \eta. \tag{A.291}$$

Next, we show that $f_{n,r,\boldsymbol{p}}$ satisfies the KL bound (A.277). We begin by splitting the integration at the points $\pm z$. By finiteness of the considered KL-divergences, we have for each $|a| \le 1$

$$D(f_{n,r,\boldsymbol{p}} \| T_{-a} f_{n,r,\boldsymbol{p}}) - D(q \| T_{-a} q) \le \int_{[-z,z]} \left(f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} - q \log \frac{q}{T_{-a} q} \right)$$

$$+\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}}$$
$$+\int_{\mathbb{R}\setminus[-z,z]} q \left|\log \frac{q}{T_{-a} q}\right|. \tag{A.292}$$

We already have a uniform bound for the last integral in (A.292): since $z \geq z_0$, the estimate in (A.254) holds and we obtain

$$\sup_{|a|\leq 1} \int_{\mathbb{R}\setminus[-z,z]} q \left|\log \frac{q}{T_a q}\right| \leq \frac{\delta}{3}. \tag{A.293}$$

We proceed to bounding the first integral in (A.292) uniformly by

$$\sup_{|a|\leq 1} \int_{[-z,z]} \left(f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} - q \log \frac{q}{T_{-a} q} \right) \leq \frac{\delta}{3}. \tag{A.294}$$

We do this via deriving an upper bound on the integrand that is uniform in both a and the variable of integration. From $w \geq \delta/(12M)$ (A.253), $\mu_w \geq e^{-\tau w^\gamma}$ (A.255), and (A.261), we have that

$$\varepsilon \leq \frac{\mu_w}{2} \cdot \min\left(1, \frac{\delta}{12Mw}\right). \tag{A.295}$$

Define the function $g : [-w, w] \to [0, \varepsilon]$ by

$$g := q - f_{n,r,\boldsymbol{p}}. \tag{A.296}$$

That the range of g is contained within $[0, \varepsilon]$ follows since $f_{n,r,\boldsymbol{p}}$ approximates q from below uniformly over $[-w, w]$ to within ε. Thus, $z = w - 1$ yields

$$\sup_{|a|\leq 1} \|T_a g\|_{L^\infty([-z,z])} \leq \varepsilon. \tag{A.297}$$

We note that, over $[-z, z]$, the inequality

$$f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} - q \log \frac{q}{T_{-a} q} \leq -q \log\left(1 - T_{-a}\frac{g}{q}\right) - g \log\left(1 - \frac{g}{q}\right) + g \log \frac{T_{-a} q}{q} \tag{A.298}$$

holds; that all the logarithms are well defined follows since $g \leq q$ over $[-w, w]$. Indeed, subtracting the left hand side from the right hand side in (A.298), we get the function

$$-q \log\left(1 - \frac{g}{q}\right) - g \log\left(1 - T_{-a}\frac{g}{q}\right), \tag{A.299}$$

which is nonnegative over $[-z, z]$ since g is nonnegative over $[-w, w]$. Now, we bound each of the terms in (A.298). It is easy to see that for $0 \leq t \leq 1/2$ one has

$$-\log(1-t) \leq 2t. \tag{A.300}$$

Now, we show that $g/q \leq 1/2$ over $[-w,w]$. Indeed, this is equivalent to $q \leq 2f_{n,r,\boldsymbol{p}}$ over $[-w,w]$. But $q - \varepsilon \leq f_{n,r,\boldsymbol{p}}$ over $[-w,w]$, which implies in view of $\varepsilon \leq \mu_w/2 \leq q/2$ (over $[-w,w]$) that $q \leq 2f_{n,r,\boldsymbol{p}}$, as desired. Thus, we obtain that over $[-z,z]$

$$-q\log\left(1 - T_{-a}\frac{g}{q}\right) \leq 2qT_{-a}\frac{g}{q} \leq \frac{2M\varepsilon}{\mu_w}, \tag{A.301}$$

and

$$-g\log\left(1-\frac{g}{q}\right) \leq \frac{2g^2}{q} \leq \frac{2\varepsilon^2}{\mu_w} \leq \varepsilon. \tag{A.302}$$

It is also clear that over $[-z,z]$

$$g\log\frac{T_{-a}q}{q} \leq \varepsilon\log\frac{M}{\mu_w} \leq \varepsilon\left(\frac{M}{\mu_w}-1\right). \tag{A.303}$$

Plugging in inequalities (A.301)–(A.303) into (A.298), we obtain the uniform bound

$$f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q} \leq \frac{3M\varepsilon}{\mu_w} \tag{A.304}$$

over $[-z,z]$. Integrating, we deduce

$$\int_{[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} - q\log\frac{q}{T_{-a}q} \leq \frac{6zM\varepsilon}{\mu_w} < \frac{\delta}{3}, \tag{A.305}$$

where the last inequality follows by (A.295).

It remains to upper bound the middle integral in (A.298), for which we also derive a uniform upper bound

$$\sup_{|a|\leq 1}\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \leq \frac{\delta}{3}. \tag{A.306}$$

We will further split the integration at the points $\pm(w+1)$. By evenness of $f_{n,r,\boldsymbol{p}}$, we have that this integral depends only on $|a|$, i.e., for each $a \in [-1,1]$

$$\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} = \int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{a}f_{n,r,\boldsymbol{p}}}. \tag{A.307}$$

Thus, it suffices for (A.306) to show that

$$\sup_{0<a\leq 1}\int_{\mathbb{R}\setminus[-z,z]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \leq \frac{\delta}{3}. \tag{A.308}$$

Consider first the integral

$$\int_{\mathbb{R}\setminus[-(w+1),w+1]} f_{n,r,\boldsymbol{p}}\log\frac{f_{n,r,\boldsymbol{p}}}{T_{-a}f_{n,r,\boldsymbol{p}}} \tag{A.309}$$

for fixed $a \in (0,1]$. From the proof of Theorem 2.7, we can write the integrand in (A.309) as follows.

Extend the definition of p_i to all $i \in \mathbb{Z}$ by

$$p_i := \begin{cases} p_{|i|}, & \text{if } -N \le i \le -1, \\ p_N r^{|i|-N}, & \text{if } |i| > N. \end{cases} \tag{A.310}$$

For each $i \in \mathbb{Z}$, there is an integer j with $|j| \le n$, such that we have

$$f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} = n p_i \log \frac{p_i}{p_{i+j}} \tag{A.311}$$

over $\mathcal{J}_{n,i}$ except possibly at a single point. By definition of w, we have that

$$\mathbb{R} \setminus [-(w+1), w+1] = \bigcup_{|i| \ge N+n} \mathcal{J}_{n,i}. \tag{A.312}$$

Further, if $|i| \le N+n$ and $|j| \le n$, then $|i+j| \ge N$. Hence, from (A.311) we have that over $\mathcal{J}_{n,i}$ with $|i| \ge N+n$

$$f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} = n p_N r^{|i|-N} (|i| - |i+j|) \log r \le n^2 p_N r^{|i|-N} \log \frac{1}{r}. \tag{A.313}$$

Summing over $|i| \ge N+n$, we obtain

$$\int_{\mathbb{R} \setminus [-(w+1), w+1]} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} = \sum_{|i| \ge N+n} \int_{\mathcal{J}_{n,i}} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} \tag{A.314}$$

$$\le n p_N \log \frac{1}{r} \sum_{|i| \ge n} r^{|i|} = \frac{2n p_N r^n \log \frac{1}{r}}{1-r}. \tag{A.315}$$

Using the upper bound on p_N in (A.287), we obtain that

$$\int_{\mathbb{R} \setminus [-(w+1), w+1]} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} \le \frac{\eta n r^n \log \frac{1}{r}}{3\beta_1 \ell_\alpha w^\alpha}. \tag{A.316}$$

As $1/e \le r \le 1$ and $\log \frac{1}{r} \le \frac{1}{r} - 1$, using the definition of r given in (A.272) and $w \ge z_4$ (see (A.253)), we have the bound

$$\frac{\eta n r^n \log \frac{1}{r}}{3\beta_1 \ell_\alpha w^\alpha} \le \frac{e\eta n(1-r)}{3\beta_1 \ell_\alpha w^\alpha} \le \frac{e\eta \theta'_\alpha}{3\beta_1 \ell_\alpha w^{\alpha+1}} \le \frac{\delta}{6}. \tag{A.317}$$

Thus, we have shown that

$$\sup_{a \in (0,1]} \int_{\mathbb{R} \setminus [-(w+1), w+1]} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}}{T_{-a} f_{n,r,\boldsymbol{p}}} \le \frac{\delta}{6}. \tag{A.318}$$

The final integral bound we need is the following:

$$\sup_{0 < a \le 1} \int_{w-1 < |x| \le w+1} f_{n,r,\boldsymbol{p}}(x) \log \frac{f_{n,r,\boldsymbol{p}}(x)}{T_{-a} f_{n,r,\boldsymbol{p}}(x)} \, dx \le \frac{\delta}{6}. \tag{A.319}$$

By evenness of $f_{n,r,\boldsymbol{p}}$, we have that

$$\int_{w-1<|x|\le w+1} f_{n,r,\boldsymbol{p}}(x) \log \frac{f_{n,r,\boldsymbol{p}}(x)}{T_{-a}f_{n,r,\boldsymbol{p}}(x)}\, dx = \int_{(w-1,w+1]} f_{n,r,\boldsymbol{p}} \log \frac{f_{n,r,\boldsymbol{p}}^2}{(T_{-a}f_{n,r,\boldsymbol{p}}) \cdot (T_a f_{n,r,\boldsymbol{p}})}. \tag{A.320}$$

Consider the function inside the logarithm in the integrand:

$$\rho(x;a) := \frac{f_{n,r,\boldsymbol{p}}(x)^2}{f_{n,r,\boldsymbol{p}}(x+a) f_{n,r,\boldsymbol{p}}(x-a)}. \tag{A.321}$$

We will prove the uniform upper bound

$$\sup_{\substack{x\in(w-1,w+1]\\ a\in(0,1]}} \rho(x;a) \le \exp\left(2w^{\gamma'}\right), \tag{A.322}$$

where $\gamma' = (\gamma+\alpha)/2 \in (\gamma,\alpha)$ is as defined in (A.248). Note that

$$(w-1, w+1] = \bigcup_{i=N-n}^{N+n} \mathcal{J}_{n,i}. \tag{A.323}$$

For each $a \in (0,1]$ and $x \in (w-1, w+1]$, there are integers $N-n \le i \le N+n$ and $0 \le j,k \le n$ such that

$$\rho(x;a) = \frac{p_i^2}{p_{i+j}p_{i-k}}. \tag{A.324}$$

Thus, it suffices to show that $\exp(w^{\gamma'})$ is an upper bound on each of the terms

$$\frac{p_i}{p_j},\ \frac{p_k}{p_N r^n},\ \frac{p_N}{p_k},\ \frac{1}{r^n} \tag{A.325}$$

for $0 \le i,j,k \le N-1$ with $|i-j| \le n$. First, for $1/r^n$, denoting $m = nw/(2\theta_\alpha)^{1/\alpha} \ge 2$, we have the bound

$$r^n = \left(\left(1-\frac{1}{m}\right)^m\right)^{(2\theta_\alpha)^{1/\alpha}/w} \ge 4^{-(2\theta_\alpha)^{1/\alpha}/w} \ge \frac{1}{2}. \tag{A.326}$$

Hence,

$$\frac{1}{r^n} \le 2 \le e^{w^{\gamma'}}. \tag{A.327}$$

For p_k/p_N with $0 \le k \le N-1$, we have the bound

$$\frac{p_k}{p_N} \le \frac{M}{np_N} = \frac{2M/(1-r)}{n\cdot(2p_N/(1-r))} = \frac{2M/(1-r)}{nP_{n,r,\boldsymbol{p}}(\mathbb{R}\setminus[-w,w])} \tag{A.328}$$

$$\le \frac{2M/(1-r)}{nQ(\mathbb{R}\setminus[-w,w])} \le \frac{M/(1-r)}{ne^{-\tau w^\gamma}} = \frac{Mwe^{\tau w^\gamma}}{\theta'_\alpha}. \tag{A.329}$$

Hence,

$$\frac{p_k}{p_N r^n} \leq \frac{2Mwe^{\tau w^\gamma}}{\theta'_\alpha} \leq e^{w^{\gamma'}}, \tag{A.330}$$

where the last inequality follows from (A.261) for all small δ, e.g., for

$$\delta \leq 3 \cdot 2^{2\gamma+2} \cdot \theta'_\alpha \cdot \chi^{-1} \tag{A.331}$$

(alternatively, we may increase the size of w at the outset). Consider next p_i/p_j for $0 \leq i, j \leq N-1$ with $|i-j| \leq n$. By definition of the p_k and uniform continuity of q, we have for $0 \leq k \leq N-2$

$$|p_k - p_{k+1}| \leq \frac{\varepsilon}{n}. \tag{A.332}$$

By the triangle inequality, we deduce

$$\left|p_i - p_j\right| \leq \frac{|i-j|\varepsilon}{n} \leq \varepsilon. \tag{A.333}$$

Thus,

$$\frac{p_i}{p_j} \leq 1 + \frac{\varepsilon}{p_j} \leq 1 + \frac{n\varepsilon}{\mu_w} \leq 1 + \frac{n}{2} \leq e^{w^{\gamma'}}. \tag{A.334}$$

The last term p_N/p_k can be bounded using (A.287) to obtain

$$\frac{p_N}{p_k} \leq \frac{\eta \cdot (1-r)/(6\beta_1 \ell_\alpha w^\alpha)}{\mu_w/n} = \frac{\eta\theta'_\alpha}{6\beta_1 \ell_\alpha \mu_w w^{\alpha+1}} \leq \frac{\eta\theta'_\alpha}{6\beta_1 \ell_\alpha} \cdot e^{\tau w^\gamma} \leq e^{w^{\gamma'}}, \tag{A.335}$$

where the last inequality follows from (A.261) for all small η, e.g., for

$$\eta \leq 24\beta_1 \ell_\alpha \cdot \chi^{-1} \cdot (\theta'_\alpha)^{-2} \tag{A.336}$$

(alternatively, we may increase the size of w at the outset). Collecting (A.327), (A.330), (A.334), and (A.335), we obtain the following upper bound on the integral in (A.320):

$$P_{n,r,\boldsymbol{p}}((w-1, w+1]) \cdot 2w^{\gamma'}. \tag{A.337}$$

Further,

$$\begin{aligned}
P_{n,r,\boldsymbol{p}}((w-1, w+1])] &\leq P_{n,r,\boldsymbol{p}}((w-1, w]) + P_{n,r,\boldsymbol{p}}((w, \infty)) && \text{(A.338)}\\
&\leq Q((w-1, w]) + \frac{1}{2} - P_{n,r,\boldsymbol{p}}([0, w]) && \text{(A.339)}\\
&\leq Q((w-1, w]) + \frac{1}{2} - (Q([0, w]) - \varepsilon w) && \text{(A.340)}\\
&= \varepsilon w + Q((z, \infty)) && \text{(A.341)}
\end{aligned}$$

$$\le \varepsilon w + \frac{\eta}{12\beta_1 \ell_\alpha z^\alpha} \tag{A.342}$$

$$\le \frac{\eta}{6\beta_1 \ell_\alpha z^\alpha}, \tag{A.343}$$

where the last inequality follows by (A.286). Hence, the integral in (A.320) is upper bounded by

$$\frac{2^\alpha}{3\beta_1 \ell_\alpha w^{\alpha-\gamma'}} \cdot \eta \le \frac{\eta}{6}, \tag{A.344}$$

where the last inequality follows since $w \ge z_{\min}$ (see (A.253)). Thus, we have shown that (A.319) holds, which when combined with (A.318) gives (A.306).

Combining (A.293), (A.294), and (A.306) gives, in view of (A.292), the desired inequality (A.277):

$$\sup_{|a|\le 1} D(f_{n,r,\boldsymbol{p}} \| T_a f_{n,r,\boldsymbol{p}}) \le \mathrm{KL}^\star + \delta. \tag{A.345}$$

Recall that we showed in (A.276) that

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \le C + \eta. \tag{A.346}$$

To sum up, we make the dependence on C explicit in the optimal values, i.e., we write $\mathrm{KL}^\star(C)$, $\mathrm{KL}^\star_{n,N,r}(C)$, and $\mathrm{KL}^\star_{\mathrm{Cactus}}(C)$. What we have shown above yields that

$$\mathrm{KL}^\star_{n,N,r}(C+\eta) \le \mathrm{KL}^\star(C) + \delta. \tag{A.347}$$

Consider the values

$$\mathrm{KL}^\circ_{\mathrm{Cactus}}(C) := \inf_{(n,N,r)\in\mathbb{N}^2\times(0,1)} \mathrm{KL}^\star_{n,N,r}(C), \tag{A.348}$$

so (as defined by (2.83) in the statement of the theorem) $\mathrm{KL}^\star_{\mathrm{Cactus}}(C) = \lim_{\eta\to 0^+} \mathrm{KL}^\circ_{\mathrm{Cactus}}(C+\eta)$. We conclude that

$$\mathrm{KL}^\star(C+\eta) \le \mathrm{KL}^\circ_{\mathrm{Cactus}}(C+\eta) \le \mathrm{KL}^\star(C) + \delta. \tag{A.349}$$

Taking $\delta \to 0^+$, we have

$$\mathrm{KL}^\star(C+\eta) \le \mathrm{KL}^\circ_{\mathrm{Cactus}}(C+\eta) \le \mathrm{KL}^\star(C). \tag{A.350}$$

Finally, being the infimum of a jointly convex function over a convex set, the function $C \mapsto \mathrm{KL}^\star(C)$ is convex. Since it is also finite, we see that $\mathrm{KL}^\star(C)$ is continuous over $(0,\infty)$. Thus, taking $\eta \to 0^+$, we see that

$$\mathrm{KL}^\star_{\mathrm{Cactus}}(C) = \mathrm{KL}^\star(C), \tag{A.351}$$

completing the proof of the theorem.

A.6 Proofs of Section 2.13.1: Maximal Shifts and Worst Shifts

A.6.1 Proof of Lemma 2.1

Let $\widetilde{p} : \mathbb{R}_+ \to \mathbb{R}_+$ be such that $p(z) = \widetilde{p}(\|z\|)$ for all $z \in \mathbb{R}^m$. Fix $a \in \mathbb{R}^m$ with $0 < \|a\| \leq s$, and let $U \in \mathbb{R}^{m \times m}$ be an orthogonal matrix such that $a = \|a\| U \boldsymbol{e}_1$. Using the change of variables $y = U^T x$, we get that $a \mapsto \mathsf{E}_\gamma(p \,\|\, T_a p)$ is spherically symmetric:

$$\begin{aligned} \mathsf{E}_\gamma(p \,\|\, T_a p) &= \int_{\mathbb{R}^m} (p(x) - \gamma p(x-a))^+ \, dx && \text{(A.352)} \\ &= \int_{\mathbb{R}^m} (\widetilde{p}(\|x\|) - \gamma \widetilde{p}(\|x-a\|))^+ \, dx && \text{(A.353)} \\ &= \int_{\mathbb{R}^m} (\widetilde{p}(\|y\|) - \gamma \widetilde{p}(\|y - \|a\| \boldsymbol{e}_1\|))^+ \, dy && \text{(A.354)} \\ &= \mathsf{E}_\gamma(p \,\|\, T_{\|a\|\boldsymbol{e}_1} p). && \text{(A.355)} \end{aligned}$$

Next, we use (A.354) to show monotonicity in $\|a\|$. Using hyperspherical coordinates, and defining

$$I(r, \phi; \|a\|) := \widetilde{p}(r) - \gamma \widetilde{p}(\sqrt{r^2 - 2\|a\| r \cos\phi + \|a\|^2}), \tag{A.356}$$

we have that

$$\mathsf{E}_\gamma(p \,\|\, T_a p) \propto \int_0^\pi \int_0^\infty I(r, \phi; \|a\|)^+ r^{m-1} \, dr \, d\phi, \tag{A.357}$$

where the proportionality constant depends only on m. Now, for $r \geq 0$, we have $2r\cos\phi \geq \|a\|$ if and only if

$$r \geq \sqrt{r^2 - 2\|a\| r \cos\phi + \|a\|^2}. \tag{A.358}$$

In particular, as $\gamma \geq 1$, and as $\widetilde{p}$ is decreasing, we have that the integrand in (A.357) vanishes whenever $2r\cos\phi \geq \|a\|$. Now, writing

$$r^2 - 2\|a\| r \cos\phi + \|a\|^2 = (\|a\| - r\cos\phi)^2 + r^2 \sin^2\phi, \tag{A.359}$$

we see that this quadratic in $\|a\|$ is strictly increasing over the region $\|a\| \in (2r\cos\phi, \infty)$. As $\widetilde{p}$ is decreasing, we conclude that the mapping

$$\|a\| \mapsto \int_{\substack{(r,\phi) \in \mathbb{R}_+ \times (0,\pi) \\ 2r\cos\phi < \|a\|}} I(r, \phi; \|a\|)^+ r^{m-1} \, dr \, d\phi \tag{A.360}$$

is increasing. This completes the proof of the lemma.

A.6.2 Proof of Proposition 2.1

Subtracting $\gamma \cdot (t-1)$ from $f(t)$, where $\gamma \in \partial f(1)$ (the subdifferential of f at 1), we may assume that f is nonnegative. Let $f^*(t) = tf(1/t)$. Then,

$$D_f(p \parallel T_a p) = D_{f^*}(T_a p \parallel p) = D_f(T_a p \parallel p) = D_{f^*}(p \parallel T_a p). \tag{A.361}$$

Let $f_1(t) = f(t) \cdot 1_{(0,1)}(t)$ and $f_2(t) = f(t) \cdot 1_{(1,\infty)}(t)$, so $f = f_1 + f_2$ and

$$D_f(p \parallel T_a p) = \int_{\mathbb{R}^m} \widetilde{p}(\|x\|)(f_1 + f_2^*)\left(\frac{\widetilde{p}(\|x - \|a\|\boldsymbol{e}_1\|)}{\widetilde{p}(\|x\|)}\right) dx \tag{A.362}$$

where $p(x) = \widetilde{p}(\|x\|)$. Then, the proof is finished as in the end of the proof of Lemma 2.1. Namely, the integrand above is non-decreasing in $\|a\|$, which can be seen from the fact that $f_1 + f_2^*$ is non-increasing and that it vanishes over $[1, \infty)$.

A.7 Proof of Theorem 2.9

Denote $f = f_{n,r,\boldsymbol{p}}$ and $\widetilde{f} = \widetilde{f}_{n,r,\boldsymbol{p}}$ for short. Using hyperspherical coordinates, we have the KL-divergence

$$\begin{aligned} D(f \parallel T_{\boldsymbol{e}_1} f) &= \int_{\mathbb{R}^m} \widetilde{f}(\|x\|) \log \frac{\widetilde{f}(\|x\|)}{\widetilde{f}(\|x - \boldsymbol{e}_1\|)}\, dx & \text{(A.363)} \\ &= (m-1)V_{m-1} \int_0^\infty \int_0^\pi I(\rho, \phi)\, d\phi\, d\rho, & \text{(A.364)} \end{aligned}$$

where we denote the integrand

$$I(\rho, \phi) := \rho^{m-1} \sin^{m-2}(\phi) \widetilde{f}(\rho) \log \frac{\widetilde{f}(\rho)}{\widetilde{f}(\sqrt{\rho^2 - 2\rho\cos\phi + 1})}. \tag{A.365}$$

With $\theta = \sqrt{\rho^2 - 2\rho\cos\phi + 1}$ for fixed $\rho > 0$,

$$\int_0^\pi I(\rho, \phi)\, d\phi = 2^{m-3} \int_{\mathbb{R}_+} \theta\rho H(1, \rho, \theta)^{m-3} \widetilde{f}(\rho) \log \frac{\widetilde{f}(\rho)}{\widetilde{f}(\theta)}\, d\theta. \tag{A.366}$$

Therefore,

$$D(f \parallel T_{\boldsymbol{e}_1} f) = A_m \int_{\mathbb{R}_+^2} \theta\rho H(1, \rho, \theta)^{m-3} \widetilde{f}(\rho) \log \frac{\widetilde{f}(\rho)}{\widetilde{f}(\theta)}\, d\theta\, d\rho. \tag{A.367}$$

Using the partition $\mathbb{R}_+ = \bigcup_{i \ge 0} \mathcal{J}_{i,n}$, we get the objective function in (2.95). From Proposition 2.1, the function $a \mapsto D(f \parallel T_a f)$ is maximized over $\|a\| \le 1$ at $a = \boldsymbol{e}_1$.

Next, note that (2.96) is the probability constraint:

$$1 = \int_{\mathbb{R}^m} f(x)\,dx = \sum_{i\geq 0} \int_{\|x\|\in\mathcal{J}_{i,n}} f(x)\,dx = \sum_{i\geq 0} p_i v_{i,n}. \tag{A.368}$$

Finally, inequality (2.97) is the cost constraint, as can be seen by expanding $\mathbb{E}_{P_{n,r,\boldsymbol{p}}}[c] = \int_{\mathbb{R}^m} f(x)c(x)\,dx$ along the $\mathcal{J}_{i,n}$.

A.8 Proof of Theorem 2.10

The proof is divided into several steps:

• *Step 1: general setup.* Fix arbitrary $\theta_1 > 0$. Let $q^\star$ be a PDF of a monotone spherically-symmetric continuous additive mechanism achieving satisfying $\mathbb{E}_{q^\star}[c] \leq C$ and

$$D(q^\star \,\|\, T_{\boldsymbol{e}_1} q^\star) \leq \mathrm{KL}^\star_{\text{monotone}}(C) + \theta_1. \tag{A.369}$$

Fix arbitrary $\theta_2 > 0$. We will construct parameters $n, N \in \mathbb{N}$, $\boldsymbol{p} \in [0,1]^{N+1}$, and $r \in (0,1)$ such that $P_{n,r,\boldsymbol{p}} \in \mathscr{F}_{n,N,r}$ and

$$D(f_{n,r,\boldsymbol{p}} \,\|\, T_{\boldsymbol{e}_1} f_{n,r,\boldsymbol{p}}) \leq \mathrm{KL}^\star_{\text{monotone}}(C) + 2\theta_1, \tag{A.370}$$

$$\mathbb{E}_{f_{n,r,\boldsymbol{p}}}[c] \leq C + \theta_2. \tag{A.371}$$

Let $\tau := \frac{1}{2}\min(1,\alpha) \in (0, \frac{1}{2}]$. Define the PDF $\psi(x) := \exp\left(-\|x\|^\tau\right)/\chi$, where χ is the normalization constant. We will use ψ as a smoothing kernel. Let $\psi_\sigma(x) := \sigma^{-m}\psi(x/\sigma)$. Let $\sigma \in (0,1)$ be small enough so that $q = q^\star * \psi_\sigma$ satisfies

$$\mathbb{E}_q[c] \leq C + \frac{\theta_2}{2}. \tag{A.372}$$

Note that the data-processing inequality implies also that

$$D(q \,\|\, T_{\boldsymbol{e}_1} q) \leq \mathrm{KL}^\star_{\text{monotone}}(C) + \theta_1. \tag{A.373}$$

Let $\widetilde{q} : \mathbb{R}_+ \to \mathbb{R}_+$ be such that $q(x) = \widetilde{q}(\|x\|)$.

Consider $n, N \in \mathbb{N}$ and $T > 2$ (to be specified later in Step 3), and set $w := N/n$, $r := 1 - 1/(TN)$. Define $\boldsymbol{p}$ by

$$p_i := \inf_{\rho\in\mathcal{J}_{i,n}} \widetilde{q}(\rho) = \widetilde{q}\left(\frac{i+1}{n}\right), \quad 0 \leq i \leq N-1, \tag{A.374}$$

$$p_N := \frac{1}{\sum_{k\geq 0} r^k v_{N+k,n}} \cdot \left(1 - \sum_{i=0}^{N-1} p_i v_{i,n}\right). \tag{A.375}$$

Denote $f = f_{n,r,\boldsymbol{p}}$ for short, and let $\widetilde{f} : \mathbb{R}_+ \to \mathbb{R}_+$ be such that $f(x) = \widetilde{f}(\|x\|)$. We will also denote $P = P_{n,r,\boldsymbol{p}}$ for short.

- *Step 2: properties of q.* By Young's inequality, q is bounded:

$$\|q\|_\infty \leq \|q^\star\|_1 \|\psi_\sigma\|_\infty = \frac{1}{\sigma^m \chi} =: M. \tag{A.376}$$

The function ψ satisfies $\psi(x)/\psi(y) \leq \exp\left(\|x-y\|^\tau\right)$ for every $x, y \in \mathbb{R}^m$. Since q is a convolution with ψ_σ, it is not hard to see that, for any fixed $\theta \in (0, 2M \log 2)$ and $x, y \in \mathbb{R}^m$, we have that $|q(x) - q(y)| < \theta$ whenever the norms of x and y satisfy $|\|x\| - \|y\|| < \left(\frac{1}{2}\sigma^{m+\tau}\chi\theta\right)^{1/\tau}$. Next, we note that by finiteness of the KL-divergence the tail of q may be ignored when considering its KL-divergence. Let $\bar{w}_0 > 1$ be such that $w > \bar{w}_0$ implies

$$\int_{\|x\| \geq w-1} \left| q(x) \log \frac{q(x)}{q(x - \boldsymbol{e}_1)} \right| dx < \frac{\theta_1}{4}. \tag{A.377}$$

Since $\mathbb{E}_q[c] < \infty$ and c is a monotone radial function,

$$\widetilde{c}(w) \cdot Q\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) \leq \int_{\mathbb{R}^m \setminus \mathcal{B}(w)} c(x) q(x)\, dx \to 0 \tag{A.378}$$

as $w \to \infty$. As $\widetilde{c}(w) \sim \beta w^\alpha$ by assumption, we see that $w^\alpha Q(\mathbb{R}^m \setminus \mathcal{B}(w)) \to 0$ as $w \to \infty$. Denote the constants

$$B' := \frac{2^{m+\alpha+2} \cdot e^{2m+1} \cdot \beta \cdot \Gamma(m+\alpha) \cdot T^\alpha}{(m-1)!} \tag{A.379}$$

$$B := 2 \max\left(1, B'\right). \tag{A.380}$$

Let $\bar{w}_1 > 0$ be such that $w > \bar{w}_1$ implies

$$Q\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) < \frac{\min(\theta_1, \theta_2)}{B w^\alpha}. \tag{A.381}$$

Let $g(w) := m V_m w^{m-1} \widetilde{q}(w)$ denote the radial distribution, and $\mu(w) := g(w)/Q(\mathbb{R}^m \setminus \mathcal{B}(w))$ the hazard function. By assumption, we have that $\int_{\mathbb{R}_+} g(w) w^\alpha\, dw < \infty$. Note that $w \mapsto \log Q(\mathbb{R}^m \setminus \mathcal{B}(w))$ is locally absolutely continuous with derivative $-\mu(w)$. Therefore, for $w > 0$,

$$Q(\mathbb{R}^m \setminus \mathcal{B}(w)) = \exp\left(-\int_0^w \mu(x)\, dx\right) \tag{A.382}$$

Taking $w \to \infty$, we get $\int_{\mathbb{R}_+} \mu(w)\, dw = \infty$. In addition, μ is continuous. Therefore, by Lemma A.8

(shown below),

$$L := \limsup_{w\to\infty}\ w\mu(w) > 0. \tag{A.383}$$

Let $\{w_k\}_{k\in\mathbb{N}}$, $w_k \nearrow \infty$, be such that $w_k\mu(w_k) \geq L/2$ for each $k \in \mathbb{N}$. If other words, for each $w \in \{w_k\}_{k\in\mathbb{N}}$, we have

$$Q(\mathbb{R}^m \setminus \mathcal{B}(w)) \leq \frac{2mV_m}{L} \cdot \widetilde{q}(w)w^m. \tag{A.384}$$

Finally, since $\psi(x)/\psi(y) \leq \exp\left(\|x-y\|^\tau\right)$, we obtain that $\widetilde{q}(w) \geq \widetilde{q}(0)e^{-(w/\sigma)^\tau}$. In addition, it is not hard to show that (for $k \geq 2$ and $w' > 0$) $\int_{w'}^{\infty} v^{k-1}e^{-v}\,dv \geq (w')^{k-1}e^{-w'}$. Thus, we obtain the lower bound

$$Q(\mathbb{R}^m \setminus \mathcal{B}(w)) \geq mV_m\widetilde{q}(0)\frac{\sigma^\tau}{\tau}w^{m-\tau}e^{-(w/\sigma)^\tau}. \tag{A.385}$$

- *Step 3: choice of parameters.* Fix any $T > 2 + 2/L$ (see (A.383)). Recall that we set $\tau = \frac{1}{2}\min(1,\alpha)$, which satisfies $\tau \in (0, \frac{1}{2}]$ and $\tau < \alpha$. Denote the constant $\tau' := \frac{\tau+\alpha}{2} \in (\tau, \alpha)$. Let $\varpi_2 > 0$ be such that $w > \varpi_2$ implies

$$\frac{2^{m+1}eT^m\tau}{\sigma^\tau} \cdot w^\tau \exp\left(\left(\frac{w}{\sigma}\right)^\tau\right) \leq \exp\left(w^{\tau'}\right). \tag{A.386}$$

Denote the constant $\varpi_3 = \frac{20eA_m}{mV_m\theta_1}$. Let $\varpi_4 > 0$ be such that $w > \varpi_4$ implies $\widetilde{c}(w) < 2\beta w^\alpha$. Let $k^* \in \mathbb{N}$ be such that $w_k > \max(\varpi_0, \varpi_1, \varpi_2, \varpi_3, \varpi_4, \sigma^{-\tau/\tau'}, 2)$ for all $k \geq k^\star$. Denote $w = w_{k^\star}$. Consider any constant $\theta_3 \in (0, 2M\log 2)$ satisfying

$$\theta_3 < \min\left(\frac{\widetilde{q}(w)}{2}, \frac{\theta_1\widetilde{q}(w)}{12MV_mw^m}, \frac{\theta_1}{2V_mw^{m+\tau'}}, \frac{\theta_2}{BV_mw^{m+\alpha}}\right). \tag{A.387}$$

With $n_0 := \left(\frac{1}{2}\sigma^{m+\tau}\chi\theta_3\right)^{-\frac{1}{\tau}}$, fix $n > \max(n_0, m)$, $N = wn$.

- *Step 4: monotonicity of f.* We only need to show $p_{N-1} \geq p_N$. Since we are choosing $n > n_0$, the continuity of q shown in Step 2 yields that $f \geq q - \theta_3$ over $\mathcal{B}(w)$. Therefore,

$$P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) \leq Q\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) + \theta_3V_mw^m. \tag{A.388}$$

As $w = w_{k^\star}$, the bound on the hazard function (A.384) yields

$$Q(\mathbb{R}^m \setminus \mathcal{B}(w)) \leq \frac{2mV_m}{L} \cdot \widetilde{q}(w)w^m. \tag{A.389}$$

In addition, our choice of θ_3 in (A.387) yields that

$$\theta_3V_mw^m < \frac{V_m}{2} \cdot \widetilde{q}(w)w^m. \tag{A.390}$$

By our choice of T in Step 3, we infer the tail bound

$$P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) < mV_m T \cdot \widetilde{q}(w) w^m. \tag{A.391}$$

Using $r = 1 - 1/(TN)$, it is elementary to derive the bound

$$\sum_{k \geq 0} r^k v_{N+k,n} \geq mV_m T w^m. \tag{A.392}$$

Combining these bounds, by definition of p_N, we get

$$p_N < \widetilde{q}(w) = p_{N-1}, \tag{A.393}$$

- Step 5: bounding the cost. As $f \leq q$ over $\mathcal{B}(w)$,

$$\mathbb{E}_f[c] \leq C + \frac{\theta_2}{2} + \mathbb{E}_f[c \cdot 1_{\mathbb{R}^m \setminus \mathcal{B}(w)}], \tag{A.394}$$

As $w > \bar{w}_4$, we have $\widetilde{c}(u) < 2\beta u^\alpha$ for $u > w$. Hence,

$$\begin{aligned}\mathbb{E}_f[c \cdot 1_{\mathbb{R}^m \setminus \mathcal{B}(w)}] &\leq 2mV_m \beta p_N \sum_{k \geq 0} r^k \int_{\mathcal{J}_{N+k,n}} \rho^{m+\alpha-1}\, d\rho \\ &\leq \frac{2emV_m \beta p_N}{n^{m+\alpha}} \sum_{k \geq 0} r^k (N+k)^{m+\alpha-1},\end{aligned}$$

where the second inequality follows from

$$(N+k+1)^\ell - (N+k)^\ell \leq e\ell(N+k)^{\ell-1} \tag{A.395}$$

for $k \geq 0$ and $\ell > 1$. By Lemma A.7, we have the upper bound

$$\sum_{k \geq 0} r^k (N+k)^{m+\alpha-1} \leq 2^{m+\alpha-1} \Gamma(m+\alpha) \cdot (TN)^{m+\alpha}. \tag{A.396}$$

In addition, refining (A.392) using Lemma A.7, we obtain

$$\sum_{k \geq 0} r^k v_{N+k,n} \geq \frac{m! \cdot V_m}{2e^{2m}} \cdot (Tw)^m \tag{A.397}$$

Combining these bounds, and recalling the definition of the constant B in (A.380), we obtain that

$$\mathbb{E}_f[c \cdot 1_{\mathbb{R}^m \setminus \mathcal{B}(w)}] \leq \frac{B}{4} \cdot P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) w^\alpha. \tag{A.398}$$

Now, by our choice of θ_3 (see (A.387)), $\theta_3 V_m w^m < \frac{\theta_2}{Bw^\alpha}$. As $w > \bar{w}_1$, the tail bound on Q in (A.381) holds. Thus, recalling the tail bound on P given by (A.388), we get $P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) \leq \frac{2\theta_2}{Bw^\alpha}$. Plugging this into (A.398), we get $\mathbb{E}_f[c \cdot 1_{\mathbb{R}^m \setminus \mathcal{B}(w)}] \le \frac{\theta_2}{2}$. Hence, from (A.394), we conclude that

$$\mathbb{E}_f[c] \le C + \theta_2. \tag{A.399}$$

• *Step 6: bounding the KL-divergence.* Next, we upper bound the difference of KL-divergences $D(f \,\|\, T_{e_1} f) - D(q \,\|\, T_{e_1} q)$ by the arbitrarily small quantity θ_1. We split the integral giving the difference of these KL-divergences to consider integrating separately over the three regions $\{x : \|x\| \le w-1\}$, $\{x : w-1 < \|x\| < w+1\}$, and $\{x : \|x\| \ge w+1\}$.

○ *Step 6.i: bounding the tail.* As $\gamma_{i,j,n} = 0$ for $|i-j| > n$,

$$\int_{\|x\| \ge w+1} f(x) \log \frac{f(x)}{f(x - \boldsymbol{e}_1)}\, dx \le A_m p_N \log \frac{1}{r} \sum_{i \ge N+n} \sum_{k=1}^{n} \gamma_{i,i+k,n}\, r^{i-N} k. \tag{A.400}$$

Now, note that for $\rho, \theta \ge 1$, we have $H(1, \rho, \theta) \le \frac{1}{2} \min(\rho, \theta)$. Therefore, for $i \ge N + n$ and $k \ge 1$, we have that

$$\gamma_{i,i+k,n} = \int_{\mathcal{J}_{i,n}} \int_{\mathcal{J}_{i+k,n}} \theta\rho \cdot H(1,\rho,\theta)^{m-3}\, d\theta\, d\rho \tag{A.401}$$

$$\le \frac{(2(i+k)+1)\left((i+1)^{m-1} - i^{m-1}\right)}{2^{m-2}(m-1)n^{m+1}}. \tag{A.402}$$

In addition, from (A.395) we see that

$$\gamma_{i,i+k,n} \le \frac{e(2(i+k)+1)i^{m-2}}{2^{m-2}n^{m+1}}. \tag{A.403}$$

Combining (A.400)–(A.403) and using $r = 1 - 1/(TN)$, it is elementary to obtain

$$\int_{\|x\| \ge w+1} f(x) \log \frac{f(x)}{f(x - \boldsymbol{e}_1)}\, dx \le \frac{eA_m p_N \log \frac{1}{r}}{2^{m-2}n^{m+1}} \sum_{i \ge N+n} \sum_{k=1}^{n} (2(i+k)+1) i^{m-2} r^{i-N} k$$

$$\le \frac{10eA_m}{mV_m} \cdot \frac{1}{Tw} \cdot p_N \sum_{\ell \ge 0} r^{\ell} v_{N+\ell,n} \tag{A.404}$$

$$= \frac{10eA_m}{mV_m} \cdot \frac{1}{Tw} \cdot P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) \le \frac{5eA_m/(mV_m)}{w}. \tag{A.405}$$

Recall that we choose $w > \bar{w}_3 = 20eA_m/(mV_m\theta_1)$. Therefore, we obtain from (A.405) the following tail bound

$$\int_{\|x\| \ge w+1} f(x) \log \frac{f(x)}{f(x - \boldsymbol{e}_1)}\, dx < \frac{\theta_1}{4}. \tag{A.406}$$

○ *Step 6.ii: bounding the approximation region.* Consider the innermost region $\{x : \|x\| \le w - 1\}$. By construction of f, we have $f(x) \le q(x)$ whenever $\|x\| \le w$. Since $n > n_0$ (see Step 3), the continuity of q shown in Step 2 yields that $f(x) \ge q(x) - \theta_3$ whenever $\|x\| \le w$. Also, as $q - \theta_3 \le f \le q$ over

$\mathcal{B}(w)$, and as $\theta_3 \leq q/2$ over $\mathcal{B}(w)$, we get that $0 < f \leq q \leq 2f$ and $0 < T_{e_1} f \leq T_{e_1} q \leq 2T_{e_1} f$ over $\mathcal{B}(w-1)$. It is not hard to derive the following elementary bound: for any numbers $0 < \phi \leq \kappa$ and $0 < \phi' \leq \kappa'$, denoting $\gamma = \kappa - \phi \geq 0$ and $\gamma' = \kappa' - \phi' \geq 0$, if $2\gamma \leq \kappa$ and $2\gamma' \leq \kappa'$ (i.e., $\kappa \leq 2\phi$ and $\kappa' \leq 2\phi'$), then we have

$$\phi \log \frac{\phi}{\phi'} - \kappa \log \frac{\kappa}{\kappa'} \leq \frac{2\kappa\gamma'}{\kappa'} + \frac{2\gamma^2}{\kappa} + \gamma \log \frac{\kappa'}{\kappa}. \tag{A.407}$$

Applying this elementary bound on the integrand

$$I := f \log \frac{f}{T_{e_1} f} - q \log \frac{q}{T_{e_1} q} \tag{A.408}$$

(with f in place of ϕ and q in place of κ) we obtain the bound

$$\sup_{x \in \mathcal{B}(w-1)} I(x) \leq \frac{2M\theta_3}{\widetilde{q}(w)} + \frac{2\theta_3^2}{\widetilde{q}(w)} + \theta_3 \log \frac{M}{\widetilde{q}(w)}. \tag{A.409}$$

Since $\theta_3 < \widetilde{q}(w)/2$ and $\log u \leq u - 1$, we infer

$$\sup_{x \in \mathcal{B}(w-1)} I(x) < \frac{3M\theta_3}{\widetilde{q}(w)} < \frac{\theta_1}{4V_m w^m}. \tag{A.410}$$

As this is a uniform bound, we conclude that

$$\int_{\|x\| \leq w-1} \left(f \log \frac{f}{T_{e_1} f} - q \log \frac{q}{T_{e_1} q} \right)(x)\, dx < \frac{\theta_1}{4}. \tag{A.411}$$

○ *Step 6.iii: bounding the intersection shell.* Consider the region $\{x : w-1 < \|x\| < w+1\}$. By monotonicity of f,

$$\sup_{w-1<\|x\|<w+1} \frac{f(x)}{f(x-\boldsymbol{e}_1)} \leq \frac{\widetilde{q}(w-1)}{p_N r^n} \leq \frac{4\widetilde{q}(0)}{p_N}. \tag{A.412}$$

Next, we derive a lower bound on p_N. By Lemma A.7,

$$\sum_{k \geq 0} r^k v_{N+k,n} \leq 2^{m-1} e(m!) V_m T^m \cdot w^m. \tag{A.413}$$

As $f \leq q$ over $\mathcal{B}(w)$, we get $P(\mathbb{R}^m \setminus \mathcal{B}(w)) \geq Q(\mathbb{R}^m \setminus \mathcal{B}(w))$. Combining (A.385), (A.386) (as $w > \bar{w}_2$), and (A.413), we deduce that

$$p_N = \frac{P(\mathbb{R}^m \setminus \mathcal{B}(w))}{\sum_{k\geq 0} r^k v_{N+k,n}} \geq 4\widetilde{q}(0) \exp\left(-w^{\tau'}\right). \tag{A.414}$$

Plugging this into (A.412) then integrating, we get

$$\int_{w-1<\|x\|<w+1} f(x) \log \frac{f(x)}{f(x-\boldsymbol{e}_1)}\, dx \leq w^{\tau'} P(\mathbb{R}^m \setminus \mathcal{B}(w)). \tag{A.415}$$

Finally, recalling the upper bound (A.388) on P, the bound $Q\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) < \frac{\theta_1}{2w^\alpha} < \frac{\theta_1}{2w^{\tau'}}$ (from (A.381) since $w > \bar{w}_1$), and that $\theta_3 V_m w^m < \frac{\theta_1}{2m^{\tau'}}$ by choice of θ_3 (see (A.387)), we infer $P\left(\mathbb{R}^m \setminus \mathcal{B}(w)\right) < \theta_1 / w^{\tau'}$. Plugging into (A.415), we get

$$\int_{w-1<\|x\|<w+1} f(x) \log \frac{f(x)}{f(x-\boldsymbol{e}_1)}\, dx < \frac{\theta_1}{4}. \tag{A.416}$$

• *Step 7: conclusion of proof.* Combining the inequalities shown in (A.377), (A.399), (A.406), (A.411), and (A.416), we get that f satisfies $\mathbb{E}_f[c] \le C + \theta_2$ and $D(f \,\|\, T_{\boldsymbol{e}_1} f) - D(q \,\|\, T_{\boldsymbol{e}_1} q) \le \theta_1$. Therefore, we have shown that

$$\mathrm{KL}^\star_{n,N,r}(C+\theta_2) \le \mathrm{KL}^\star_{\text{monotone}}(C) + 2\theta_1. \tag{A.417}$$

Define

$$\widetilde{\mathrm{KL}}(C) := \inf_{(n,N,r)\in\mathbb{N}^2\times(0,1)} \mathrm{KL}^\star_{n,N,r}(C), \tag{A.418}$$

so $\mathrm{KL}^\star_{\text{isotropic}}(C) = \lim_{\theta_2\to 0^+} \widetilde{\mathrm{KL}}(C+\theta_2)$ as defined in (2.98). Then, we have that

$$\mathrm{KL}^\star_{\text{monotone}}(C+\theta_2) \le \widetilde{\mathrm{KL}}(C+\theta_2) \le \mathrm{KL}^\star_{\text{monotone}}(C) + 2\theta_1. \tag{A.419}$$

Note that $C \mapsto \mathrm{KL}^\star(C)$ is continuous over $(0,\infty)$, since it is finite and convex there (indeed, it is the infimum of a convex function over a convex set). Taking $\theta_1 \to 0^+$ then $\theta_2 \to 0^+$, we conclude that $\mathrm{KL}^\star_{\text{isotropic}} = \mathrm{KL}^\star_{\text{monotone}}$, and the proof is complete.

Lemma A.7. *Let* $\mathrm{Li}_{-u}(r) := \sum_{k\ge 1} k^u r^k$ *be the polylogarithm function. For each* $r \in (e^{-e/4}, 1)$ *and* $u \ge 1$, *we have*

$$\frac{1}{2e^{2(u+1)}} \le \frac{(1-r)^{u+1}}{\Gamma(u+1)} \cdot \mathrm{Li}_{-u}(r) \le \frac{3}{2}. \tag{A.420}$$

Proof. Let $f(z) := z^u r^z$, so $\mathrm{Li}_{-u} = \sum_{k\ge1} f(k)$. Then, f increases over $[0, z^\star]$ and decreases over $[z^\star, \infty)$, where $z^\star = u/\log\frac{1}{r}$. Therefore, by non-negativity of f over $\mathbb{R}_+$, we have

$$\left| \mathrm{Li}_{-u}(r) - \int_{\mathbb{R}_+} f(z)\, dz \right| \le 2f(z^\star). \tag{A.421}$$

The proof is completed by computing the integral of f and applying simple bounds on $f(z^\star)$. □

Lemma A.8. *With* $\mathbb{R}_+ = [0,\infty)$, *let* $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ *be a continuous nonnegative function with* $\int_{\mathbb{R}_+} \mu(w)\, dw = \infty$ *and*

$$\int_{\mathbb{R}_+} \frac{w^\alpha \mu(w)}{\exp\left(\int_0^w \mu(x)\, dx\right)}\, dw < \infty, \tag{A.422}$$

for some $\alpha > 0$. *Then,* $\limsup_{w\to\infty}\, w\mu(w) > 0$.

Proof. By continuity and non-integrability, $\int_A^\infty \mu(w)\,dw = \infty$ for every $A > 0$. This immediately implies, by (A.422), that

$$\liminf_{w\to\infty} \frac{w^\alpha}{\exp\left(\int_0^w \mu(x)\,dx\right)} = 0. \tag{A.423}$$

Fix $\{w_i\}_{i\in\mathbb{N}}$, $w_i \nearrow \infty$, with $w_i^\alpha \exp\left(-\int_0^{w_i} \mu(x)\,dx\right) \to 0$ as $i \to \infty$. Then, for all large i, we have that $\int_0^{w_i} \mu(x)\,dx / \log w_i \geq \alpha$. Suppose for the sake of contradiction that $\lim_{w\to\infty} w\mu(w) = 0$. Then, by l'Hôpital's rule,

$$\lim_{w\to\infty} \frac{\int_0^w \mu(x)\,dx}{\log w} = \lim_{w\to\infty} w\mu(w) = 0, \tag{A.424}$$

contradicting the existence of the sequence $\{w_i\}_{i\in\mathbb{N}}$. □

A.9 Auxiliary Results for Minimzing the Fisher Information

We prove in this appendix Lemma 2.2 and Proposition 2.2, and we also introduce and prove the following lemma, which will be useful in the proof of Theorem 2.14 in the next appendix.

Lemma A.9. *With $\mathcal{P}_0 \subset \mathcal{P}$ denoting the set of strictly positive PDFs, we have that*

$$\inf_{\substack{p\in\mathcal{P}_0 \\ \mathbb{E}_p[c]\leq C}} I(p) = \inf_{\substack{p\in\mathcal{P} \\ \mathbb{E}_p[c]\leq C}} I(p). \tag{A.425}$$

Proof. See Appendix A.9.3. □

A.9.1 Proof of Lemma 2.2

By [BS91, Chapter 2, Theorems 3.1 and 3.5], there is a minimal eigenvalue E_0 of $\mathcal{H}_{\theta c}$, which corresponds to a 1-dimensional eigenspace $\{\gamma y\}_{\gamma\in\mathbb{R}} \subset L^2(\mathbb{R})$ where $y \in L^2(\mathbb{R})$ has no zeros. Then, there is a unique $\gamma \in \mathbb{R}$ such that $\|\gamma y\|_2 = 1$ and $\gamma y(x) > 0$ for all $x \in \mathbb{R}$, namely, $\gamma := \operatorname{sgn}(y(0))/\|y\|_2$. Setting $y_{\theta,c} = \gamma y$ yields the desired uniqueness result. Further, this uniqueness yields that $y_{\theta,c}$ is even since $y_{\theta,c}(-x)$ also satisfies the same differential equation, so a normalized version of $y_{\theta,c}(x) + y_{\theta,c}(-x)$ does too.

A.9.2 Proof of Proposition 2.2

We will use the following asymptotic of $y_{\theta,c}$.

Theorem A.1 ([BS91, Chapter 2, Theorem 4.6]). *Fix $\theta > 0$, and let E_0 be the eigenvalue associated with $y_{\theta,c}$. As $x_1, x - x_1 \to \infty$ or $x_1, x - x_1 \to -\infty$, we have the asymptotic*

$$y_{\theta,c}(x) \sim \frac{\exp\left(-\int_{x_1}^{x} \sqrt{\theta c(t) - E_0}\, dt\right)}{(\theta c(x))^{1/4}}. \tag{A.426}$$

We denote $y = y_{\theta,c}$ for readability. Denote $f = -y'/y$ and $g = \theta c - E_0$, and note that f satisfies the Riccati equation

$$-f' + f^2 = g. \tag{A.427}$$

With this notation, the eigenvalue equation for y is $y'' = gy$. Since c grows without bound, g is eventually strictly positive. Since y is strictly positive and $y'' = gy$, we conclude that y'' is eventually positive a.e., i.e., there is an N such that $\lambda(\{x \in (N, \infty)\ ;\ y''(x) < 0\}) = 0$. Since y' is absolutely continuous,

$$y'(t_1) - y'(t_2) = \int_{t_2}^{t_1} y''(t)\, dt \geq 0 \tag{A.428}$$

for all large t_1 and t_2 with $t_1 > t_2$, i.e., y' is eventually increasing. As y decays to zero at infinity, and as y' eventually increases, we infer that y' is eventually negative. Thus, f is eventually positive. We will show that, for all large x,

$$f(x) \leq \sqrt{2g(x)}, \tag{A.429}$$

which is equivalent to

$$\left|\frac{y'(x)}{y(x)}\right| \leq \sqrt{2\left(\theta c(x) - E_0\right)}. \tag{A.430}$$

This is enough to finish the proof of the lemma by evenness of y and c. Now, we show that (A.429) holds.

Set $h = \sqrt{2g}$, so we want to show that $f \leq h$ is eventually satisfied. Denote $z = f - h$. Differentiating and using $-f' + f^2 = g$ (see (A.427)), we obtain

$$-z' + z^2 + 2zh = h' - g. \tag{A.431}$$

Now, we note that $h' < g$ eventually holds. Indeed, as $x \to \infty$,

$$\frac{h'(x)}{g(x)} = \frac{1}{\sqrt{2}} \frac{g'(x)}{g(x)^{3/2}} \propto \frac{c'(x)}{c(x)^{3/2}} \to 0 \tag{A.432}$$

by assumption on c. Thus, by (A.431), we eventually have $-z' < 0$, i.e., z is strictly increasing over (x_0, ∞) for some $x_0 > 0$.

Suppose, for the sake of contradiction, that there is an $x_1 > x_0$ such that $f(x_1) > h(x_1)$, i.e.,

$z(x_1) > 0$. Then, as z is strictly increasing over (x_0, ∞), we have that $z(x) > 0$ for all $x \geq x_1$. In other words,

$$-\frac{y'(x)}{y(x)} > \sqrt{2g(x)} \tag{A.433}$$

for all $x \geq x_1$. Increase x_1 if necessary so that $y(x) < 1$ for $x \geq x_1$. Then,

$$y(x) \leq \exp\left(-\int_{x_2}^{x} \sqrt{2g(t)}\, dt\right) \tag{A.434}$$

for all $x > x_2 \geq x_1$. Let x_3 and x_4 satisfying $x_4 > x_3 > x_1$ be such that

$$y(x) \geq \frac{1}{2g(x)^{1/4}} \exp\left(-\int_{x_3}^{x} \sqrt{g(t)}\, dt\right) \tag{A.435}$$

for every $x > x_4$. Then, for all $x > x_4$,

$$(\sqrt{2} - 1) \int_{x_3}^{x} \sqrt{g(t)}\, dt \leq \log\left(2g(x)^{1/4}\right). \tag{A.436}$$

Denote

$$w(t) = \sqrt{(\sqrt{2} - 1)\sqrt{g(t)}}, \tag{A.437}$$

so (A.436) can be rewritten as

$$\frac{\int_{x_3}^{x} w(t)^2\, dt}{\log\left(\gamma \cdot w(x)\right)} \leq 1, \tag{A.438}$$

where $\gamma := 2\sqrt{1 + \sqrt{2}}$ is an absolute constant. To arrive at a contradiction, we take $x \to \infty$ and use L'Hôpital's rule:

$$\lim_{x\to\infty} \frac{\int_{x_3}^{x} w(t)^2\, dt}{\log\left(\gamma \cdot w(x)\right)} = \lim_{x\to\infty} \frac{w(x)^3}{w'(x)} = \infty. \tag{A.439}$$

To see the last limit diverges, note that

$$\frac{w(x)^3}{w'(x)} \propto \frac{c(x)^{3/2}}{c'(x)} \to \infty. \tag{A.440}$$

The limit in (A.439) contradicts inequality (A.438). Thus, there is no $x_1 > x_0$ such that $f(x_1) > h(x_1)$. Hence, (A.429) eventually holds, and the proof is complete.

In the course of this proof of Proposition 2.2, we have shown the following useful property of $y_{\theta,c}$ that will be used later.

Lemma A.10. *For c satisfying Assumption 2.2 and any $\theta > 0$, the function $y_{\theta,c}$ is eventually decreasing.*

A.9.3 Proof of Lemma A.9

For each $p \in \mathcal{P}$ and $\sigma > 0$, denote $p^{\sigma}(x) = p(x/\sigma)/\sigma$. Let ϕ denote the Gaussian density $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$.

We begin by noting that the limit

$$\lim_{\sigma \to 0^+} \mathbb{E}_{p*\phi^{\sigma}}[c] = \mathbb{E}_p[c] \tag{A.441}$$

holds for every PDF p that satisfies $\mathbb{E}_p[c] < \infty$. This limit can be proved in the same way Lemma A.2 is proved. Indeed, by the assumed additive and multiplicative regularity of c, it is not hard to see that, for the random variables $Z_{\sigma} \sim p * \phi^{\sigma}$, the set $\{c(Z_{\sigma})\}_{0<\sigma\leq 1}$ is uniformly bounded by an integrable random variable. In particular, the set $\{c(Z_{\sigma})\}_{0<\sigma\leq 1}$ is uniformly integrable, so the Lebesgue-Vitali theorem [Bog07, Theorem 4.5.4] yields that the limit (A.441) holds.

Now, we show that the function $I_0^{\star} : \mathbb{R} \to [0,\infty]$ defined by

$$I_0^{\star}(C) := \inf_{\substack{p \in \mathcal{P}_0 \\ \mathbb{E}_p[c] \leq C}} I(p) \tag{A.442}$$

is continuous at C. We may write

$$I_0^{\star}(C) = \inf_{p \in \mathcal{P}_0} I(p) + \mathbb{I}_{(-\infty,C]}\left(\mathbb{E}_p[c]\right). \tag{A.443}$$

Being the infimum of a jointly convex function over a convex set, $I_0^{\star}$ is convex. Further, this function is finite over $(0,\infty)$. To see that $I_0^{\star}(C)$ is finite, we only need to take p the Gaussian PDF with small enough variance. Hence, being convex and finite, $I_0^{\star}$ is continuous over $(0,\infty)$.

Define

$$I^{\star}(C) := \inf_{\substack{p \in \mathcal{P} \\ \mathbb{E}_p[c] \leq C}} I(p) \tag{A.444}$$

Now, fix $\varepsilon, \eta > 0$, and let $p \in \mathcal{P}$ be a PDF such that $\mathbb{E}_p[c] \leq C$ and

$$I(p) \leq I^{\star}(C) + \varepsilon. \tag{A.445}$$

Since the Fisher information satisfies the convolution inequality, we have

$$I(p * \phi^{\sigma}) \leq I(p) \tag{A.446}$$

for every $\sigma > 0$. By the limit in (A.441), there is a $\sigma = \sigma(\eta)$ such that

$$\mathbb{E}_{p*\phi^\sigma}[c] \le \mathbb{E}_p[c] + \eta \le C + \eta. \tag{A.447}$$

Note that $p * \phi^\sigma \in \mathcal{P}_0$ by strict positivity of ϕ. Therefore,

$$I_0^\star(C+\eta) \le I(p*\phi^\sigma) \le I(p) \le I^\star(C) + \varepsilon. \tag{A.448}$$

By continuity of $I_0^\star$ at C, we may take $\eta \to 0^+$ to obtain

$$I_0^\star(C) \le I^\star(C) + \varepsilon. \tag{A.449}$$

By arbitrariness of ε, we deduce

$$I_0^\star(C) \le I^\star(C). \tag{A.450}$$

But the reverse inequality is trivial, thus equality is attained in (A.450), completing the proof of the lemma.

A.10 Proof of Theorem 2.13

We use the integration shorthand

$$\int_A f := \int_A f(x)\,dx. \tag{A.451}$$

We will let $L^2(\mathbb{R}, c)$ denote the space of Lebesgue measurable functions $f : \mathbb{R} \to \mathbb{R}$ such that $\|f\|_{2,c} := (\int_{\mathbb{R}} f^2 c)^{1/2} < \infty$. First, we note that $C = \|y\|_{2,c}^2$ is indeed finite as can be deduced from the expansion of y in Theorem A.1.

Denote the space of absolutely continuous functions on $\mathbb{R}$ by $\mathrm{AC}(\mathbb{R})$, and those that are locally absolutely continuous over $\mathbb{R}$ by $\mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Consider the vector space

$$V := L^2(\mathbb{R}) \cap L^2(\mathbb{R}, c) \cap \mathrm{AC}_{\mathrm{loc}}(\mathbb{R}). \tag{A.452}$$

Let E be the eigenvalue of y, so

$$y'' = (\theta c - E)y. \tag{A.453}$$

Consider the modified Dirichlet energy $\mathcal{E} : V \to \mathbb{R} \cup \{\infty\}$ defined by

$$\mathcal{E}(w) := \|w'\|_2^2 + \theta\|w\|_{2,c}^2 - E\|w\|_2^2. \tag{A.454}$$

We start by showing that y is a global minimizer of $\mathcal{E}$, and

$$0 = \mathcal{E}(y) = \inf_{w \in V} \mathcal{E}(w). \tag{A.455}$$

Note that $y \in V$ since $y \in \mathcal{C}^1(\mathbb{R})$.

Fix an arbitrary $w \in V$, and we will show that $\mathcal{E}(w) \geq 0$. Since w is a.e. differentiable, we have $(y \cdot (w/y)')^2 \geq 0$ a.e. Rearranging this inequality, and noting the eigenvalue equation (A.453) satisfied by y, we obtain that a.e.

$$\begin{aligned}
(w')^2 &\geq \frac{2y'ww'}{y} - \frac{(y')^2w^2}{y^2} && \text{(A.456)}\\
&= \left(\frac{y'w^2}{y}\right)' - \frac{y''w^2}{y} && \text{(A.457)}\\
&= \left(\frac{y'w^2}{y}\right)' - (\theta c - E)w^2. && \text{(A.458)}
\end{aligned}$$

Note that $y'w^2/y \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Thus, integrating (A.458) over any $[-t,t]$ with $t > 0$, we obtain

$$\|w'1_{[-t,t]}\|_2^2 \geq \left.\frac{y'w^2}{y}\right|_{-t}^{t} - \theta\|w1_{[-t,t]}\|_{2,c}^2 + E\|w1_{[-t,t]}\|_2^2. \tag{A.459}$$

Next, we show that there exists a sequence $t_n \nearrow \infty$ such that

$$\liminf_{n\to\infty} \left.\frac{y'w^2}{y}\right|_{-t_n}^{t_n} \geq 0. \tag{A.460}$$

This would readily yield $\mathcal{E}(w) \geq 0$ from inequality (A.459). By assumption, $w \in L^2(\mathbb{R}, c)$, so symmetry of c implies

$$\int_0^\infty (w(x)^2 + w(-x)^2)c(x)\,dx = \int_{\mathbb{R}} w^2 c < \infty. \tag{A.461}$$

In particular, there is a sequence $\{t_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ such that, as $n \to \infty$, we have $t_n \nearrow \infty$ and

$$\left(w(t_n)^2 + w(-t_n)^2\right) c(t_n) \to 0. \tag{A.462}$$

In addition, by the upper bound (2.114) in Proposition 2.2, there is an $A \in (0,\infty)$ such that

$$\left|\frac{y'(x)}{y(x)}\right| \leq A \cdot c(|x|) \tag{A.463}$$

holds for all large $|x|$. Then, for all large n,

$$\left.\frac{y'w^2}{y}\right|_{-t_n}^{t_n} = \frac{y'(t_n)w(t_n)^2}{y(t_n)} - \frac{y'(-t_n)w(t_n)^2}{y(-t_n)} \tag{A.464}$$

$$\geq -\left|\frac{y'(t_n)}{y(t_n)}\right| w(t_n)^2 - \left|\frac{y'(-t_n)}{y(-t_n)}\right| w(-t_n)^2 \tag{A.465}$$

$$\geq -Ac(t_n)\left(w(t_n)^2 + w(-t_n)^2\right). \tag{A.466}$$

Taking the limit inferior in (A.466) we obtain, in view of (A.462), that

$$\liminf_{n\to\infty} \left.\frac{y'w^2}{y}\right|_{-t_n}^{t_n} \geq 0. \tag{A.467}$$

In addition, by the assumption that $w \in L^2(\mathbb{R})$, the monotone convergence theorem implies

$$\lim_{n\to\infty} \theta\|w1_{[-t_n,t_n]}\|_{2,c}^2 - E\|w1_{[-t_n,t_n]}\|_2^2 = \theta\|w\|_{2,c}^2 - E\|w\|_2^2. \tag{A.468}$$

Taking the limit inferior of (A.459) along the t_n, and using (A.467) and (A.468) we conclude that

$$\|w'\|_2^2 \geq -\theta\|w\|_{2,c}^2 + E\|w\|_2^2. \tag{A.469}$$

As $w \in L^2(\mathbb{R}, c) \cap L^2(\mathbb{R})$, (A.469) is equivalent to $\mathcal{E}(w) \geq 0$.

We have just shown that

$$\inf_{w\in V} \mathcal{E}(w) \geq 0. \tag{A.470}$$

On the other hand, we may show that $\mathcal{E}(y) = 0$. Indeed, as $y \in \mathcal{C}^1(\mathbb{R})$ and $y' \in \mathrm{AC}(\mathbb{R})$, we have that $yy' \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Note that $yy' = O(y^2c)$ by the upper bound in Proposition 2.2. As $y \in L^2(\mathbb{R}, c)$, we get $yy' \in L^1(\mathbb{R})$. Thus, there exist sequences $a_n, b_n \nearrow \infty$ such that $y(-a_n)y'(-a_n), y(b_n)y'(b_n) \to 0$. Therefore, we have that

$$\begin{aligned}
\mathcal{E}(y) &= \|y'\|_2^2 + \theta\|y\|_{2,c}^2 - E\|y\|_2^2 && \text{(A.471)}\\
&= \lim_{n\to\infty} \|y'1_{[-a_n,b_n]}\|_2^2 + \theta\|y1_{[-a_n,b_n]}\|_{2,c}^2 - E\|y1_{[-a_n,b_n]}\|_2^2 && \text{(A.472)}\\
&= \lim_{n\to\infty} \int_{-a_n}^{b_n} (y')^2 + (\theta c - E)y^2 && \text{(A.473)}\\
&= \lim_{n\to\infty} \int_{-a_n}^{b_n} (y')^2 + yy'' && \text{(A.474)}\\
&= \lim_{n\to\infty} \int_{-a_n}^{b_n} (yy')' && \text{(A.475)}\\
&= \lim_{n\to\infty} y(b_n)y'(b_n) - y(-a_n)y'(-a_n) && \text{(A.476)}\\
&= 0, && \text{(A.477)}
\end{aligned}$$

where (A.472) follows by the monotone convergence theorem as $y \in L^2(\mathbb{R}) \cap L^2(\mathbb{R}, c)$. Thus, y globally minimizes $\mathcal{E}$ over V.

Next, we show that the already shown properties of y imply that p (which we defined by $p = y^2$ in the beginning of this proof) minimizes the Fisher information. For that, we consider first a couple of important quantities.

Define, for $\gamma \in \mathbb{R}$,

$$I_\gamma^\star := \inf_{\substack{w \in V \\ \|w\|_{2,c}^2 \le \gamma,\ \|w\|_2 = 1}} 4\|w'\|_2^2. \tag{A.478}$$

It is not hard to see that V is closed under positive dilation, so in particular $u(x) = w(x/\sigma)/\sqrt{\sigma}$ is in V if $w \in V$. This in turn yields (via choosing σ large enough if necessary) that the inequality $\|w\|_{2,c}^2 \le \gamma$ in the definition of $I_\gamma^\star$ can be replaced with an equality, i.e.,

$$I_\gamma^\star = \inf_{\substack{w \in V \\ \|w\|_{2,c}^2 = \gamma,\ \|w\|_2 = 1}} 4\|w'\|_2^2. \tag{A.479}$$

Our next goal is to show that $E \le E^\star$, where we define

$$E^\star := \inf_{\gamma \in \mathbb{R}} I_\gamma^\star + \theta\gamma. \tag{A.480}$$

Indeed, by (A.479), we use (A.480) to deduce that $E^\star$ satisfies

$$E^\star = \inf_{\substack{w \in V \\ \|w\|_2 = 1}} 4\|w'\|_2^2 + \theta\|w\|_{2,c}^2 \tag{A.481}$$

$$= \inf_{w \in V \setminus \{0\}} \frac{4\|w'\|_2^2 + \theta\|w\|_{2,c}^2}{\|w\|_2^2} \tag{A.482}$$

$$= E + \inf_{w \in V \setminus \{0\}} \frac{\mathcal{E}(w)}{\|w\|_2^2} \tag{A.483}$$

$$\ge E, \tag{A.484}$$

where (A.482) follows since V is a vector space, and (A.484) since $\inf_{w \in V} \mathcal{E}(w) \ge 0$ (see (A.470)).

Next, we deduce that $I(p) = I_C^\star$. Note that $p = y^2$ implies $(p')^2/p = 4(y')^2$. Thus, $I(p) = 4\|y'\|_2^2$. From $E \le E^\star$ and the definition of $E^\star$ in (A.480), we obtain

$$E\|y\|_2^2 = E \le E^\star \le I_C^\star + \theta C = I_C^\star + \theta\|y\|_{2,c}^2. \tag{A.485}$$

Adding $4\|y'\|_2^2 - E\|y\|_2^2$ to both sides, we obtain

$$I(p) \le I_C^\star + \mathcal{E}(y). \tag{A.486}$$

As $\mathcal{E}(y) = 0$ (see (A.477)), we conclude that $I(p) \le I_C^\star$. The reverse inequality also holds since

$\|y\|_2 = 1$ and $\|y\|_{2,c}^2 = C$, so we conclude that

$$I(p) = I_C^\star. \tag{A.487}$$

Finally, we are ready to show that p globally minimizes the Fisher information, i.e., with $\mathcal{P}$ denoting the set of all possible PDFs, we show that

$$I(p) = \inf_{\substack{q \in \mathcal{P} \\ \mathbb{E}_q[c] \le C}} I(q). \tag{A.488}$$

We start by showing that $I(p)$ is minimal among strictly positive PDFs. Denote the set of strictly positive PDFs by $\mathcal{P}_0$,

$$\mathcal{P}_0 := \{q \in \mathcal{P} \; ; \; q(x) > 0 \text{ for every } x \in \mathbb{R}\}. \tag{A.489}$$

Note that, by definition of the Fisher information, $q \in \mathrm{AC}(\mathbb{R})$ if $I(q) < \infty$. Further, if $q \in \mathrm{AC}(\mathbb{R})$, then $\sqrt{q} \in \mathrm{AC}_{\mathrm{loc}}(\mathbb{R})$. Then, for every $q \in \mathcal{P}_0$ such that $I(q) < \infty$, setting $w = \sqrt{q}$, we get

$$I(q) = 4\|w'\|_2^2. \tag{A.490}$$

Thus, we conclude from $I(p) = I_C^\star$ (see (A.487)) that

$$I(p) = \inf_{\substack{q \in \mathcal{P}_0 \\ \mathbb{E}_q[c] \le C}} I(q). \tag{A.491}$$

However, the same argument cannot be applied to a PDF q that has zeros. For this, we apply Lemma A.9, to obtain from (A.491) that

$$I(p) = \inf_{\substack{q \in \mathcal{P}_0 \\ \mathbb{E}_q[c] \le C}} I(q) = \inf_{\substack{q \in \mathcal{P} \\ \mathbb{E}_q[c] \le C}} I(q), \tag{A.492}$$

which is the global optimality of p claimed in (A.488). Since p is strictly positive, and since it minimizes the Fisher information among all possible PDFs, we conclude that it is the unique minimizer of the Fisher information over all possible PDFs (see, e.g., [HR09, Proposition 4.5]), and the proof of the theorem is complete.

A.11 Proof of Proposition 2.3: Regularity of the Schrödinger PDF

We need the following well-known differentiation under the integral sign result.

Theorem A.2. *Let $U \subset \mathbb{R}$ be open, and (X, Σ, μ) be a measure space. Suppose $f : U \times X \to \mathbb{R}$ satisfies:*

1. *For each $a \in U$, we have $f(a,\,\cdot\,) \in L^1(\mu)$.*

2. *For μ-almost every $x \in X$, the function $f(\,\cdot\,, x)$ is differentiable over U.*

3. *The function $x \mapsto \sup_{a_0 \in U} \left| \frac{\partial f}{\partial a}(a_0, x) \right|$ is μ-integrable.*

Then, over U,

$$\frac{d}{da} \int_X f(a,x)\, d\mu(x) = \int_X \frac{\partial f}{\partial a}(a,x)\, d\mu(x). \tag{A.493}$$

We use Theorem A.2 to differentiate $a \mapsto D(p \,\|\, T_a p)$ twice, then conclude using Taylor's theorem. We have that

$$D(p \,\|\, T_a p) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{p(x-a)}\, dx. \tag{A.494}$$

Denote the function $f_1(a,x) := \log \frac{p(x)}{p(x-a)}$. For each $a \in \mathbb{R}$, we have that $f_1(a,\,\cdot\,)$ is continuous; indeed, it is differentiable by differentiability and strict positivity of p (recall that $p = y_{\theta,c}^2$). We consider the Borel space $(X, \Sigma, \mu) = (\mathbb{R}, \mathcal{B}, p(x)\, dx)$. Hence, for the sake of showing the integrability $f_1(a,\,\cdot\,) \in L^1(p(x)\, dx)$, we may ignore any bounded interval. By the asymptotic expansion of $y_{\theta,c}$ in Theorem A.1, we have the following asymptotic formula. Let E denote the eigenvalue of $y_{\theta,c}$. For each $a \in \mathbb{R}$, as $x \to \infty$ we have

$$\frac{p(x)}{p(x-a)} \sim \sqrt{\frac{c(x-a)}{c(x)}} \cdot \exp\left(-2 \int_{x-a}^{x} \sqrt{\theta c(t) - E}\, dt\right). \tag{A.495}$$

By Assumption 2.2, we have that for all large x

$$\frac{1}{2\rho(-a)} \leq \frac{c(x-a)}{c(x)} \leq 2\rho(a). \tag{A.496}$$

Further, for $|a| < 1$, we have that for all large x

$$\int_{x-a}^{x} \sqrt{\theta c(t) - E}\, dt \leq \sqrt{\theta c(x)} \leq \sqrt{\theta} \cdot c(x). \tag{A.497}$$

Thus, we conclude the integrability $f_1(a,\,\cdot\,) \in L^1(p(x)\, dx)$.

For each $x \in \mathbb{R}$, $f_1(\,\cdot\,, x)$ is differentiable with derivative

$$\frac{\partial f_1}{\partial a}(a_0, x) = \frac{-p'(x-a_0)}{p(x-a_0)} = \frac{-2y'_{\theta,c}(x-a_0)}{y_{\theta,c}(x-a_0)}. \tag{A.498}$$

We consider $a_0 \in U = (-1,1)$. From Proposition 2.2, there is some $z_0 = z_0(\theta, c)$ such that $z > z_0(\theta, c)$ implies

$$\left| \frac{y'_{\theta,c}(z)}{y_{\theta,c}(z)} \right| \leq 2\sqrt{2\theta} \cdot c(z). \tag{A.499}$$

Hence, for $x > z_0 + 1$, we have for all $|a_0| < 1$

$$\left| \frac{y'_{\theta,c}(x - a_0)}{y_{\theta,c}(x - a_0)} \right| \leq 2\sqrt{2\theta} \cdot c(x - a_0). \tag{A.500}$$

Using Assumption 2.2, we have

$$\sup_{|a_0|<1} c(x - a_0) \leq \left(\sup_{|a_0|<1} \rho(a_0) \right) \cdot (c(x) + 1), \tag{A.501}$$

where $A := \sup_{|a_0|<1} \rho(a_0)$ is finite. Combining these inequalities, we conclude that

$$\sup_{|a_0|<1} \left| \frac{\partial f_1}{\partial a}(a_0, x) \right| \leq 4A\sqrt{2\theta} \cdot (c(x) + 1) \tag{A.502}$$

for all large x. As $(a, x) \mapsto \frac{\partial f_1}{\partial a}(a, x)$ is continuous, we conclude that $\sup_{|a_0|<1} \left| \frac{\partial f_1}{\partial a}(a_0, x) \right| \in L^1(p(x)\, dx)$.

Therefore, we may apply Theorem A.2 to differentiate the KL-divergence and obtain

$$\frac{d}{da} D(p \,\|\, T_a p) = -\int_{\mathbb{R}} p(x) \cdot \frac{p'(x - a)}{p(x - a)}\, dx \tag{A.503}$$

over $|a| < 1$. Performing a change of variable, we obtain that

$$\frac{d}{da} D(p \,\|\, T_a p) = -\int_{\mathbb{R}} p(x + a) \cdot \frac{p'(x)}{p(x)}\, dx. \tag{A.504}$$

Next, we apply Theorem A.2 to differentiate the KL-divergence a second time. This time, we use the usual Lebesgue space $(\mathbb{R}, \mathcal{B}, \lambda)$. Consider the function $f_2(a, x) := p(x + a) \cdot \frac{p'(x)}{p(x)}$. Inequality (A.502) shows that $f_2(a, \cdot) \in L^1(\lambda)$ for each $a \in (-1, 1)$. Further, for each $x \in \mathbb{R}$, $f_2(\cdot, x)$ is differentiable over $(-1, 1)$ with derivative

$$\frac{\partial f_2}{\partial a}(a, x) = p'(x + a) \cdot \frac{p'(x)}{p(x)}. \tag{A.505}$$

We write

$$\left| \frac{\partial f_2}{\partial a}(a, x) \right| = \left| \frac{p'(x + a)}{p(x + a)} \right| \cdot \left| \frac{p'(x)}{p(x)} \right| \cdot p(x + a). \tag{A.506}$$

Via the same derivation of inequality (A.502), but using the full power of Proposition 2.2 this time (i.e., $\sqrt{c}$ as an upper bound instead of c), and applying Lemma A.10 (i.e., that p is eventually decreasing), we obtain the bound

$$\sup_{|a|<1} \left| \frac{\partial f_2}{\partial a}(a, x) \right| \leq 8\sqrt{2A\theta} \cdot c(x) p(x - 1) \tag{A.507}$$

for all large x. Therefore, $\sup_{|a|<1} \left| \frac{\partial f_2}{\partial a}(a, x) \right| \in L^1(\lambda)$.

Hence, we may apply Theorem A.2 again to obtain that

$$\frac{d^2}{da^2} D(p \,\|\, T_a p) = -\int_{\mathbb{R}} p'(x+a) \cdot \frac{p'(x)}{p(x)} \, dx \tag{A.508}$$

over $(-1, 1)$. Setting $a = 0$ in $D(p \,\|\, T_a p)$ and its first two derivatives, we obtain from Taylor's theorem that

$$D(p \,\|\, T_a p) = \frac{a^2}{2} I(p) + o(a^2) \tag{A.509}$$

as $a \to 0$, i.e., expansion (2.100) holds.

Next, we show that $\sup_{|a| \le s} \mathrm{V}(p \,\|\, T_a p) < \infty$ for some $s > 0$. For this, it suffices to show that

$$\int_{\mathbb{R}} p(x) \cdot \sup_{|a| \le 1} \left| \log \frac{p(x)}{p(x-a)} \right|^2 dx < \infty \tag{A.510}$$

Using the asymptotic formula for $y_{\theta,c}$ in Theorem 2.11, we have

$$\frac{p(x)}{p(x-a)} \sim \sqrt{\frac{c(x-a)}{c(x)}} \cdot \exp\left(-2 \int_{x-a}^{x} \sqrt{\theta c(t) - E \, dt} \right). \tag{A.511}$$

Hence, the same method showing integrability of $f_1(a, \cdot\,)$ shows the desired result.

A.12 Proofs of Section 2.14.5

A.12.1 Proof of Proposition 2.4

According to Theorem 2.14, we need to solve the following differential equation

$$y''(x) = \left(\theta x^2 - E \right) y(x), \tag{A.512}$$

for some $\theta > 0$ and eigenvalue E. It can be easily verified that

$$y_1(x) = \left(\frac{\sqrt{\theta}}{\pi} \right)^{1/4} e^{-x^2 \cdot \sqrt{\theta}/2}, \tag{A.513}$$

solves (A.512) with the corresponding eigenvalue

$$E_0 = \sqrt{\theta}. \tag{A.514}$$

Thus, by the uniqueness property in Lemma 2.2, the unit-norm, strictly-positive, ground-state eigenfunction, denoted by $y_{\theta,c}$ in Lemma 2.2, is given by y_1. Therefore, the optimal density is given by $y_{\theta,c}^2$. The corresponding cost for this density is therefore equal to

$$\left(\frac{\sqrt{\theta}}{\pi}\right)^{1/2}\int_{\mathbb{R}} x^2 e^{-x^2\sqrt{\theta}}\,dx = \frac{1}{2\sqrt{\theta}}. \tag{A.515}$$

To ensure that the incurred cost is equal to C, we thus need to choose

$$\theta = \frac{1}{4C^2}. \tag{A.516}$$

Plugging this into (A.513), we obtain

$$y_{\theta,c}(x)^2 = \frac{1}{\sqrt{2\pi C}} e^{-x^2/(2C)}, \tag{A.517}$$

which, according to Theorem 2.14, is the optimal density $p_{c,C}^{\star}$. Thus, Gaussian density is optimal in the small-sensitivity regime.

A.12.2 Proof of Lemma 2.3

Denote

$$\alpha = \frac{1}{3C\cdot\mathsf{Ai}(a_1')^2}, \quad \beta = \frac{-2a_1'}{3C}, \tag{A.518}$$

so we have $p_{\mathsf{Ai},C}(x) = \alpha\mathsf{Ai}(\beta|x|+a_1')^2$. Recall that the Airy function Ai satisfies the differential equation (2.74), i.e.,

$$\mathsf{Ai}''(x) = x\mathsf{Ai}(x). \tag{A.519}$$

We now verify that $p_{\mathsf{Ai},C}$ is a PDF. By evenness of $p_{\mathsf{Ai},C}$, a change of variable yields the integral

$$\int_{\mathbb{R}} p_{\mathsf{Ai},C} = 2\alpha\int_0^\infty \mathsf{Ai}(\beta x + a_1')^2\,dx = \frac{2\alpha}{\beta}\int_{a_1'}^\infty \mathsf{Ai}(x)^2\,dx. \tag{A.520}$$

Using $\mathsf{Ai}''(x) = x\mathsf{Ai}(x)$, one obtains

$$\left(x\mathsf{Ai}(x)^2 - \mathsf{Ai}'(x)^2\right)' = \mathsf{Ai}(x)^2. \tag{A.521}$$

Using this antiderivative of $\mathsf{Ai}(x)^2$ in (A.520), one obtains

$$\int_{\mathbb{R}} p_{\mathsf{Ai},C} = \frac{2\alpha}{\beta}\int_{a_1'}^\infty \mathsf{Ai}(x)^2\,dx = \frac{2\alpha\cdot(-a_1')\mathsf{Ai}(a_1')^2}{\beta} = 1. \tag{A.522}$$

Hence, $p_{\mathsf{Ai},C}$ is indeed a PDF.

Next, we show that the first absolute moment of $p_{\mathsf{Ai},C}$ is C. Beginning as in (A.520), then using

$x\mathsf{Ai}(x) = \mathsf{Ai}''(x)$, we get that

$$\int_{\mathbb{R}} |x| p_{\mathsf{Ai},C}(x)\,dx = \frac{2\alpha}{\beta^2}\int_{a_1'}^{\infty}(x-a_1')\mathsf{Ai}(x)^2\,dx = \frac{2\alpha}{\beta^2}\left(\int_{a_1'}^{\infty}\mathsf{Ai}''(x)\mathsf{Ai}(x)\,dx - a_1'\int_{a_1'}^{\infty}\mathsf{Ai}(x)^2\,dx\right). \tag{A.523}$$

Using integration by parts, we obtain

$$\int_{a_1'}^{\infty}\mathsf{Ai}''(x)\mathsf{Ai}(x)\,dx = -\int_{a_1'}^{\infty}\mathsf{Ai}'(x)^2\,dx. \tag{A.524}$$

Using $\mathsf{Ai}''(x) = x\mathsf{Ai}(x)$, one can verify that

$$\frac{1}{3}\left(-x^2\mathsf{Ai}(x)^2 + x\mathsf{Ai}'(x)^2 + 2\mathsf{Ai}(x)\mathsf{Ai}'(x)\right)' = \mathsf{Ai}'(x)^2. \tag{A.525}$$

Hence, we have the integral

$$\int_{a_1'}^{\infty}\mathsf{Ai}'(x)^2\,dx = \frac{(a_1')^2}{3}\mathsf{Ai}(a_1')^2. \tag{A.526}$$

In sum, we obtain that

$$\int_{\mathbb{R}} |x| p_{\mathsf{Ai},C}(x)\,dx = \frac{2\alpha}{\beta^2}\cdot\frac{2(a_1')^2\mathsf{Ai}(a_1')^2}{3} = C, \tag{A.527}$$

as desired.

Finally, we check the variance formula for $p_{\mathsf{Ai},C}$. We start with rewriting the integral as

$$\int_{\mathbb{R}} x^2 p_{\mathsf{Ai},C}(x)\,dx = \frac{2\alpha}{\beta^3}\left(\int_{a_1'}^{\infty}x^2\mathsf{Ai}(x)^2\,dx - 2a_1'\int_{a_1'}^{\infty}x\mathsf{Ai}(x)^2\,dx + (a_1')^2\int_{a_1'}^{\infty}\mathsf{Ai}(x)^2\,dx\right) \tag{A.528}$$

$$= \frac{2\alpha}{\beta^3}\left(\int_{a_1'}^{\infty}\mathsf{Ai}''(x)^2\,dx + \frac{(-a_1')^3\mathsf{Ai}(a_1')^2}{3}\right). \tag{A.529}$$

It can be directly verified that

$$\frac{1}{5}\left((x^3-1)\mathsf{Ai}(x)^2 - x^2\mathsf{Ai}'(x)^2 + 2x\mathsf{Ai}(x)\mathsf{Ai}'(x)\right)' = \mathsf{Ai}''(x)^2. \tag{A.530}$$

Hence,

$$\int_{a_1'}^{\infty}\mathsf{Ai}''(x)^2\,dx = \frac{(-a_1')^3+1}{5}. \tag{A.531}$$

Therefore,

$$\int_{\mathbb{R}} x^2 p_{\mathsf{Ai},C}(x)\,dx = \frac{2\alpha}{\beta^3}\left(\frac{8}{15}\cdot(-a_1')^3 + \frac{1}{5}\right)\cdot\mathsf{Ai}(a_1')^2 = \frac{9}{4}\left(\frac{8}{15} + \frac{1}{5\cdot(-a_1')^3}\right)\cdot C^2, \tag{A.532}$$

and the proof is complete.

A.12.3 Proof of Proposition 2.5

First, we notice that, according to Theorem 2.14, we need to solve the differential equation (2.117). Let

$$y_1 = \sqrt{p_{\mathsf{Ai},C}}. \tag{A.533}$$

Thus, from Definition 2.7, we have

$$y_1(x) = \gamma \cdot \mathsf{Ai}\left(\theta^{1/3}|x| + a_1'\right), \tag{A.534}$$

where

$$\gamma := \frac{1}{\sqrt{3C} \cdot \mathsf{Ai}(a_1')}, \tag{A.535}$$

and

$$\theta = \left(\frac{-2a_1'}{3C}\right)^3. \tag{A.536}$$

Differentiating separately for $x < 0$, $x = 0$, and $x > 0$, we obtain

$$y_1'(x) = \theta^{1/3}\gamma \operatorname{sgn}(x)\mathsf{Ai}'(\theta^{1/3}|x| + a_1'), \tag{A.537}$$

for every $x \in \mathbb{R}$ (where $\operatorname{sgn}(x) = x/|x|$ for $x \neq 0$, and $\operatorname{sgn}(0) = 0$). Thus, y_1' is absolutely continuous. Differentiating again, we obtain for every $x \in \mathbb{R}$

$$y_1''(x) = \theta^{2/3}\gamma\mathsf{Ai}''\left(\theta^{1/3}|x| + a_1'\right). \tag{A.538}$$

Since Ai is a solution of the differential equation of (2.74), it follows that $\mathsf{Ai}''(z) = z\mathsf{Ai}(z)$ for every $z \in \mathbb{R}$ and thus

$$y_1''(x) = \left(\theta|x| + \theta^{2/3}a_1'\right) y_1(x), \tag{A.539}$$

and hence y_1 solves the equation (2.117). Therefore, we conclude from Lemma 2.2 that $y_{\theta,c} = y_1$ is the ground-state eigenfunction of $\mathcal{H}_{\theta c}$. Moreover, since $\int_{\mathbb{R}} cy_{\theta,c}^2 = \int_{\mathbb{R}} cp_{\mathsf{Ai},C} = C$, Theorem 2.14 implies that $p_{c,C}^\star = p_{\mathsf{Ai},C}$, as desired.

A.13 Subsampling: Proof of Lemma 2.4

Subsampling is a fundamental tool in the analysis of differentially-private mechanisms. Informally, subsampling entails applying a differentially-private mechanism to a small set of randomly sampled datapoints from a given dataset. There are several ways of formally defining the subsampling operator, see, e.g., [BBG18]. The most well-known one, Poisson subsampling, is parameterized by the subsampling rate $\lambda \in (0,1]$ which indicates the probability of selecting a datapoint. More formally, the subsampled datapoints from a dataset D can be expressed as $\{x \in D \,:\, B_x = 1\}$, where B_x is a Bernoulli random variable with parameter λ independent for each $x \in D$. Given any mechanism $\mathcal{M}$, we define the subsampled mechanism $\mathcal{M}_\lambda$ as the composition of $\mathcal{M}$ and the Poisson subsampling operator. Characterizing the privacy guarantees of subsampled mechanisms is the subject of "privacy amplification by subsampling" principle [KLN+11]. This principle is well-studied particularly for characterizing the privacy guarantees of subsampled Gaussian mechanisms in the context of a variant of differential privacy, namely, Rényi differential privacy [ZW19, ACG+16, MTZ19]. We can mirror their formulation to characterize ε and δ for the subsampled Gaussian mechanisms. Recall that a Gaussian mechanism satisfies $\mathcal{M}(D) = \mathcal{N}(f(D), \sigma^2 I_d)$ where f is a query function with ℓ_2-sensitivity 1. For the *subsampled* Gaussian, the optimal privacy curve (of a single composition) is

$$\delta_{\mathcal{M}_\lambda}(\varepsilon) = \max\left\{\mathsf{E}_{e^\varepsilon}(P\|Q), \mathsf{E}_{e^\varepsilon}(Q\|P)\right\}, \tag{A.540}$$

where $P = \mathcal{N}(0, \sigma^2 I_d)$ and $Q = (1-\lambda)P + \lambda P'$, and $P' \sim \mathcal{N}(\boldsymbol{e}_1, \sigma^2 I_d)$ where $\boldsymbol{e}_1$ is the first standard basis vector. In Lemma 2.4 (restated below for convenience), we show that the above maximum is always attained by $\mathsf{E}_{e^\varepsilon}(Q\|P)$ for any $\varepsilon \geq 0$, and that it holds for a larger family of DP mechanisms (including Gaussian and Laplace mechanisms). A similar ordering bound was proved by [MTZ19, Theorem 5] for the Rényi divergence.

Lemma A.11. *Fix a Borel probability measure P over $\mathbb{R}^n$ that is symmetric around the origin (i.e., $P(\mathcal{A}) = P(-\mathcal{A})$ for every Borel $\mathcal{A} \subset \mathbb{R}^n$), and fix constants $(s, \lambda, \gamma) \in \mathbb{R}^n \times [0,1] \times [1,\infty)$. Let $T_s P$ be the probability measure given by $(T_s P)(\mathcal{A}) = P(\mathcal{A} - s)$, and let $Q = (1-\lambda)P + \lambda T_s P$. We have the inequality $\mathsf{E}_\gamma(P\|Q) \leq \mathsf{E}_\gamma(Q\|P)$, with equality if and only if $(\gamma - 1)\,\lambda\,\|s\|\,\mathsf{E}_\gamma(Q\|P) \;= 0$.*

Proof. The case $\lambda = 0$ is clear, so assume $\lambda \in (0,1]$. Suppose for now that $\gamma \cdot (1-\lambda) < 1$. Denote $R := T_s P$, and consider the function $G : (0,\infty) \to [0,\infty)$ defined by

$$G(t) := t \cdot \mathsf{E}_{1+\frac{\gamma-1}{t}}(P\|R). \tag{A.541}$$

Since $\gamma' \mapsto \mathsf{E}_{\gamma'}(P\|R)$ is monotonically decreasing, we have that G is monotonically increasing. Note that $0 < \gamma\lambda + 1 - \gamma \leq \lambda$. Thus, plugging $t \in \{\gamma\lambda + 1 - \gamma, \lambda\}$ into G, we obtain

$$(\gamma\lambda + 1 - \gamma) \cdot \mathsf{E}_{\frac{\gamma\lambda}{\gamma\lambda+1-\gamma}}(P\|R) \leq \lambda \cdot \mathsf{E}_{\frac{\lambda-(1-\gamma)}{\lambda}}(P\|R). \tag{A.542}$$

Now, note that

$$(\gamma\lambda + 1 - \gamma) \cdot \mathsf{E}_{\frac{\gamma\lambda}{\gamma\lambda+1-\gamma}}(P\|R) = (\gamma\lambda + 1 - \gamma) \cdot \sup_{\mathcal{A}} P(\mathcal{A}) - \frac{\gamma\lambda}{\gamma\lambda + 1 - \gamma} \cdot R(\mathcal{A}) \tag{A.543}$$

$$= \sup_{\mathcal{A}} P(\mathcal{A}) - \gamma \cdot ((1-\lambda)P(\mathcal{A}) + \lambda R(\mathcal{A})) \tag{A.544}$$

$$= \mathsf{E}_{\gamma}(P\|Q), \tag{A.545}$$

where the suprema are taken over all Borel sets $\mathcal{A} \subset \mathbb{R}^n$. In addition, by symmetry of P around the origin, we have that

$$\mathsf{E}_{\gamma'}(P\|R) = \sup_{\mathcal{A}} P(\mathcal{A}) - \gamma' P(\mathcal{A} - s) \tag{A.546}$$

$$= \sup_{\mathcal{A}} P(-\mathcal{A}) - \gamma' P(-\mathcal{A} - s) \tag{A.547}$$

$$= \sup_{\mathcal{A}} P(\mathcal{A}) - \gamma' P(\mathcal{A} + s) \tag{A.548}$$

$$= \sup_{\mathcal{A}} P(\mathcal{A} - s) - \gamma' P(\mathcal{A}) \tag{A.549}$$

$$= \mathsf{E}_{\gamma'}(R\|P). \tag{A.550}$$

Therefore,

$$\lambda \cdot \mathsf{E}_{\frac{\lambda-(1-\gamma)}{\lambda}}(P\|R) = \lambda \cdot \mathsf{E}_{\frac{\lambda-(1-\gamma)}{\lambda}}(R\|P) \tag{A.551}$$

$$= \lambda \cdot \sup_{\mathcal{A}} R(\mathcal{A}) - \frac{\lambda - (1-\gamma)}{\lambda} \cdot P(\mathcal{A}) \tag{A.552}$$

$$= \sup_{\mathcal{A}} \ ((1-\lambda)P(\mathcal{A}) + \lambda R(\mathcal{A})) - \gamma P(\mathcal{A}) \tag{A.553}$$

$$= \mathsf{E}_{\gamma}(Q\|P). \tag{A.554}$$

We conclude from (A.542) the desired inequality $\mathsf{E}_{\gamma}(P\|Q) \leq \mathsf{E}_{\gamma}(Q\|P)$. In addition, the case $\gamma \cdot (1-\lambda) \geq 1$ follows immediately since then $\mathsf{E}_{\gamma}(P\|Q) = 0 \leq \mathsf{E}_{\gamma}(Q\|P)$. □

In light of this lemma, the privacy guarantee of a subsampled Gaussian mechanism is fully characterized by computing only $\mathsf{E}_{e^{\varepsilon}}((1-\lambda)P + \lambda T_s P\|P)$, where $P = \mathcal{N}(0, \sigma^2 I_d)$. Based on this result, for our numerical experiments, we only compute the saddle-point accountant with this order of P and Q.

A.14 The Method of Steepest Descent

We describe the general approach for the method of steepest descent. Our task is to compute the contour integral

$$I_n = \frac{1}{2\pi i} \int_{t-i\infty}^{t+i\infty} e^{F_n(z)}\, dz. \tag{A.555}$$

What we will obtain is an asymptotic expansion

$$I_n \overset{\text{as. ex.}}{\sim} \frac{e^{F_n(t_0)}}{\sqrt{2\pi F_n''(t_0)}} \left(1 + \sum_{m=2}^{\infty} \beta_{n,m}\right). \tag{A.556}$$

In a nutshell, the method of steepest descent is a powerful tool for choosing the best parameter t that renders the computation of I_n easiest. In particular, this choice of t is called the *saddle-point*, which is found as follows.

Here, F_n is holomorphic over a strip $(0, T) + i\mathbb{R}$ in the complex plane, the parameter $n \in \mathbb{N}$ is growing without bound, and $t \in (0, T) \subset \mathbb{R}$ is a free parameter. In particular, the value of the integral I_n is assumed to be independent of the parameter t. This could be satisfied for certain choices of F_n by virtue of its analyticity and in view of Cauchy's integral theorem. As we show in Theorem 2.15, computing the above contour integral amount to exactly computing the privacy parameter $\delta_{L^{(n)}}(\varepsilon)$ if we choose the function

$$F_n(z) = K_{L^{(n)}}(z) - z\varepsilon - \log z - \log(1 + z). \tag{A.557}$$

Suppose that $F_n''(t) > 0$ over $t \in (0, T)$—in particular, F_n is strictly convex over the real interval $(0, T)$—and that there is a value $t_0 \in (0, T)$ solving the equation $F_n'(t_0) = 0$, which is then necessarily unique. Then, a second order Taylor expansion around t_0 yields that

$$F_n(z) = F_n(t_0) + \frac{(z - t_0)^2}{2} F_n''(t_0) + o(|z - t_0|^3). \tag{A.558}$$

Looking at the the values of the approximating quadratic $F_n(t_0) + \frac{(z-t_0)^2}{2} F_n''(t_0)$ for z near t_0 along the real axis (so $z = t$ for $t_0 \approx t \in \mathbb{R}$) and along the axis $t_0 + i\mathbb{R}$ (so $z = t_0 + is$ for $0 \approx s \in \mathbb{R}$), we see that this approximation has a local minimum at t_0 along the real axis and it has a local maximum at t_0 along the axis $t_0 + i\mathbb{R}$. Hence, t_0 is a saddle-point for the approximating quadratic. Further, as the integral we are concerned with runs along the contour $t + i\mathbb{R}$, we expect the value of I_n to come primarily from values $z \approx t_0$.

Now, consider the function

$$G_n(z) = F_n(t_0 + z) - F_n(t_0) - \frac{z^2}{2} F_n''(t_0). \tag{A.559}$$

We have that G_n is holomorphic over some vertical strip centered at the origin, and

$$G_n^{(k)}(0) = \begin{cases} 0, & 0 \leq k \leq 2 \\ F_n^{(k)}(t_0), & k \geq 3. \end{cases} \tag{A.560}$$

We assume for the next steps that G_n is an entire function. Thus, G_n has the expansion

$$G_n(z) = \sum_{k \geq 3} \frac{F_n^{(k)}(t_0)}{k!} z^k. \tag{A.561}$$

Furthermore, $e^{G_n(z)}$ has the power series expansion

$$e^{G_n(z)} = 1 + \sum_{k \geq 3} \alpha_{n,k} z^k, \tag{A.562}$$

where

$$\alpha_{n,k} = \frac{1}{k!} B_k(0, 0, F_\varepsilon^{(3)}(t_0), \ldots, F_\varepsilon^{(k)}(t_0)). \tag{A.563}$$

As we may write

$$F_n(t_0 + is) = F_n(t_0) + G_n(is) - \frac{F_n''(t_0)}{2} s^2, \tag{A.564}$$

we get the exact value of the integral

$$I_n = \frac{e^{F_n(t_0)}}{2\pi} \int_{-\infty}^{\infty} e^{-s^2 F_n''(t_0)/2} \left(1 + \sum_{k \geq 3} \alpha_{n,k} (is)^k \right) ds. \tag{A.565}$$

The derived steps thus far have all been justified rigorously. The final step, however, is a heuristic, where we truncate the power series expansion to obtain possible estimates of I_n. The point is that the derived expressions through this heuristic have the potential of being proved by other means to be indeed close approximations of I_n.

For instance, dropping the whole series beyond the constant term yields the basic saddle-point approximation

$$I_{n,1} := \frac{e^{F_n(t_0)}}{2\pi} \int_{-\infty}^{\infty} e^{-s^2 F_n''(t_0)/2} \, ds = \frac{e^{F_n(t_0)}}{\sqrt{2\pi F_n''(t_0)}}. \tag{A.566}$$

Note that this approximation is in fact exact if F_n is a quadratic, i.e., for computing the Gaussian integral. Keeping the terms $k \in \{3, \cdots, 2k^\star\}$, it is not hard to see that one obtains the $k^\star$-th estimate

$$I_{n,k^\star} := \frac{e^{F_n(t_0)}}{\sqrt{2\pi F_n''(t_0)}} \left(1 + \sum_{m=2}^{k^\star} \beta_{n,m}\right), \tag{A.567}$$

where we denote the constants

$$\beta_{n,m} := \frac{(-1)^m B_{2m}(0, 0, F_n^{(3)}(t_0), \ldots, F_n^{(2m)}(t_0))}{2^m m! F_n''(t_0)^m}. \tag{A.568}$$

Then one might say that I_n has the "asymptotic expansion"

$$I_n \overset{\text{as. ex.}}{\sim} \frac{e^{F_n(t_0)}}{\sqrt{2\pi F_n''(t_0)}} \left(1 + \sum_{m=2}^{\infty} \beta_{n,m}\right). \tag{A.569}$$

Recall that this does not mean that the above equation holds with equality for any particular n. Rather, it is a heuristic indicating the potential for the truncated expansion to give close approximations for the intended integral I_n.

A.15 Satisfiability of the Assumptions

We explain here how Assumption 2.6 is satisfied by the subsampled Gaussian and Laplace mechanisms. Note that by the Lebesgue decomposition theorem, the probability measure of the PLRV can always be decomposed into a sum of an absolutely continuous measure, a discrete measure, and a singular measure (such as the Cantor distribution). Thus, Assumption 2.6 requires the exclusion of singular components. This can be easily seen to be satisfied by the subsampled Gaussian and Laplace mechanisms. Further, Assumption 2.6 does not impose any requirement on the discrete part. Thus, we consider the continuous part here.

Note that the PLRV for the subsampled Gaussian mechanism (with subsampling rate λ, variance σ^2, and sensitivity s) is given by

$$L = \log\left(1 - \lambda + \lambda e^{(2sX - s^2)/(2\sigma^2)}\right), \tag{A.570}$$

where $X \sim (1-\lambda)\mathcal{N}(0, \sigma^2) + \lambda \mathcal{N}(s, \sigma^2)$. Hence,

$$\mathbb{P}[L \le z] = \mathbb{P}\left[X \le \frac{s}{2} + \frac{\sigma^2}{s} \log\left(\frac{e^z - (1-\lambda)}{\lambda}\right)\right] \tag{A.571}$$

if $z > \log(1-\lambda)$, and $\mathbb{P}[L \leq z] = 0$ otherwise. So, L is continuous with PDF

$$p_L(z) = A\frac{e^{2z}}{g(z)^{3/2}} \cdot \left(\frac{g(z)}{\lambda}\right)^{-\frac{\sigma^2}{2s^2}\log\frac{g(z)}{\lambda}} \cdot 1_{(\log(1-\lambda),\infty)}(z), \tag{A.572}$$

where $g(z) = e^z - (1-\lambda)$ and we have the constant $A = \frac{\sigma}{s} \cdot \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{s^2}{8\sigma^2}\right)$. From this, we see that $p_L(z)$ decays superexponentially as $z \to \infty$. Further, it is continuous. Indeed, we only need to check continuity at $z = \log(1-\lambda)$. But this is immediate using, e.g., $y = \log\frac{g(z)}{\lambda}$ and taking $y \to -\infty$. These properties imply that Assumption 2.6 is satisfied by the subsampled Gaussian mechanism.

Finally, we note that the case of the subsampled Laplace mechanism is simpler. Indeed, taking the analogous expression for L as in (A.570), we see that L has only a discrete component and a continuous component. Further, the continuous part comes from values of X between 0 and s. This boundedness translates into the fact that the PDF of the continuous part of L is compactly supported, so Assumption 2.6 is satisfied in this case too.

A.16 Well-Definedness of the Saddle-Point

The well-definedness of the saddle-point, given $\varepsilon < \operatorname{ess\,sup} L$, follows from convexity of F_ε over the positive reals. Namely, we show that F_ε is *complex-differentiable* and that there is a unique positive real t_0 such that $F_\varepsilon'(t_0) = 0$. Let $K_L|_{\mathbb{R}}$ be the restriction of the CGF to the real axis. We have that $K_L|_{\mathbb{R}}$ is convex over $(0,\infty)$, and thus, $F_\varepsilon|_{\mathbb{R}}$ is strictly convex there. Thus, the minimum of F_ε over the positive reals is unique; further, the *real* derivative at this minimum vanishes. Nevertheless, finiteness of M_L over $(0,\infty)$ implies its analyticity over the half-plane $(0,\infty)+i\mathbb{R}$; in particular, the *complex* derivative of F_ε exists in the same half-plane. Hence, the function F_ε is complex-differentiable at t_0, and its derivative vanishes there, as required.

A.17 Proof of Theorem 2.15

Before proving Theorem 2.15, we show the following general Parseval identity. For $f \in L^1(\mathbb{R})$, we denote the Fourier transform by

$$\hat{f}(\xi) := \int_{\mathbb{R}} f(x)e^{-ix\xi}\,dx. \tag{A.573}$$

Lemma A.12. *Let $P = Q + R$ be a Borel probability measure on $\mathbb{R}$, where Q is absolutely continuous with respect to the Lebesgue measure whose PDF is square-integrable and R is discrete. For any continuous function*

$f : \mathbb{R} \to \mathbb{R}$ such that $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $\hat{f} \in L^1(\mathbb{R})$, and $\mathbb{E}_{X \sim P}[|f(X)|] < \infty$, we have the Parseval identity

$$\int_{\mathbb{R}} f(x)\, dP(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi) \phi_P(\xi)\, d\xi, \tag{A.574}$$

where $\phi_P(\xi) := \mathbb{E}_{X \sim P}[e^{i\xi X}]$ is the characteristic function.

Proof. Let q denote the PDF of Q. Suppose R is supported over $\{x_j\}_{j \in J}$, where J is at most countable, and write $r_j = R(\{x_j\})$. Then, we may write

$$\int_{\mathbb{R}} f(x)\, dP(x) = \int_{\mathbb{R}} f(x)\, dQ(x) + \int_{\mathbb{R}} f(x)\, dR(x) \tag{A.575}$$

$$= \int_{\mathbb{R}} f(x) q(x)\, dx + \sum_{j \in J} f(x_j) r_j. \tag{A.576}$$

Since $f, q \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, we have the Parseval identity

$$\int_{\mathbb{R}} f(x) q(x)\, dx = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi) \phi_Q(\xi)\, d\xi. \tag{A.577}$$

As we also have continuity of f and integrability of $\hat{f}$, we also have the Fourier inversion

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi) e^{ix\xi}\, d\xi \tag{A.578}$$

for *every* $x \in \mathbb{R}$. In particular, we have that

$$\sum_{j \in J} f(x_j) r_j = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi) \phi_R(\xi)\, d\xi. \tag{A.579}$$

The desired result follows by $\phi_P = \phi_Q + \phi_R$. □

Now, we apply Lemma A.12 to derive Theorem 2.15.

Proof of Theorem 2.15. Expectations of functions of $\widetilde{L}$ can be written in terms of L as $\mathbb{E}[f(\widetilde{L})] = \mathbb{E}[e^{tL} f(L)] / M_L(t)$. Thus, the MGF of the tilted variable $\widetilde{L}$ is given by

$$M_{\widetilde{L}}(z) = \mathbb{E}[e^{z\widetilde{L}}] = \frac{\mathbb{E}[e^{tL} e^{zL}]}{M_L(t)} = \frac{M_L(t+z)}{M_L(t)}. \tag{A.580}$$

Similarly, expectations of functions of L can be written in terms of $\widetilde{L}$ as $\mathbb{E}[f(L)] = M_L(t)\, \mathbb{E}[e^{-t\widetilde{L}} f(\widetilde{L})]$. Thus, we can write the privacy curve δ_L in terms of the tilted variable $\widetilde{L}$ as

$$\delta_L(\varepsilon) = \mathbb{E}\left[\left(1 - e^{\varepsilon - L}\right)^+ \right] \tag{A.581}$$

$$= M_L(t)\, \mathbb{E}\left[e^{-t\widetilde{L}} \left(1 - e^{\varepsilon - \widetilde{L}}\right)^+ \right]. \tag{A.582}$$

In other words, the formula in (2.131) holds.

Next, we use Assumption 2.6 to apply Parseval's identity (Lemma A.12) to the expectation in (A.582) to get the contour-integral formula in (2.132). Specifically, consider the function

$$f(x) = e^{-tx}\left(1 - e^{\varepsilon - x}\right)^{+}, \tag{A.583}$$

Note that f is bounded, continuous, and $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Further, we have the Fourier transform

$$\hat{f}(s) = \frac{e^{-(t+is)\varepsilon}}{(t+is)(t+1+is)} \in L^1(\mathbb{R}). \tag{A.584}$$

In addition, by Assumption 2.6, the probability measure P_L induced by L may be written as $P_L = Q_L + R_L$, where Q_L is absolutely continuous with respect to the Lebesgue measure whose PDF q_L satisfies that $x \mapsto e^{\tau x} q_L(x)^2$ is integrable for every $\tau > 0$ and R_L is discrete. Suppose R_L is supported over $\{x_j\}_{j\in J}$ with J at most countable, and write $r_{L,j} = R(\{x_j\})$. Then, by definition of exponential tilting, for every Borel set $B \subset \mathbb{R}$, we have that

$$\begin{aligned}
P_{\widetilde{L}}(B) &= \frac{1}{M_L(t)} \int_B e^{tx}\, dP_L(x) &\text{(A.585)}\\
&= \frac{1}{M_L(t)} \int_B e^{tx}\, dQ_L(x) + \frac{1}{M_L(t)} \int_B e^{tx}\, dR_L(x) &\text{(A.586)}\\
&= \frac{1}{M_L(t)} \int_B e^{tx} q_L(x)\, dx + \frac{1}{M_L(t)} \sum_{\substack{j\in J\\ x_j \in B}} e^{tx_j} r_{L,j} &\text{(A.587)}\\
&= \widetilde{Q}(B) + \widetilde{R}(B), &\text{(A.588)}
\end{aligned}$$

where we define the Borel measures

$$\begin{aligned}
\widetilde{Q}(B) &:= \frac{1}{M_L(t)} \int_B e^{tx} q_L(x)\, dx, &\text{(A.589)}\\
\widetilde{R}(B) &:= \frac{1}{M_L(t)} \sum_{\substack{j\in J\\ x_j \in B}} e^{tx_j} r_{L,j}. &\text{(A.590)}
\end{aligned}$$

From these definitions, it is clear that $\widetilde{R}$ is discrete and $\widetilde{Q}$ is absolutely continuous with respect to the Lebesgue measure with PDF $\widetilde{q}(x) := e^{tx} q_L(x)/M_L(t)$. Furthermore, by assumption on q_L, we have that $\widetilde{q} \in L^2(\mathbb{R})$. Therefore, we may apply Parseval's identity (Lemma A.12) on f and $P_{\widetilde{L}}$ to obtain

$$\mathbb{E}\left[f(\widetilde{L})\right] = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(s) \phi_{P_{\widetilde{L}}}(s)\, ds. \tag{A.591}$$

Next, applying the formula for $M_{\widetilde{L}}$ in (A.580), we see that

$$\phi_{P_{\widetilde{L}}}(s) = \mathbb{E}[e^{is\widetilde{L}}] = M_{\widetilde{L}}(is) = \frac{M_L(t+is)}{M_L(t)}. \tag{A.592}$$

Therefore, combining formulas (A.582) and (A.591), we get

$$\delta_L(\varepsilon) = M_L(t)\mathbb{E}[f(\widetilde{L})] = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(s) M_L(t+is)\, ds. \tag{A.593}$$

Now, using the contour $\{z = t + is \;:\; -\infty < s < \infty\}$ oriented counter-clockwise, we see that (A.593) may be rewritten as the contour integral

$$\delta_L(\varepsilon) = \frac{1}{2\pi i}\int_{t-i\infty}^{t+i\infty} \hat{f}((z-t)/i) M_L(z)\, dz. \tag{A.594}$$

Finally, using the formula for $\hat{f}$ in (A.584), we deduce

$$\begin{aligned} \delta_L(\varepsilon) &= \frac{1}{2\pi i}\int_{t-i\infty}^{t+i\infty} \frac{e^{-z\varepsilon}}{z(z+1)} M_L(z)\, dz && \text{(A.595)} \\ &= \frac{1}{2\pi i}\int_{t-i\infty}^{t+i\infty} e^{F_\varepsilon(z)}\, dz, && \text{(A.596)} \end{aligned}$$

where we define

$$F_\varepsilon(z) := K_L(z) - \varepsilon z - \log(z) - \log(1+z) \tag{A.597}$$

and we take the principal branch for the complex logarithm. This is precisely the desired formula for δ_L stated in (2.132), and the proof of the theorem is therefore complete. □

A.18 Proof of Theorem 2.16: The Large-Composition Regime

We may show this result using the standard Berry-Esseen approach. By the Berry-Esseen theorem, we have for $Z \sim \mathcal{N}(\mathbb{E}[L], \sigma_L^2)$ that

$$\begin{aligned} \delta_L(\varepsilon) &= \mathbb{E}\left[\left(1 - e^{\varepsilon - L}\right)^+\right] && \text{(A.598)} \\ &= \int_0^1 \mathbb{P}\left[L > \varepsilon - \log(1-u)\right] du && \text{(A.599)} \\ &= \delta_Z(\varepsilon) + \theta \cdot \frac{0.56\ \mathrm{P}_0}{\sigma_L^3} && \text{(A.600)} \end{aligned}$$

where $|\theta| \le 1$. A direct computation yields that, for any $\varepsilon \ge \mathbb{E}[L]$, with $Z \sim \mathcal{N}(\mathbb{E}[L], \sigma_L^2)$,

$$\delta_Z(\varepsilon) = \Phi\left(\frac{\mathbb{E}[L] - \varepsilon}{\sigma_L}\right) - e^{\varepsilon - \mathbb{E}[L] + \sigma_L^2/2}\, \Phi\left(\frac{\mathbb{E}[L] - \sigma_L^2 - \varepsilon}{\sigma_L}\right). \tag{A.601}$$

Plugging in $\varepsilon = \mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L$, we obtain that

$$\delta_Z(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L) = \delta - e^{-\Phi^{-1}(\delta)\sigma_L + \sigma_L^2/2}\Phi\left(\Phi^{-1}(\delta) - \sigma_L\right). \tag{A.602}$$

Using $\Phi(-x) = Q(x) = \frac{q(x)e^{-x^2/2}}{\sqrt{2\pi}}$ for $x > 0$, we obtain

$$\begin{aligned}\delta_Z(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L) &= \delta - \frac{q(\sigma_L - \Phi^{-1}(\delta))}{\sqrt{2\pi}}e^{-\Phi^{-1}(\delta)^2/2} && \text{(A.603)}\\ &= \delta - \frac{q(\sigma_L - \Phi^{-1}(\delta))}{\sqrt{2\pi}}e^{-(-\Phi^{-1}(\delta))^2/2} && \text{(A.604)}\\ &= \delta - \frac{q(\sigma_L - \Phi^{-1}(\delta))}{q(-\Phi^{-1}(\delta))}\Phi(\Phi^{-1}(\delta)) && \text{(A.605)}\\ &= \delta \cdot \left(1 - \frac{q(\sigma_L - \Phi^{-1}(\delta))}{q(-\Phi^{-1}(\delta))}\right). && \text{(A.606)}\end{aligned}$$

Note that $q(x) \sim 1/x$ as $x \to \infty$. Since $\sigma_L/(-\Phi^{-1}(\delta)) \to \infty$ by assumption, we also have $\sigma_L - \Phi^{-1}(\delta)) \to \infty$. Thus, we obtain

$$\frac{q(\sigma_L - \Phi^{-1}(\delta))}{q(-\Phi^{-1}(\delta))} \sim \frac{1}{-\Phi^{-1}(\delta)q(-\Phi^{-1}(\delta))} \cdot \frac{1}{\frac{\sigma_L}{-\Phi^{-1}(\delta)} - 1}. \tag{A.607}$$

As $\limsup \delta < 1/2$, we get that the term $-\Phi^{-1}(\delta)q(-\Phi^{-1}(\delta))$ is bounded away from 0. Therefore, we get that

$$\delta_Z(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L) = \delta \cdot (1 + o(1)). \tag{A.608}$$

From (A.600), and since $\mathrm{P}_0 = o(\sigma_L^3)$ by assumption, we conclude that

$$\delta_L(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L) = \delta \cdot (1 + o(1)). \tag{A.609}$$

In other words, $\mathcal{M}$ is $(\mathbb{E}[L] - \Phi^{-1}(\delta)\sigma_L, \delta \cdot (1 + o(1))$-DP, as desired.

A.19 Proof of Theorem 2.17: Asymptotic of the Saddle-Point

We write $K = K_L$ for short. Consider the saddle-point equation (2.134):

$$K'(t) = \varepsilon + \frac{1}{t} + \frac{1}{1+t}. \tag{A.610}$$

The left-hand side strictly increases from $\mathbb{E}[L]$ to $\operatorname{ess\,sup} L$ over $t \in [0, \infty)$, whereas the right-hand side strictly decreases from ∞ to ε over the same interval. Hence, there exists a unique solution $t = t_0 > 0$, which we call the saddle-point.

We show first that $t_0 \to 0$ as $n \to \infty$. Suppose, for the sake of contradiction, that $t^* := \limsup_{n\to\infty} t_0 > 0$, and let $n_k \nearrow \infty$ be a sequence of indices such that the sequence of the n_k-th saddle points, denoted $t_0^{(k)}$, converge to t^*. Let $\rho_2 : (0,\infty) \to (0,\infty)$ be defined by $\rho_2(t) := (K'(t) - \mathbb{E}[L])/(t\sigma_L^2)$, so $\rho_2(t) \to 1$ as $t \to 0^+$ and

$$K'(t) = \mathbb{E}[L] + \sigma_L^2 t \rho_2(t). \tag{A.611}$$

Note that ρ_2 is a continuous function. Noting that $\varepsilon = \mathbb{E}[L] + b\sigma_L$, rearranging the saddle-point equation yields that

$$\frac{1 + \frac{\sigma_L^2}{\mathbb{E}[L]} t\rho_2(t)}{1 + b\frac{\sigma_L}{\mathbb{E}[L]}} = 1 + \frac{1}{\varepsilon t} + \frac{1}{\varepsilon \cdot (1+t)}. \tag{A.612}$$

Taking $t \in \{t_0^{(k)}\}_{k\in\mathbb{N}}$, letting $k \to \infty$, and recalling the assumptions that $(\mathbb{E}[L], \sigma_L^2) \sim n \cdot (\mathrm{KL}, \mathrm{V})$ for $\mathrm{KL}, \mathrm{V} > 0$ and that $b = o(\sqrt{n})$, we infer from (A.612) that

$$\frac{\mathrm{V} t^* \rho_2(t^*)}{\mathrm{KL}} = 0. \tag{A.613}$$

Equality (A.613) contradicts that $\mathrm{V}, t^*, \rho_2(t^*), \mathrm{KL} > 0$. Thus, we must have that $t^* = 0$.

Consider the reparametrization $t = d/\sigma_L$, so d is a variable over $(0,\infty)$. The saddle-point equation can be rewritten as

$$\left(\rho_2(t) - \frac{b}{\sigma_L}\right) d^2 - \left(b + \frac{2}{\sigma_L}\right) d - \left(1 - \frac{\rho_2(t) d^3}{\sigma_L}\right) = 0. \tag{A.614}$$

We rewrite the saddle-point equation in this "quadratic" form since it closely approximates the quadratic $d^2 - bd - 1 = 0$ at the saddle-point. Indeed, let $d_0 > 0$ be such that $t_0 = d_0/\sigma_L$. We obtain from (A.614) the inequality $\frac{1}{2}d_0^2 - (b+1)d_0 - 1 \le 0$ for all large n. This latter inequality yields that

$$d_0 \le b + 1 + \sqrt{(b+1)^2 + 2} = o(n^{1/6}). \tag{A.615}$$

Hence, $\rho_2(t_0) d_0^3/\sigma_L \to 0$ as $n \to \infty$, i.e., the "constant" term in (A.614) approaches 1. Thus, for all large n, completing the square in (A.614) yields (denoting $t = t_0$, $\rho = \rho_2$, and $\sigma = \sigma_L$ for short)

$$d_0 = \frac{b + \frac{2}{\sigma} + \sqrt{\left(b + \frac{2}{\sigma}\right)^2 + 4\left(1 - \frac{\rho(t) d_0^3}{\sigma}\right)\left(\rho(t) - \frac{b}{\sigma}\right)}}{2\left(\rho(t) - \frac{b}{\sigma}\right)}. \tag{A.616}$$

Taking $n \to \infty$, we obtain

$$d_0 \sim \frac{b + \sqrt{b^2 + 4}}{2}, \tag{A.617}$$

which gives the desired asymptotic formula for the saddle-point $t_0 = d_0/\sigma_L$.

A.20 Contrast between SPA and the Standard CLT

To illustrate the advantage of our tilting approach, we compare the asymptotic behavior of the error in Theorem 2.19 to that obtainable from non-tilted Berry-Esseen. Let $L = L_1 + \cdots + L_n$ for independent PLRVs $L_1, \cdots, L_n$ that satisfy Assumption 2.5. Suppose that Assumption 2.7 holds too.

By the Berry-Esseen theorem, we have for a Gaussian $Z \sim \mathcal{N}(\mathbb{E}[L], \sigma_L^2)$ that[2]

$$\begin{aligned}\delta_L(\varepsilon) &= \mathbb{E}\left[\left(1 - e^{\varepsilon - L}\right)^+\right] && \text{(A.618)}\\ &= \int_0^1 \mathbb{P}\left[L > \varepsilon - \log(1-u)\right] du && \text{(A.619)}\\ &= \delta_Z(\varepsilon) + \theta \cdot \frac{0.56\ \mathrm{P}_0}{\sigma_L^3} && \text{(A.620)}\end{aligned}$$

where $|\theta| \le 1$. By Assumption 2.7, the error term in the standard Berry-Esseen approach shown above satisfies

$$\mathrm{err}_{\mathrm{Standard}}(\varepsilon) := \frac{0.56\ \mathrm{P}_0}{\sigma_L^3} \sim \frac{0.56\ \mathrm{P}}{\mathrm{V}^{3/2} \cdot \sqrt{n}}. \tag{A.621}$$

Thus, the improvement our approach yields is asymptotically (see Theorem 2.19 for the definitions of $C(b)$ and τ)

$$\frac{\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0)}{\mathrm{err}_{\mathrm{Standard}}(\varepsilon)} \sim \frac{2\sqrt{e}}{C(b)^\tau}. \tag{A.622}$$

Even for moderate values of b, the above ratio is very small (recall that we denote $\varepsilon = \mathbb{E}[L] + b\sigma_L$). For example, if $b \approx 6.4$ (so $\delta \approx 10^{-10}$ in the limit; see Theorem 2.16 on the high-composition regime), we obtain the limit of the ratio

$$\lim_{n\to\infty} \frac{\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0)}{\mathrm{err}_{\mathrm{Standard}}(\varepsilon)} \approx 3 \times 10^{-9}. \tag{A.623}$$

In addition, in the complementary regime of $\delta \to 0$, e.g., when $\varepsilon = \mathbb{E}[L] + b\sigma_L$ with $b \ge \sqrt{\log n}$ (and still $b = o(n^{1/6})$), one has that the error term in the standard CLT *dominates* the approximation of δ:

$$\delta_Z(\varepsilon) = o\left(\mathrm{err}_{\mathrm{Standard}}(\varepsilon)\right). \tag{A.624}$$

In contrast, in the same regime, our error term $\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0)$ is always vanishingly smaller than the

[2] Note that Z is not necessarily a PLRV associated to a Gaussian mechanism, since in general $\sigma_L^2 \ne 2\mathbb{E}[L]$.

approximation itself, i.e.,

$$\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0) = o\left(\delta_{L,\,\text{SP-CLT}}(\varepsilon)\right). \tag{A.625}$$

A.21 Proofs of Section 2.19.1

A.21.1 Proof of Proposition 2.6

Denote $K = K_L$ for short. The Gaussian expectation may be computed as

$$\begin{aligned}\mathbb{E}\left[\bar{f}\left(Z-\varepsilon,t\right)\right] = \exp\left(\frac{K''(t)t^2}{2} - (K'(t)-\varepsilon)t\right)\cdot Q\left(\sqrt{K''(t)}\,t - \frac{K'(t)-\varepsilon}{\sqrt{K''(t)}}\right)\\ - \exp\left(\frac{K''(t)(t+1)^2}{2} - (K'(t)-\varepsilon)(t+1)\right)\cdot Q\left(\sqrt{K''(t)}\,(t+1) - \frac{K'(t)-\varepsilon}{\sqrt{K''(t)}}\right).\end{aligned} \tag{A.626}$$

Using $Q(z) = \frac{q(z)}{\sqrt{2\pi}}e^{-z^2/2}$ and the definitions of α, β, γ, we get

$$\mathbb{E}\left[\bar{f}\left(Z-\varepsilon,t\right)\right] = \frac{q(\alpha)-q(\beta)}{\sqrt{2\pi}}\, e^{-\gamma^2/2}. \tag{A.627}$$

Plugging this into the definition of $\delta_{L,\,\text{SP-CLT}}$ completes the proof.

A.21.2 Proof of Proposition 2.7

Let $Z \sim \mathcal{N}\left(K_L'(t), K_L''(t)\right)$ be the variable in the expectation in (2.145). Its PDF is upper bounded by $p_Z(z) \le \frac{1}{\sqrt{2\pi K_L''(t)}}$. Thus

$$\mathbb{E}\left[e^{-t(Z-\varepsilon)}\left(1-e^{-(Z-\varepsilon)}\right)^+\right] = \int_\varepsilon^\infty p_Z(z)e^{-t(z-\varepsilon)}\left(1-e^{-(z-\varepsilon)}\right)dz \tag{A.628}$$

$$\le \frac{1}{\sqrt{2\pi K_L''(t)}}\int_\varepsilon^\infty e^{-t(z-\varepsilon)}\left(1-e^{-(z-\varepsilon)}\right)dz \tag{A.629}$$

$$= \frac{1}{\sqrt{2\pi K_L''(t)}\;t(t+1)}. \tag{A.630}$$

Applying this bound to the definition of $\delta_{L,\,\text{SP-CLT}}(\varepsilon)$ in (2.145) completes the proof.

A.22 Proof of Theorem 2.18

A simple key feature of exponential tilting, stated here without proof, is that it respects addition and independence.

Lemma A.13. *For independent L_j, the exponential tilting of $L = L_1 + \cdots + L_n$ with parameter t is $\widetilde{L} = \widetilde{L}_1 + \cdots + \widetilde{L}_n$, where $\widetilde{L}_j$ is the exponential tilting of L_j with parameter t for each j. Further, $\widetilde{L}_1, \ldots, \widetilde{L}_n$ are independent too.*

Fix $t > 0$. Recall from (2.143) that

$$\delta_L(\varepsilon) = e^{K_L(t)-\varepsilon t}\ \mathbb{E}\left[\bar{f}\left(\widetilde{L}-\varepsilon, t\right)\right] \tag{A.631}$$

where $\widetilde{L}$ is the exponential tilting of L with parameter t, and

$$\bar{f}(x,t) = e^{-xt}(1-e^{-x})^{+} \tag{A.632}$$

Note that $K_L'(t) = \mathbb{E}[\widetilde{L}]$ and $K_L''(t) = \mathrm{Var}[\widetilde{L}]$. We consider the function $\bar{f}(x,t)$. We show next that, for fixed t, $x \mapsto \bar{f}(x,t)$ is a unimodal function with a maximal value of $t^t/(t+1)^{t+1}$. Certainly $\bar{f}(x,t) \geq 0$ for all x. For $x > 0$ the derivative (with respect to x) is

$$\bar{f}'(x,t) = -te^{-tx}(1-e^{-x}) + e^{-tx}e^{-x} \tag{A.633}$$

$$= e^{-tx}\left[-t+(t+1)e^{-x}\right]. \tag{A.634}$$

Note that $-t+(t+1)e^{-x}$ is monotonically decreasing in x, which means that $\bar{f}(x,t)$ is increasing until $-t+(t+1)e^{-x} = 0$, and is subsequently decreasing. In particular, the maximal value of $\bar{f}$ is attained when

$$x = x_0 = -\log\frac{t}{t+1}. \tag{A.635}$$

Note that $x_0 > 0$. Thus, the maximal value of $\bar{f}$ is

$$f_{\max} := \bar{f}(x_0,t) = \bar{f}\left(-\log\frac{t}{t+1}, t\right) \tag{A.636}$$

$$= \left(\frac{t}{t+1}\right)^t\left(1-\frac{t}{t+1}\right) = \frac{t^t}{(t+1)^{t+1}}. \tag{A.637}$$

Thus, between $x = 0$ and $x = x_0$, $\bar{f}(x,t)$ is monotonically increasing from 0 to $f_{\max}$; then from $x = x_0$ to $x = \infty$, $\bar{f}(x,t)$ is monotonically decreasing from $f_{\max}$ to 0. Thus, there exist functions $f_1^{-1}(z)$,

$f_2^{-1}(z)$ such that, for any $z \in (0, f_{\max})$, $\bar{f}(x,t) > z$ if and only if

$$f_1^{-1}(z) < x < f_2^{-1}(z).$$

Therefore,

$$\mathbb{E}[\bar{f}(\widetilde{L}-\varepsilon,t)] = \int_0^{f_{\max}} \mathbb{P}\left[\bar{f}(\widetilde{L}-\varepsilon,t) > z\right] dz \tag{A.638}$$

$$= \int_0^{f_{\max}} \mathbb{P}\left[f_1^{-1}(z) < \widetilde{L}-\varepsilon < f_2^{-1}(z)\right] dz. \tag{A.639}$$

In addition, we may apply the Berry-Esseen theorem to write

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left[\widetilde{L} > x\right] - \mathbb{P}[Z > x]\right| \le \frac{0.56\,\mathrm{P}_t}{K_L''(t)^{3/2}} \tag{A.640}$$

where $Z \sim \mathcal{N}\left(K_L'(t), K_L''(t)\right)$ and P_t is defined in the beginning of Section 2.19. Thus we have the upper bound

$$\delta_L(\varepsilon) = e^{K_L(t)-\varepsilon t}\ \mathbb{E}\left[\bar{f}\left(\widetilde{L}-\varepsilon,t\right)\right] \tag{A.641}$$

$$= e^{K_L(t)-\varepsilon t}\int_0^{f_{\max}} \mathbb{P}\left[f_1^{-1}(z) < \widetilde{L}-\varepsilon < f_2^{-1}(z)\right] dz \tag{A.642}$$

$$\le e^{K_L(t)-\varepsilon t}\left(\frac{1.12 f_{\max}\mathrm{P}_t}{K_L''(t)^{3/2}} + \int_0^{f_{\max}} \mathbb{P}\left[f_1^{-1}(z) < Z-\varepsilon < f_2^{-1}(z)\right] dz\right) \tag{A.643}$$

$$= e^{K_L(t)-\varepsilon t}\ \left(\mathbb{E}\left[\bar{f}\left(Z-\varepsilon,t\right)\right] + \frac{1.12 f_{\max}\mathrm{P}_t}{K_L''(t)^{3/2}}\right) \tag{A.644}$$

Similarly, we have the lower bound

$$\delta_L(\varepsilon) \ge e^{K_L(t)-\varepsilon t}\ \left(\mathbb{E}\left[\bar{f}\left(Z-\varepsilon,t\right)\right] - \frac{1.12 f_{\max}\mathrm{P}_t}{K_L''(t)^{3/2}}\right). \tag{A.645}$$

This completes the proof of the theorem.

A.23 Proof of Theorem 2.19: Asymptotic of the SPA-CLT Approximation Error

We write $K = K_L$ for short. Recall the definition of the error term in (2.150)

$$\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0) = e^{K(t_0)-\varepsilon t_0}\frac{t_0^{t_0}}{(1+t_0)^{1+t_0}} \cdot \frac{1.12\,\mathrm{P}_{t_0}}{K''(t_0)^{3/2}}. \tag{A.646}$$

From the characterization of the saddle-point in Theorem 2.17, we have that

$$t_0 \sim \frac{b+\sqrt{b^2+4}}{2\sigma_L}. \tag{A.647}$$

By Assumption 2.7, we have that $\sigma_L^2 = K''(0) \sim n\mathrm{V}$ as $n \to \infty$. Hence, $t_0 \sim c/\sqrt{n}$ for $c = (b+\sqrt{b^2+4})/(2\mathrm{V}) = o(n^{1/6})$. Thus, by Assumption 2.7 again, $(K''(t_0), \mathrm{P}_{t_0}) \sim n \cdot (\mathrm{V}, \mathrm{P})$. As we also have that $t_0 \to 0$, we conclude that

$$\frac{t_0^{t_0}}{(1+t_0)^{1+t_0}} \cdot \frac{1.12\,\mathrm{P}_{t_0}}{K''(t_0)^{3/2}} \sim \frac{1.12\,\mathrm{P}}{\mathrm{V}^{3/2}\cdot\sqrt{n}}. \tag{A.648}$$

Thus, it only remains to analyze the asymptotic of $\exp\left(K(t_0) - \varepsilon t_0\right)$.

We use the following Taylor expansion of K around 0:

$$K(t_0) = t_0 \cdot \mathbb{E}[L] + \frac{t_0^2}{2}\cdot\sigma_L^2 + \frac{t_0^3}{6}\cdot K'''(\xi), \tag{A.649}$$

where $0 \le \xi \le t_0$. Using $\varepsilon = \mathbb{E}[L] + b\sigma_L$, and writing $t_0 = d_0/\sigma_L$ (so $d_0 \sim (b+\sqrt{b^2+4})/2$ by (A.647)), we obtain

$$K(t_0) - \varepsilon t_0 = \frac{d_0^2}{2} - bd_0 + \frac{d_0^3 K'''(\xi)}{6\sigma_L^3}. \tag{A.650}$$

Now, note that $K'''(\xi) = \sum_{j=1}^n K_{L_j}'''(\xi)$. Thus, applying the triangle inequality, we obtain that $|K'''(\xi)| \le \mathrm{P}_\xi$. As $0 \le \xi \le t_0$, Assumption 2.7 yields that $|K'''(\xi)| = O(n)$. As $\sigma_L = \Theta(\sqrt{n})$, and $d_0 = o(n^{1/6})$, we infer that

$$\frac{d_0^3 K'''(\xi)}{6\sigma_L^3} \to 0 \tag{A.651}$$

as $n \to \infty$. Hence,

$$\exp\left(K(t_0) - \varepsilon t_0\right) \sim \exp\left(\frac{d_0^2}{2} - bd_0\right). \tag{A.652}$$

Writing $d_0 = \tau_0 \cdot (b+\sqrt{b^2+4})/2$, so $\tau_0 > 0$ and $\tau_0 \to 1$ by (A.647), then collecting terms, we obtain

$$\frac{d_0^2}{2} - bd_0 = \frac{\tau_0^2}{2} - (2-\tau_0)\tau_0 \cdot \frac{b^2 + b\sqrt{b^2+4}}{4}. \tag{A.653}$$

Therefore, we obtain that

$$\exp\left(K(t_0) - \varepsilon t_0\right) \sim \frac{\sqrt{e}}{C(b)^\tau} \tag{A.654}$$

where $\tau := (2-\tau_0)\tau_0 \to 1$. Putting the asymptotics shown above together, we conclude that

$$\mathrm{err}_{\mathrm{SP}}(\varepsilon; t_0) \sim \frac{1.12\sqrt{e}\,\mathrm{P}}{\mathrm{V}^{3/2}\cdot C(b)^\tau \cdot \sqrt{n}}, \tag{A.655}$$

as desired.

A.24 Instantiation of the Saddle-point Accountant

The algorithm SaddlePointAccountant (Algorithm 2), giving the workflow of the versions of the SPA, is presented here.

Algorithm 2 : SaddlePointAccountant (SPA)

```
Input: A finite set $\mathcal{E} \subset [0,\infty)$ (values of $\varepsilon$), and tightly dominating distributions $(P_1,Q_1),\dots,(P_n,Q_n)$.
Output: Four approximations $\delta^{(k)}_{L,\text{SP-MSD}}$, $1 \le k \le 3$, and $\delta_{L,\text{SP-CLT}}$ of the privacy curve $\delta_L$, and an error bound so that $|\delta_L(\varepsilon) - \delta_{L,\text{SP-CLT}}(\varepsilon)| \le \text{err}_{\text{SP}}(\varepsilon)$.
$L_j \leftarrow \log \frac{dP_j}{dQ_j}(X_j)$ where $X_j \sim P_j$        $j \in [n]$
$K_{L_j}(t) \leftarrow \log \mathbb{E}\left[e^{tL_j}\right]$        $j \in [n]$
$L \leftarrow L_1 + \cdots + L_n$
$K_L \leftarrow K_{L_1} + \cdots + K_{L_n}$
for $\varepsilon \in \mathcal{E}$ do
    $t_0 \leftarrow$ positive solution to $K_L'(t_0) = \varepsilon + \frac{1}{t_0} + \frac{1}{t_0+1}$
    $F_\varepsilon(t) \leftarrow K_L(t) - \varepsilon t - \log t - \log(t+1)$
    $\beta_{\varepsilon,2} \leftarrow \frac{1}{8}\frac{F_\varepsilon^{(4)}(t_0)}{F_\varepsilon''(t_0)^2}$
    $\beta_{\varepsilon,3} \leftarrow -\frac{5}{24}\frac{F_\varepsilon^{(3)}(t_0)^2}{F_\varepsilon''(t_0)^3} - \frac{1}{48}\frac{F_\varepsilon^{(6)}(t_0)}{F_\varepsilon''(t_0)^3}$
    $\delta^{(1)}_{L,\text{SP-MSD}}(\varepsilon) \leftarrow \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F_\varepsilon''(t_0)}}$
    $\delta^{(2)}_{L,\text{SP-MSD}}(\varepsilon) \leftarrow \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F_\varepsilon''(t_0)}}(1+\beta_{\varepsilon,2})$
    $\delta^{(3)}_{L,\text{SP-MSD}}(\varepsilon) \leftarrow \frac{e^{F_\varepsilon(t_0)}}{\sqrt{2\pi F_\varepsilon''(t_0)}}(1+\beta_{\varepsilon,2}+\beta_{\varepsilon,3})$
    $\gamma \leftarrow \frac{K_L'(t_0)-\varepsilon}{\sqrt{K_L''(t_0)}}$
    $(\alpha,\beta) \leftarrow \left(\sqrt{K_L''(t_0)}\; t_0 - \gamma, \sqrt{K_L''(t_0)}\;(t_0+1) - \gamma\right)$
    $\delta_{L,\text{SP-CLT}}(\varepsilon) \leftarrow e^{K_L(t_0)-\varepsilon t_0-\gamma^2/2}\frac{q(\alpha)-q(\beta)}{\sqrt{2\pi}}$
    $\widetilde{L}_j \leftarrow$ exp. tilt of $L_j$ with parameter $t_0$,        $j \in [n]$
    $\mathrm{P}_{t_0} \leftarrow \sum_{j\in[n]} \mathbb{E}\left[\left|\widetilde{L}_j - K_{L_j}'(t_0)\right|^3\right]$
    $\text{err}_{\text{SP}}(\varepsilon) \leftarrow e^{K_L(t_0)-\varepsilon t_0}\frac{t_0^{t_0}}{(1+t_0)^{1+t_0}} \cdot \frac{1.12\,\mathrm{P}^{(n)}_{t_0}}{K_L''(t_0)^{3/2}}$
end for
Return: $\delta^{(k)}_{L,\text{SP-MSD}}, 1 \le k \le 3, \delta_{L,\text{SP-CLT}}$, $\text{err}_{\text{SP}}$.
```

A.25 Ground-Truth Curve Computation

We explain here how the ground-truth curve in Figure 2.7 is computed. Since the setting there is for self-composition, we employ that here too. So, let $L_1, \cdots, L_n$ be i.i.d. PLRVs for the subsampled Gaussian mechanism, and consider the PLRV $L = L_1 + \cdots + L_n$ for the composed mechanism.

Recall that the saddle-point accountant gives various approximations to the contour integral given in Theorem 2.15, which we copy here:

$$\delta_{L_1}(\varepsilon) = \frac{1}{2\pi i} \int_{t-i\infty}^{t+i\infty} e^{F_\varepsilon(z)}\, dz \tag{A.656}$$

where the function F_ε is defined as:

$$F_\varepsilon(z) = K_{L_1}(z) - \varepsilon z - \log z - \log(1+z). \tag{A.657}$$

After n compositions, the contour integral becomes:

$$\delta_L(\varepsilon) = \frac{1}{2\pi i} \int_{t-i\infty}^{t+i\infty} e^{nK_{L_1}(z) - \varepsilon z - \log z - \log(1+z)}\, dz. \tag{A.658}$$

Recall that this formula holds for any value of $t > 0$.

The ground-truth in (A.658) is then computed via standard numerical integration, which evidently is a time-consuming process, yet it is one that can produce a reference value to relatively compare accountants' accuracies.

Let $P = \mathcal{N}(0, \sigma^2)$, $Q = (1-\lambda)\mathcal{N}(0,\sigma^2) + \lambda\mathcal{N}(s,\sigma^2)$. The composed subsampled Gaussian has the PLRV $L = L_1 + \cdots + L_n$, where the L_j are independent and (see Lemma 2.4)

$$\begin{aligned} L_j &= \log \frac{dQ}{dP}(X) = \log\left(1 - \lambda + \lambda e^{s(2X-s)/(2\sigma^2)}\right), \\ X &\sim (1-\lambda)\mathcal{N}(0,\sigma^2) + \lambda\mathcal{N}(s,\sigma^2). \end{aligned} \tag{A.659}$$

In addition, the MGF of L_1 may be written as

$$\begin{aligned} M_{L_1}(z) &= \mathbb{E}[e^{zL_1}] && \text{(A.660)} \\ &= \mathbb{E}_{X\sim Q}\left[\left(\frac{dQ}{dP}(X)\right)^z\right] && \text{(A.661)} \\ &= \mathbb{E}_{X\sim P}\left[\left(\frac{dQ}{dP}(X)\right)^{z+1}\right] && \text{(A.662)} \\ &= \int_{-\infty}^{\infty} \left(1 - \lambda + \lambda e^{s(2x-s)/(2\sigma^2)}\right)^{z+1} dP(x). && \text{(A.663)} \end{aligned}$$

Recall that the CGF is given by

$$K_{L_1}(z) = \log M_{L_1}(z). \tag{A.664}$$

Plugging in the log integral (A.664) into the contour integral (A.658), the contour integral can be directly computed using standard numerical libraries. We note that this calculation is very slow, as the integrand in (A.658) itself involves an integral over $\mathbb{R}$. Moreover, we numerically invert this function via bisection to obtain the curve described in Figure 2.7. This ground-truth curve was computed on a 64-core cluster using multi-processing to distribute the workload, and took a wall-time of 45 minutes. This amounts to a runtime of 48 CPU hours. In contrast, all other accountants run in the order of seconds on a commercial laptop.

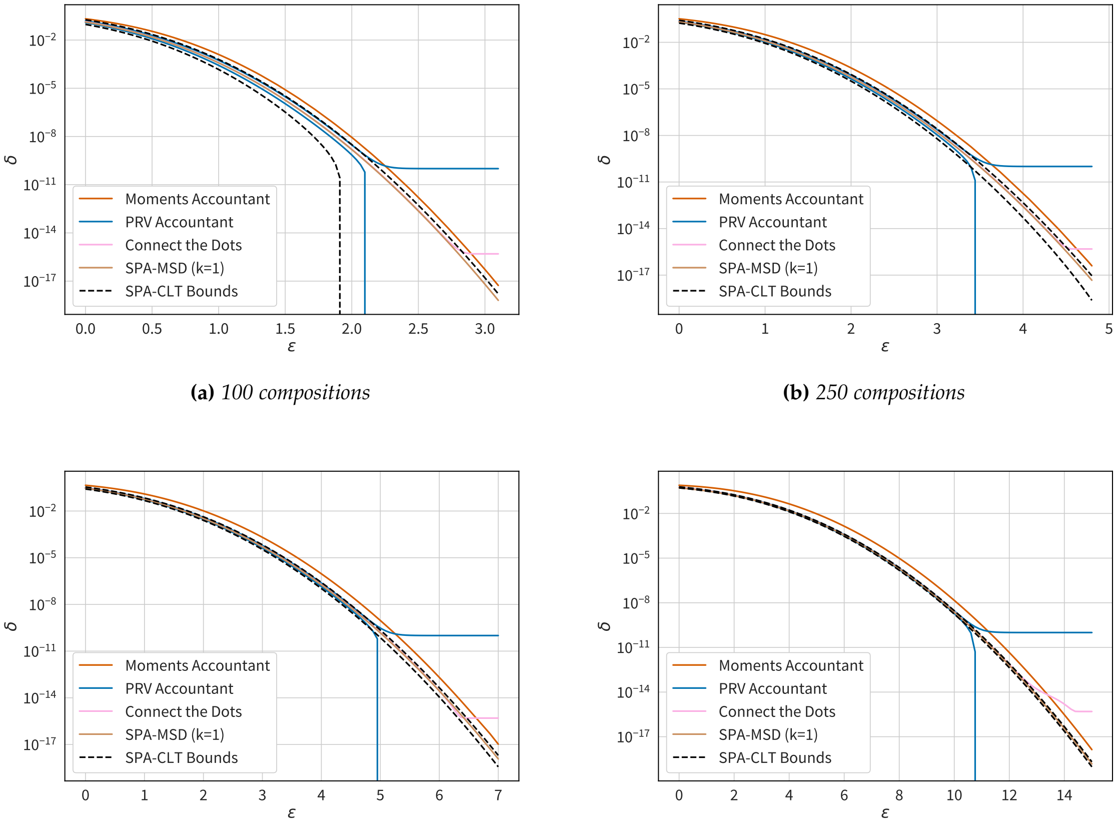
A.26 Additional Numerical Experiments

We provide further experiments exploring the flexibility of the saddle-point accountant. We show that the SPA-MSD approximations can be accurate even in the moderate-composition regime, though the SPA-CLT bounds become loose for a small number of compositions. We demonstrate this using parameters used by a real-world application of DP on the image classification SGD algorithm in [DBH+22], which uses the subsampled Gaussian as the DP mechanism. In particular, we use the noise scale $\sigma = 9.4$ and subsampling rate $\lambda = 2^{14}/50000$, as these were the values that allowed a 40-layer Wide-ResNet to achieve a new SOTA accuracy of 81.4% on CIFAR-10 under $(\varepsilon = 8, \delta = 10^{-5})$-DP. This algorithm went up to $n = 2000$ compositions to achieve this SOTA.

First, we plot the (ε, δ)-curves at $n \in \{100, 250, 500, 2000\}$ compositions in Figure A.1. We observe that the CLT bounds get tighter as the number of compositions increases, but the order-1 SPA-MSD remains consistently accurate for all presented compositions and values of δ.

Second, we demonstrate the accuracy of the order-1 SPA-MSD for all compositions less than 2000 in Figure A.2, where we fix $\delta = 10^{-5}$, vary the number of compositions, and plot the resulting value of ε.

These two plots verify that the order-1 SPA-MSD is much more accurate than the CLT bounds suggest.



(a) *100 compositions*

(b) *250 compositions*

(c) *500 compositions*

(d) *2000 compositions*

Figure A.1: *Accounting for the composition of $n \in \{100, 250, 500, 2000\}$ subsampled Gaussian mechanisms, with noise scale $\sigma = 9.4$ and subsampling rate $\lambda = 2^{14}/50000$. The PRV Accountant [GLW21] discretization parameters are $\varepsilon_{\text{error}} = 0.1, \delta_{\text{error}} = 10^{-10}$. The Connect the Dots [DGK⁺22] discretization interval length is 0.005.*

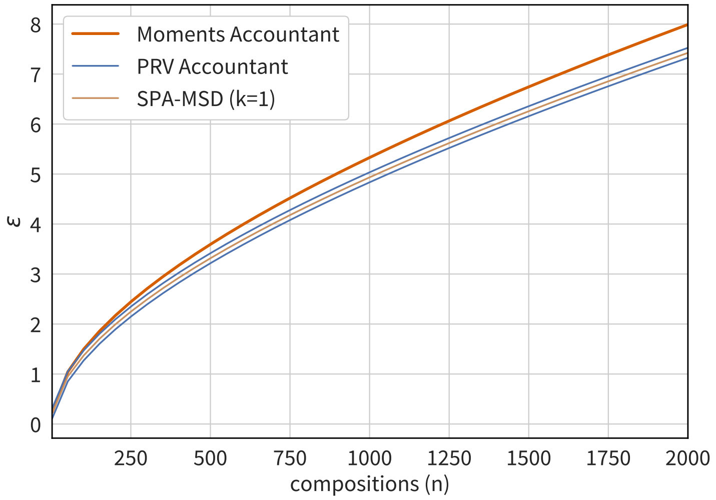


Figure A.2: *Privacy budget ε of the subsampled Gaussian mechanism after* $1 \leq n \leq 2000$ *compositions using the order-*1 *SPA-MSD, the Moments Accountant, and the PRV Accountant [GLW21]. We use subsampling* $\lambda = 2^{14}/50000$*, noise scale* $\sigma = 9.4$*, and* $\delta = 10^{-5}$*. The discretization parameters for the PRV Accountant are* $\varepsilon_{\text{error}} = 0.1, \delta_{\text{error}} = 10^{-10}$.

Appendix B

Appendix to Chapter 3

The theoretical details of Chapter 3 are included in this appendix. The outline is as follows:

- Appendix B.1: we present the proofs of Section 3.4. Specifically, we prove Theorems 3.1 and 3.2.
- Appendix B.2: we prove the strong duality stated in Theorem 3.3 from Section 3.5.
- Appendix B.3: we prove the theoretical properties of Algorithm 1 stated in Section 3.6.

B.1 Proofs of Section 3.4: Existence, Uniqueness, and the Formula for Model Projection

We prove Theorems 3.1 and 3.2 in this appendix. Namely, we show that under Assumption 3.1 model projection exists and is unique, and we also derive its formula. The proof is lengthy, so we divide it into several subsections. The outline of this appendix is as follows:

- In Appendix B.1.1, we lay the groundwork for the proof. We set up the generalized optimization problem (B.3) over general Banach spaces $\mathcal{C}(\mathcal{X})$ of continuous functions, and introduce the relevant notation and assumptions.
- In Appendix B.1.2, we state general results on the general optimization problem (B.3).
- In Appendix B.1.3, we prove Theorems 3.1–3.2 using the general results stated in Appendix B.1.2.
- In Appendices B.1.4–B.1.6, we prove the general results stated previously in Appendix B.1.2.

B.1.1 Notation and Setup

For the starting points below, we assume only that $\mathcal{X}$ is a topological space (i.e., we do not assume $\mathcal{X} = \mathbb{R}^m$ yet). Let $\mathcal{C}(\mathcal{X})$ denote the Banach space of continuous and bounded functions

$$\mathcal{C}(\mathcal{X}) := \left\{ \boldsymbol{h} : \mathcal{X} \to \mathbb{R}^C \mid \boldsymbol{h} \text{ continuous and } \sup_{x \in \mathcal{X}} \|\boldsymbol{h}(x)\|_1 < \infty \right\}, \tag{B.1}$$

which is Banach when equipped with the sup norm, i.e., for $\boldsymbol{h} \in \mathcal{C}(\mathcal{X})$

$$\|\boldsymbol{h}\|_\infty := \sup_{x \in \mathcal{X}} \|\boldsymbol{h}(x)\|_1. \tag{B.2}$$

Consider the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{C}(\mathcal{X})} \quad & \mathbb{E}\left[F(X, \boldsymbol{h}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] \leq 0, \qquad k \in [K], \end{aligned} \tag{B.3}$$

where F and $G_1, \cdots, G_K$ are functions defined on $\mathcal{X} \times \mathbb{R}^C$ and taking values in $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. We denote by $\mathcal{C}(\mathcal{X}, \mathcal{Z}) \subset \mathcal{C}(\mathcal{X})$, for $\mathcal{Z} \subset \mathbb{R}^C$, the subset of functions taking values in $\mathcal{Z}$, i.e.,

$$\mathcal{C}(\mathcal{X}, \mathcal{Z}) := \{\boldsymbol{h} \in \mathcal{C}(\mathcal{X}) \mid \boldsymbol{h}(x) \in \mathcal{Z} \text{ for every } x \in \mathcal{X}\}. \tag{B.4}$$

Note that $\mathcal{C}(\mathcal{X}, \mathcal{Z})$ is closed or convex if $\mathcal{Z}$ is closed (in $\mathbb{R}^C$) or convex, respectively. Therefore, $\mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C)$ is a convex complete metric space (for any $\mathcal{X}$). However, it is not compact in general. Therefore, it might not be straightforward to tackle the optimization problem (B.3) even when restricted to only $\mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C)$. Therefore, we tackle (B.3) indirectly via solving a much more restricted problem of the form

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{K}} \quad & \mathbb{E}\left[F(X, \boldsymbol{h}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] \leq 0, \qquad k \in [K] \end{aligned} \tag{B.5}$$

for a compact subset $\mathcal{K} \subset \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C)$ then showing that the problem (B.5) produces a global optimizer. We also consider ε-truncations of the simplex

$$\boldsymbol{\Delta}_C^\varepsilon := \boldsymbol{\Delta}_C \cap [\varepsilon, 1]^C \tag{B.6}$$

and the corresponding space

$$\mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C) := \bigcup_{\varepsilon > 0} \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C^\varepsilon). \tag{B.7}$$

We set

$$\mathbf{\Delta}_C^+ := \mathbf{\Delta}_C \cap (0,1]^C. \tag{B.8}$$

We let $\mathcal{F}$ denote the feasibility region in (B.3), i.e.,

$$\mathcal{F} := \left\{ \boldsymbol{h} \in \mathcal{C}(\mathcal{X}) \mid \max_{k \in [K]} \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] \leq 0 \right\}. \tag{B.9}$$

We denote by $\mathcal{S}$ the strict-feasibility region, i.e.,

$$\mathcal{S} := \left\{ \boldsymbol{h} \in \mathcal{C}(\mathcal{X}) \mid \max_{k \in [K]} \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] < 0 \right\}. \tag{B.10}$$

We let $\mathcal{D}$ be the set of functions in $\mathcal{C}(\mathcal{X})$ at which the objective function and the constraints are integrable[1]

$$\mathcal{D} := \left\{ \boldsymbol{h} \in \mathcal{C}(\mathcal{X}) \mid \max\left(\mathbb{E}\left[|F(X, \boldsymbol{h}(X))|\right], \max_{k \in [K]} \mathbb{E}\left[|G_k(X, \boldsymbol{h}(X))|\right] \right) < \infty \right\}. \tag{B.11}$$

For a function $\psi : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the domain of ψ is the set of points at which ψ is defined and finite

$$\mathbf{dom}\ \psi := \{ v \in \mathcal{V} \mid \psi(v) < \infty \}. \tag{B.12}$$

We define the intersection of domains

$$D := \bigcap_{x \in \mathcal{X}} \left\{ \boldsymbol{p} \in \mathbb{R}^C \mid \max(F(x, \boldsymbol{p}), G_1(x, \boldsymbol{p}), \cdots, G_K(x, \boldsymbol{p})) < \infty \right\}. \tag{B.13}$$

We denote the convex hull and closure of a set $\mathcal{A}$ by $\mathrm{co}(\mathcal{A})$ and $\overline{\mathcal{A}}$, respectively. Abusing notation, we will also denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. We denote the indicator function of a set $\mathcal{U} \subset \mathcal{C}(\mathcal{X})$ by $\mathbb{I}_{\mathcal{U}}$

$$\mathbb{I}_{\mathcal{U}}(h) := \begin{cases} 0 & \text{if } h \in \mathcal{U}, \\ \infty & \text{otherwise.} \end{cases} \tag{B.14}$$

We define extended functionals $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_K : \mathcal{C}(\mathcal{X}) \to \overline{\mathbb{R}}$ by

$$\mathcal{A}(\boldsymbol{h}) := \mathbb{E}\left[F(X, \boldsymbol{h}(X))\right] + \mathbb{I}_{\mathcal{D}}(\boldsymbol{h}), \tag{B.15}$$

$$\mathcal{B}_k(\boldsymbol{h}) := \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] + \mathbb{I}_{\mathcal{D}}(\boldsymbol{h}), \quad k \in [K]. \tag{B.16}$$

In these definitions, it is understood that the value ∞ is assigned outside the set $\mathcal{D}$ regardless of whether the original function is defined and regardless of its value if it is defined there, e.g., if

[1] We say that a function $V : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is integrable if $\mathbb{E}\left[|V(X)|\right] < \infty$.

$\boldsymbol{h} \in \mathcal{C}(\mathcal{X})$ is such that $F(\,\cdot\,, \boldsymbol{h}(\,\cdot\,))$ is not integrable or if its integral is $-\infty$ then $\mathcal{A}(\boldsymbol{h})$ is defined to be ∞ because $\boldsymbol{h} \notin \mathcal{D}$. For $\beta : \mathbb{R}^n \to \mathbb{R}$ and $\ell \in [n]$, the notation $\partial_\ell \beta$ will refer to the partial derivative of β with respect to its ℓ-th input.

Recall the definition of the convex conjugate (see equation (3.14)).

Definition B.1 (Convex Conjugate)**.** The convex conjugate of a proper[2] function $W : \boldsymbol{\Delta}_C \to \overline{\mathbb{R}}$ is the function $W^{\mathrm{conj}} : \mathbb{R}^C \to \overline{\mathbb{R}}$ defined by

$$W^{\mathrm{conj}}(\boldsymbol{v}) := \sup_{\boldsymbol{q} \in \boldsymbol{\Delta}_C} \langle \boldsymbol{v}, \boldsymbol{q} \rangle - W(\boldsymbol{q}). \tag{B.17}$$

The convex conjugate of an f-divergence $D_f(\,\cdot\, \parallel \boldsymbol{p})$ is denoted by $D_f^{\mathrm{conj}}(\,\cdot\,, \boldsymbol{p})$. If, for a fixed $\boldsymbol{v}$, the maximum in (B.17) is attained at a unique point, then we denote that point by $\boldsymbol{q}^{\mathrm{conj}}(\boldsymbol{v})$.

We will prove results under some subset of assumptions that we introduce here and in the beginning of the following section. The first set of assumptions has to do with the well-definedness of our optimization problem, and it will be sufficient to develop the general theory.

Assumption B.1. *The functions $\{F, G_1, \cdots, G_K\}$ in* (B.3)*, and the feasibility set $\mathcal{D}$ in* (B.11) *satisfy:*

(a) the set $\mathcal{D}$ is nonempty,

(b) for $J \in \{F, G_1, \cdots, G_K\}$, $\inf_{\boldsymbol{h} \in \mathcal{D}} J(\,\cdot\,, \boldsymbol{h}(\,\cdot\,))$ is lower bounded by an integrable function,

(c) for $J \in \{F, G_1, \cdots, G_K\}$ and $x \in \mathcal{X}$, the function $J(x, \,\cdot\,)$ is lower-semicontinuous,

(d) the functions $\boldsymbol{q} \mapsto F(x, \boldsymbol{q})$ are strictly convex, and the functions $\boldsymbol{q} \mapsto G_k(x, \boldsymbol{q})$ are convex.

Remark B.1. Note that item (b) of Assumption B.1 is satisfied if, e.g., the functions $F, G_1, \cdots, G_K$, are lower bounded by a constant.

Next, we show how a unique minimizer of (B.5) can be obtained from the dual problem. This procedure is possible thanks to Sion's minimax theorem. It will be useful to introduce the following quantities. First, the following term will bound the norm of optimal dual variables corresponding to the dual of the optimization problem (B.3).

Definition B.2. For $\boldsymbol{q} \in \mathcal{S} \cap \mathcal{D}$, we define

$$\theta_{\boldsymbol{q}} := \frac{\mathcal{A}(\boldsymbol{q}) - \inf_{\boldsymbol{h} \in \mathcal{D}} \mathcal{A}(\boldsymbol{h})}{-\max_{k \in [K]} \mathcal{B}_k(\boldsymbol{q})}. \tag{B.18}$$

[2] We say W is proper if **dom** W is nonempty.

Under item (b) of Assumption B.1, $\theta_q \in [0, \infty)$. Next, we define the Lagrangian of the optimization problem (B.3).

Definition B.3. Define the Lagrangian function $\mathcal{L} : \mathcal{D} \times \mathbb{R}_+^K \to \mathbb{R}$ by

$$\mathcal{L}(\boldsymbol{h}, \boldsymbol{\lambda}) := \mathbb{E}\left[F(X, \boldsymbol{h}(X)) + \sum_{k \in [K]} \lambda_k G_k(X, \boldsymbol{h}(X)) \right] = \mathcal{A}(\boldsymbol{h}) + \sum_{k \in [K]} \lambda_k \mathcal{B}_k(\boldsymbol{h}). \tag{B.19}$$

We use the following notation for what will be shown to be a class of models that contains the optimal model.

Definition B.4. For fixed $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ and $\mathcal{Z} \subset D$, define $\boldsymbol{q}_{\boldsymbol{\lambda}}^{\mathcal{Z}} : \mathcal{X} \to \mathcal{Z}$ by

$$\boldsymbol{q}_{\boldsymbol{\lambda}}^{\mathcal{Z}}(x) := \operatorname*{argmin}_{\boldsymbol{p} \in \mathcal{Z}} \; F(x, \boldsymbol{p}) + \sum_{k \in [K]} \lambda_k G_k(x, \boldsymbol{p}), \qquad \text{for every} \;\; x \in \mathcal{X}, \tag{B.20}$$

if the minimization in (B.20) has a unique solution for every $x \in \mathcal{X}$.

Remark B.2. One way to guarantee the well-definedness of $\boldsymbol{q}_{\boldsymbol{\lambda}}^{\mathcal{Z}}$, for any fixed $\boldsymbol{\lambda} \in \mathbb{R}_+^K$, is to ensure $\mathcal{Z}$ is a nonempty convex and compact set, each $F(x, \cdot)$ is lower-semicontinuous and strictly convex, and each $G_k(x, \cdot)$ is lower-semicontinuous and convex. Indeed, under such assumptions, each mapping $\boldsymbol{p} \mapsto F(x, \boldsymbol{p}) + \sum_{k \in [K]} \lambda_k G_k(x, \boldsymbol{p})$ is lower-semicontinuous and strictly convex, which is then uniquely minimized over the convex and compact set $\mathcal{Z}$.

B.1.2 A Generalized Result and Proof Technique

We present in this section generalized results on the general optimization problem (B.3). These general results will be combined to prove Theorems 3.1 and 3.2 in the next subsection (Appendix B.1.3). Specifically, in this subsection, we:

- Derive a general duality result in Theorem B.1 for the problem (B.3) *conditioned* on the precompactness of the set $\mathcal{Q}$ of potentially optimal models. The proof of this theorem is relegated to Appendix B.1.4.

- Prove in Theorem B.2 that the set $\mathcal{Q}$ of potentially optimal models is indeed precompact. The proof of this theorem is relegated to Appendix B.1.5.

- Derive formulas for the convex conjugate in Lemma B.2. The proof of this lemma is relegated to Appendix B.1.6.

Then, we combine Theorems B.1–B.2 and Lemma B.2 to prove Theorems 3.1–3.2 in Appendix B.1.3.

The main theorem underlying our results is the following general result on optimizers of the problem (B.3).

Theorem B.1. *Suppose Assumption B.1 holds. Let $\mathcal{Z} \subset D$ be convex and compact such that $\mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{S}$ is nonempty, say $\boldsymbol{p} \in \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{S}$, and set $\Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}_+^K \mid \|\boldsymbol{\lambda}\|_1 \leq \theta_{\boldsymbol{p}}\}$. If*

$$\mathcal{Q} := \left\{ \boldsymbol{q}_{\boldsymbol{\lambda}}^{\mathcal{Z}} \mid \boldsymbol{\lambda} \in \Lambda \right\} \tag{B.21}$$

is precompact and $\mathcal{Q} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z}) \subset \mathcal{D}$, then the problem

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{C}(\mathcal{X}, \mathcal{Z})} \quad & \mathbb{E}\left[F(X, \boldsymbol{h}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[G_k(X, \boldsymbol{h}(X))\right] \leq 0, \qquad k \in [K] \end{aligned} \tag{B.22}$$

has a unique solution, and this solution is $\boldsymbol{q}_{\boldsymbol{\lambda}^\star}^{\mathcal{Z}}$ where $\boldsymbol{\lambda}^\star$ is any solution of

$$\sup_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}(\boldsymbol{q}_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}). \tag{B.23}$$

Proof. See Appendix B.1.4. □

We apply Theorem B.1 to the problem of model projection. An intermediate step is that in which separability of the objective function F and linearity of the constraining functions G_k are assumed. We will be interested in the situation $\mathcal{Z} \subset [0,1]^C$. We introduce the following assumptions.

Assumption B.2. *The functions F and $G_1, \cdots, G_K$ satisfy the following:*

(a) For each $x \in \mathcal{X}$, the function $F(x, \cdot)$ is separable and can be written as

$$F(x, \boldsymbol{p}) = \sum_{c \in [C]} f_c(x, p_c) \tag{B.24}$$

for $\mathcal{C}^1(\mathbb{R})$ strictly convex functions $f_c(x, \cdot)$ satisfying $\lim_{t_0 \to 0^+} \frac{\partial f_c}{\partial t}(x, t_0) = -\infty$.

(b) For each fixed $(k, x) \in [K] \times \mathcal{X}$ the function $G_k(x, \cdot)$ is linear, i.e.,

$$G_k(x, \boldsymbol{q}) = \boldsymbol{q}^T \boldsymbol{g}_k(x). \tag{B.25}$$

Further, for each $k \in [K]$ the function $\boldsymbol{g}_k : \mathcal{X} \to \mathbb{R}^C$ is continuous. We denote

$$\boldsymbol{G} = \left(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_K\right)^T. \tag{B.26}$$

Note that item (a) of Assumption B.2 implies that $t_0 \mapsto (\partial f_c/\partial t)(x, t_0)$ is strictly increasing for fixed $(c, x) \in [C] \times \mathcal{X}$, so it is invertible. We let φ_c denote the inverse, i.e., define $\varphi_c(x, \cdot\,) : (-\infty, \partial_{m+1} f_c(x, 1^-)) \to \mathbb{R}^C_{>0}$ by

$$\left(\partial_{m+1} f_c\right)(x, \varphi_c(x, u)) = u. \tag{B.27}$$

Each $\varphi_c(x, \cdot\,)$ is continuous and strictly increasing, so it is a bijection from its domain to $(0, 1)$. Therefore, fixing $x \in \mathcal{X}$, for any $\boldsymbol{a} \in \mathbb{R}^C$ the mapping

$$\gamma \mapsto \sum_{c \in [C]} \varphi_c\left(x, \gamma + a_c\right) \tag{B.28}$$

is a strictly increasing continuous bijection from an interval $\mathcal{I}_1 = (-\infty, \tau_1)$ to another $\mathcal{I}_2 = (0, \tau_2)$ where $\tau_2 > 1$. We define $\gamma : \mathcal{X} \times \mathbb{R}^K \to \mathbb{R}$ implicitly by

$$\sum_{c \in [C]} \varphi_c\left(x, \gamma(x, \boldsymbol{\lambda}) + v_c(x; \boldsymbol{\lambda})\right) = 1. \tag{B.29}$$

Note that we allow $\boldsymbol{\lambda}$ with negative coordinates in the definition of $\gamma(x, \boldsymbol{\lambda})$. Recall that we set $\boldsymbol{v}(x; \boldsymbol{\lambda}) = -\boldsymbol{G}(x)^T \boldsymbol{\lambda}$.

In the sequel, we will take the f_j to be the following functions. For any $(c, x, t) \in [C] \times \mathcal{X} \times [0, 1]$,

$$f_c(x, t) := h_c^{\text{base}}(x) f\left(\frac{t}{h_c^{\text{base}}(x)}\right). \tag{B.30}$$

Then, $F(x, \boldsymbol{p}) = \sum_{c \in [C]} f_c(x, p_c)$ satisfies

$$F(x, \boldsymbol{p}) = D_f(\boldsymbol{p} \,\|\, \boldsymbol{h}^{\text{base}}(x)). \tag{B.31}$$

We denote, under the assumption $f'(0^+) = -\infty$, the inverse of f' by ϕ. Then,

$$\varphi_c(x, t) = h_c^{\text{base}}(x) \phi(t). \tag{B.32}$$

We extend the definition of the f-divergence so that for any $\boldsymbol{p}, \boldsymbol{q} \in [0, 1]^C$ with $\boldsymbol{q} > \boldsymbol{0}$

$$D_f(\boldsymbol{p} \| \boldsymbol{q}) = \sum_{c \in [C]} q_c f\left(\frac{p_c}{q_c}\right). \tag{B.33}$$

We will repeatedly use the following bound on the values of φ_c.

Lemma B.1. *Fix $\boldsymbol{y} \in \boldsymbol{\Delta}_C^+$, and let $f : (0, \infty) \to \mathbb{R}$ be strictly convex and continuously differentiable over $(0, \infty)$ such that $f'(0^+) = -\infty$ and denote the inverse of its derivative by ϕ. For each $c \in [C]$, define $f_c : [0, 1] \to \overline{\mathbb{R}}$ by $f_c(t) = y_c f(t/y_c)$, and let $\varphi_c : (-\infty, f'(1/y_c)] \to (0, 1]$ be the inverse of f_c'. Let*

$v \in \mathbb{R}^C$ and $\theta \in \mathbb{R}_+$ be such that $\|v\|_\infty \leq \theta$, and let $\gamma \in \mathbb{R}$ be the unique real number such that $\sum_{c \in [C]} \varphi_c(\gamma + v_c) = 1$. Then,

$$\min_{c \in [C]} \varphi_c(\gamma + v_c) \geq \phi\left(f'\left(\frac{1}{C}\right) - 2\theta\right) \cdot \min_{c \in [C]} y_c. \tag{B.34}$$

Proof. For each $c \in [C]$, let $\beta_c = \gamma + v_c$. Then, $|\beta_i - \beta_j| \leq 2\theta$ for every $(i, j) \in [C]^2$. Since $\sum_{c \in [C]} \varphi_c(\beta_c) = 1$, there exists at least one $a \in [C]$ such that

$$\varphi_a(\beta_a) \geq \frac{1}{C}. \tag{B.35}$$

Therefore, $\beta_a \geq f'_a(1/C)$. Furthermore, $f'_a(1/C) = f'(1/(Cy_a)) \geq f'(1/C)$. Then,

$$\min_{c \in [C]} \beta_c \geq f'\left(\frac{1}{C}\right) - 2\theta. \tag{B.36}$$

Finally, we have

$$\min_{c \in [C]} \varphi_c(\beta_c) \geq \min_{c \in [C]} \varphi_c\left(\min_{i \in [C]} \beta_i\right) \geq \min_{c \in [C]} \varphi_c\left(f'\left(\frac{1}{C}\right) - 2\theta\right) = \min_{c \in [C]} y_c \phi\left(f'\left(\frac{1}{C}\right) - 2\theta\right), \tag{B.37}$$

where the last step is because $\varphi_c(t) = y_c \phi(t)$. □

In view of this result, we will employ the following notation. Write

$$y_{\min} := \inf_{x,c} \; h_c^{\text{base}}(x), \tag{B.38}$$

and, for $\theta > 0$, let

$$t_{\min}(\theta) := \phi\left(f'\left(\frac{1}{C}\right) - 2\theta - 1\right) y_{\min} \tag{B.39}$$

and

$$u_{\min}(\theta) := f'\left(\frac{1}{C}\right) - 2\theta - 1. \tag{B.40}$$

We use $2\theta + 1$ instead of 2θ to obtain a strict inequality

$$\varphi_c(x, \gamma(x, \boldsymbol{\lambda}) + v_c(x; \boldsymbol{\lambda})) > t_{\min}(\|\boldsymbol{\lambda}\|) \tag{B.41}$$

The following regularity conditions guarantee that an optimizer over a compact set $\mathcal{K} \subset \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C)$ is also a global optimizer. Note that we introduce the following definition only for the case $\mathcal{X} = \mathbb{R}^m$.

Definition B.5. Assume $\mathcal{X} = \mathbb{R}^m$. We call the functions f_c and $\boldsymbol{G}$ *regular* if

(a) every function $f_c(x, \cdot\,)$ is twice continuously differentiable and, for every $\varepsilon > 0$,

$$\inf_{(c,x,t)\in[C]\times\mathbb{R}^m\times(\varepsilon,1)} \partial^2_{m+1} f_c(x,t) > 0, \tag{B.42}$$

(b) the partial derivatives $\partial_\ell \partial_{m+1} f_c(x,t)$ and $\partial_\ell G_{k,c}(x)$ exist and are continuous, and for every $\varepsilon > 0$,

$$\sup_{(\ell,k,c,x,t)\in[m]\times[K]\times[C]\times\mathbb{R}^m\times(\varepsilon,1)} \max\left(|\partial_\ell\partial_{m+1} f_c(x,t)|, |G_{k,c}(x)|, |\partial_\ell G_{k,c}(x)|\right) < \infty, \tag{B.43}$$

(c) the functions $\partial_{m+1} f_c(\,\cdot\,, t)$ are continuous for every $t \in (0,1]$ and $c \in [C]$.

We show that the regularity conditions on the f_c and $\boldsymbol{G}$ yield Lipschitzness of φ_c and local Lipschitzness of γ. This in turn will yield precompactness of the set $\mathcal{Q}$ given in equation (B.21) in Theorem B.1. The key tool we employ is utilizing a simplified version of the implicit function theorem, where the simplicity is due to the triviality of gluing.

Theorem B.2. *Under Assumptions 3.1 and B.2, for any $\theta \in \mathbb{R}_+$ the set*

$$\mathcal{Q} = \left\{ \boldsymbol{q}^{\Delta_C}_{\boldsymbol{\lambda}} \;\middle|\; \boldsymbol{\lambda} \in \mathbb{R}^K_+,\ \|\boldsymbol{\lambda}\|_1 \le \theta \right\} \tag{B.44}$$

is a precompact subset of $C(\mathbb{R}^m, \boldsymbol{\Delta}_C)$.

Proof. See Appendix B.1.5. □

The explicit formula for $\boldsymbol{h}^{\text{opt}}$ is a direct consequence of the formula for $\boldsymbol{q}^{\Delta_C}_{\boldsymbol{\lambda}}$.

Lemma B.2. *Let $f : [0,\infty) \to \overline{\mathbb{R}}$ be a strictly convex continuously differentiable function over*[3] *$(0,\infty)$ such that $f(1) = 0$ and $f'(0^+) = -\infty$, and let ϕ be the inverse of f'. Fix $\boldsymbol{q} \in \boldsymbol{\Delta}^+_C$, and define $F : [0,1]^C \to \overline{\mathbb{R}}$ by*

$$F(\boldsymbol{p}) = \mathbb{E}_{c\sim\boldsymbol{q}}\left[f\left(\frac{p_c}{q_c}\right)\right]. \tag{B.45}$$

Then, the convex conjugate of F is defined over all of $\mathbb{R}^C$ and satisfies

$$F^{\text{conj}}(\boldsymbol{v}) = \mathbb{E}_{c\sim\boldsymbol{q}}\left[v_c\phi(\gamma(\boldsymbol{v}) + v_c) - f(\phi(\gamma(\boldsymbol{v}) + v_c))\right] \tag{B.46}$$

where $\gamma : \mathbb{R}^C \to \mathbb{R}$ is the unique function satisfying

$$\mathbb{E}_{c\sim\boldsymbol{q}}\left[\phi(\gamma(\boldsymbol{v}) + v_c)\right] = 1, \qquad \boldsymbol{v} \in \mathbb{R}^C. \tag{B.47}$$

[3] We set $f(0) := f(0^+)$.

Further, for any $\boldsymbol{v} \in \mathbb{R}^C$, we have that

$$q_c^{\mathrm{conj}}(\boldsymbol{v}) = q_c \phi(\gamma(\boldsymbol{v}) + v_c), \qquad \textit{for every } \ c \in [C]. \tag{B.48}$$

Proof. See Appendix B.1.6. □

Corollary B.1. *Under Assumptions 3.1 and B.2, the c-th coordinate of $\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_C}(x)$ (see (B.20)) is given by $\varphi_c\left(x, \gamma(x, \boldsymbol{\lambda}) + v_c(x; \boldsymbol{\lambda})\right)$.*

The final ingredient in the proof is a direct consequence of Lemma B.1.

Corollary B.2. *Under Assumptions 3.1 and B.2, for any $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\varepsilon \in [0, t_{\min}(\|\boldsymbol{\lambda}\|))$*

$$\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c^\varepsilon} = \boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c}. \tag{B.49}$$

B.1.3 Proof of Theorems 3.1 and 3.2

We are now ready to derive the model projection formula. We operate under Assumption 3.1, and note that the model projection problem we consider will satisfy Assumption B.2. We apply the general results stated in Appendix B.1.2 above with $\mathcal{Z} = \boldsymbol{\Delta}_C^\varepsilon$ for all small enough ε.

By continuity of f,

$$D \supset \boldsymbol{\Delta}_C^+. \tag{B.50}$$

Further, for any $\varepsilon \in (0,1)$,

$$\mathcal{D} \supset \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C^\varepsilon), \tag{B.51}$$

so $\mathcal{D} \supset \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$. Fix $\widetilde{\boldsymbol{h}} \in \mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$ such that $\mathbb{E}\left[v\widetilde{\boldsymbol{h}}(X)\right] < \mathbf{0}$, i.e., $\widetilde{\boldsymbol{h}} \in \mathcal{S}$. Let ε be small enough that $\widetilde{\boldsymbol{h}} \in \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C^\varepsilon)$. Denote $\widetilde{\theta} = \theta_{\widetilde{\boldsymbol{h}}}$. Fix $\theta \geq \widetilde{\theta}$. Decrease, if necessary, the value of ε so that $\varepsilon < t_{\min}(\theta)$. Then, by Corollary B.2,

$$\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c^\varepsilon} = \boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_c} \tag{B.52}$$

for all $\boldsymbol{\lambda}$ with $\|\boldsymbol{\lambda}\| \leq \theta$.

By Theorem B.2, we have precompactness of the set

$$\mathcal{Q} := \left\{\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_C} \mid \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 \leq \theta\right\} \tag{B.53}$$

and that $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C)$. But, by (B.52),

$$\mathcal{Q} = \{\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_C^\varepsilon} \mid \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 \leq \theta\}. \tag{B.54}$$

Then, $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C^{\varepsilon})$. Precompactness of $\mathcal{Q}$, then, implies by Theorem B.1 (using $\mathcal{Z} = \boldsymbol{\Delta}_C^{\varepsilon}$) that the problem

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{C}(\mathcal{X}, \boldsymbol{\Delta}_C^{\varepsilon})} \quad & \mathbb{E}\left[D_f(\boldsymbol{h}(X) \,\|\, \boldsymbol{h}^{\text{base}}(X))\right], \\ \text{s.t.} \quad & \mathbb{E}\left[\boldsymbol{G}(X)\boldsymbol{h}(X)\right] \leq \mathbf{0} \end{aligned} \tag{B.55}$$

has the unique solution $\boldsymbol{q}_{\boldsymbol{\lambda}^\star}^{\boldsymbol{\Delta}_C}$ for any $\boldsymbol{\lambda}^\star$ solving

$$\inf_{\boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 \leq \widetilde{\theta}} \mathbb{E}\left[D_f^{\text{conj}}(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X))\right] \tag{B.56}$$

where we used the fact that

$$\mathcal{L}(\boldsymbol{q}_{\boldsymbol{\lambda}}^{\boldsymbol{\Delta}_C}, \boldsymbol{\lambda}) = -\mathbb{E}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X)\right)\right]. \tag{B.57}$$

By Corollary B.3, we may remove the condition $\|\boldsymbol{\lambda}\| \leq \widetilde{\theta}$. As the solution $\boldsymbol{q}_{\boldsymbol{\lambda}^\star}^{\boldsymbol{\Delta}_C}$ does not depend on ε, and as ε is arbitrary, we may extend the optimization to be over all of $\mathcal{C}_+(\mathcal{X}, \boldsymbol{\Delta}_C)$. Finally, the proof is complete in view of the equation of $\boldsymbol{q}_{\boldsymbol{\lambda}^\star}^{\boldsymbol{\Delta}_C}$ as given by Corollary B.1.

B.1.4 Proof of Theorem B.1: a Generalized Result

We start by proving intermediate results on the general optimization problem (B.3)–(B.5), then we combine these component results to derive Theorem B.1 at the end of this subsection.

First, we have the following basic result on the existence and uniqueness of solutions to the general optimization problem (B.5) over compact subsets of $\mathcal{C}(\mathcal{X})$.

Lemma B.3. *Suppose items (a)–(c) of Assumption B.1 all hold. For a compact set $\mathcal{K} \subset \mathcal{D}$ such that $\mathcal{K} \cap \mathcal{F}$ is nonempty, the following optimization problem has a minimizer*

$$\begin{aligned} \min_{\boldsymbol{h} \in \mathcal{K}} \ & \mathcal{A}(\boldsymbol{h}), \\ \text{s.t.} \ & \mathcal{B}_k(\boldsymbol{h}) \leq 0, \ \ k \in [K]. \end{aligned} \tag{B.58}$$

If, in addition, $\mathcal{K}$ is convex and item (d) holds, then the minimizer is unique.

Proof. We prove the existence of a minimizer first. Then we treat uniqueness. Suppose that items (a)–(c) of Assumption B.1 all hold, and fix a compact set $\mathcal{K} \subset \mathcal{D}$. We show that the objective function is lower-semicontinuous on $\mathcal{K}$ and that the feasibility set $\mathcal{K} \cap \mathcal{F}$ is compact, which together yield via the extreme value theorem the existence of a minimizer. We start by showing that the mappings

$\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_K$ are lower-semicontinuous on $\mathcal{K}$. Lower-semicontinuity of the $\mathcal{B}_i$ will yield that the feasibility set $\mathcal{K} \cap \mathcal{F}$ of (B.58) is compact.

Fix $J \in \{F, G_1, \cdots, G_K\}$, and we will show that the mapping $h \mapsto \mathbb{E}[J(X, h(X))] + \mathbb{I}_{\mathcal{D}}(h)$ is lower-semicontinuous when restricted to $\mathcal{K}$. As $\mathcal{K} \subset \mathcal{D}$ by assumption, this mapping is just $h \mapsto \mathbb{E}[J(X, h(X))]$. As $\mathcal{K}$ is a metric space, lower-semincontinuity on $\mathcal{K}$ is equivalent to sequential-lower-semicontinuity [KZ06, Theorem 7.1.2]. Fix a convergent sequence $h_n \to h$ in $\mathcal{K}$ (i.e., $\sup_{x \in \mathcal{X}} \|h_n(x) - h(x)\|_1 \to 0$ as $n \to \infty$). By item (b) of Assumption B.1, we may apply Fatou's lemma to obtain

$$\liminf_{n\to\infty} \mathbb{E}[J(X, h_n(X))] \geq \mathbb{E}\left[\liminf_{n\to\infty} J(X, h_n(X))\right]. \tag{B.59}$$

Uniform convergence $h_n \to h$ implies, in particular, pointwise convergence: $h_n(x) \to h(x)$ for every $x \in \mathcal{X}$. Therefore, by lower-semicontinuity of each $J(x, \cdot)$ (item (c) of Assumption B.1)

$$\mathbb{E}\left[\liminf_{n\to\infty} J(X, h_n(X))\right] \geq \mathbb{E}[J(X, h(X))]. \tag{B.60}$$

Therefore,

$$\liminf_{n\to\infty} \mathbb{E}[J(X, h_n(X))] \geq \mathbb{E}[J(X, h(X))], \tag{B.61}$$

and lower-semicontinuity of $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_K$ on $\mathcal{K}$ follows. In particular, the lower-level sets

$$\mathcal{V}_k := \{h \in \mathcal{K} \mid \mathbb{E}\left[G_k(X, h(X))\right] \leq 0\} \tag{B.62}$$

are closed[4] [KZ06, Theorem 7.1.1]. Therefore, the feasibility set $\mathcal{F} \cap \mathcal{K} = \bigcap_{k \in [K]} \mathcal{V}_k$ is closed. By compactness of $\mathcal{K}$, the feasibility set $\mathcal{F} \cap \mathcal{K}$ is compact too. Finally, lower-semicontinuity of $\mathcal{A}$ on $\mathcal{K}$ and compactness of the nonempty (by assumption) feasibility set $\mathcal{F} \cap \mathcal{K}$ yield the existence of a minimizer [KZ06, Theorem 7.3.1].

Finally, we show uniqueness of the minimizer. Suppose that $\mathcal{K}$ is also convex, and that item (d) of Assumption B.1 holds too. Since expectation is a linear operator, $h \mapsto \mathbb{E}[F(X, h(X))]$ is strictly convex, and each $h \mapsto \mathbb{E}[G_k(X, h(X))]$ is convex. Hence, the lower-level sets (B.62) are convex which implies that the feasibility set $\mathcal{K} \cap \mathcal{F}$ is convex. Thus, the optimization problem (B.58) has a unique minimizer, and the proof of the lemma is complete. □

It will be useful to introduce the following notation.

Definition B.6. For a given $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ and a subset $\mathcal{K} \subset \mathcal{D}$, define the function in $\mathcal{K}$ that achieves the

[4] The $\mathcal{V}_i$ are closed both in $\mathcal{K}$ and in $\mathcal{C}(\mathcal{X})$, as the compact set $\mathcal{K}$ is closed in the Hausdorff space $\mathcal{C}(\mathcal{X})$.

minimal value of the Lagrangian by

$$h_{\boldsymbol{\lambda}}^{\mathcal{K}} := \operatorname*{argmin}_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{B.63}$$

if there is such a unique function.

We need the following intermediate result, which expresses the solutions to the general optimization problem (B.5) in terms of $h_{\boldsymbol{\lambda}}^{\mathcal{K}}$ we just defined above.

Theorem B.3. *Suppose Assumption B.1 holds, and fix a nonempty compact and convex $\mathcal{K} \subset \mathcal{D}$. For every $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^{k}$, the function $\mathcal{L}(\,\cdot\,, \boldsymbol{\lambda})$ has a unique minimizer over $\mathcal{K}$, i.e., $h_{\boldsymbol{\lambda}}^{\mathcal{K}}$ in (B.63) is well-defined. In addition, if $\boldsymbol{\lambda}^{\star}$ satisfies*

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^{\star}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{k}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{B.64}$$

then $h_{\boldsymbol{\lambda}^{\star}}^{\mathcal{K}}$ is the unique solution for problem (B.5).

Proof. Since $\mathcal{K}$ is compact and $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is strictly convex and lower-semicontinuous for any fixed $\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}$, there is a unique minimizer of $\mathcal{L}(h, \boldsymbol{\lambda})$ over $\mathcal{K}$. Hence, $h_{\boldsymbol{\lambda}}^{\mathcal{K}}$ is well-defined and satisfies

$$\mathcal{L}(h_{\boldsymbol{\lambda}}^{\mathcal{K}}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{B.65}$$

Next, we prove strong duality for (B.5). Again, the mapping $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is strictly convex and lower-semicontinuous for each fixed $\boldsymbol{\lambda}$. Also, $\boldsymbol{\lambda} \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$ is concave for each fixed h (as it is affine). Therefore, by Sion's minimax theorem and the compactness of $\mathcal{K}$,

$$\inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{B.66}$$

Let $h^{\star}$ denote the unique solution of (B.5), whose existence and uniqueness are guaranteed by Lemma B.3. We have that

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \mathcal{L}(h^{\star}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{B.67}$$

Combining (B.67), (B.66), and (B.64) together, we have

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \mathcal{L}(h^{\star}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}_{+}^{K}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^{\star}). \tag{B.68}$$

Furthermore, since

$$\mathcal{L}(h^\star, \boldsymbol{\lambda}^\star) \leq \sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \mathcal{L}(h^\star, \boldsymbol{\lambda}) \quad \text{and} \quad \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^\star) \leq \mathcal{L}(h^\star, \boldsymbol{\lambda}^\star), \tag{B.69}$$

then we have

$$\mathcal{L}(h^\star, \boldsymbol{\lambda}^\star) \leq \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^\star) \leq \mathcal{L}(h^\star, \boldsymbol{\lambda}^\star) \tag{B.70}$$

which implies $\mathcal{L}(h^\star, \boldsymbol{\lambda}^\star) = \inf\limits_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}^\star)$. Therefore, by strict convexity of $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda}^\star)$, $h^\star = h_{\boldsymbol{\lambda}^\star}^{\mathcal{K}}$. □

Next, we prove the existence of a $\boldsymbol{\lambda}^\star$ satisfying (B.64) in Theorem B.3 whenever $\mathcal{K} \cap \mathcal{S} \neq \emptyset$. It will be convenient to introduce the following quantity, which will be used to bound the searching space of dual variable.

Definition B.7. For a subset $\mathcal{K} \subset \mathcal{D}$, we define

$$\theta(\mathcal{K}) := \inf_{q \in \mathcal{K} \cap \mathcal{S}} \frac{\mathcal{A}(q) - \inf\limits_{h \in \mathcal{K}} \mathcal{A}(h)}{- \max\limits_{k \in [K]} \mathcal{B}_k(q)}. \tag{B.71}$$

We note that under items (a)–(b) of Assumption 3.1, if $\mathcal{K} \subset \mathcal{D}$ is such that $\mathcal{K} \cap \mathcal{S}$ is nonempty, then $\theta(\mathcal{K}) \in \mathbb{R}_{\geq 0}$. Indeed, fix an integrable $L : \mathcal{X} \to \mathbb{R}$ such that

$$L(x) \leq \inf_{h \in \mathcal{D}} F(x, h(x)) \tag{B.72}$$

for every $x \in \mathcal{X}$. Then, for any $q \in \mathcal{K} \cap \mathcal{S}$

$$-\infty < \mathbb{E}\left[L(X)\right] \leq \inf_{h \in \mathcal{D}} \mathcal{A}(h) \leq \inf_{h \in \mathcal{K}} \mathcal{A}(h) \leq \mathcal{A}(q) < \infty. \tag{B.73}$$

Thus, $\inf_{h \in \mathcal{K}} \mathcal{A}(h) \in \mathbb{R}$. Hence, by definition of $\mathcal{D}$ and because $\mathcal{K} \cap \mathcal{S} \subset \mathcal{D}$, we obtain $\theta(\mathcal{K}) \in \mathbb{R}_+$.

Theorem B.4. *Suppose items (a)–(b) of Assumption B.1 both hold, and fix $\mathcal{K} \subset \mathcal{D}$. If $\mathcal{K} \cap \mathcal{S}$ is nonempty, then*

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 \leq \theta(\mathcal{K})}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{B.74}$$

there exists a $\boldsymbol{\lambda}^\star$ that achieves the supremum in the left-hand-side in (B.74)*, and any such maximizer satisfies $\|\boldsymbol{\lambda}^\star\|_1 \leq \theta(\mathcal{K})$.*

Proof. If the equality

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \mathbf{0}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) \tag{B.75}$$

holds, then the desired equality (B.74) also holds and $\boldsymbol{\lambda} = \mathbf{0}$ achieves the supremum. Thus, for the remainder of the proof, we assume that (B.75) does not hold, i.e.,

$$\inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0}) < \sup_{\boldsymbol{\lambda}\in\mathbb{R}_+^K} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}). \tag{B.76}$$

For any $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ and $q \in \mathcal{S}$, by the definition of $\mathcal{S}$ (see (B.10)), $\mathbb{E}\left[G_k(X, q(X))\right] < 0$ for all $k \in [K]$. Then, for any $q \in \mathcal{K} \cap \mathcal{S}$

$$\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \inf_{h\in\mathcal{K}\cap\mathcal{S}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \mathcal{L}(q,\boldsymbol{\lambda}) = \mathcal{A}(q) + \sum_{k\in[K]} \lambda_k \mathcal{B}_k(q) \leq \mathcal{A}(q) + \|\boldsymbol{\lambda}\|_1 \max_{k\in[K]} \mathcal{B}_k(q) \tag{B.77}$$

where we used the fact that $q \in \mathcal{K} \cap \mathcal{S} \subset \mathcal{K} \subset \mathcal{D}$. Thus, we have

$$\|\boldsymbol{\lambda}\|_1 \leq \frac{\mathcal{A}(q) - \inf\limits_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})}{-\max\limits_{k\in[K]} \mathcal{B}_k(q)}. \tag{B.78}$$

Now, if $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ satisfies both $\|\boldsymbol{\lambda}\|_1 > \theta(\mathcal{K})$ and $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \geq \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0})$, then, we must have (because $\mathcal{L}(h,\mathbf{0}) = \mathbb{E}\left[F(X,h(X))\right] = \mathcal{A}(h)$ for $h \in \mathcal{D}$)

$$\theta(\mathcal{K}) < \|\boldsymbol{\lambda}\|_1 \leq \frac{\mathcal{A}(q) - \inf\limits_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})}{-\max\limits_{k\in[K]} \mathcal{B}_k(q)} \leq \frac{\mathcal{A}(q) - \inf\limits_{h\in\mathcal{K}} \mathcal{A}(h)}{-\max\limits_{k\in[K]} \mathcal{B}_k(q)} \tag{B.79}$$

for every $q \in \mathcal{K} \cap \mathcal{S}$. Taking the infimum over all $q \in \mathcal{K} \cap \mathcal{S}$, we obtain

$$\theta(\mathcal{K}) < \|\boldsymbol{\lambda}\|_1 \leq \theta(\mathcal{K}), \tag{B.80}$$

which is absurd. Thus, every $\boldsymbol{\lambda}$ that satisfies $\|\boldsymbol{\lambda}\|_1 > \theta(\mathcal{K})$ must have $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) < \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0})$. Taking the supremum over all such $\boldsymbol{\lambda}$ implies

$$\sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 > \theta(\mathcal{K})}} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}) \leq \inf_{h\in\mathcal{K}} \mathcal{L}(h,\mathbf{0}) < \sup_{\boldsymbol{\lambda}\in\mathbb{R}_+^K} \inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda}). \tag{B.81}$$

In particular, the desired equality (B.74) holds.

Finally, being the pointwise infimum of linear (in particular, upper-semicontinuous) functions in $\boldsymbol{\lambda}$, $\inf_{h\in\mathcal{K}} \mathcal{L}(h,\boldsymbol{\lambda})$ is upper-semicontinuous. Hence, having $\theta(\mathcal{K}) < \infty$ would imply that at least one $\boldsymbol{\lambda}^\star$ maximizing the dual optimization problem (B.74) exists. By inequality (B.81), $\|\boldsymbol{\lambda}^\star\|_1 \leq \theta(\mathcal{K})$ for any such maximizer $\boldsymbol{\lambda}^\star$. □

Though Theorem B.4 gives a way to bound the value of the dual parameter $\boldsymbol{\lambda}$, the upper bound $\theta(\mathcal{K})$ might not be computable. In particular, computing $\theta(\mathcal{K})$ requires global information about

$\mathcal{K}$. Nevertheless, note that removing the outer infimum in the definition of $\theta(\mathcal{K})$ still yields a finite upper bound. Further, relaxing the inner infimum to be over the domain $\mathcal{D}$ also gives a finite upper bound (under item (b) of Assumption 3.1).

Under item (b) of Assumption 3.1, θ_q is always finite. Also, $\theta(\mathcal{K}) \leq \theta_q$ whenever $\mathcal{K} \subset \mathcal{D}$ and $q \in \mathcal{K} \cap \mathcal{S}$. Thus, Theorem B.4 immediately implies the following result.

Corollary B.3. *Suppose items (a)–(b) of Assumption B.1 both hold, and fix $\mathcal{K} \subset \mathcal{D}$. If $\mathcal{K} \cap \mathcal{S}$ is nonempty and $q \in \mathcal{K} \cap \mathcal{S}$, then*

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 \leq \theta_q}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{B.82}$$

and the supremum is achievable. Furthermore, all maximizers have 1*-norm at most θ_q.*

Next, we give a more tractable way of expressing $h_{\boldsymbol{\lambda}}^{\mathcal{K}}$.

Theorem B.5. *Suppose Assumption B.1 holds. Fix a nonempty convex and compact subset $\mathcal{Z} \subset D$, and a nonempty convex and compact subset $\mathcal{K} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{D}$. For any $\boldsymbol{\lambda} \in \mathbb{R}_+^K$, if $q_{\boldsymbol{\lambda}}^{\mathcal{Z}} \in \mathcal{K}$, then $h_{\boldsymbol{\lambda}}^{\mathcal{K}} = q_{\boldsymbol{\lambda}}^{\mathcal{Z}}$.*

Proof. For each $x \in \mathcal{X}$, let $\mathcal{R}_x \subset \mathbb{R}^C$ denote the image of $\mathcal{K}$ under the mapping $h \mapsto h(x)$, i.e.,

$$\mathcal{R}_x := \{h(x) \mid h \in \mathcal{K}\}. \tag{B.83}$$

We have, by assumption, $\bigcup_{x \in \mathcal{X}} \mathcal{R}_x \subset \mathcal{Z}$. Fix $\boldsymbol{\lambda} \in \mathbb{R}_+^K$, and write

$$L(x, q) = F(x, q) + \sum_{k \in [K]} \lambda_k G_k(x, q) \tag{B.84}$$

for short. Then, for any $(x, h) \in \mathcal{X} \times \mathcal{K}$

$$L(x, h(x)) \geq \inf_{p \in \mathcal{K}} L(x, p(x)) \geq \inf_{r \in \mathcal{R}_x} L(x, r) \geq \inf_{q \in \mathcal{Z}} L(x, q) = L(x, q_{\boldsymbol{\lambda}}^{\mathcal{Z}}(x)). \tag{B.85}$$

Assume that $q_{\boldsymbol{\lambda}}^{\mathcal{Z}} \in \mathcal{K}$. Then, taking the expectation of the two far ends of (B.85) then the infimum for $h \in \mathcal{K}$ we get

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) \geq \mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}). \tag{B.86}$$

However, it is also true that

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) \leq \mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}). \tag{B.87}$$

Therefore, we get the equality

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}). \tag{B.88}$$

By strict convexity of $h \mapsto \mathcal{L}(h, \boldsymbol{\lambda})$, and by definition of $h_{\boldsymbol{\lambda}}^{\mathcal{K}}$, we have $h_{\boldsymbol{\lambda}}^{\mathcal{K}} = q_{\boldsymbol{\lambda}}^{\mathcal{Z}}$. □

Finally, we are ready to prove our generalized result in Theorem B.1.

Proof of Theorem B.1. Write $\theta = \theta_{\boldsymbol{p}}$, and note that $\theta \in \mathbb{R}_+$. Let $\boldsymbol{u} \in \mathcal{C}(\mathcal{X}, \mathcal{Z}) \cap \mathcal{F}$ be arbitrary. Consider the two sets

$$\mathcal{K} = \overline{\operatorname{co}(\mathcal{H} \cup \{\boldsymbol{p}\})}, \tag{B.89}$$

$$\mathcal{K}' = \overline{\operatorname{co}(\mathcal{H} \cup \{\boldsymbol{p}, \boldsymbol{u}\})}. \tag{B.90}$$

The sets $\mathcal{K}$ and $\mathcal{K}'$ are convex and compact, and they satisfy $\mathcal{K}, \mathcal{K}' \subset \mathcal{C}(\mathcal{X}, \mathcal{Z})$ because $\mathcal{C}(\mathcal{X}, \mathcal{Z})$ is convex and closed and $\mathcal{H} \subset \mathcal{C}(\mathcal{X}, \mathcal{Z})$ by assumption. If $\boldsymbol{\lambda} \in \Lambda$, then by definition $q_{\boldsymbol{\lambda}}^{\mathcal{Z}}$ is an element in both $\mathcal{K}$ and $\mathcal{K}'$, hence by Theorem B.5

$$h_{\boldsymbol{\lambda}}^{\mathcal{K}} = q_{\boldsymbol{\lambda}}^{\mathcal{Z}} = h_{\boldsymbol{\lambda}}^{\mathcal{K}'}. \tag{B.91}$$

By Corollary B.3,

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}), \tag{B.92}$$

and the same is true for $\mathcal{K}'$

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}) = \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{B.93}$$

By definition, $\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(h_{\boldsymbol{\lambda}}^{\mathcal{K}}, \boldsymbol{\lambda})$ and $\inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(h_{\boldsymbol{\lambda}}^{\mathcal{K}'}, \boldsymbol{\lambda})$.

Therefore, for any $\boldsymbol{\lambda} \in \Lambda$

$$\inf_{h \in \mathcal{K}} \mathcal{L}(h, \boldsymbol{\lambda}) = \mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}) = \inf_{h \in \mathcal{K}'} \mathcal{L}(h, \boldsymbol{\lambda}). \tag{B.94}$$

Thus, the problems (B.92) and (B.93) are equivalent to each other, and they are equivalent to

$$\sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}_+^K \\ \|\boldsymbol{\lambda}\|_1 \leq \theta}} \mathcal{L}(q_{\boldsymbol{\lambda}}^{\mathcal{Z}}, \boldsymbol{\lambda}). \tag{B.95}$$

Furthermore, there is a $\boldsymbol{\lambda}^\star$ achieving this supremum. In addition, by Theorem B.3, for any such $\boldsymbol{\lambda}^\star$ we have that $q_{\boldsymbol{\lambda}^\star}^{\mathcal{Z}}$ is the unique solution to both $\inf_{h \in \mathcal{K} \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right]$ and $\inf_{h \in \mathcal{K}' \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right]$. Now,

$$\mathbb{E}\left[F(X, q_{\boldsymbol{\lambda}^\star}^{\mathcal{Z}}(X))\right] = \inf_{h \in \mathcal{K}' \cap \mathcal{F}} \mathbb{E}\left[F(X, h(X))\right] \leq \mathbb{E}\left[F(X, \boldsymbol{u}(X))\right]. \tag{B.96}$$

Therefore, by arbitrariness of $\boldsymbol{u}$,

$$\mathbb{E}\left[F(X, q^{\mathcal{Z}}_{\boldsymbol{\lambda}^\star}(X))\right] = \inf_{\boldsymbol{u} \in \mathcal{C}(\mathcal{X},\mathcal{Z}) \cap \mathcal{F}} \mathbb{E}\left[F(X, \boldsymbol{u}(X))\right]. \tag{B.97}$$

Finally, uniqueness follows by convexity of the set $\mathcal{F}$ and strict convexity of the function $\mathcal{A}|_{\mathcal{C}(\mathcal{X},\mathcal{Z})}$. □

B.1.5 Proof of Theorem B.2: Precompactness of Potentially Optimal Models

We note that Assumption 3.1 implies regularity of $f_c(x,t) = h^{\text{base}}_c(x) f(t/h^{\text{base}}_c(x))$ and $\boldsymbol{G}$ as defined by Definition B.5. To see this, note that $\partial^2_{m+1} f_c(x,t) = f''(t/h^{\text{base}}_c(x))/h^{\text{base}}_c(x)$. By continuity of f'', item (a) is satisfied. Also,

$$\partial_\ell \partial_{m+1} f_c(x,t) = \frac{-t\partial_\ell h^{\text{base}}_c(x)}{h^{\text{base}}_c(x)^2} f''\left(\frac{t}{h^{\text{base}}_c(x)}\right) \tag{B.98}$$

and again continuity of f'' implies that item (b) is also satisfied.

We employ the following version of the implicit function theorem.

Theorem B.6 (Implicit Function Theorem). *Let $\Omega \subset \mathbb{R}^e \times \mathbb{R}$ be an open set, denote by $U \subset \mathbb{R}^e$ and $V \subset \mathbb{R}$ its projections, and let $C : \Omega \to \mathbb{R}$ be a differentiable function. If there exists a unique function $c : U \to V$ satisfying both $(\boldsymbol{a}, c(\boldsymbol{a})) \in \Omega$ and $C(\boldsymbol{a}, c(\boldsymbol{a})) = 0$ for every $\boldsymbol{a} \in U$, and if $\partial_{e+1} C(\boldsymbol{a}, c(\boldsymbol{a})) \neq 0$ for every $\boldsymbol{a} \in U$, then c is differentiable and $\partial_i c(\boldsymbol{a}) = (-\partial_i C/\partial_{e+1} C)|_{(\boldsymbol{a}, c(\boldsymbol{a}))}$ for every $(i, \boldsymbol{a}) \in [e] \times U$.*

We begin by deriving upper bounds on the partial derivatives of the φ_c and γ. Then, we conclude from Lipshcitzness of the φ_c and γ total boundedness of $\mathcal{Q}$ via compactness of $\boldsymbol{\Delta}_C$. As a by-product, it will follow that $\mathcal{Q}$ consists of continuous functions, i.e., that $\mathcal{Q} \subset \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C)$. For convenience of notation, we will show precompactness when $\|\boldsymbol{\lambda}\|_1$ is restricted to be at most $\theta - 1$ for some $\theta > 1$.

Fix $c \in [C]$, and we will show an upper bound on the partial derivatives of φ_c. Set (see (B.40) for the definition of u_{min})

$$\Omega_c := \{(\boldsymbol{x}, u) \in \mathbb{R}^m \times \mathbb{R} \mid u_{\text{min}}(\theta) < u < \partial_{m+1} f_c(\boldsymbol{x}, 1)\}. \tag{B.99}$$

By the assumption of continuity of $\partial_{m+1} f_c(\,\cdot\,, 1)$, the set Ω_c is open; indeed, Ω_c is the intersection of the preimage of the open set $(0, \infty)$ under the continuous map $(\boldsymbol{x}, u) \mapsto \partial_{m+1} f_c(\boldsymbol{x}, 1) - u$ with the open set $\mathbb{R}^m \times (u_{\text{min}}(\theta), \infty)$. The set Ω_c is nonempty. Indeed, for any $x \in \mathbb{R}^m$, we have by monotonicity of f' that

$$u_{\text{min}}(\theta) = f'\left(\frac{1}{C}\right) - 2\theta - 1 < f'\left(\frac{1}{C}\right) \leq f'\left(\frac{1}{h^{\text{base}}_c(\boldsymbol{x})}\right) = \partial_{m+1} f_c(\boldsymbol{x}, 1). \tag{B.100}$$

Define $\rho_c : \Omega_c \times (0,1) \to \mathbb{R}$ by

$$\rho_c(\boldsymbol{x}, u, t) = \partial_{m+1} f_c(\boldsymbol{x}, t) - u. \tag{B.101}$$

For any $(\boldsymbol{x}, u) \in \Omega_c$, there exists a unique $t \in (0,1)$ such that $\rho_c(\boldsymbol{x}, u, t) = 0$, namely, $t = \varphi_c(\boldsymbol{x}, u)$. In other words, $\varphi_c(\boldsymbol{x}, u)$ is defined via

$$\rho_c(\boldsymbol{x}, u, \varphi_c(\boldsymbol{x}, u)) = 0. \tag{B.102}$$

By assumption on f_c, all partial derivative of ρ_c exist and are continuous. Therefore, ρ_c is differentiable. Further, by regularity of f_c, $\partial_{m+2}\rho_c(\boldsymbol{x}, u, t) \neq 0$. Hence, by the implicit function theorem, φ_c is differentiable and its partial derivatives are given by

$$\partial_{m+1}\varphi_c(\boldsymbol{x}, u) = -\frac{\partial_{m+1}\rho_c(\boldsymbol{x}, u, \varphi_c(\boldsymbol{x}, u))}{\partial_{m+2}\rho_c(\boldsymbol{x}, u, \varphi_c(\boldsymbol{x}, u))} = \frac{1}{\partial^2_{m+1} f_c(\boldsymbol{x}, \varphi_c(\boldsymbol{x}, u))}, \tag{B.103}$$

$$\partial_\ell \varphi_c(\boldsymbol{x}, u) = -\frac{\partial_\ell \rho_c(\boldsymbol{x}, u, \varphi_c(\boldsymbol{x}, u))}{\partial_{m+2}\rho_c(\boldsymbol{x}, u, \varphi_c(\boldsymbol{x}, u))} = \frac{-\partial_{\ell, m+1} f_c(\boldsymbol{x}, \varphi_c(\boldsymbol{x}, u))}{\partial^2_{m+1} f_c(\boldsymbol{x}, \varphi_c(\boldsymbol{x}, u))}, \tag{B.104}$$

for every $(\boldsymbol{x}, u) \in \Omega_c$, where $\ell \leq m$. Because φ_c is differentiable, it is also continuous. Further, by assumption of regularity, we have the bound

$$\max_{r \in [m+1]} \max_{c \in [C]} \sup_{(\boldsymbol{x}, u) \in \Omega_c} |\partial_r \varphi_c(\boldsymbol{x}, u)| \leq A \tag{B.105}$$

for some positive constant A.

Next, we show an upper bound on partial derivative of γ. Consider the function $\tau : \mathbb{R}^m \times \mathcal{B}_1(\mathbf{0}, \theta) \times \mathbb{R}_{>0} \to \mathbb{R}$ defined by

$$\tau(\boldsymbol{x}, \boldsymbol{\lambda}, \varepsilon) := \min_{j \in [c]} \left(\partial_{m+1} f_c(\boldsymbol{x}, 1^-) + \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x}) \right) - \max_{c \in [C]} \left(\partial_{m+1} f_c(\boldsymbol{x}, \varepsilon) + \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x}) \right). \tag{B.106}$$

We may lower bound τ uniformly over $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathcal{B}_1(\mathbf{0}, \theta)$ by

$$\tau(\boldsymbol{x}, \boldsymbol{\lambda}, \varepsilon) \geq f'(1^-) - f'\left(\frac{\varepsilon}{y_{\min}}\right) - 2\theta E \tag{B.107}$$

where $E = \sup_{k,c,x} |G_{k,c}(\boldsymbol{x})|$ is finite by assumption. The uniformity of this lower bound implies that

$$\inf_{x, \|\boldsymbol{\lambda}\|_1 \leq \theta} \tau(\boldsymbol{x}, \boldsymbol{\lambda}, \varepsilon) \geq f'(1^-) - f'\left(\frac{\varepsilon}{y_{\min}}\right) - 2\theta E. \tag{B.108}$$

Taking $\varepsilon \to 0^+$, the lower bound in (B.108) approaches ∞. In particular, we get that

$$\inf_{\boldsymbol{x}, \|\boldsymbol{\lambda}\|_1 \le \theta} \tau(\boldsymbol{x}, \boldsymbol{\lambda}, \varepsilon) > 0 \tag{B.109}$$

for every small enough ε. Fix $\varepsilon \in (0, t_{\text{min}}(\theta))$ such that (B.109) is satisfied. Define the set

$$\begin{aligned}\Omega := \Bigg\{ (\boldsymbol{x}, \boldsymbol{\lambda}, u) \in \mathbb{R}^{m+k+1} \mid \max_{c \in [C]} \partial_{m+1} f_c(\boldsymbol{x}, \varepsilon) + \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x}) &< u \\ < \min_{c \in [C]} \partial_{m+1} f_c(\boldsymbol{x}, 1^-) + \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x}) \Bigg\}. \end{aligned} \tag{B.110}$$

The set Ω is nonempty by inequality (B.109). Further, similarly to the Ω_c, the set Ω is open. Note that for any $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathbb{R}^k$ with $\|\boldsymbol{\lambda}\|_1 \le \theta$, we have $(\boldsymbol{x}, \boldsymbol{\lambda}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda})) \in \Omega$. For each $c \in [C]$, define $\psi_c : \Omega \to (0, 1)$ by

$$\psi_c(\boldsymbol{x}, \boldsymbol{\lambda}, u) = \varphi_c \left(\boldsymbol{x}, u - \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x}) \right). \tag{B.111}$$

Define $\eta : \Omega \to (-1, C)$ by

$$\eta(\boldsymbol{x}, \boldsymbol{\lambda}, u) = -1 + \sum_{c \in [C]} \psi_c(\boldsymbol{x}, \boldsymbol{\lambda}, u). \tag{B.112}$$

Then, $\gamma(\boldsymbol{x}, \boldsymbol{\lambda})$ is defined by

$$\eta(\boldsymbol{x}, \boldsymbol{\lambda}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda})) = 0. \tag{B.113}$$

As we have shown that each φ_c is differentiable, and as each partial derivative $\partial_\ell G_{k,c}$ is assumed to exist and be continuous, the function η is differentiable. Further, we may compute the partial derivatives of η by the chain rule

$$\partial_{m+K+1} \eta(\boldsymbol{x}, \boldsymbol{\lambda}, u) = \sum_c \partial_{m+K+1} \psi_c(\boldsymbol{x}, \boldsymbol{\lambda}, u) = \sum_c \partial_{m+1} \varphi_c \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \tag{B.114}$$

$$= \sum_c \frac{1}{\partial^2_{m+1} f_c \left(\boldsymbol{x}, \varphi \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \right)}, \tag{B.115}$$

$$\begin{aligned}\partial_\ell \eta(\boldsymbol{x}, \boldsymbol{\lambda}, u) \overset{\ell \le m}{=} \sum_c \Bigg(\partial_\ell \varphi_c \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \\ - \left(\sum_k \lambda_k \partial_\ell G_{k,c}(\boldsymbol{x}) \right) \partial_{m+1} \varphi_c \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \Bigg) \end{aligned} \tag{B.116}$$

$$= - \sum_c \frac{\partial_{\ell, m+1} f_c \left(\boldsymbol{x}, \varphi_c \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \right) + \sum_k \lambda_k \partial_\ell G_{k,c}(\boldsymbol{x})}{\partial^2_{m+1} f_c \left(\boldsymbol{x}, \varphi \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \right)}, \tag{B.117}$$

$$\partial_{m+\ell} \eta(\boldsymbol{x}, \boldsymbol{\lambda}, u) \overset{1 \le \ell \le K}{=} \sum_c -G_{\ell,c}(\boldsymbol{x}) \partial_{m+1} \varphi_c \left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x}) \right) \tag{B.118}$$

$$= \sum_c \frac{-G_{\ell,c}(\boldsymbol{x})}{\partial^2_{m+1} f_c\left(\boldsymbol{x}, \varphi\left(\boldsymbol{x}, u - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right)}. \tag{B.119}$$

Therefore, by the implicit function theorem, we have that γ is differentiable and

$$\partial_\ell \gamma(\boldsymbol{x}, \boldsymbol{\lambda}) \overset{\ell \leq m}{=} \frac{-\partial_\ell \eta(\boldsymbol{x}, \boldsymbol{\lambda}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda}))}{\partial_{m+K+1} \eta(\boldsymbol{x}, \boldsymbol{\lambda}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda}))} = \frac{\sum_c \frac{\partial_{\ell,m+1} f_c\left(\boldsymbol{x}, \varphi_c\left(\boldsymbol{x}, \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right) + \sum_k \lambda_k \partial_\ell G_{k,c}(\boldsymbol{x})}{\partial^2_{m+1} f_c\left(\boldsymbol{x}, \varphi\left(\boldsymbol{x}, \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right)}}{\sum_c \frac{1}{\partial^2_{m+1} f_c\left(\boldsymbol{x}, \varphi\left(\boldsymbol{x}, \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right)}}, \tag{B.120}$$

$$\partial_{m+\ell} \gamma(\boldsymbol{x}, \boldsymbol{\lambda}) \overset{1 \leq \ell \leq K}{=} \frac{\sum_c \frac{G_{\ell,c}(\boldsymbol{x})}{\partial^2_{m+1} f_c\left(\boldsymbol{x}, \varphi\left(\boldsymbol{x}, \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right)}}{\sum_c \frac{1}{\partial^2_{m+1} f_c\left(\boldsymbol{x}, \varphi\left(\boldsymbol{x}, \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \sum_k \lambda_k G_{k,c}(\boldsymbol{x})\right)\right)}}. \tag{B.121}$$

Thus, by assumption of regularity

$$\sup_{r,\boldsymbol{x}} |\partial_r \gamma(\boldsymbol{x}, \boldsymbol{\lambda})| \leq B \cdot (2 + \|\boldsymbol{\lambda}\|_1) \tag{B.122}$$

for some positive constant B.

Define functions $\boldsymbol{p}_{\boldsymbol{\lambda}} \in \mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C)$, one for each $\boldsymbol{\lambda} \in \mathbb{R}^K$, as follows. For each $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathbb{R}^K$, let $\boldsymbol{p}_{\boldsymbol{\lambda}}(\boldsymbol{x}) \in \boldsymbol{\Delta}_C$ be the probability vector whose c-th coordinate is

$$\varphi_c\left(\boldsymbol{x}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda}) - \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x})\right). \tag{B.123}$$

When $\boldsymbol{\lambda} \geq \boldsymbol{0}$, we get $q^{\boldsymbol{\Delta}_c}_{\boldsymbol{\lambda}} = \boldsymbol{p}_{\boldsymbol{\lambda}}$. Let $\mathcal{Q}' = \{\boldsymbol{p}_{\boldsymbol{\lambda}} \mid \|\boldsymbol{\lambda}\|_1 \leq \theta\}$.

We have the Lipshitz conditions

$$|\varphi_c(\boldsymbol{x}, u) - \varphi_c(\boldsymbol{x}, u')| \leq A\sqrt{m+1}|u - u'| \tag{B.124}$$

$$|\gamma(\boldsymbol{x}, \boldsymbol{\lambda}) - \gamma(\boldsymbol{x}, \boldsymbol{\lambda}')| \leq B(2+\theta)\sqrt{m+K}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_1^2 \tag{B.125}$$

for every $\boldsymbol{x} \in \mathbb{R}^m$, u, u' such that $(\boldsymbol{x}, u), (\boldsymbol{x}, u') \in \Omega_c$, and $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathcal{B}_1(\boldsymbol{0}, \theta)$. Let

$$L = \max\left(A\sqrt{m+1}, B(2+\theta)\sqrt{m+K}\right). \tag{B.126}$$

Fix $\nu > 0$, and set $\delta = \min(1, \nu/(LC(L+E)))$. Let $N \in \mathbb{N}$ and $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_N \in \mathcal{B}_1(\boldsymbol{0}, \theta)$ be such that the balls $\mathcal{B}_1(\boldsymbol{\lambda}_r, \delta)$ cover $\mathcal{B}_1(\boldsymbol{0}, \theta)$. Fix $\boldsymbol{p}_{\boldsymbol{\lambda}} \in \mathcal{Q}'$. Let $r \in [N]$ be such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_r\|_1 \leq \delta$. Then, for every $\boldsymbol{x} \in \mathbb{R}^m$,

$$\|\boldsymbol{p}_{\boldsymbol{\lambda}}(\boldsymbol{x}) - \boldsymbol{p}_{\boldsymbol{\lambda}_r}(\boldsymbol{x})\|_1 = \sum_{c \in [C]} \left| \varphi_c\left(\boldsymbol{x}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda}) - \sum_{k \in [K]} \lambda_k G_{k,c}(\boldsymbol{x})\right) - \varphi_c\left(\boldsymbol{x}, \gamma(\boldsymbol{x}, \boldsymbol{\lambda}_r) - \sum_{k \in [K]} \lambda_{r,k} G_{k,c}(\boldsymbol{x})\right) \right| \tag{B.127}$$

$$\leq L \sum_{c\in[C]} \left| \gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \gamma(\boldsymbol{x},\boldsymbol{\lambda}_r) + \sum_{k\in[K]} (\lambda_{r,k} - \lambda_k) G_{k,c}(\boldsymbol{x}) \right| \tag{B.128}$$

$$\leq LC \left(|\gamma(\boldsymbol{x},\boldsymbol{\lambda}) - \gamma(\boldsymbol{x},\boldsymbol{\lambda}_r)| + E\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_r\|_1 \right) \tag{B.129}$$

$$\leq LC(L\delta^2 + E\delta) \tag{B.130}$$

$$\leq \nu. \tag{B.131}$$

Therefore, $\mathcal{Q}'$ is totally bounded. Hence, $\mathcal{Q}$ is totally bounded too. As $\mathcal{C}(\mathbb{R}^m, \boldsymbol{\Delta}_C)$ is a complete metric space, $\mathcal{Q}$ is precompact, and the proof is complete.

B.1.6 Proof of Lemma B.2: the Convex Conjugate Formula

By definition of the convex conjugate (Definition B.1), for any $\boldsymbol{v} \in \mathbb{R}^C$

$$F^{\mathrm{conj}}(\boldsymbol{v}) = \sup_{\boldsymbol{p}\in\boldsymbol{\Delta}_C} \boldsymbol{v}^T\boldsymbol{p} - F(\boldsymbol{p}) = -\inf\{F(\boldsymbol{p}) - \boldsymbol{v}^T\boldsymbol{p} \mid \boldsymbol{p} \in [0,1]^C, \mathbf{1}^T\boldsymbol{p} = 1\}. \tag{B.132}$$

Fix $\boldsymbol{v}$. Let $\eta_{\boldsymbol{v}} := \min_{c\in[C]} f'(1/q_c) - v_c$. For any $\gamma \in (-\infty, \eta_{\boldsymbol{v}})$, define $\boldsymbol{p}(\gamma) \in (0,1)^C$ by

$$p_c(\gamma) := q_c\phi(\gamma + v_c). \tag{B.133}$$

Note that both f' and ϕ are strictly increasing and continuous functions, so for any $\gamma \in (-\infty, \eta_{\boldsymbol{v}})$,

$$0 = \lim_{t\to-\infty} p_c(t) < p_c(\gamma) < q_c\phi(\eta_{\boldsymbol{v}} + v_c) \leq q_c\phi(f'(1/q_c)) = 1 \tag{B.134}$$

for every $c \in [C]$. Let $a \in [C]$ be such that $\eta_{\boldsymbol{v}} = f'(1/q_a) - v_a$. We have that

$$\lim_{\gamma\to\eta_{\boldsymbol{v}}} p_a(\gamma) = q_a \lim_{u\to f'(1/q_a)} \phi(u) = 1, \tag{B.135}$$

so

$$\lim_{\gamma\to\eta_{\boldsymbol{v}}} \sum_{c\in[C]} p_c(\gamma) > 1. \tag{B.136}$$

On the other hand,

$$\lim_{\gamma\to-\infty} \sum_{c\in[C]} p_c(\gamma) = 0. \tag{B.137}$$

The intermediate value theorem implies that $\gamma(\boldsymbol{v})$ as given in (B.47) is well-defined.

Introducing a Lagrange multiplier η

$$F^{\mathrm{conj}}(\boldsymbol{v}) = -\inf_{\boldsymbol{p}\in[0,1]^C} \sup_{\eta\in\mathbb{R}} F(\boldsymbol{p}) - \boldsymbol{v}^T\boldsymbol{p} - \eta(\mathbf{1}^T\boldsymbol{p} - 1). \tag{B.138}$$

Define $g_{\boldsymbol{v}} : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ by

$$g_{\boldsymbol{v}}(\eta) := \inf_{\boldsymbol{p} \in [0,1]^C} F(\boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{p} - \eta(\mathbf{1}^T \boldsymbol{p} - 1). \tag{B.139}$$

Note that

$$g_{\boldsymbol{v}}(\gamma(\boldsymbol{v})) = F(\boldsymbol{p}(\boldsymbol{v})) - \boldsymbol{v}^T \boldsymbol{p}(\boldsymbol{v}). \tag{B.140}$$

Indeed, we have $(0,1]^C \subset \mathbf{dom}\, F$ and

$$\nabla \left(F(\boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{p} - \gamma(\boldsymbol{v})(\mathbf{1}^T \boldsymbol{p} - 1) \right) \Big|_{\boldsymbol{p} = \boldsymbol{p}(\gamma(\boldsymbol{v}))} = \left(f' \left(\frac{p_c(\gamma(\boldsymbol{v}))}{q_c} \right) - v_c - \gamma(\boldsymbol{v}) \right)_{c \in [C]} = \mathbf{0}, \tag{B.141}$$

so (B.140) follows by convexity of F. Then,

$$\begin{aligned}
F^{\mathrm{conj}}(\boldsymbol{v}) &= - \inf_{\boldsymbol{p} \in [0,1]^C} \sup_{\eta \in \mathbb{R}} F(\boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{p} - \eta(\mathbf{1}^T \boldsymbol{p} - 1) && \text{(B.142)} \\
&\leq - \sup_{\eta \in \mathbb{R}} \inf_{\boldsymbol{p} \in [0,1]^C} F(\boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{p} - \eta(\mathbf{1}^T \boldsymbol{p} - 1) && \text{(B.143)} \\
&= - \sup_{\eta \in \mathbb{R}} g_{\boldsymbol{v}}(\eta) && \text{(B.144)} \\
&\leq -g_{\boldsymbol{v}}(\gamma(\boldsymbol{v})) && \text{(B.145)} \\
&= \boldsymbol{v}^T \boldsymbol{p}(\boldsymbol{v}) - F(\boldsymbol{p}(\boldsymbol{v})). && \text{(B.146)}
\end{aligned}$$

Therefore, formula (B.46) holds.

Further, by strict convexity of F, $\boldsymbol{p}(\boldsymbol{v})$ is the unique minimizer of $F(\boldsymbol{h}) - \boldsymbol{v}^T \boldsymbol{h}$ for $\boldsymbol{h} \in \boldsymbol{\Delta}_C^+$. We show that $\boldsymbol{q}^{\mathrm{conj}}(\boldsymbol{v}) = \boldsymbol{p}(\boldsymbol{v})$. If $f(0^+) = \infty$, then F takes the value ∞ on the relative boundary $\boldsymbol{\Delta}_C \setminus \boldsymbol{\Delta}_C^+$ of $\boldsymbol{\Delta}_C$, so $\boldsymbol{p}(\boldsymbol{v})$ is the unique minimizer of $F(\boldsymbol{h}) - \boldsymbol{v}^T \boldsymbol{h}$ over $\boldsymbol{h} \in \boldsymbol{\Delta}_C$, i.e., $\boldsymbol{q}^{\mathrm{conj}}(\boldsymbol{v}) = \boldsymbol{p}(\boldsymbol{v})$. Assume $f(0^+) < \infty$. Then, F is convex over $\boldsymbol{\Delta}_C$. Let $G(\boldsymbol{h}) = F(\boldsymbol{h}) - \boldsymbol{v}^T \boldsymbol{h}$. For $\boldsymbol{h} \in \boldsymbol{\Delta}_C$ such that $G(\boldsymbol{h}) \leq G(\boldsymbol{p}(\boldsymbol{v}))$, the point $\frac{1}{2}(\boldsymbol{p}(\boldsymbol{v}) + \boldsymbol{h})$ lies in $\boldsymbol{\Delta}_C^+$ and satisfies

$$G\left(\frac{1}{2}(\boldsymbol{p}(\boldsymbol{v}) + \boldsymbol{h}) \right) \leq \frac{1}{2}(G(\boldsymbol{p}(\boldsymbol{v})) + G(\boldsymbol{h})) \leq G(\boldsymbol{p}(\boldsymbol{v})), \tag{B.147}$$

so by uniqueness of $\boldsymbol{p}(\boldsymbol{v})$, we must have $\boldsymbol{h} = \boldsymbol{p}(\boldsymbol{v})$. Therefore, $\boldsymbol{p}(\boldsymbol{v})$ is the unique minimizer of G over $\boldsymbol{\Delta}_C$ when $f(0^+) < \infty$ too, and $\boldsymbol{q}^{\mathrm{conj}}(\boldsymbol{v}) = \boldsymbol{p}(\boldsymbol{v})$, completing the proof of equation (B.48) and the lemma.

B.2 Proof of Theorem 3.3: Strong Duality

We use the following minimax theorem, which is a generalization of Sion's minimax theorem.

Theorem B.7 ([ET99, Chapter VI, Prop. 2.2])**.** *Let V and Z be two reflexive Banach spaces, and fix two convex, closed, and non-empty subsets $\mathcal{A} \subset V$ and $\mathcal{B} \subset Z$. Let $L : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be a function such that for each $u \in \mathcal{A}$ the function $p \mapsto L(u, p)$ is concave and upper semicontinuous, and for each $p \in \mathcal{B}$ the function $u \mapsto L(u, p)$ is convex and lower semicontinuous. Suppose that there exist points $u_0 \in \mathcal{A}$ and $p_0 \in \mathcal{B}$ such that $\lim_{p\in\mathcal{B},\|p\|\to\infty} L(u_0, p) = -\infty$ and $\lim_{u\in\mathcal{A},\|u\|\to\infty} L(u, p_0) = \infty$. Then, L has at least one saddle-point $(\overline{u}, \overline{p})$, and*

$$L(\overline{u}, \overline{p}) = \min_{u\in\mathcal{A}} \sup_{p\in\mathcal{B}} L(u, p) = \max_{p\in\mathcal{B}} \inf_{u\in\mathcal{A}} L(u, p). \tag{B.148}$$

In particular, in (B.148), there exists a minimizer in $\mathcal{A}$ of the outer minimization, and a maximizer in $\mathcal{B}$ of the outer maximization.

Denote $\boldsymbol{h}_i := \boldsymbol{h}(X_i)$, $\boldsymbol{p}_i := \boldsymbol{h}^{\text{base}}(X_i)$, $\boldsymbol{a}_i := \boldsymbol{a}(X_i)$, and $\boldsymbol{G}_i := \boldsymbol{G}(X_i)$, and let the matrix $\boldsymbol{\mathcal{G}}_N := \left(\boldsymbol{G}_1/\sqrt{N}, \cdots, \boldsymbol{G}_N/\sqrt{N}, \boldsymbol{I}_K\right) \in \mathbb{R}^{K\times(NC+K)}$ be as in the theorem statement. We may rewrite the optimization (3.27) as

$$\begin{aligned} \underset{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]}{\text{minimize}} \quad & \frac{1}{N}\sum_{i\in[N]} D_f\left(\boldsymbol{h}_i \| \boldsymbol{p}_i\right) + \tau_1 \cdot \left(\|\boldsymbol{a}_i\|_2^2 + \|\boldsymbol{b}\|_2^2\right) \\ \text{subject to} \quad & \frac{1}{N}\sum_{i\in[N]} \boldsymbol{G}_i\boldsymbol{h}_i + \tau_2 \cdot (\boldsymbol{G}_i\boldsymbol{a}_i - \boldsymbol{b}) \leq \boldsymbol{0}. \end{aligned} \tag{B.149}$$

We define f at 0 by the right limit $f(0) := f(0+)$. Assume for now that $f(0+) < \infty$, and we will explain at the end of this proof how to treat the case $f(0+) = \infty$. For the optimization problem (B.149), the Lagrangian $L : \boldsymbol{\Delta}_C^N \times \mathbb{R}^{NC} \times \mathbb{R}^K \times \mathbb{R}_+^K \to \mathbb{R}$ is given by

$$\begin{aligned} L\left((\boldsymbol{h}_i)_{i\in[N]}, (\boldsymbol{a}_i)_{i\in[N]}, \boldsymbol{b}, \boldsymbol{\lambda}\right) := \frac{1}{N}\sum_{i\in[N]} & D_f\left(\boldsymbol{h}_i\|\boldsymbol{p}_i\right) + \tau_1\left(\|\boldsymbol{a}_i\|_2^2 + \|\boldsymbol{b}\|_2^2\right) \\ & + \boldsymbol{\lambda}^T\left(\boldsymbol{G}_i\boldsymbol{h}_i + \tau_2\left(\boldsymbol{G}_i\boldsymbol{a}_i - \boldsymbol{b}\right)\right). \end{aligned} \tag{B.150}$$

With $\boldsymbol{v}(x;\boldsymbol{\lambda}) := -\boldsymbol{G}(x)^T\boldsymbol{\lambda}$ as in the theorem statement, and denoting $\boldsymbol{v}_i := \boldsymbol{v}(X_i;\boldsymbol{\lambda}) = -\boldsymbol{G}_i^T\boldsymbol{\lambda}$, we may rewrite the Lagrangian as

$$\begin{aligned} L\left((\boldsymbol{h}_i)_{i\in[N]}, (\boldsymbol{a}_i)_{i\in[N]}, \boldsymbol{b}, \boldsymbol{\lambda}\right) = \frac{1}{N}\sum_{i\in[N]} & D_f\left(\boldsymbol{h}_i\|\boldsymbol{p}_i\right) - \boldsymbol{v}_i^T\boldsymbol{h}_i + \tau_1\|\boldsymbol{a}_i\|_2^2 - \tau_2\boldsymbol{v}_i^T\boldsymbol{a}_i \\ & + \tau_1\|\boldsymbol{b}\|_2^2 - \tau_2\boldsymbol{\lambda}^T\boldsymbol{b}. \end{aligned} \tag{B.151}$$

The optimization problem (B.149) can be written as

$$\inf_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^K_+} L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right). \tag{B.152}$$

We check that the Lagrangian L satisfies the conditions in Theorem B.7. First, any Euclidean space $\mathbb{R}^M$ (for $M\in\mathbb{N}$) is a reflexive Banach space since it is finite-dimensional. In addition, the convex nonempty sets $\boldsymbol{\Delta}_C^N\times\mathbb{R}^{NC}\times\mathbb{R}^K$ and $\mathbb{R}^K_+$ are closed in their respective ambient Euclidean spaces. By continuity and convexity of f, and linearity of L in $\boldsymbol{\lambda}$, we have that L satisfies all the convexity, concavity, and semicontinuity conditions in Theorem B.7. Further, fixing any $\boldsymbol{h}_i\in\boldsymbol{\Delta}_C$, $i\in[N]$, and letting $\boldsymbol{a}_i=\boldsymbol{0}$, $i\in[N]$, and $\boldsymbol{b}=\frac{1}{\tau_2}\left(\boldsymbol{1}+\frac{1}{N}\sum_{i\in[N]}\boldsymbol{G}_i\boldsymbol{h}_i\right)$, we would get that

$$L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right)=-\boldsymbol{\lambda}^T\boldsymbol{1}+\frac{1}{N}\sum_{i\in[N]}D_f\left(\boldsymbol{h}_i\|\boldsymbol{p}_i\right)+\tau_1\|\boldsymbol{b}\|_2^2\to-\infty \quad \text{as} \quad \|\boldsymbol{\lambda}\|_2\to\infty. \tag{B.153}$$

In addition, choosing $\boldsymbol{\lambda}=\boldsymbol{0}$, we have the Lagrangian

$$L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right)=\frac{1}{N}\sum_{i\in[N]}D_f\left(\boldsymbol{h}_i\|\boldsymbol{p}_i\right)+\tau_1\|\boldsymbol{a}_i\|_2^2+\tau_1\|\boldsymbol{b}\|_2^2\to\infty \tag{B.154}$$

as $\|\boldsymbol{b}\|_2+\sum_{i\in[N]}\|\boldsymbol{h}_i\|_2+\|\boldsymbol{a}_i\|_2\to\infty$. Thus, we may apply the minimax result in Theorem B.7 to obtain the existence of a saddle-point of L and that

$$\begin{aligned}\min_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} &\sup_{\boldsymbol{\lambda}\in\mathbb{R}^K_+} L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right)\\ &=\max_{\boldsymbol{\lambda}\in\mathbb{R}^K_+}\inf_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right).\end{aligned} \tag{B.155}$$

In particular, there exists a minimizer $(\boldsymbol{h}_i^{\text{opt},N},\boldsymbol{a}_i^{\text{opt},N},\boldsymbol{b}^{\text{opt},N})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]$, of the outer minimization in the left-hand side in (B.155), and a maximizer $\boldsymbol{\lambda}^\star\in\mathbb{R}^K_+$ of the outer maximization in the right-hand side of (B.155). By strict convexity of the objective function in (B.149) (and convexity of the feasibility set), we obtain that the minimizer $(\boldsymbol{h}_i^{\text{opt},N},\boldsymbol{a}_i^{\text{opt},N},\boldsymbol{b}^{\text{opt},N})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]$, is unique. We show next that the optimizer $\boldsymbol{\lambda}^\star$ is unique too, which we will denote by $\boldsymbol{\lambda}^\star_{\zeta,N}$ as in the theorem statement. We also show that, for each fixed $\boldsymbol{\lambda}\in\mathbb{R}^K_+$, there is a unique minimizer $(\boldsymbol{h}_i^{\boldsymbol{\lambda}},\boldsymbol{a}_i^{\boldsymbol{\lambda}},\boldsymbol{b}^{\boldsymbol{\lambda}})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]$, of the *inner* minimization in the right-hand side of (B.155); by strict convexity of f, this would imply that $\boldsymbol{h}_i^{\text{opt},N}=\boldsymbol{h}_i^{\boldsymbol{\lambda}^\star_{\zeta,N}}$.

Now, fix $\boldsymbol{\lambda}\in\mathbb{R}^K_+$, and consider the inner minimization in (B.155). We have that

$$\inf_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} L\left((\boldsymbol{h}_i)_{i\in[N]},(\boldsymbol{a}_i)_{i\in[N]},\boldsymbol{b},\boldsymbol{\lambda}\right)$$

$$= \inf_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} \frac{1}{N}\sum_{i\in[N]} D_f\left(\boldsymbol{h}_i\|\boldsymbol{p}_i\right) - \boldsymbol{v}_i^T\boldsymbol{h}_i + \tau_1\|\boldsymbol{a}_i\|_2^2 - \tau_2\boldsymbol{v}_i^T\boldsymbol{a}_i + \tau_1\|\boldsymbol{b}\|_2^2 - \tau_2\boldsymbol{\lambda}^T\boldsymbol{b} \tag{B.156}$$

$$= \frac{1}{N}\sum_{i\in[N]} \inf_{\boldsymbol{h}_i\in\boldsymbol{\Delta}_C} D_f(\boldsymbol{h}_i\|\boldsymbol{p}_i) - \boldsymbol{v}_i^T\boldsymbol{h}_i + \inf_{\boldsymbol{a}_i\in\mathbb{R}^C} \tau_1\|\boldsymbol{a}_i\|_2^2 - \tau_2\boldsymbol{v}_i^T\boldsymbol{a}_i + \inf_{\boldsymbol{b}\in\mathbb{R}^K} \tau_1\|\boldsymbol{b}\|_2^2 - \tau_2\boldsymbol{\lambda}^T\boldsymbol{b} \tag{B.157}$$

$$= \frac{1}{N}\sum_{i\in[N]} -D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i) - \frac{1}{2}\zeta\|\boldsymbol{v}_i\|_2^2 - \frac{1}{2}\zeta\|\boldsymbol{\lambda}\|_2^2 \tag{B.158}$$

$$= -\frac{\zeta}{2}\left\|\boldsymbol{\mathcal{G}}_N^T\boldsymbol{\lambda}\right\|_2^2 - \frac{1}{N}\sum_{i\in[N]} D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i) \tag{B.159}$$

where $\zeta := \tau_2^2/(2\tau_1)$. Here, the minimizers are $\boldsymbol{a}_i^{\boldsymbol{\lambda}} := \frac{\tau_2}{2\tau_1}\boldsymbol{v}_i$ and $\boldsymbol{b}_i^{\boldsymbol{\lambda}} := \frac{\tau_2}{2\tau_1}\boldsymbol{\lambda}$, and $\boldsymbol{h}_i^{\boldsymbol{\lambda}}$ is the unique probability vector in $\boldsymbol{\Delta}_C$ for which $D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i) = D_f(\boldsymbol{h}_i^{\boldsymbol{\lambda}}\|\boldsymbol{p}_i) - \boldsymbol{v}_i^T\boldsymbol{h}_i^{\boldsymbol{\lambda}}$; the existence and uniqueness of $\boldsymbol{h}_i^{\boldsymbol{\lambda}}$ is guaranteed since $\boldsymbol{q} \mapsto D_f(\boldsymbol{q}\|\boldsymbol{p}_i) - \boldsymbol{v}_i^T\boldsymbol{q}$ is lower semicontinuous and strictly convex, and $\boldsymbol{\Delta}_C$ is compact. Rewriting it in the form (B.159), the function

$$\boldsymbol{\lambda} \mapsto \inf_{(\boldsymbol{h}_i,\boldsymbol{a}_i,\boldsymbol{b})\in\boldsymbol{\Delta}_C\times\mathbb{R}^C\times\mathbb{R}^K, i\in[N]} L\left((\boldsymbol{h}_i)_{i\in[N]}, (\boldsymbol{a}_i)_{i\in[N]}, \boldsymbol{b}, \boldsymbol{\lambda}\right) \tag{B.160}$$

can be seen to be strictly concave. Indeed, the function $\boldsymbol{\lambda} \mapsto \left\|\boldsymbol{\mathcal{G}}_N^T\boldsymbol{\lambda}\right\|_2^2$ is strictly convex. Also, each function $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i)$ is convex as it is a pointwise supremum of linear functions: recalling that $\boldsymbol{v}_i = -\boldsymbol{G}_i^T\boldsymbol{\lambda}$, we have the formula

$$D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i) = \sup_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} -\boldsymbol{q}^T\boldsymbol{G}_i^T\boldsymbol{\lambda} - D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}_i). \tag{B.161}$$

Hence, the outer maximizer $\boldsymbol{\lambda}^\star$ in (B.155) is indeed unique, which we denote by $\boldsymbol{\lambda}_{\zeta,N}^\star$. Note that $\boldsymbol{\lambda}_{\zeta,N}^\star$ is the unique solution to the *minimization* (3.29), i.e.,

$$\boldsymbol{\lambda}_{\zeta,N}^\star = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}_+^K} \frac{1}{N}\sum_{i\in[N]} D_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i) + \frac{\zeta}{2}\left\|\boldsymbol{\mathcal{G}}_N^T\boldsymbol{\lambda}\right\|_2^2, \tag{B.162}$$

as stated by the theorem.

Since $\boldsymbol{h}^{\text{opt},N} = \boldsymbol{h}^{\boldsymbol{\lambda}_{\zeta,N}^\star}$, the following formula for $\boldsymbol{h}^{\boldsymbol{\lambda}}$ (for a general $\boldsymbol{\lambda} \in \mathbb{R}_+^K$) yields the desired functional form (3.28) for $\boldsymbol{h}^{\text{opt},N}$ in terms of $\boldsymbol{\lambda}_{\zeta,N}^\star$. Specifically, we restate here the relevant part of Lemma B.2 below.

Lemma B.4. *Let $f : [0,\infty) \to \mathbb{R} \cup \{\infty\}$ be a strictly convex function that is continuously differentiable over $(0,\infty)$ and satisfying $f(0) = f(0+)$, $f(1) = 0$, and $f'(0+) = -\infty$. Let ϕ denote the inverse of f'. Fix $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$ and $\boldsymbol{v} \in \mathbb{R}^C$. Then, the unique minimizer of $\boldsymbol{q} \mapsto D_f(\boldsymbol{q}\|\boldsymbol{p}) - \boldsymbol{v}^T\boldsymbol{q}$ over $\boldsymbol{q} \in \boldsymbol{\Delta}_C$ is given by $q_c^\star = p_c \cdot \phi(\gamma + v_c)$, $c \in [C]$, where $\gamma \in \mathbb{R}$ is the unique number satisfying $\mathbb{E}_{c\sim\boldsymbol{p}}[\phi(\gamma + v_c)] = 1$.*

From Lemma B.4, and using $\boldsymbol{v}(x;\boldsymbol{\lambda}^\star_{\zeta,N}) = -\boldsymbol{G}(x)^T\boldsymbol{\lambda}^\star_{\zeta,N}$ and $\phi = (f')^{-1}$, we get that there exists a uniquely defined function $\gamma : \mathbb{X} \times \mathbb{R}^K \to \mathbb{R}$ for which

$$\mathbb{E}_{c\sim\boldsymbol{h}^{\text{base}}(x)}\left[\phi\left(\gamma(x;\boldsymbol{\lambda}^\star_{\zeta,N}) + v_c(x;\boldsymbol{\lambda}^\star_{\zeta,N})\right)\right] = 1 \tag{B.163}$$

for every $x \in \mathbb{X}$. For this γ, we know from Lemma B.4 that

$$h_c^{\boldsymbol{\lambda}^\star_{\zeta,N}}(x) = h_c^{\text{base}}(x) \cdot \phi\left(\gamma(x;\boldsymbol{\lambda}^\star_{\zeta,N}) + v_c(x;\boldsymbol{\lambda}^\star_{\zeta,N})\right) \tag{B.164}$$

for every $c \in [C]$ and $x \in \mathbb{X}$. Since $\boldsymbol{h}^{\text{opt},N} = \boldsymbol{h}^{\boldsymbol{\lambda}^\star_{\zeta,N}}$, we obtain formula (3.28) for $\boldsymbol{h}^{\text{opt},N}$ in terms of $\boldsymbol{\lambda}^\star_{\zeta,N}$, and the proof of Theorem 3.3 is complete in the case $f(0+) < \infty$.

Finally, we note how the case $f(0+) = \infty$ is treated, so assume $f(0) = f(0+) = \infty$. The only difference in this case is that the Lagrangian L might attain the value ∞, whereas we need it to be $\mathbb{R}$-valued to apply the minimax result in Theorem B.7. Nevertheless, the only way L can be infinite is if some classifier $\boldsymbol{h}_i$ has an entry equal to 0, in which case the objective function in (3.27) (or (B.149)) will also be infinite, so such a classifier can be thrown out without affecting the optimization problem. More precisely, we still have strict convexity and lower semicontinuity of the objective function in (B.149). Thus, there is a unique minimizer $\boldsymbol{h}^{\text{opt},N}$ of (B.149). For this optimizer, there must be an $\varepsilon_1 > 0$ such that $\boldsymbol{h}^{\text{opt},N}(x) \geq \varepsilon_1\mathbf{1}$ for *every* $x \in \mathbb{X}$. Thus, the optimization problem (B.149) remains unchanged if $\boldsymbol{\Delta}_C$ is restricted to classifiers bounded away from 0 by ε_1. Moreover, by the same reasoning, the optimization problem (B.161) for finding D_f^{conj} also remains unchanged if $\boldsymbol{\Delta}_C$ is replaced by the set of classifiers bounded away from 0 by some $\varepsilon_2 > 0$ that is *independent* of the X_i. Hence, choosing $\varepsilon = \min(\varepsilon_1, \varepsilon_2) > 0$, and replacing $\boldsymbol{\Delta}_C$ by $\widetilde{\boldsymbol{\Delta}}_C := \{\boldsymbol{q} \in \boldsymbol{\Delta}_C\, ;\, \boldsymbol{q} \geq \varepsilon\mathbf{1}\}$ in the above proof, we attain the same results for the case $f(0+) = \infty$.

B.3 Proofs of Section 3.6: Theoretical Results for `FairProjection`

The theoretical details for `FairProjection` stated in Section 3.6 are proved in this appendix. The outline is as follows:

- Algorithm 1 is derived in Appendix B.3.1.
- The inner iterations of Algorithm 1 are further developed in Appendix B.3.2.
- The $\frac{1}{2}$-Lipschitzness of the softmax function (Proposition 3.1) is proved in Appendix B.3.3.

- The convergence rate result in Theorem 3.4 is proved in Appendix B.3.4, and an extension of it (to general f-divergences) is discussed in Appendix B.3.5.
- The performance of `FairProjection` for the population problem (3.19) as stated in Theorem 3.5 is proved in Appendix B.3.6.

B.3.1 Algorithm 1: derivation of the ADMM iterations

ADMM is applicable to problems taking the form

$$\begin{aligned} \underset{(\boldsymbol{V},\boldsymbol{\lambda})\in\mathbb{R}^V\times\mathbb{R}^K}{\text{minimize}} \quad & F(\boldsymbol{V})+\psi(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{V}+\boldsymbol{B}\boldsymbol{\lambda}=\boldsymbol{m}, \end{aligned} \tag{B.165}$$

where $F:\mathbb{R}^V\to\mathbb{R}\cup\{\infty\}$ and $\psi:\mathbb{R}^K\to\mathbb{R}\cup\{\infty\}$ are closed proper convex functions, and $\boldsymbol{A}\in\mathbb{R}^{U\times V}$, $\boldsymbol{B}\in\mathbb{R}^{U\times K}$, and $\boldsymbol{m}\in\mathbb{R}^U$ are fixed.

We rewrite the convex problem (3.29) into the ADMM form (B.165) as follows. With the samples $X_1,\cdots,X_N \overset{\text{i.i.d.}}{\sim} P_X$ fixed, we denote the following fixed vectors and matrices: for each $i\in[N]$, set

$$\boldsymbol{p}_i := \boldsymbol{h}^{\text{base}}(X_i)\in\boldsymbol{\Delta}_C^+ = \{\boldsymbol{q}\in\boldsymbol{\Delta}_C\ ;\ \boldsymbol{q}>\boldsymbol{0}\}, \tag{B.166}$$

$$\boldsymbol{G}_i := \boldsymbol{G}(X_i)\in\mathbb{R}^{K\times C}. \tag{B.167}$$

We introduce a variable $\boldsymbol{V}:=(\boldsymbol{v}_i)_{i\in[N]}\in\mathbb{R}^{NC}$ (with components $\boldsymbol{v}_i\in\mathbb{R}^C$), and consider the objective functions

$$F(\boldsymbol{V}) := \frac{1}{N}\sum_{i\in[N]} D_f^{\text{conj}}(\boldsymbol{v}_i,\boldsymbol{p}_i)+\frac{\zeta}{2}\|\boldsymbol{V}\|_2^2, \tag{B.168}$$

$$\psi(\boldsymbol{\lambda}) := \mathbb{I}_{\mathbb{R}_+^K}(\boldsymbol{\lambda})+\frac{\zeta}{2}\|\boldsymbol{\lambda}\|_2^2. \tag{B.169}$$

Then, setting[5]

$$\boldsymbol{A}=\frac{1}{\sqrt{N}}\boldsymbol{I}_{NC},\quad \boldsymbol{B}=\frac{1}{\sqrt{N}}(\boldsymbol{G}_i)_{i\in[N]}^T,\ \text{and}\quad \boldsymbol{m}=\boldsymbol{0}_{NC}, \tag{B.170}$$

our finite-sample problem (3.29) takes the ADMM form (B.165).

In addition, this reparametrization allows us to parallelize the ADMM iterations, which we briefly review next. One starts with forming the augmented Lagrangian for problem (B.165), $L_\rho:\mathbb{R}^V\times\mathbb{R}^K\times\mathbb{R}^U\to\mathbb{R}\cup\{\infty\}$, where $\rho>0$ is a fixed *penalty parameter* and $\boldsymbol{U}\in\mathbb{R}^U$ denotes a *dual*

[5] The prefactor $1/\sqrt{N}$ is unnecessary since $\boldsymbol{m}=\boldsymbol{0}$, but we introduce it to simplify the ensuing expressions.

variable, by

$$L_\rho(\boldsymbol{V},\boldsymbol{\lambda},\boldsymbol{U}) := F(\boldsymbol{V}) + \psi(\boldsymbol{\lambda}) + \boldsymbol{U}^T\left(\boldsymbol{A}\boldsymbol{V} + \boldsymbol{B}\boldsymbol{\lambda} - \boldsymbol{m}\right) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{V} + \boldsymbol{B}\boldsymbol{\lambda} - \boldsymbol{m}\right\|_2^2. \tag{B.171}$$

The ADMM iterations then repeatedly update the triplet after the t-th iteration $(\boldsymbol{V}^{(t)},\boldsymbol{\lambda}^{(t)},\boldsymbol{U}^{(t)})$ into a triplet $(\boldsymbol{V}^{(t+1)},\boldsymbol{\lambda}^{(t+1)},\boldsymbol{U}^{(t+1)})$ that is given by

$$\boldsymbol{V}^{(t+1)} \in \operatorname*{argmin}_{\boldsymbol{V}\in\mathbb{R}^V} L_\rho(\boldsymbol{V},\boldsymbol{\lambda}^{(t)},\boldsymbol{U}^{(t)}), \tag{B.172}$$

$$\boldsymbol{\lambda}^{(t+1)} \in \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^V} L_\rho(\boldsymbol{V}^{(t+1)},\boldsymbol{\lambda},\boldsymbol{U}^{(t)}), \tag{B.173}$$

$$\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} + \rho\cdot\left(\boldsymbol{A}\boldsymbol{V}^{(t+1)} + \boldsymbol{B}\boldsymbol{\lambda}^{(t+1)}\right). \tag{B.174}$$

We next instantiate the ADMM iterations to our problem, and we note that we will consider the scaled dual variable $\boldsymbol{W} = \sqrt{N}\boldsymbol{U}$.

In our case, the augmented Lagrangian splits into non-interacting components along the $\boldsymbol{v}_i$. This splitting allows parallelizability of the $\boldsymbol{V}$-update step, which is the most computationally intensive step. Consider a conforming decomposition $\boldsymbol{U} = (\boldsymbol{u}_i)_{i\in[N]}$ for $\boldsymbol{u}_i \in \mathbb{R}^C$, and let $\boldsymbol{W} = \sqrt{N}\boldsymbol{U}$. With some algebra, one can show that the ADMM iterations for the ADMM problem specified by (B.168)–(B.170) are expressible by[6]

$$\boldsymbol{v}_i^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{v}\in\mathbb{R}^C} D_f^{\text{conj}}(\boldsymbol{v},\boldsymbol{p}_i) + \mathcal{R}_i^{(t)}(\boldsymbol{v}), \qquad i\in[N], \tag{B.175}$$

$$\boldsymbol{\lambda}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}_+^K} \boldsymbol{\lambda}^T\boldsymbol{Q}\boldsymbol{\lambda} + \boldsymbol{q}^{(t)T}\boldsymbol{\lambda}, \tag{B.176}$$

$$\boldsymbol{w}_i^{(t+1)} = \boldsymbol{w}_i^{(t)} + \rho\cdot\left(\boldsymbol{v}_i^{(t+1)} + \boldsymbol{G}_i^T\boldsymbol{\lambda}^{(t+1)}\right), \qquad i\in[N], \tag{B.177}$$

where $\mathcal{R}_i^{(t)} : \mathbb{R}^C \to \mathbb{R}$ is the quadratic form

$$\mathcal{R}_i^{(t)}(\boldsymbol{v}) := \frac{\rho+\zeta}{2}\|\boldsymbol{v}\|_2^2 + \left(\boldsymbol{w}_i^{(t)} + \rho\boldsymbol{G}_i^T\boldsymbol{\lambda}^{(t)}\right)^T\boldsymbol{v}, \tag{B.178}$$

and the fixed matrix $\boldsymbol{Q}\in\mathbb{R}^{K\times K}$ and vectors $\boldsymbol{q}^{(t)}\in\mathbb{R}^K$ are given by

$$\boldsymbol{Q} := \frac{\zeta}{2}\boldsymbol{I}_K + \frac{\rho}{2N}\sum_{i\in[N]}\boldsymbol{G}_i\boldsymbol{G}_i^T, \tag{B.179}$$

[6] Note also that in these specific ADMM iterations, unlike in the general ADMM iterations, we write "= argmin" as opposed to "∈ argmin" since strict convexity and coercivity guarantee that a unique minimizer exists (see [CST17] for a case where argmin is empty). Also, we write here $\boldsymbol{q}^{(t)T}$ instead of $\left(\boldsymbol{q}^{(t)}\right)^T$ for readability.

$$\boldsymbol{q}^{(t)} := \frac{1}{N}\sum_{i\in[N]} \boldsymbol{G}_i \cdot \left(\boldsymbol{w}_i^{(t)} + \boldsymbol{v}_i^{(t+1)}\right). \tag{B.180}$$

Note that both the first (B.175) and last (B.177) steps can be carried out for each sample $i \in [N]$ in parallel.

B.3.2 The inner iterations: minimizing the convex conjugate of f-divergence

Only updating the primal-variable $\boldsymbol{v}_i$ in Algorithm 1, i.e., solving

$$\min_{\boldsymbol{v}\in\mathbb{R}^C} D_f^{\mathrm{conj}}(\boldsymbol{v},\boldsymbol{p}) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T\boldsymbol{v} \tag{B.181}$$

for fixed $(\boldsymbol{p},\xi,\boldsymbol{a}) \in \boldsymbol{\Delta}_C^+ \times (0,\infty) \times \mathbb{R}^C$, is a nonstandard task. We propose in this section two approaches to execute this step, which aim at re-expressing the required minimization as either a fixed-point or a root-finding problem. In more detail, if one has access to an explicit formula for the gradient of D_f^{conj}, then one can transform (B.181) into a fixed-point equation. This case applies for the KL-divergence, for which $\nabla D_{\mathsf{KL}}^{\mathrm{conj}}$ is the softmax function (Appendix B.3.2). Furthermore, for the convergence of the fixed-point iterations, we derive an improved Lipschitz constant for the softmax function in Appendix B.3.3. On the other hand, if one does not have a tractable formula for $\nabla D_f^{\mathrm{conj}}$, we propose the reduction provided in Lemma 3.2, whose proof is provided in Appendix B.3.2. We specialize the reduction provided by Lemma 3.2 to the cross-entropy case in Appendix B.3.2. Finally, we include in Appendix B.3.2 a general formula for $\nabla D_f^{\mathrm{conj}}$ that can be used for the $\boldsymbol{v}_i$-update step for a general f-divergence, and we also utilize it in Appendices B.3.4–B.3.6 to prove the convergence rate of Algorithm 1 stated in Theorems 3.4–3.5.

Primal update for KL-divergence

Consider the case when the f-divergence of choice is the KL-divergence, i.e., $f(t) = t\log t$. Then, the convex conjugate D_f^{conj} is given by the log-sum-exp function [DV75], namely, for $(\boldsymbol{p},\boldsymbol{v}) \in \boldsymbol{\Delta}_C^+ \times \mathbb{R}^C$ we have

$$D_f^{\mathrm{conj}}(\boldsymbol{v},\boldsymbol{p}) = \log\sum_{c\in[C]} p_c e^{v_c}. \tag{B.182}$$

Thus, the first step in a given ADMM iteration, as in (B.175) (see also the beginning of the for-loop in Algorithm 1), amounts to solving

$$\min_{\boldsymbol{v}\in\mathbb{R}^C} \log\sum_{c\in[C]} p_c e^{v_c} + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T\boldsymbol{v} \tag{B.183}$$

Algorithm 3 : $\operatorname*{argmin}_{v \in \mathbb{R}^C} D_{\mathsf{KL}}^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) + \xi \|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T \boldsymbol{v}$

```
Input: $\xi > 0$, $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$, $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^C$.
$z_c \leftarrow v_c + \log p_c$                                  $c \in [C]$
$b_c \leftarrow a_c - 2\xi \log p_c$                             $c \in [C]$
repeat
    $\boldsymbol{z} \leftarrow -\frac{1}{2\xi} \left(\sigma(\boldsymbol{z}) + \boldsymbol{b}\right)$
until convergence
Output: $v_c := z_c - \log p_c$                                 $c \in [C]$
```

for $\xi := \frac{\rho+\zeta}{2} > 0$ and some fixed vectors $(\boldsymbol{p}, \boldsymbol{a}) \in \boldsymbol{\Delta}_C^+ \times \mathbb{R}^C$; see (B.166), (B.175) and (B.178) for explicit expressions. The problem (B.183) is strictly convex. Further, we may recast this problem, via introducing the variable $\boldsymbol{z} \in \mathbb{R}^C$ by $z_c := v_c + \log p_c$, as

$$\min_{\boldsymbol{z} \in \mathbb{R}^C} \log \sum_{c \in [C]} e^{z_c} + \xi \|\boldsymbol{z}\|_2^2 + \boldsymbol{b}^T \boldsymbol{z}, \tag{B.184}$$

where $b_c = a_c - 2\xi \log p_c$ is fixed. To solve this latter problem, it suffices to find a zero of the gradient, which is given by

$$\nabla_{\boldsymbol{z}} \left(\log \sum_{c \in [C]} e^{z_c} + \xi \|\boldsymbol{z}\|_2^2 + \boldsymbol{b}^T \boldsymbol{z} \right) = \sigma(\boldsymbol{z}) + 2\xi \boldsymbol{z} + \boldsymbol{b} \tag{B.185}$$

where $\sigma : \mathbb{R}^C \to \boldsymbol{\Delta}_C^+$ denotes the softmax function $\sigma(\boldsymbol{z}) := \left(\frac{e^{z_{c'}}}{\sum_{c \in [C]} e^{z_c}} \right)_{c' \in [C]}$. Thus, we arrive at the fixed-point problem $\theta(\boldsymbol{z}) = \boldsymbol{z}$ for the function

$$\theta(\boldsymbol{z}) := -\frac{1}{2\xi} \left(\sigma(\boldsymbol{z}) + \boldsymbol{b} \right). \tag{B.186}$$

We solve $\theta(\boldsymbol{z}) = \boldsymbol{z}$ using a fixed-point-iteration method, i.e., with some initial $\boldsymbol{z}_0$, we iteratively compute the compositions $\theta^{(m)}(\boldsymbol{z}_0)$ for $m \in \mathbb{N}$. This procedure is summarized in Algorithm 3.

The exponentially-fast convergence of Algorithm 3 is guaranteed in view of Lipschitzness of θ as defined in (B.186). Indeed, it is known that the softmax function is 1-Lipschitz (see, e.g., [GP17, Prop. 4]); we improve this Lipschitz constant to $1/2$ in Appendix B.3.3. This improvement yields a better guarantee on the convergence speed of `FairProjection`. Indeed, as a lower value of the ADMM penalty ρ correlates with a faster convergence, lowering the Lipschitz constant of the softmax function allows us to speed up `FairProjection` by choosing $\rho > \frac{1}{2} - \zeta$ instead of $\rho > 1 - \zeta$.

Proof of Lemma 3.2: primal update for general f-divergences

The lemma follows by the following sequence of steps:

$$\begin{aligned}
\min_{\boldsymbol{v}\in\mathbb{R}^C} D_f^{\text{conj}}(\boldsymbol{v},\boldsymbol{p}) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T\boldsymbol{v} &\overset{\text{(I)}}{=} \min_{\boldsymbol{v}\in\mathbb{R}^C}\max_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} \boldsymbol{q}^T\boldsymbol{v} - D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) + \boldsymbol{a}^T\boldsymbol{v} + \xi\|\boldsymbol{v}\|_2^2 && \text{(B.187)}\\
&\overset{\text{(II)}}{=} \max_{\boldsymbol{q}\in\boldsymbol{\Delta}_C}\min_{\boldsymbol{v}\in\mathbb{R}^C} \boldsymbol{q}^T\boldsymbol{v} - D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) + \boldsymbol{a}^T\boldsymbol{v} + \xi\|\boldsymbol{v}\|_2^2 && \text{(B.188)}\\
&\overset{\text{(III)}}{=} \max_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} -D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) - \frac{1}{4\xi}\|\boldsymbol{a}+\boldsymbol{q}\|_2^2 && \text{(B.189)}\\
&= -\min_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) + \frac{1}{4\xi}\|\boldsymbol{a}+\boldsymbol{q}\|_2^2 && \text{(B.190)}\\
&= -\min_{\boldsymbol{q}\in\mathbb{R}_+^C}\sup_{\theta\in\mathbb{R}} D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) + \frac{1}{4\xi}\|\boldsymbol{a}+\boldsymbol{q}\|_2^2 + \theta\cdot\left(\mathbf{1}^T\boldsymbol{q}-1\right) && \text{(B.191)}\\
&\overset{\text{(IV)}}{=} -\sup_{\theta\in\mathbb{R}}\min_{\boldsymbol{q}\in\mathbb{R}_+^C} D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) + \frac{1}{4\xi}\|\boldsymbol{a}+\boldsymbol{q}\|_2^2 + \theta\cdot\left(\mathbf{1}^T\boldsymbol{q}-1\right) && \text{(B.192)}\\
&\overset{\text{(V)}}{=} -\sup_{\theta\in\mathbb{R}} -\theta + \sum_{c\in[C]}\min_{q_c\geq 0}\; p_c f\left(\frac{q_c}{p_c}\right) + \frac{1}{4\xi}(a_c+q_c)^2 + \theta q_c, && \text{(B.193)}
\end{aligned}$$

where (I) holds by definition of D_f^{conj} (see (3.14)), (II) by Sion's minimax theorem, (III) since the inner minimization occurs at $\boldsymbol{v} = -\frac{1}{2\xi}(\boldsymbol{q}+\boldsymbol{a})$, (IV) by generalized minimax theorems [ET99, see, e.g., Chapter VI, Proposition 2.2] (restated as Theorem B.7 herein for convenience), and (V) by separability.

Primal update for cross-entropy

In the cross-entropy (CE) case, i.e., $f(t) = -\log t$, instead of using an explicit formula for D_f^{conj} (which would yield unwieldy expressions), we utilize the reduction shown in Lemma 3.2. Thus, we have the equality

$$\min_{\boldsymbol{v}\in\mathbb{R}^C} D_f^{\text{conj}}(\boldsymbol{v},\boldsymbol{p}) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T\boldsymbol{v} = -\sup_{\theta\in\mathbb{R}} -\theta + \sum_{c\in[C]}\min_{q_c\geq 0}\; p_c f\left(\frac{q_c}{p_c}\right) + \frac{1}{4\xi}(a_c+q_c)^2 + \theta q_c. \quad \text{(B.194)}$$

As per (B.194), we focus next on solving the inner single-variable minimization

$$\min_{q\geq 0} -p\log q + \frac{1}{4\xi}(a+q)^2 + \theta q. \quad \text{(B.195)}$$

It is easily seen that the solution to this minimization is the unique point making the objective's derivative vanish, i.e., it is $q^\star \in (0,\infty)$ for which

$$-\frac{p}{q^\star} + \frac{q^\star}{2\xi} + \theta + \frac{a}{2\xi} = 0. \quad \text{(B.196)}$$

```
**Algorithm 4 :** $\operatorname*{argmin}_{\boldsymbol{v}\in\mathbb{R}^C} D_{\mathsf{CE}}^{\mathsf{conj}}(\boldsymbol{v},\boldsymbol{p}) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}^T\boldsymbol{v}$
Input: $\xi > 0$, $z \in \mathbb{R}$, $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$, $\boldsymbol{a} \in \mathbb{R}^C$.
repeat
    $g(z) \leftarrow -1 + \sum_{c\in[C]} \sqrt{\left(z + \frac{a_c}{2}\right)^2 + 2p_c\xi} - \left(z + \frac{a_c}{2}\right)$
    $g'(z) \leftarrow -C + \sum_{c\in[C]} \frac{2z + a_c}{\sqrt{\left(z + \frac{a_c}{2}\right)^2 + 2p_c\xi}}$
    $z \leftarrow z - \frac{g(z)}{g'(z)}$
until convergence
Output: $v_c := \frac{1}{2\xi}\left(z - \frac{a_c}{2} - \sqrt{\left(z + \frac{a_c}{2}\right)^2 + 2p_c\xi}\right)$
```

This is easily solvable as a quadratic, yielding

$$q^\star = \sqrt{\left(\theta\xi + \frac{a}{2}\right)^2 + 2p\xi} - \left(\theta\xi + \frac{a}{2}\right). \tag{B.197}$$

Therefore, solving (B.194) amounts to finding the constant $\theta \in \mathbb{R}$ that yields a probability vector $\boldsymbol{q} \in \boldsymbol{\Delta}_C$, where

$$q_c := \sqrt{\left(\theta\xi + \frac{a_c}{2}\right)^2 + 2p_c\xi} - \left(\theta\xi + \frac{a_c}{2}\right). \tag{B.198}$$

Consider the function

$$g(z) := -1 + \sum_{c\in[C]} \sqrt{\left(z + \frac{a_c}{2}\right)^2 + 2p_c\xi} - \left(z + \frac{a_c}{2}\right), \tag{B.199}$$

so we simply are looking for a root of g (then set $\theta = z/\xi$ and $\boldsymbol{v} = -\frac{1}{2\xi}(\boldsymbol{q} + \boldsymbol{a})$). This can be efficiently accomplished via Newton's method. Namely, we compute

$$g'(z) = -C + \sum_{c\in[C]} \frac{2z + a_c}{\sqrt{\left(z + \frac{a_c}{2}\right)^2 + 2p_c\xi}}, \tag{B.200}$$

then, starting from $z^{(0)}$, we form the sequence

$$z^{(t+1)} := z^{(t)} - \frac{g\left(z^{(t)}\right)}{g'\left(z^{(t)}\right)}. \tag{B.201}$$

This procedure is summarized in Algorithm 4.

On the gradient of the convex conjugate of f-divergence

The following general result on the differentiability of D_f^{conj} can be used to carry out the $\boldsymbol{v}_i$-update step for a general f-divergence, and it will also be useful in Appendices B.3.4–B.3.6 for proving the convergence rate of Algorithm 1 as stated in Theorems 3.4–3.5.

Lemma B.5. *Suppose $f : (0, \infty) \to \mathbb{R}$ is strictly convex. For any fixed $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$, the function $\boldsymbol{v} \mapsto D_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p})$ is differentiable, and its gradient is given by*

$$\nabla_{\boldsymbol{v}} D_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) = \boldsymbol{q}_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) \in \boldsymbol{\Delta}_C, \tag{B.202}$$

where

$$\boldsymbol{q}_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) := \operatorname*{argmin}_{\boldsymbol{q} \in \boldsymbol{\Delta}_C} D_f(\boldsymbol{q} \,\|\, \boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{q}. \tag{B.203}$$

Proof. From [Roc09, Proposition 11.3], since $\boldsymbol{q} \mapsto D_f(\boldsymbol{q} \,\|\, \boldsymbol{p})$ is a lower semicontinuous proper convex function, the subgradient of its convex conjugate $\boldsymbol{v} \mapsto D_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p})$ is given by

$$\partial_{\boldsymbol{v}} D_f^{\mathrm{conj}}(\boldsymbol{v}, \boldsymbol{p}) = \operatorname*{argmin}_{\boldsymbol{q} \in \boldsymbol{\Delta}_C} D_f(\boldsymbol{q} \,\|\, \boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{q}. \tag{B.204}$$

Recall also that a function is differentiable at a point if and only if its subgradient there consists of a singleton [BFG87]. Thus, it only remains to show that the right-hand side in (B.204) is a singleton. For this, we note that $\boldsymbol{q} \mapsto D_f(\boldsymbol{q} \,\|\, \boldsymbol{p}) - \boldsymbol{v}^T \boldsymbol{q}$ is lower semicontinuous and strictly convex, and $\boldsymbol{\Delta}_C$ is compact. □

B.3.3 Proof of Proposition 3.1: ½-Lipschitzness of the Softmax Function

As stated in Section 3.6 and Appendix B.3.2, the convergence speed of the inner iteration (the $\boldsymbol{v}_i$ update step) of `FairProjection` can be guaranteed to be faster if the Lipschitz constant of the softmax function is lowered from 1 (which is proved in [GP17, Prop. 4]). By Lipschitzness here, we mean ℓ_2-norm Lipschitzness. We prove 1-Lipschitzness in this appendix.

We will need the following result.

Lemma B.6 (Theorem 2.1.6 in [Nes04]). *A twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and has an L-Lipschitz continuous gradient if and only if its Hessian is positive semidefinite with maximal eigenvalue at most L.*

Since the softmax function is the gradient of the log-sum-exp function, and since the spectral norm is upper bounded by the Frobenius norm, it suffices to upper bound the Frobenius norm of the Jacobian of σ by $1/2$. Suppose that σ is operating on n symbols. Consider the sum of powers functions $s_k(\boldsymbol{x}) := \sum_{i\in[n]} x_i^k$ for $\boldsymbol{x} \in \mathbb{R}^n$. For any $\boldsymbol{v} \in \mathbb{R}^n$, denoting $\boldsymbol{x} = \sigma(\boldsymbol{v})$, the square of the Frobenius norm of the Jacobian of σ at $\boldsymbol{v}$ is given by

$$w(\boldsymbol{x}) := s_2(\boldsymbol{x})^2 + s_2(\boldsymbol{x}) - 2s_3(\boldsymbol{x}). \tag{B.205}$$

We show that $w(\boldsymbol{x}) \leq \frac{1}{4}$ for any $n \in \mathbb{N}$ and $\boldsymbol{x} \in \boldsymbol{\Delta}_n$.

The approach we take is via reduction to the case $n \leq 3$, which one can directly verify. Namely, assuming, without loss of generality, that $x_1 \leq x_2 \leq \cdots \leq x_n$, we show that if $x_1 + x_2 \leq 1/2$ then $w(\boldsymbol{y}) \geq w(\boldsymbol{x})$ where $\boldsymbol{y} \in \boldsymbol{\Delta}_{n-1}$ is given by $\boldsymbol{y} = (x_1 + x_2, x_3, \cdots, x_n)$. Note that if $n \geq 4$ then we must have $x_1 + x_2 \leq 1/2$, because $x_1 + x_2 \leq x_3 + x_4$ and $x_1 + x_2 + x_3 + x_4 \leq 1$. Thus, we will have reduced the problem from an $n \geq 4$ to $n - 1$, which iteratively reduces the problem to $n \leq 3$. Fix $n \geq 4$.

Denote $\boldsymbol{z} = (x_3, \cdots, x_n)$. A direct computation yields that

$$w(\boldsymbol{y}) - w(\boldsymbol{x}) = 2x_1x_2 \cdot (2s_2(\boldsymbol{z}) + g(x_1, x_2)) \tag{B.206}$$

with the quadratic

$$g(a, b) := 2a^2 + 2b^2 + 2ab - 3a - 3b + 1. \tag{B.207}$$

By assumption, $x_i \geq \max(x_1, x_2)$ for each $i \geq 3$, so $2s_2(\boldsymbol{z}) \geq (n-2)x_1^2 + (n-2)x_2^2 \geq x_1^2 + x_2^2$. Then,

$$w(\boldsymbol{y}) - w(\boldsymbol{x}) \geq 2x_1x_2 \cdot h(x_1, x_2) \tag{B.208}$$

with

$$h(a, b) := 3a^2 + 3b^2 + 2ab - 3a - 3b + 1. \tag{B.209}$$

Now, we show that h is nonnegative for every $a, b \geq 0$ with $a + b \leq 1/2$. With $c = a + b$, we may write

$$h(a, b) = 3c^2 - (3 + 4a)c + 4a^2 + 1. \tag{B.210}$$

This quadratic in c has its vertex at $c_{\min} = (3 + 4a)/6$. As $a \geq 0$, $c_{\min} \geq 1/2$. As $a + b \leq 1/2$, we see that the minimum of h is attained for $c = 1/2$. Substituting $b = 1/2 - a$, we obtain

$$h(a, b) = \left(2a - \frac{1}{2}\right)^2, \tag{B.211}$$

which is nonnegative, as desired.

B.3.4 Proof of Theorem 3.4: Convergence Rate of Algorithm 1

We recall a general result on the R-linear convergence rate for ADMM, which corresponds to case 1 in scenario 1 in [DY16]; see Tables 1 and 2 therein. Recall that a sequence $\{z^{(t)}\}_{t\in\mathbb{N}}$ is said to converge R-linearly to $z^\star$ if there is a constant $\eta \in (0,1)$ and a sequence $\{\beta^{(t)}\}_{t\in\mathbb{N}}$ such that $\|z^{(t)} - z^\star\| \le \beta^{(t)}$ and $\sup_t \left(\beta^{(t+1)}/\beta^{(t)}\right) \le \eta$. In particular, one has exponentially small errors:

$$\|z^{(t)} - z^\star\| \le \beta^{(0)} \cdot \eta^t. \tag{B.212}$$

The following theorem is used in our proof of Theorem 3.4.

Theorem B.8 ([DY16])**.** *Suppose that problem* (B.165) *has a saddle point, F is strongly convex and differentiable with Lipschitz-continuous gradient, $\boldsymbol{A}$ has full row-rank, and $\boldsymbol{B}$ has full column-rank. Then, the ADMM iterations* (B.172)–(B.174) *converge R-linearly to a global optimizer.*

In Appendix B.3.1, we show that the dual (3.29) of our fairness optimization problem (3.27) can be written in the ADMM general form (B.165) with the choices

$$F(\boldsymbol{V}) = \frac{1}{N}\sum_{i\in[N]} D_f^{\text{conj}}\left(\boldsymbol{v}_i, \boldsymbol{p}_i\right) + \frac{\zeta}{2}\left\|\boldsymbol{V}\right\|_2^2 \tag{B.213}$$

and

$$\boldsymbol{A} = \frac{1}{\sqrt{N}}\boldsymbol{I}_{NC}, \quad \boldsymbol{B} = \frac{1}{\sqrt{N}}(\boldsymbol{G}_i)_{i\in[N]}^T. \tag{B.214}$$

Recall from Theorem 3.3 (see also the proof in Appendix B.2) that our problem (3.29) has a saddle point. Further, the function $F : \mathbb{R}^{NC} \to \mathbb{R}$ is ζ-strongly convex and differentiable. Indeed, each $\boldsymbol{v} \mapsto D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i)$ is convex, and the term $\frac{\zeta}{2}\|\boldsymbol{V}\|_2^2$ is ζ-strongly convex, so F is ζ-strongly convex too. In addition, by the formula for ∇D_f^{conj} in Lemma B.5, the gradient of F is

$$\nabla F(\boldsymbol{V}) = \frac{1}{N}\boldsymbol{q}_f^{\text{conj}}(\boldsymbol{V}) + \zeta\boldsymbol{V}, \tag{B.215}$$

where

$$\boldsymbol{q}_f^{\text{conj}}(\boldsymbol{V}) := \left(\boldsymbol{q}_f^{\text{conj}}(\boldsymbol{v}_i, \boldsymbol{p}_i)\right)_{i\in[N]}, \tag{B.216}$$

with $\boldsymbol{q}_f^{\text{conj}}(\boldsymbol{v}_i)$ as defined in (B.203).

In the KL-divergence case, i.e., $f(t) = t\log t$, the gradient of D_f^{conj} is given by the softmax function (see Appendix B.3.2)

$$\boldsymbol{q}_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}) = \sigma\left(\boldsymbol{v} + \log\boldsymbol{p}\right) = \left(\frac{p_c e^{v_c}}{\sum_{c'\in[C]} p_{c'} e^{v_{c'}}}\right)_{c\in[C]}. \tag{B.217}$$

Therefore, we have that

$$\nabla F(\boldsymbol{V}) = \frac{1}{N} \left(\sigma\left(\boldsymbol{v}_i + \log \boldsymbol{p}_i\right)\right)_{i\in[N]} + \zeta \boldsymbol{V}. \tag{B.218}$$

By Proposition 3.1, the softmax function σ is $\frac{1}{2}$-Lipschitz. Hence, ∇F is $\left(\frac{1}{2N} + \zeta\right)$-Lipschitz.

Therefore, the general ADMM convergence rate in Theorem B.8 yields that there is a constant $r > 0$ such that

$$\left\|\boldsymbol{\lambda}^{(t)}_{\zeta,N} - \boldsymbol{\lambda}^{\star}_{\zeta,N}\right\|_2 \leq \beta \cdot e^{-rt} \tag{B.219}$$

where $\beta := \left\|\boldsymbol{\lambda}^{(0)}_{\zeta,N} - \boldsymbol{\lambda}^{\star}_{\zeta,N}\right\|_2$. (Although Theorem B.8 guarantees exponentially-fast convergence of $\boldsymbol{\lambda}^{(t)}_{\zeta,N}$ to *a* global optimizer, recall that $\boldsymbol{\lambda}^{\star}_{\zeta,N}$ is the *unique* optimizer of (3.29), as Theorem 3.3 shows.)

Finally, it remains to bound the distance between $\boldsymbol{h}^{\text{opt},N}$ and the output classifier $\boldsymbol{h}^{(t)}$ after the t-th iteration of Algorithm 1. Note that $\phi(u) = (f')^{-1}(u) = e^{u-1}$, so γ may be obtained explicitly, and equation (3.28) becomes

$$h^{\text{opt},N}_{c'}(x) = \frac{h^{\text{base}}_{c'}(x) \cdot e^{v_{c'}(x;\boldsymbol{\lambda}^{\star}_{\zeta,N})}}{\sum_{c\in[C]} h^{\text{base}}_{c}(x) \cdot e^{v_{c}(x;\boldsymbol{\lambda}^{\star}_{\zeta,N})}}. \tag{B.220}$$

Thus, using $\boldsymbol{\lambda}^{(t)} := \boldsymbol{\lambda}^{(t)}_{\zeta,N}$ in place of $\boldsymbol{\lambda}^{\star}_{\zeta,N}$, we obtain that the t-th classifier obtained by Algorithm 1 is

$$h^{(t)}_{c'}(x) = \frac{h^{\text{base}}_{c'}(x) \cdot e^{v_{c'}(x;\boldsymbol{\lambda}^{(t)})}}{\sum_{c\in[C]} h^{\text{base}}_{c}(x) \cdot e^{v_{c}(x;\boldsymbol{\lambda}^{(t)})}}. \tag{B.221}$$

Therefore, we have the ratios

$$\frac{h^{(t)}_{c'}(x)}{h^{\text{opt},N}_{c'}(x)} = \frac{\sum_{c\in[C]} h^{\text{base}}_{c}(x) e^{v_{c}(x;\boldsymbol{\lambda}^{\star}_{\zeta,N})}}{\sum_{c\in[C]} h^{\text{base}}_{c}(x) e^{v_{c}(x;\boldsymbol{\lambda}^{(t)})}} \cdot \exp\left(v_{c'}(x;\boldsymbol{\lambda}^{(t)}) - v_{c'}(x;\boldsymbol{\lambda}^{\star}_{\zeta,N})\right). \tag{B.222}$$

By definition of $\boldsymbol{v}$, $\boldsymbol{v}(x;\boldsymbol{\lambda}) = -\boldsymbol{G}(x)^T\boldsymbol{\lambda}$. Thus, we obtain from (B.219) and boundedness of $\boldsymbol{G}$ that

$$\left\|\boldsymbol{v}(x;\boldsymbol{\lambda}^{(t)}) - \boldsymbol{v}(x;\boldsymbol{\lambda}^{\star}_{\zeta,N})\right\|_\infty = e^{-\Omega(t)}, \tag{B.223}$$

where the implicit constant is independent of x. Applying (B.223) in (B.222), and noting that $e^{\pm e^{-\Omega(t)}} = 1 \pm e^{-\Omega(t)}$ as $t \to \infty$, we conclude that

$$\left|\frac{h^{(t)}_{c'}(x)}{h^{\text{opt},N}_{c'}(x)} - 1\right| = e^{-\Omega(t)}, \quad c' \in [C], \tag{B.224}$$

uniformly in x. We may rewrite (B.224) as

$$\boldsymbol{h}^{(t)}(x) = \boldsymbol{h}^{\text{opt},N}(x) \cdot \left(1 \pm e^{-\Omega(t)}\right), \tag{B.225}$$

which is the desired convergence rate in the theorem statement, and the proof is complete.

B.3.5 Extension of Theorem 3.4

Though Theorem 3.4 is shown for the KL-divergence, the proof directly extends to general f-divergences satisfying Assumption 3.1. In fact, Lipschitz continuity of the gradient of $D_{\mathsf{KL}}^{\mathsf{conj}}$ is the only specific property that we apply to derive the KL-divergence case. For a general f-divergence, Lipschitz continuity of $\nabla D_f^{\mathsf{conj}}$ may be derived as follows. Combining Lemmas B.4–B.5 reveals the formula $\nabla_{\boldsymbol{v}} D_f^{\mathsf{conj}}(\boldsymbol{v}, \boldsymbol{p}) = (p_c \cdot \phi\,(\gamma(\boldsymbol{v}) + v_c))_{c\in[C]}$, where $\phi = (f')^{-1}$ and $\gamma(\boldsymbol{v})$ is uniquely defined by $\mathbb{E}_{c\sim \boldsymbol{p}}\,[\phi(\gamma(\boldsymbol{v}) + v_c)] = 1$, with $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$ fixed. Since $\phi' = 1/(f'' \circ \phi)$, we have that ϕ is locally Lipschitz. From the proof of Theorem B.2, we know that $\boldsymbol{v} \mapsto \gamma(\boldsymbol{v})$ is locally Lipschitz. Thus, $\boldsymbol{v} \mapsto \nabla_{\boldsymbol{v}} D_f^{\mathsf{conj}}(\boldsymbol{v}, \boldsymbol{p})$ is locally Lipschitz. Further, $\boldsymbol{\lambda} \mapsto \nabla_{\boldsymbol{v}} D_f^{\mathsf{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}), \boldsymbol{p})$ is then also locally Lipschitz. Note that we may restrict $\boldsymbol{\lambda}$ *a priori* to be within some finite ball (see Lemma B.7). Thus, if, e.g., X is compactly-supported, we would obtain the desired Lipschitzness properties of the gradient of D_f^{conj}, and the proof of Theorem 3.4 carries through for D_f in place of D_{KL}.

B.3.6 Proof of Theorem 3.5: Convergence Rate to the Population Problem

The proof is divided in this appendix into several lemmas. We note first that, in the course of the proof of Theorems 3.1–3.2, it was shown that at least one minimizer $\boldsymbol{\lambda}^\star$ of (3.19) exists. Further, any such minimizer satisfies the following bound. Denote the constraint function by $\boldsymbol{\mu}(\boldsymbol{h}) := \mathbb{E}_{P_X}[\boldsymbol{G}\boldsymbol{h}]$. Throughout this proof, we set $\mathcal{X} := \mathbb{R}^d$.

Lemma B.7. *Suppose Assumption 3.1 holds, and fix a strictly feasible classifier $\boldsymbol{h} \in \mathcal{H}$, i.e., $\boldsymbol{\mu}(\boldsymbol{h}) < \boldsymbol{0}$. Every minimizer $\boldsymbol{\lambda}^\star \in \mathbb{R}_+^K$ of (3.19) must satisfy the inequality*

$$\|\boldsymbol{\lambda}^\star\|_1 \le \lambda_{\max} := \frac{D_f\left(\boldsymbol{h} \,\|\, \boldsymbol{h}^{\mathrm{base}} \mid P_X\right)}{\min\limits_{k\in[K]} -\mu_k(\boldsymbol{h})}. \tag{B.226}$$

We note that for the fairness metrics specified in Table 3.2, one valid choice of a strictly feasible $\boldsymbol{h}$ (i.e., one for which $\boldsymbol{\mu}(\boldsymbol{h}) < \boldsymbol{0}$) is the uniform classifier $\boldsymbol{h}(x) \equiv \frac{1}{C}\boldsymbol{1}$. In any case, we have that $\lambda_{\max} < \infty$ since both $\boldsymbol{h}$ and $\boldsymbol{h}^{\mathrm{base}}$ are assumed to belong to $\mathcal{H}$ and f is continuous over $(0, \infty)$; e.g., one bound on $\lambda_{\max}$ is $\lambda_{\max} \le \max_{m\le t\le M} f(t)/\min_{k\in[K]} -\mu_k(\boldsymbol{h})$ where $m = \inf_{c,x} h_c(x)$ and $M = 1/\inf_{x,c} h_c^{\mathrm{base}}(x)$.

We will also need the following constants for the convergence analysis:

$$g_{\text{mean}} := \mathbb{E}\left[\|\boldsymbol{G}(X)\|_2^2\right], \tag{B.227}$$

$$g_{\max} := \sup_{x \in \mathcal{X}} \|\boldsymbol{G}(x)\|_2^2. \tag{B.228}$$

Clearly, $g_{\text{mean}} \le g_{\max}$. By the boundedness of $\boldsymbol{G}$ in the second item in Assumption 3.1, $g_{\max}$ is finite.

Remark B.3. Although the results for `FairProjection` are stated to hold under Assumption 3.1, we note that those conditions do not essentially restrict applicability of `FairProjection`. Indeed, we focus on the CE and KL cases, for which f satisfies the imposed conditions. We also note that only boundedness of $\boldsymbol{G}$ is required for Theorem 3.3, which is true for the fairness metrics in Table 3.2 in non-degenerate cases (e.g., no empty groups). The condition on $\boldsymbol{h}^{\text{base}}$ being bounded away from zero can be made to hold by perturbing it if necessary with negligible noise. The condition on $\boldsymbol{h}^{\text{base}}$ being continuous is automatically satisfied if its domain is a finite set (as is the case for Theorem 3.3). Finally, the strict feasibility condition is verified by the uniform classifier.

Now, consider a form of ℓ_2 regularization of (3.19):

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^K} \mathbb{E}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X)\right) + \frac{\zeta}{2}\left\|\widetilde{\boldsymbol{G}}(X)^T\boldsymbol{\lambda}\right\|_2^2\right] \tag{B.229}$$

where $\widetilde{\boldsymbol{G}}(x) := (\boldsymbol{G}(x), \boldsymbol{I}_K) \in \mathbb{R}^{K\times(K+C)}$. We show now that there is a unique minimizer $\boldsymbol{\lambda}_\zeta^\star$ of (B.229).

Lemma B.8. *Under Assumption 3.1, there exists a unique minimizer $\boldsymbol{\lambda}_\zeta^\star$ of the regularized problem* (B.229).

Proof. Denote the function $A : \mathbb{R}_+^K \to \mathbb{R}$ by

$$A(\boldsymbol{\lambda}) := \mathbb{E}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X)\right) + \frac{\zeta}{2}\left\|\widetilde{\boldsymbol{G}}(X)^T\boldsymbol{\lambda}\right\|_2^2\right]. \tag{B.230}$$

That the range of A falls within $\mathbb{R}$ follows by Assumption 3.1, since then the function $x \mapsto D_f^{\text{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(x))$ is P_X-integrable. We will show that A is lower semicontinuous and ζ-strongly convex.

By Lemma B.5, $\boldsymbol{v} \mapsto D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p})$ is differentiable for any fixed $\boldsymbol{p} \in \boldsymbol{\Delta}_C^+$, implying that it is also continuous. Thus, $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(x))$ is continuous for each $x \in \mathcal{X}$. Hence, by Fatou's lemma and boundedness of $\boldsymbol{G}$, A is lower semicontinuous.

Next, to show strong convexity, we note that $\boldsymbol{\lambda} \mapsto D_f^{\text{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(x))$ is convex for each $x \in \mathcal{X}$. Indeed, this function is the supremum of affine functions. Further, the regularization term is

ζ-strongly convex, as its Hessian is given by

$$\zeta \cdot \left(\mathbb{E}\left[\widetilde{\boldsymbol{G}}(X)\widetilde{\boldsymbol{G}}(X)^T \right] + \boldsymbol{I} \right), \tag{B.231}$$

which is positive definite with minimal eigenvalue at least ζ.

Now, for each fixed $\theta > 0$, consider the compact set $\Lambda_\theta := \{\boldsymbol{\lambda} \in \mathbb{R}_+^K \; ; \; \|\boldsymbol{\lambda}\|_2^2 \leq \theta\}$. By what we have shown thus far, there is a unique minimizer $\boldsymbol{\lambda}_\theta$ of A over Λ_θ. By strong convexity, if A has a global minimizer then it is unique. We will show that $\boldsymbol{\lambda}_\theta$ is a global minimizer of A, where $\theta = 2(A(\mathbf{0}) - D^\star)/\zeta$. Suppose that $\mathbf{0}$ is not a global minimzer. Fix $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ such that $A(\mathbf{0}) > A(\boldsymbol{\lambda})$. Then,

$$A(\mathbf{0}) > A(\boldsymbol{\lambda}) \geq D^\star + \frac{\zeta}{2}\left(\mathbb{E}\left[\left\| \boldsymbol{G}(X)^T\boldsymbol{\lambda} \right\|_2^2 \right] + \|\boldsymbol{\lambda}\|_2^2 \right) \geq D^\star + \frac{\zeta}{2}\|\boldsymbol{\lambda}\|_2^2. \tag{B.232}$$

Thus, $\|\boldsymbol{\lambda}\|_2^2 < \theta$. This implies that $\boldsymbol{\lambda}_\theta$ is a global minimizer of A, hence it is the unique global minimizer of A. The proof of the lemma is thus complete. □

The following bound shows that $\boldsymbol{\lambda}_\zeta^\star$ is within $O(\zeta)$ of achieving $D^\star$ (see (3.19)).

Lemma B.9. *Suppose Assumption 3.1 holds, fix $\zeta \geq 0$, and denote the unique solution and the optimal objective value of* (B.229) *by $\boldsymbol{\lambda}_\zeta^\star$ and $D_\zeta^\star$, respectively. We have the bounds*

$$\mathbb{E}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X; \boldsymbol{\lambda}_\zeta^\star), \boldsymbol{h}^{\text{base}}(X) \right) \right] \leq D_\zeta^\star \leq D^\star + \theta_{\text{reg}} \cdot \zeta, \tag{B.233}$$

where we define the constant $\theta_{\text{reg}} := \lambda_{\max}^2 \cdot (1 + g_{\text{mean}})/2$.

Proof. The first bound is trivial. Using Lemma B.7, we may fix a $\boldsymbol{\lambda}^\star \in \mathbb{R}_+^K$ with $\|\boldsymbol{\lambda}^\star\|_1 \leq \lambda_{\max}$ such that $\boldsymbol{\lambda}^\star$ achieves $D^\star$. By definition of $D_\zeta^\star$,

$$D_\zeta^\star \leq \mathbb{E}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X; \boldsymbol{\lambda}^\star), \boldsymbol{h}^{\text{base}}(X) \right) + \frac{\zeta}{2}\left\| \widetilde{\boldsymbol{G}}(X)^T\boldsymbol{\lambda}^\star \right\|_2^2 \right] \leq D^\star + \theta_{\text{reg}} \cdot \zeta,$$

where the last inequality follows since for the 2-matrix norm, $\|\boldsymbol{M}\boldsymbol{\lambda}\|_2 \leq \|\boldsymbol{M}\|_2\|\boldsymbol{\lambda}\|_2$ and $\|\boldsymbol{M}^T\|_2 = \|\boldsymbol{M}\|_2$. □

Next, we derive a sample-complexity bound for the finite-sample problem (3.29) via generalizing the proofs of Theorem 3 in [AAW+20] and Theorem 13.2 in [HR19].

Lemma B.10. *Suppose Assumption 3.1 holds, and let $\lambda_{\max}$ and $g_{\max}$ be as defined in Lemma B.7 and equation* (B.228)*. For any $\delta \in (0, 1)$, with $\boldsymbol{\lambda}_{\zeta,N}^\star$ denoting the unique solution to* (3.29)*, it holds with* *probability at least* $1-\delta$ *that*

$$\mathbb{E}_X\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}_{\zeta,N}^{\star}),\boldsymbol{h}^{\text{base}}(X)\right)\right] \leq D_{\zeta}^{\star} + \frac{2g_{\max}\cdot(1+\zeta\cdot\lambda_{\max})^2}{\delta\zeta N}. \tag{B.234}$$

Proof. Let $\Lambda := \{\boldsymbol{\lambda}\in\mathbb{R}_+^K\ ;\ \|\boldsymbol{\lambda}\|_1 \leq \lambda_{\max}\}$, and consider the function $\ell : \Lambda\times\mathcal{X}\to\mathbb{R}$ defined by

$$\ell(\boldsymbol{\lambda},x) := D_f^{\text{conj}}\left(\boldsymbol{v}(x;\boldsymbol{\lambda}),\boldsymbol{h}^{\text{base}}(x)\right) + \frac{\zeta}{2}\left\|\widetilde{\boldsymbol{G}}(x)^T\boldsymbol{\lambda}\right\|_2^2. \tag{B.235}$$

Note that the regularized problem (B.229) can be written as

$$D_{\zeta}^{\star} := \min_{\boldsymbol{\lambda}\in\mathbb{R}_+^K}\ \mathbb{E}\left[\ell(\boldsymbol{\lambda},X)\right], \tag{B.236}$$

and the finite-sample version of it (3.29) can also be written as

$$D_{\zeta,N}^{\star} := \min_{\boldsymbol{\lambda}\in\mathbb{R}_+^K}\ \frac{1}{N}\sum_{i\in[N]}\ell(\boldsymbol{\lambda},X_i). \tag{B.237}$$

We show first that, for each fixed $x\in\mathcal{X}$, the function $\boldsymbol{\lambda}\mapsto\ell(\boldsymbol{\lambda},x)$ is ζ-strongly convex over Λ. The gradient of the regularization term is $\zeta\widetilde{\boldsymbol{G}}(x)^T\boldsymbol{\lambda}$, and its Hessian is given by

$$\nabla_{\boldsymbol{\lambda}}^2\,\frac{\zeta}{2}\left\|\widetilde{\boldsymbol{G}}(x)^T\boldsymbol{\lambda}\right\|_2^2 = \zeta\boldsymbol{G}(x)\boldsymbol{G}(x)^T + \zeta\boldsymbol{I}_K. \tag{B.238}$$

Further, the function $\boldsymbol{\lambda}\mapsto D_f^{\text{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}),\boldsymbol{h}^{\text{base}}(x))$ is convex as it is a pointwise supremum of linear functions. Indeed, for any $\boldsymbol{p}\in\boldsymbol{\Delta}_C$, recalling that $\boldsymbol{v}(x;\boldsymbol{\lambda}) = -\boldsymbol{G}(x)^T\boldsymbol{\lambda}$, we have the formula

$$D_f^{\text{conj}}(\boldsymbol{v}(x;\boldsymbol{\lambda}),\boldsymbol{p}) = \sup_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} -\boldsymbol{q}^T\boldsymbol{G}(x)^T\boldsymbol{\lambda} - D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}). \tag{B.239}$$

Next, we show Lipschitzness of $\boldsymbol{\lambda}\mapsto\ell(\boldsymbol{\lambda},x)$. For any fixed $\boldsymbol{v}\in\mathbb{R}^C$ and $\boldsymbol{p}\in\boldsymbol{\Delta}_C^+$, we have the gradient (see Lemma B.5)

$$\nabla_{\boldsymbol{v}}D_f^{\text{conj}}(\boldsymbol{v},\boldsymbol{p}) = \boldsymbol{q}^{\text{conj}}(\boldsymbol{v}) \in \boldsymbol{\Delta}_C, \tag{B.240}$$

where

$$\boldsymbol{q}^{\text{conj}}(\boldsymbol{v}) := \operatorname*{argmin}_{\boldsymbol{q}\in\boldsymbol{\Delta}_C} D_f(\boldsymbol{q}\,\|\,\boldsymbol{p}) - \boldsymbol{v}^T\boldsymbol{q}. \tag{B.241}$$

Thus, we have the gradient

$$\nabla_{\boldsymbol{\lambda}}D_f^{\text{conj}}\left(\boldsymbol{v}(x;\boldsymbol{\lambda}),\boldsymbol{h}^{\text{base}}(x)\right) = -\boldsymbol{G}(x)\boldsymbol{q}^{\text{conj}}\left(\boldsymbol{v}(x;\boldsymbol{\lambda})\right). \tag{B.242}$$

Hence, the gradient of $\boldsymbol{\lambda}\mapsto\ell(\boldsymbol{\lambda},x)$ is

$$\nabla_{\boldsymbol{\lambda}}\ell(\boldsymbol{\lambda},x) = -\boldsymbol{G}(x)\boldsymbol{q}^{\text{conj}}\left(\boldsymbol{v}(x;\boldsymbol{\lambda})\right) + \zeta\widetilde{\boldsymbol{G}}(x)^T\boldsymbol{\lambda}, \tag{B.243}$$

which therefore satisfies the bound

$$\|\nabla_{\boldsymbol{\lambda}} \ell(\boldsymbol{\lambda}, x)\|_2 \leq \|\boldsymbol{G}(x)\|_2 \left(1 + \zeta \cdot \lambda_{\max}\right). \tag{B.244}$$

Therefore, each $\boldsymbol{\lambda} \mapsto \ell(\boldsymbol{\lambda}, x)$ is A-Lipschitz with

$$A = (1 + \zeta \cdot \lambda_{\max}) \cdot \sup_{x \in \mathcal{X}} \|\boldsymbol{G}(x)\|_2. \tag{B.245}$$

Thus, by Theorem 13.1 in [HR19], with probability $1 - \delta$ we have the bound

$$\mathbb{E}_X \left[\ell\left(\boldsymbol{\lambda}^\star_{\zeta,N}, X\right)\right] \leq D^\star_\zeta + \frac{2A^2}{\delta\zeta N}. \tag{B.246}$$

With probability one, we have the bound

$$\mathbb{E}_X \left[D_f^{\text{conj}}\left(\boldsymbol{v}\left(X; \boldsymbol{\lambda}^\star_{\zeta,N}\right), \boldsymbol{h}^{\text{base}}(X)\right)\right] \leq \mathbb{E}_X \left[\ell\left(\boldsymbol{\lambda}^\star_{\zeta,N}, X\right)\right]. \tag{B.247}$$

This completes the proof of the lemma. □

Now, we are ready to finish the proof of Theorem 3.5 by specializing the above lemmas to the KL-divergence case. So, we set $f(t) = t \log t$ for the rest of the proof. By Lemmas B.9–B.10, we have with probability $1 - \delta$

$$\mathbb{E}_X \left[D_f^{\text{conj}}\left(\boldsymbol{v}(X; \boldsymbol{\lambda}^\star_{\zeta,N}), \boldsymbol{h}^{\text{base}}(X)\right)\right] \leq D^\star + \theta_{\text{reg}} \cdot \zeta + \frac{2g_{\max} \cdot (1 + \zeta \cdot \lambda_{\max})^2}{\delta\zeta N}. \tag{B.248}$$

Thus, by Lipschitzness (Proposition 3.1) and (B.219)

$$\mathbb{E}_X \left[D_f^{\text{conj}}\left(\boldsymbol{v}(X; \boldsymbol{\lambda}^{(t)}_{\zeta,N}), \boldsymbol{h}^{\text{base}}(X)\right)\right] \leq D^\star + \frac{1}{2}\sqrt{g_{\text{mean}}}\beta e^{-rt} + \theta_{\text{reg}} \cdot \zeta + \frac{2g_{\max} \cdot (1 + \zeta \cdot \lambda_{\max})^2}{\delta\zeta N}. \tag{B.249}$$

Here, we are choosing the constant β independently of N (as the optimal values of $\boldsymbol{\lambda}$ are bounded), and r of order $\sqrt{\frac{\zeta}{\frac{1}{2N} + \zeta}}$ (as can be guaranteed from Corollary 3.1 and Theorem 3.4 in [DY16]).

Choose $\zeta = \Theta(N^{-1/2})$. Collecting the constants in (B.249), we obtain that

$$\mathbb{E}_X \left[D_f^{\text{conj}}\left(\boldsymbol{v}(X; \boldsymbol{\lambda}^{(t)}_{\zeta,N}), \boldsymbol{h}^{\text{base}}(X)\right)\right] \leq D^\star + \frac{1}{2}\sqrt{g_{\text{mean}}}\beta e^{-rt} + \frac{\ell}{\delta\sqrt{N}} \tag{B.250}$$

for some constant ℓ that is completely determined by θ_{reg}, $g_{\max}$, and $\lambda_{\max}$. This bound can be further upper bounded by $D^\star + O(N^{-1/2})$ by choosing $t \geq \frac{1}{2r}\log N = \Theta(\log N)$, thereby completing the proof of the theorem.

B.4 Additional experiments and more details on the experimental setup

B.4.1 Numerical Benchmark Details

Datasets

The HSLS dataset is collected from 23,000+ participants across 944 high schools in the USA, and it includes thousands of features such as student demographic information, school information, and students' academic performance across several years. We preprocessed the dataset (e.g., dropping rows with a significant number of missing entries, performing k-NN imputation, normalization), and the number of samples reduced to 14,509.

The ENEM dataset, collected from the 2020 Brazilian high school national exam and made available by the Brazilian Government [INE20], is comprised of student demographic information, socio-economic questionnaire answers (e.g., parents education level, if they own a computer) and exam scores. We preprocess the dataset by removing missing values, repeated exam takers, and students taking the exam before graduation ("treineiros") and obtain ∼1.4 million samples with 138 features.

Hyperparameters

For logistic regression and gradient boosting, we use the default parameters given by Scikit-learn. For random forest, we set the number of trees and the minimum number of samples per leaf to 10. For all classifiers, we fixed the random state to 42. When running `FairProjection` (cf. Algorithm 1), we set the hyperparameters $\zeta = 1/\sqrt{N}$ (see Theorem 3.5) and $\rho = 2$ (see Appendix B.3.2), where N is the number of samples.

Benchmark Methods

For binary classification, we compare with six different benchmark methods:

- EqOdds [HPS+16]: We use AIF360 implementation of EqOddsPostprocessing and we use 50% of the test set as a validation set, i.e., 70% training set, 15% validation set, 15% test set.

- CalEqOdds [PRW+17]: We use AIF360 implementation of CalibratedEqOddsPostprocessing and we use 50% of the test set as a validation set, i.e., 70% training set, 15% validation set, 15% test set.

- Reduction [ABD$^+$18]: We use AIF360 implementation of ExponentiatedGradientReduction, and we use 10 different epsilon values as follows: $[0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2]$. We used EqualizedOdds constraint for MEO experiments and DemographicParity for statistical parity experiments.

- Rejection [KKZ12]: We use AIF360 implementation of RejectOptionClassification. We use the default parameters except metric_ub and metric_lb, namely, low_class_thresh $= 0.01$, high_class_thresh $= 0.99$, num_class_thresh $= 100$, num_ROC_margin $= 50$. We set the values metric_ub $= \varepsilon$ and metric_lb $= -\varepsilon$ to obtain trade-off curves. Epsilon values we used are: $[0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2]$.

- LevEqOpp [CDH$^+$19]: We used the code provided in the Github repo, originally programmed in R. We converted it into Python, and verified that the Python version achieved similar accuracy/fairness performance to their R version on UCI Adult dataset. We follow the same hyperparameters setup in [CDH$^+$19].

The following four methods, despite being mentioned in Table 3.1, are not included in the experiments:

- FACT [KCT20]: We used the code provided on the Github repo. We did not include the results in the main text as we found that:

  (i) This method is not directly comparable because they find post-processing parameters on the entire test set and apply them on the test set. This is different from all other methods we are comparing including our method, which use training set or a separate validation set to fit the post-processing mechanism. For this reason, FACT often has a point that lies above all other curves on the accuracy-fairness plot. However, this is not a fair comparison. We include the results of FACT in the COMPAS plots for the sake of demonstration.

  (ii) We found the results produced by this method inconsistent. Partial reason is due to the problem of finding mixing rates—probability of flipping $\widehat{Y} = 1$ to 0 (i.e., $P(\widetilde{Y} = 0|\widehat{Y} = 1)$) and vice-versa—which have to be between 0 and 1. But there are cases where these values lie outside $[0, 1]$, which leads to erroneous and inconsistent results.

For the results we present in the COMPAS plots, we used 20 epsilon values from 1 to 10^{-4}, equidistant in log space. We used 10 different train/test splits as we do in all other experiments. If certain splits does not produce a feasible solution, we drop those results. If none of the 10 splits produce a feasible solution, we drop the epsilon value. At the end, we had 19 epsilon values.

- Identifying [JN20]: Their optimization formulation is a special case of our formulation when f-divergence is KL divergence, but their algorithm requires retraining a classifier multiple times to solve the optimization problem, which results in a much slower runtime compared to ours (see Lines 1037–1046 in Appendix B.4). Nevertheless, we will add experiments for binary classification using [JN20] in the final version.
- FST [WRC20, WRC21]: Codes are not available publicly.
- Overlapping [YCK20]: We did not include this method for binary classification experiments as it reduces to the Reductions [ABD+18] approach for the binary class, binary protected group case. We could not benchmark for multi-class experiments with the code available online as it was assuming binary class (even though multiple protected groups).

For multi-class comparison, we compare with Adversarial [ZLM18]. In theory, the adversarial debiasing method is applicable to multi-class labels and groups, but its AIF360 implementation works only for binary labels and binary groups. We adapted their implementation to work on multi-class labels by changing the last layer of the classifier model from one-neuron sigmoid activation to multi-neuron soft-max activation. We varied adversary_loss_weight to obtain a trade-off curve, values taken from $[0.001, 0.01, 0.1, 0.2, 0.35, 0.5, 0.75]$. For all other parameters, we used the default values: num_epochs = 50, batch_size = 128, classifier_num_hidden_units = 200.

There are some methods that are relevant to our work but we could not benchmark in our experiments due to the lack of publicly available codes, including [WRC21], [MW18], [JSW22].

B.4.2 Additional experiments on runtime of FairProjection

We preform an ablation study on the runtime to illustrate that the parallelizability of `FairProjection` can significantly reduce the runtime, especially when the dataset contains hundreds of thousands of samples. We report the runtime of `FairProjection-KL` on ENEM with 2 classes, 2 groups, and with different sizes. In Table B.1, we observe that when the number of samples exceeds 200k, parallelization leads to 10.1× to 15.5× speedup of the runtime.

Method	# of Samples (in thousands)					
	20	50	100	200	500	∼1400
Non-Parallel	0.37±0.00	0.87±0.01	1.72±0.01	3.53±0.01	9.09±0.01	25.26±0.02
Parallel (GPU)	0.18±0.00	0.22±0.01	0.25±0.01	0.32±0.01	0.64±0.01	1.63±0.05
Speedup	2.00×	3.92×	7.21×	10.97×	14.23×	15.46×

Table B.1: *Execution time of parallel (on GPU) and non-parallel (on CPU) versions of the `FairProjection-KL` ADMM algorithm on the ENEM datasets with different sizes (time shown in minutes) with gradient boosting base classifiers.*

B.4.3 Additional Explanation on runtime comparison

The theoretical analysis below contrasts the runtimes of both `FairProjection` and Reduction [ABD+18], which is in line with our numerically observed comparison in Table 3.3. Two key factors make `FairProjection` faster than Reduction:

1. `FairProjection` needs a much lower number of iterations than Reduction does (logarithmic vs. polynomial).
2. Each iteration for FairProjection is less computationally expensive than its counterpart in Reduction. In fact, it is independent of the underlying model being projected, whereas Reduction requires retraining.

In more detail, one can obtain from [ABD+18, Theorem 3] that the Reductions approach converges in $O(N^2)$ iterations (where N is the number of samples and we use the suggested $\alpha = 1/2$ in [ABD+18, Theorem 3] according to the discussion at the top of page 6 therein). Taking the runtime of each iteration into consideration, one cannot hope for a runtime faster than $O(N^4)$ for Reduction. In fact, the runtime for Reduction must be higher than $O(N^4)$, since each of its iterations performs the subroutine $\text{BEST}_h(\boldsymbol{\lambda})$, which is a 'cost-sensitive classification' problem (i.e., numerically solving for an optimal classifier), and the $O(N^4)$ estimate would hold only if this *retraining* procedure can be done in *constant time* (which might be overly optimistic). In contrast, `FairProjection` does not require this retraining procedure at all, runs in $O(\log N)$ iterations, has $O(N)$ runtime for each iteration, and can perform much of each iteration in a parallel way.

For the dependence of the runtime of `FairProjection` on the number of groups, we note that there is a *linear* dependence on the number of constraints K when the number of samples N is much larger than K (which is the case for all datasets we consider), so one can say that the runtime is at most $\gamma KN \log N$ for an *absolute* constant γ. Note that there are $K = 2AC$ constraints for statistical parity, where A is the number of sensitive groups, and C is the number of classes; e.g., for the ENEM-1.4M-2C dataset that is used in Table 3.3, we get $K = 8$ for statistical parity. The K factor in the $O(KN \log N)$ rate comes from the creation of the vector $\boldsymbol{q}$ in Algorithm 1. If one does not parallelize, still one gets a runtime of $O(CKN \log N)$. Interestingly, the v_i-update step runtime in Algorithm 1 is $O(C)$ for a fixed $i \in [N]$ for both KL-divergence and Cross Entropy (see Appendices B.3.2 and B.3.3).

B.4.4 Omitted Experimental Results on Accuracy-Fairness Trade-off

Accuracy-fairness trade-off in binary classification

We include the results of benchmark methods and Fair Projection on 4 datasets (HSLS, ENEM-50k, Adult, and COMPAS) and 3 base classifiers (Logistic regression, Random forest, and GBM) in Figures B.1-B.8. For equalized odds experiments, we have six benchmark methods (`EqOdds`, `Rejection`, `Reduction`, `CalEqOdds`, `FACT`, `LevEqOpp`). For statistical parity experiments, we have `Rejection` and `Reduction`. We plot Fair Projection with both cross entropy and KL divergence.

When a method performs significantly worse than others, we did not plot its results. We did not include `Rejection` in the Adult plots as it did not produce consistent and reliable results on this dataset. `CalEqOdds` is included only in COMPAS as its performance was significantly worse and the point was too far away from other curves in all other datasets. `FACT` is also included only in the COMPAS plots and the reasons for this are explained in Appendix B.4.1.

We observe that Fair Projection performs consistently well in all four datasets. `FairProjection-CE` and `FairProjection-KL` have similar performance (i.e., overlapping curves) in most cases. The performance of Fair Projection is often comparable with `Reduction`. `Rejection` has competitive performance in ENEM-50k and HSLS, but its performance falters in COMPAS and Adult. `EqOdds` produces a point with very low MEO but with a substantial loss in accuracy. `LevEqOpp` also yields a point with low MEO but with a much smaller accuracy drop. Even though `LevEqOpp` only optimizes for FNR difference between two groups, it performs surprisingly well in terms of MEO in all four datasets. However, we note that `LevEqOpp` can only produce a point, not a curve, and it does not enjoy the generality of Fair Projection as it is specifically designed for binary-class, binary-group predictions and minimizing Equalized Opportunity difference.

Accuracy-fairness trade-off in multi-class/multi-group classification

In the main text, we showed the performance of `FairProjection-CE` on multi-class prediction with 5 classes and 2 groups (see Figure 3.2). We include results under a few different multi-class settings here. First, we show results on ENEM-50k-5-5 which has 5 classes and 5 groups in Figure B.9 and B.10 . We obtain 5 groups by not binarizing the race feature. Then, we show results on binary classification with 5 groups in Figure B.11 and B.12. Finally, we include the extended version of Figure 3.2 that include both `FairProjection-CE` and `FairProjection-KL` in Figure B.13.

To measure multi-class performance, we extend the definition of mean equalized odds (MEO) and statistical parity as follows:

$$\mathsf{MEO} = \max_{i \in \mathcal{Y}} \max_{s_1, s_2 \in \mathcal{S}} (|\mathsf{TPR}_i(s_1) - \mathsf{TPR}_i(s_2)| + |\mathsf{FPR}_i(s_1) - \mathsf{FPR}_i(s_2)|)/2 \tag{B.251}$$

$$\mathsf{Statistical\ Parity} = \max_{i \in \mathcal{Y}} \max_{s_1, s_2 \in \mathcal{S}} |\mathsf{Rate}_i(s_1) - \mathsf{Rate}_i(s_2)| \tag{B.252}$$

where we denote $\mathsf{TPR}_i(s) = P(\widehat{Y} = i \mid Y = i, S = s)$, $\mathsf{FPR}_i(s) = P(\widehat{Y} = i \mid Y \neq i, S = s)$, and $\mathsf{Rate}_i(s) = P(\widehat{Y} = i \mid S = s)$.

In all experiments, `FairProjection` reduces MEO and statistical parity significantly (e.g., 0.22 to 0.14) with a negligible sacrifice in accuracy.

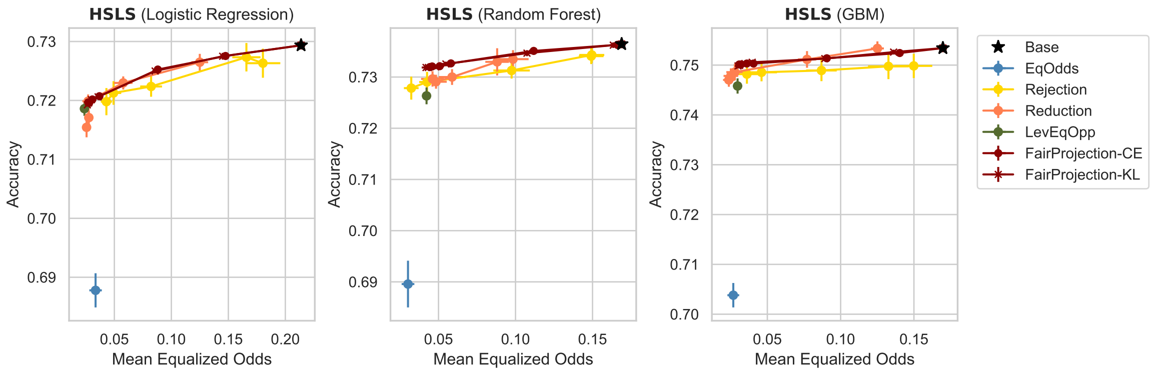


Figure B.1: *Accuracy-fairness curves of FairProjection and benchmark methods on the HSLS dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is MEO.*

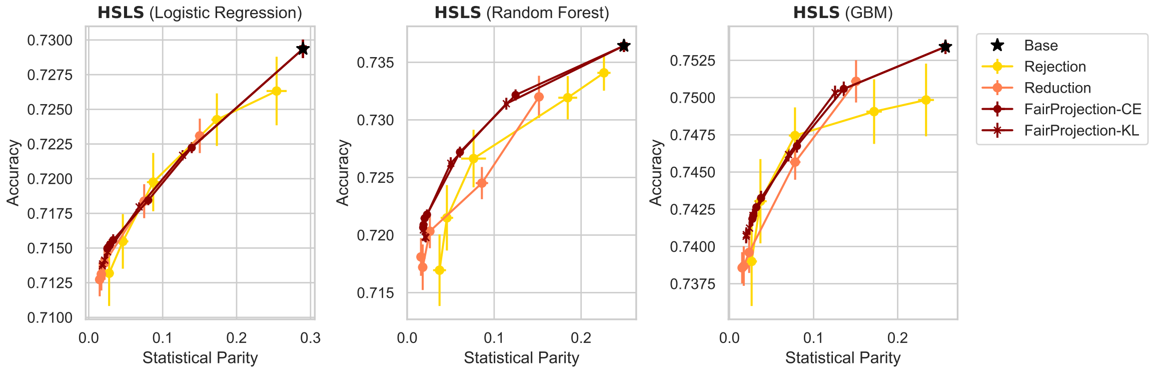


Figure B.2: *Accuracy-fairness curves of FairProjection and benchmark methods on the HSLS dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is statistical parity.*

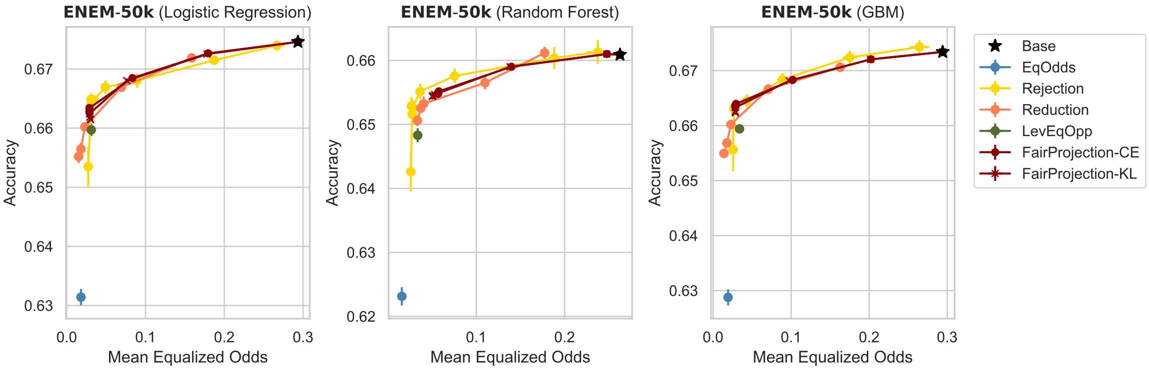


Figure B.3: *Accuracy-fairness curves of FairProjection and benchmark methods on the ENEM-50k-2C dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is MEO.*

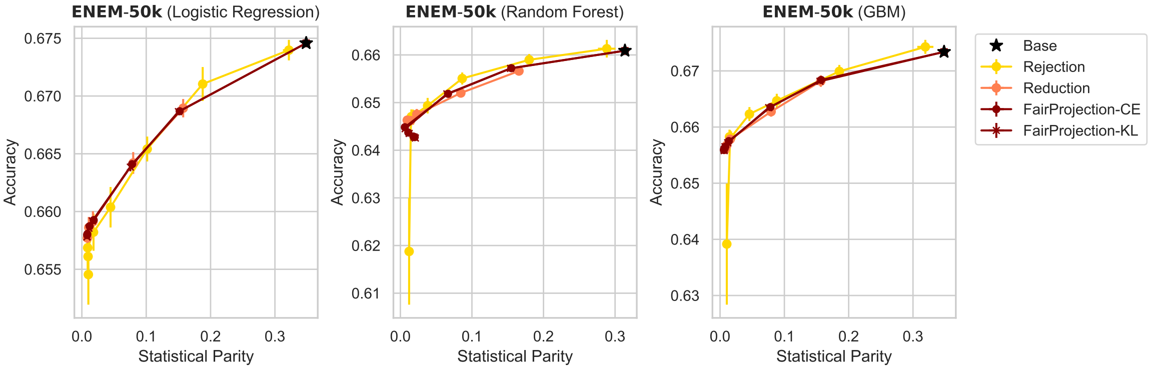


Figure B.4: *Accuracy-fairness curves of FairProjection and benchmark methods on the ENEM-50k-2C dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is statistical parity.*

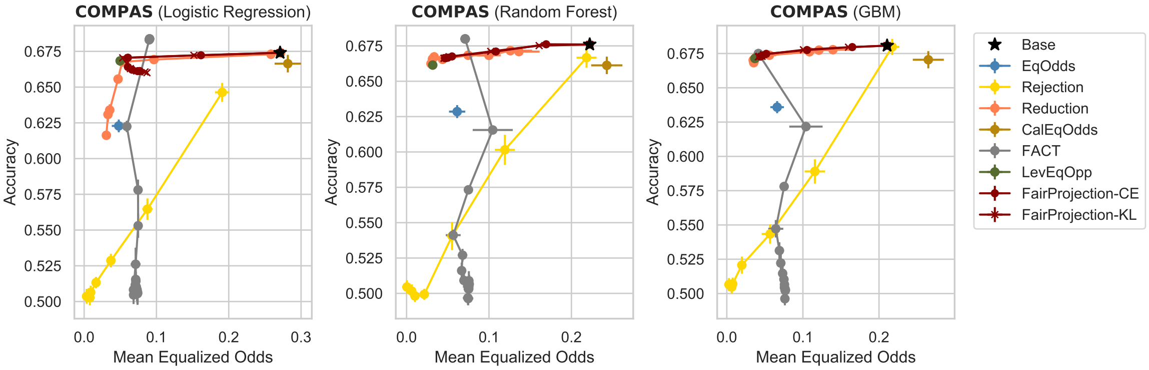


Figure B.5: *Accuracy-fairness curves of FairProjection and benchmark methods on COMPAS with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is MEO.*

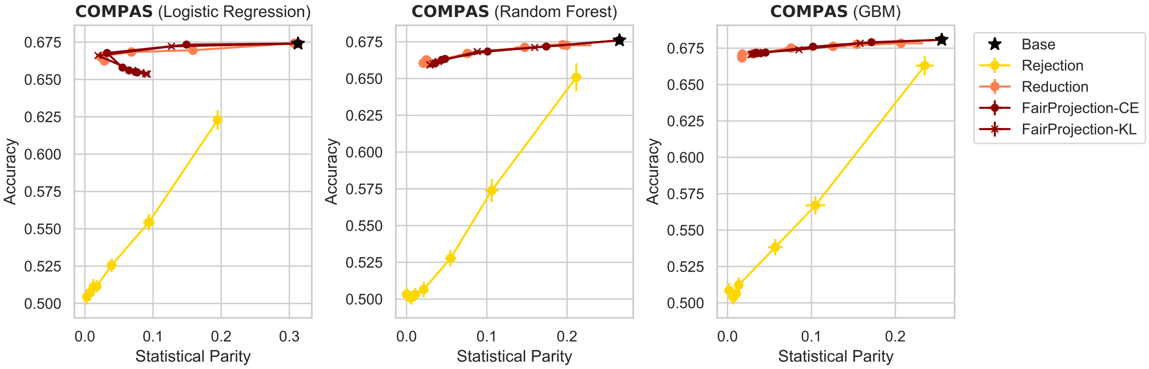


Figure B.6: *Accuracy-fairness curves of FairProjection and benchmark methods on COMPAS with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is statistical parity.*

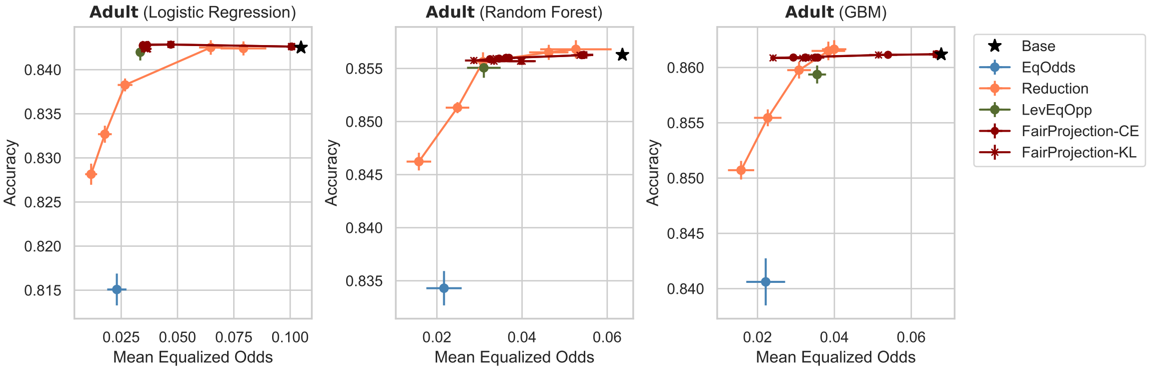


Figure B.7: *Accuracy-fairness curves of FairProjection and benchmark methods on the Adult dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is MEO.*

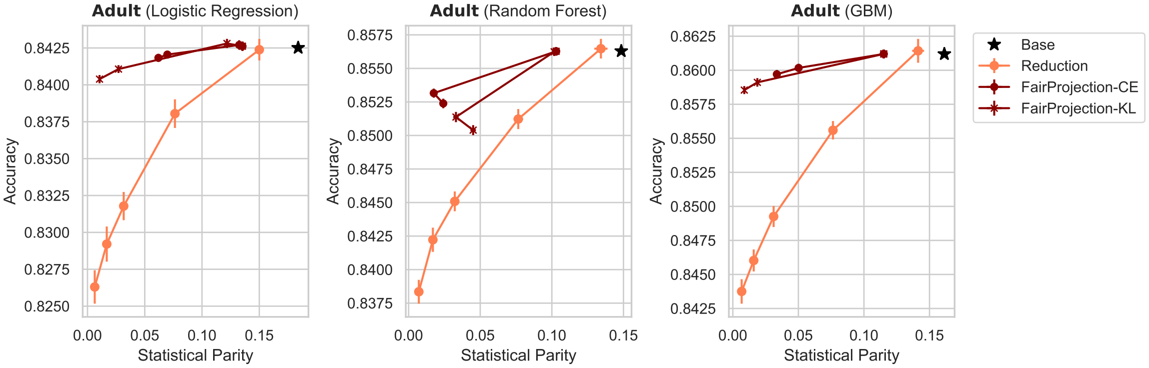


Figure B.8: *Accuracy-fairness curves of FairProjection and benchmark methods on the Adult dataset with 3 different models (Logistic regression, Random forest, GBM). The fairness constraint is statistical parity.*

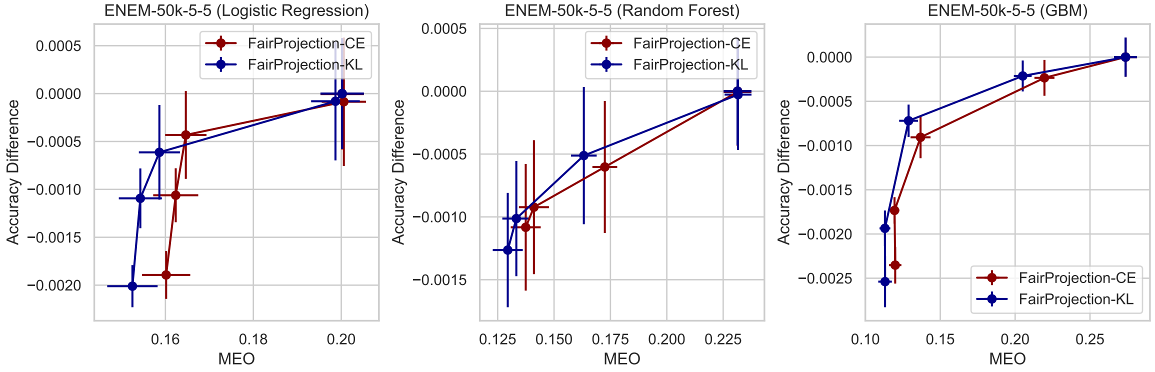


Figure B.9: *Accuracy-fairness curves of* `FairProjection-CE` *and* `FairProjection-KL` *on ENEM-50k with with 5 labels, 5 groups and different base classifiers base classifiers. The fairness constraint is MEO.*

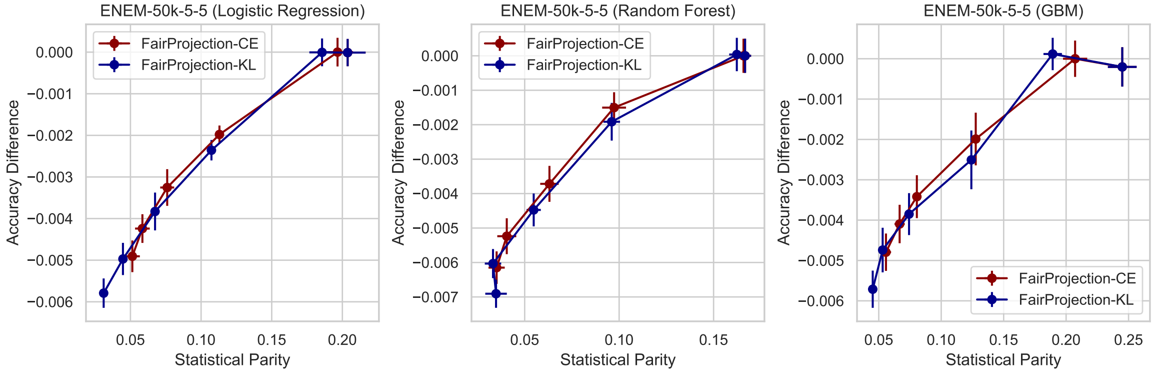


Figure B.10: *Accuracy-fairness curves of* `FairProjection-CE` *and* `FairProjection-KL` *on ENEM-50k with with 5 labels, 5 groups and different base classifiers base classifiers. The fairness constraint is SP.*

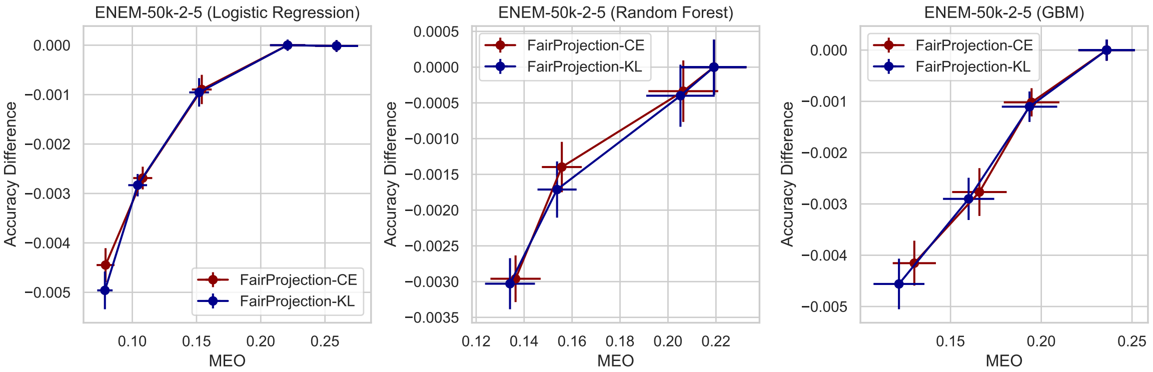


Figure B.11: *Accuracy-fairness curves of* `FairProjection-CE` *and* `FairProjection-KL` *on ENEM-50k with with 2 labels, 5 groups and different base classifiers base classifiers. The fairness constraint is MEO.*

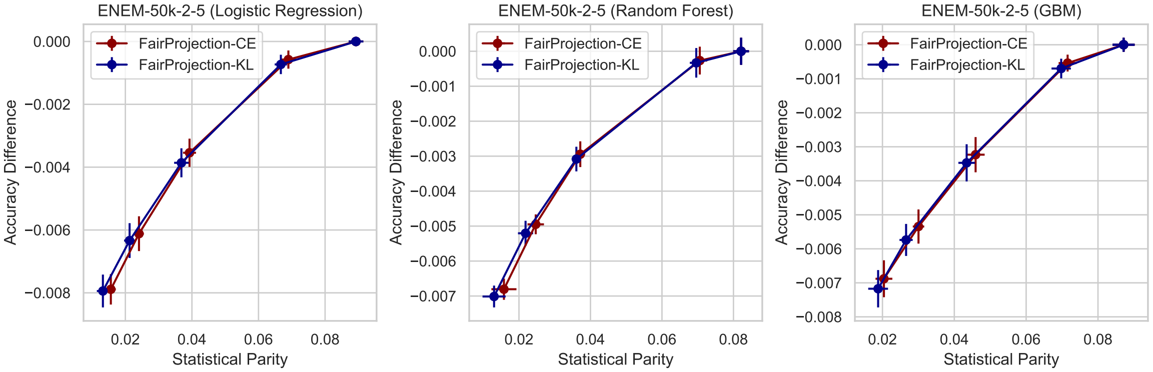


Figure B.12: *Accuracy-fairness curves of* `FairProjection-CE` *and* `FairProjection-KL` *on ENEM-50k with with 2 labels, 5 groups and different base classifiers base classifiers. The fairness constraint is SP.*

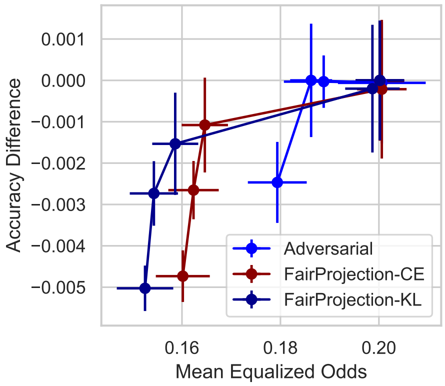


Figure B.13: *Comparison of* `FairProjection-CE` *and* `FairProjection-KL` *with* `Adversarial` *on ENEM-50k-5-2, meaning 5 labels, 2 groups. The reason for the difference comparing to Fig. 3.2 is that we resampled 50k data points from ENEM.*

Method			Feature				
	Multiclass	Multigroup	Scores	Curve	Parallel	Rate	Metric
Reductions [ABD+18]	✗	✓	✓	✓	✗	✓	SP, (M)EO
Reject-option [KKZ12]	✗	✓	✗	✓	✗	✗	SP, (M)EO
EqOdds [HPS+16]	✗	✓	✗	✗	✗	✓	EO
LevEqOpp [CDH+19]	✗	✗	✗	✗	✗	✗	FNR
CalEqOdds [PRW+17]	✗	✗	✓	✗	✗	✓	MEO
FACT [KCT20]	✗	✗	✗	✓	✗	✗	SP, (M)EO
Identifying [JN20]	✓✗	✓	✓	✓	✗	✗	SP, (M)EO
FST [WRC20, WRC21]	✗	✓	✓	✓	✗	✓	SP, (M)EO
Overlapping [YCK20]	✓	✓	✓	✓	✗	✗	SP, (M)EO
Adversarial [ZLM18]	✓	✓	N/A	✓	✓	✗	SP, (M)EO
`FairProjection` (ours)	✓	✓	✓	✓	✓	✓	SP, (M)EO

Copy of Table 3.1. *Comparison between benchmark methods.* ***Multiclass/multigroup****: implementation takes datasets with multiclass/multigroup labels;* ***Scores****: processes raw outputs of probabilistic classifiers;* ***Curve****: outputs fairness-accuracy tradeoff curves (instead of a single point);* ***Parallel****: parallel implementation (e.g., on GPU) is available;* ***Rate****: convergence rate or sample complexity guarantee is proved;* ***Metric****: applicable fairness metric, with SP↔Statistical Parity, EO↔Equalized Odds, MEO↔Mean EO. Since* `FairProjection` *is a post-processing method, we focus our comparison on post-processing fairness intervention methods, except for Reductions [ABD+18], which is a representative in-processing method, and Adversarial [ZLM18], which we use to benchmark multi-class prediction. For comparing in-processing methods, see [LPB+21, Table 1].*

B.4.5 More on related work

Our method is a model-agnostic post-processing method, so we focus our comparison on such post-processing fairness intervention methods. In the above table, the only exception is Adversarial [ZLM18], which we use to benchmark multi-class prediction. Adversarial [ZLM18] is an in-processing method based on generative-adversarial network (GAN) where the adversary tries to guess the sensitive group attribute S from Y and $\widehat{Y}$. Even though this GAN-based approach is applicable to multi-class, multi-group prediction, it cannot be universally applied to any pre-trained classifier like our method.

EqOdds [HPS+16], CalEqOdds [PRW+17] and LevEqOpp [CDH+19] are post-processing methods designed for binary prediction with binary groups. They find different decision thresholds for each group that equalize FNR and FPR of two groups. CalEqOdds [PRW+17] has an additional constraint that the post-processed classifier must be well-calibrated, and we observe in our experiments that this stringent constraint leads to a low-accuracy classifier especially when there is a big gap in the base rate between the two groups. FACT [KCT20] follows a similar approach but generalizes this to an optimization framework that can have both equalized odds and statistical parity constraints and flexible accuracy-fairness trade-off. The optimization formulation finds a desired confusion matrix, and their proposed post-processing method flips the predictions to match the desired confusion matrix. Reject-option [KKZ12] is similar in that it flips predictions near the decision threshold. In [KKZ12], instead of finding the optimal confusion matrix, it performs grid search to find the optimal margin around the decision threshold that can minimize either equalized odds or statistical parity. For these methods that center around modifying decision thresholds, it is not straightforward to extend to multi-class and multi-group as one will have to consider $\binom{|\mathcal{Y}|}{2} \cdot \binom{|\mathcal{S}|}{2}$ boundaries.

FST [WRC20, WRC21] tackles fairness intervention via minimizing cross-entropy for binary classes. Their method is inherently tailored to binary classification *and* only a cross-entropy objective function, and our `FairProjection`-CE reduces to FST for the case of CE and binary classification tasks. Identifying [JN20] is a method for minimizing KL-divergence for group-fairness intervention, which changes the label weights (via a convex combination) between unweighted and weighted samples, but it is not clear that this would navigate a good fairness-accuracy trade-off curve. Their method can be extended to non-binary prediction with non-binary groups by an appropriate choice of base classifier and fairness constraints, which is a non-trivial extension of the accompanying code, and we chose not to pursue this. Note that [JN20] and `FairProjection` solve the KL-divergence minimization in very different ways. In particular, the runtime of [WRC20, WRC21] on a 350k training dataset is longer than 30 minutes using logistic regression as a base classifier (in comparison, the runtime of `FairProjection` for a 500k dataset is less than 1 minute). This is because they require reweighing the data and retraining a large number of times. Hence, it is inherently non-parallelizable.

Fairness in Multi-Class Prediction

Methods that are based on optimization with a fairness regularizer often can be easily extended to multi-class prediction as it only requires a small change in the regularizer. For example, instead of using $|\mathrm{FNR}_0(x) - \mathrm{FNR}_1(x)|$, one can replace this with

$$\sum_{i \in \mathcal{Y}} \sum_{j \neq i \in \mathcal{Y}} |P(\widehat{Y} = j \mid Y = i, S = 0) - P(\widehat{Y} = j \mid Y = i, S = 1)|. \tag{B.253}$$

FERM [DOBD$^+$18] mentions how their method can be extended to multi-class sensitive attribute. Similarly, we believe that their method can be used for multi-class labels as well. The reductions approach [ABD$^+$18] assumes binary labels but is has natural extension to multi-class, which is explored in [YCK20]. In-processing methods proposed in [CHS20] and [ZLM18] allow for both multi-class labels and multi-class group attributes. [ZLM18] aims to achieve the independence between the sensitive attribute S and $\widehat{Y}$ or $\widehat{Y}$ given Y by training an adversary who tries to figure out $\widehat{S}$. [CHS20] directly estimates the fairness loss (e.g., B.253) using kernel density estimation. They also demonstrate the empirical performance in a three-class classification using synthetic data. Another in-processing method is [AAV19] where the authors propose a way to incorporate multi-class fairness constraints into decision tree training. The preprocessing method suggested in [CKV20] is conceptually similar to our methods in that it aims to minimize the KL-divergence between the original distribution and preprcoessed distribution while satisfying fairness constraints. Their method, however, requires all feature vectors to be binary, and applies only to demographic parity or representation rate. There exist other notions of fairness, which is different from commonly-used group fairness metrics such as envy-freeness [BDNP19] or best-effort [KJW$^+$21], which can be applied to multi-class prediction tasks.

Finally, there are unpublished works [DEHH21, YX20] that could handle multi-class classification. Specifically, [DEHH21] presents a post-processing method that selects different thresholds for each group to achieve demographic parity. [YX20] formulates SVM training as a mixed-integer program and integrates fairness regularizer in the objective, which can also deal with multi-class.

B.5 Datasheet for ENEM 2020 dataset

Questions

The questions below are derived from [GMV$^+$21] and aim to provide context about the ENEM-2020 dataset. We highlight that we did not create the dataset nor collect the data included in it. Instead, we simply provide a link to the ENEM-2020 data at [INE20]. At the time of writing, the ENEM-2020 dataset is open and made freely available by the Brazilian Government at [INE20] under a Creative Commons Attribution-NoDerivs 3.0 Unported License [Com]. We provide the datasheet below to clarify certain aspects of the dataset (e.g., motivation, composition, etc.) since the original information is available in Portuguese at [INE20], thus limiting its access to a broader audience. The website [INE20] contains a link to download a `.zip` file which contains the ENEM-2020 data in `.csv` format and extensive accompanying documentation.

The datasheet below is **not** a substitute for the explanatory files that are downloaded together with the dataset at [INE20], and we emphatically recommend the user to familiarize themselves with associated documentation prior to usage. We also strongly recommend the user to carefully read the "Leia-Me" (readme) file `Leia_Me_Enem_2020.pdf` available in the same `.zip` folder that contains the dataset. The answers in the datasheet below are based on an English translation of information available at [INE20] and may be incomplete or inaccurate. The datasheet below is based on our own independent analysis and in no way represents or attempts to represent the opinion or official position of the Brazilian Government and its agencies.

We also note that we do not distribute the ENEM-2020 dataset directly nor host the dataset ourselves. Instead, we provide a link to download the data from a public website hosted by the Brazilian Government. The dataset may become unavailable in case the link in [INE20] becomes inaccessible.

Motivation

- **For what purpose was the dataset created?** According to the "Leia-me" (Read Me) file that accompanies the data, the dataset was made available to fufill the mission of the Instituto

Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) of developing and disseminating data about exams and evaluations of basic education in Brazil.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The dataset was developed by INEP, which is a government agency connected to the Brazilian Ministry of Education.
- **Who funded the creation of the dataset?** The data is made freely available by the Brazilian Government.

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** The instances of the dataset are information about individual students who took the Exame Nacional do Ensino Médio (ENEM). The ENEM is the capstone exam for Brazilian students who are graduating or have graduated high school.
- **How many instances are there in total (of each type, if appropriate)?** The raw data provided in at [INE20] has approximately 5.78 million entries. The processed version we use in our experiments has approximately 1.4 million entries.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** The data provided is the lowest level of aggregation of data collected from ENEM exam-takers made available by INEP.
- **What data does each instance consist of?** We provide a brief description of the features available in the raw public data provided at [INE20]. Upon downloading the data, a detailed description of features and their values are available (in Portuguese) in the file titled

  `Dicionário_Mircrodados_ENEM_2020.xsls.`

  The features include:

  - **Information about exam taker:** exam registration number (masked), year the exam was taken (2020), age range, sex, marriage status, race, nationality, status of high school graduation, year of high school graduation, type of high school (public, private, n/a), if they are a "treineiro" (i.e., taking the exam as practice).

  - **School data**: city and state of participant's school, school administration type (private, city, state, or federal), location (urban or rural), and school operation status.
  - **Location where exam was taken**: city and state.
  - **Data on multiple-choice questions**: The exam is divided in 4 parts (translated from Portuguese): natural sciences, human sciences, languages and codes, and mathematics. For each part there is data if the participant attended the corresponding portion of the exam, the type of exam book they received, their overall grade, answers to exam questions, and the answer sheet for the exam.
  - **Data on essay question**: if participant took the exam, grade on different evaluation criteria, and overall grade.
  - **Data on socio-economic questionnaire answers:** the data include answers to 25 socio-economic questions (e.g., number of people who live in your house, family average income, if the your house has a bathroom, etc.).

- **Is there a label or target associated with each instance?** No, there is no explicit label. In our fairness benchmarks, we use grades in various components of the exam as a predicted label.

- **Is any information missing from individual instances?** Yes, certain instances have missing values.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** No explicit relationships identified.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** No.

- **Are there any errors, sources of noise, or redundancies in the dataset?** The data contains missing values and, according to INEP, was collected from individual exam takers. The information is self-reported and collected at the time of the exam.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** Self-contained.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** According to the *Leia-me* (readme) file (in Portuguese) that accompanies the dataset and our own inspection, the dataset does not contain any feature that allows direct identification of exam takers such as name, email, ID number, birth date, address, etc. The exam registration number has been substituted by a sequentially generated mask. INEP states that the released data is aligned with the Brazilian *Lei Geral de Proteção dos Dados* (LGPD, General Law for Data Protection). We emphatically recommend the user to view the Readme file prior to usage.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** The official terminology used by the Brazilian Government to denote race can be viewed as offensive. Specifically, the term used to describe the race of exam takers of Asian heritage is "Amarela," which is the Portuguese word for the color yellow. Moreover, the term "Pardo," which roughly translates to brown, is used to denote individuals of multiple or mixed ethnicity. This outdated and inappropriate terminology is still in official use by the Brazilian Government, including in its population census. The dataset itself includes integers to denote race, which are mapped to specific categories through the variable dictionary.

- **Does the dataset relate to people?** Yes.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** Yes. Information about age, sex, and race are included in the dataset.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** The *Leia-me* (readme) file notes that the individual exam-takers cannot be directly identified from the data. However, in the same file, INEP recognizes that the Brazilian data protection law (LGPD) does not clearly define what constitutes a reasonable effort of de-identification. Thus, INEP adopted a cautious approach: this dataset is a simplified/abbreviated version of the ENEM micro-data compared to prior releases and aims to remove any features that may allow identification of the exam-taker.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms**

of government identification, such as social security numbers; criminal history)? The data includes race information and socio-economic questionnaire answers.

Collection Process

Since we did not produce the data, we cannot speak directly about the collection process. Our understanding is that the data contains self-reported answers from exam-takers of the ENEM collected at the time of the exam. The exam was applied on 17 and 24 of January 2021 (delayed due to COVID). The data was aggregated and made publicly available by INEP at [INE20]. After consulting the IRB office at our institution, no specific IRB was required to use this data since it is anonymized and publicly available.

Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Some mild pre-processing was done on the data to ensure anonymity, as indicated in the "Leia-me" file. This includes aggregating participant ages, masking exam registration numbers, and removing additional information that could allow de-anonymization.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** The raw data is not publicly available.

Uses

- **Has the dataset been used for any tasks already?** We have used this dataset to benchmark fairness interventions in ML in the present paper. ENEM microdata has also been widely used in studies ranging from public policy in Brazil to item response theory in high school exams.
- **Are there tasks for which the dataset should not be used?** INEP does not clearly define tasks that should not be used on this dataset. However, no attempt should be made to de-anonymize the data.

Distribution and Maintenance

The ENEM-2020 dataset is open and made freely available by the Brazilian Government at [INE20] under a Creative Commons Attribution-NoDerivs 3.0 Unported License [Com] at the time of writing. The dataset may become unavailable in case the link in [INE20] becomes inaccessible.

Appendix C

Appendix to Chapter 4

C.1 Proofs of Section 4.2

C.1.1 PMMSE Formula: Proof of Lemma 4.1

The matrix $\boldsymbol{M}_{Y,n}$ is symmetric. We show that it is positive-semidefinite, and that it is positive-definite if and only if $|\mathrm{supp}(Y)| > n$. For any $\boldsymbol{d} \in \mathbb{R}^{n+1}$, we have the inequality

$$\boldsymbol{d}^T \boldsymbol{M}_{Y,n} \boldsymbol{d} = \boldsymbol{d}^T \, \mathbb{E}\left[\boldsymbol{Y}^{(n)} \left(\boldsymbol{Y}^{(n)} \right)^T \right] \boldsymbol{d} \tag{C.1}$$

$$= \mathbb{E}\left[\boldsymbol{d}^T \boldsymbol{Y}^{(n)} \left(\boldsymbol{Y}^{(n)} \right)^T \boldsymbol{d} \right] \tag{C.2}$$

$$= \mathbb{E}\left[\left| \boldsymbol{d}^T \boldsymbol{Y}^{(n)} \right|^2 \right] \geq 0, \tag{C.3}$$

so $\boldsymbol{M}_{Y,n}$ is positive-semidefinite. Furthermore, the equality case $\mathbb{E}\left[\left| \boldsymbol{d}^T \boldsymbol{Y}^{(n)} \right|^2 \right] = 0$ holds if and only if $\left| \boldsymbol{d}^T \boldsymbol{Y}^{(n)} \right|^2 = 0$, and this latter relation holds if and only if $\boldsymbol{d}^T \boldsymbol{Y}^{(n)} = 0$. Therefore, $\boldsymbol{M}_{Y,n}$ is positive-definite if and only if $\boldsymbol{d}^T \boldsymbol{Y}^{(n)} = 0$ implies $\boldsymbol{d} = \boldsymbol{0}$, i.e., $\boldsymbol{Y}^{(n)}$ does not lie almost surely in a hyperplane in $\mathbb{R}^{n+1}$. Finally, $\boldsymbol{Y}^{(n)}$ lies almost surely in a hyperplane in $\mathbb{R}^{n+1}$ if and only if $|\mathrm{supp}(Y)| \leq n$. Therefore, the desired result that $\boldsymbol{M}_{Y,n}$ is invertible if and only if $|\mathrm{supp}(Y)| > n$ follows.

Next, assume that $|\mathrm{supp}(Y)| > n$, so by what we have shown above, $\boldsymbol{M}_{Y,n}$ is invertible. Let $\boldsymbol{M}_{Y,n}^{1/2}$ denote the lower-triangular matrix in the Cholesky decomposition of $\boldsymbol{M}_{Y,n}$, i.e., $\boldsymbol{M}_{Y,n}^{1/2}$ is the unique lower-triangular matrix with positive diagonal entries that satisfies $\boldsymbol{M}_{Y,n}^{1/2} \left(\boldsymbol{M}_{Y,n}^{1/2} \right)^T = \boldsymbol{M}_{Y,n}$, and denote $\boldsymbol{M}_{Y,n}^{-1/2} := \left(\boldsymbol{M}_{Y,n}^{1/2}\right)^{-1}$. We show that the entries of the vector $\boldsymbol{V} = \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}$ comprise an orthonormal basis for $\mathscr{P}_n(Y)$. We have that

$$\mathbb{E}\left[\boldsymbol{V}\boldsymbol{V}^T\right] = \mathbb{E}\left[\boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}\left(\boldsymbol{Y}^{(n)}\right)^T\left(\boldsymbol{M}_{Y,n}^{-1/2}\right)^T\right] \tag{C.4}$$

$$= \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{M}_{Y,n}\left(\boldsymbol{M}_{Y,n}^{-1/2}\right)^T = \boldsymbol{I}_{n+1}. \tag{C.5}$$

Hence, the entries of the vector $\boldsymbol{V}$ form an orthonormal subset of $\mathscr{P}_n(Y)$. Since $\{1, Y, \cdots, Y^n\}$ spans $\mathscr{P}_n(Y)$, and $\boldsymbol{M}_{Y,n}^{-1/2}$ is invertible, we conclude that the entries of $\boldsymbol{V}$ also span $\mathscr{P}_n(Y)$. Hence, the entries of $\boldsymbol{V}$ form an orthonormal basis of $\mathscr{P}_n(Y)$.

Then, the general expansion of orthogonal projections yields the formula $E_n[X \mid Y] = \mathbb{E}\left[X\boldsymbol{V}^T\right]\boldsymbol{V}$. Substituting $\boldsymbol{V} = \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}$ we obtain (4.28). Then, expanding the PMMSE formula $\text{pmmse}_n(X \mid Y) = \mathbb{E}[(X - E_n[X \mid Y])^2]$, we obtain (4.29). The proof of the lemma is thus complete.

We note that an alternative proof of this lemma, once one obtains the invertibility of $\boldsymbol{M}_{Y,n}$, is via differentiation under the integral sign with respect to the polynomial coefficients in $E_n[X \mid Y]$ in the same way as the LMMSE is usually derived.

C.1.2 PMMSE for Symmetric random variables: Proof of Lemma 4.4

We may assume that X and Z are symmetric around 0, since $E_m[X + a \mid X + Z + b] = a + E_m[X \mid X + Z]$ for every $m \in \mathbb{N}$ and $a, b \in \mathbb{R}$. Then, $\mathbb{E}[X^j] = \mathbb{E}[Z^j] = 0$ for every odd $j \in \mathbb{N}$. Set $Y = X + Z$ and $n = 2k$. Then, $\mathbb{E}[Y^j] = 0$ for every odd $j \in \mathbb{N}$, and $\mathbb{E}[XY^\ell] = 0$ for every even $\ell \in \mathbb{N}$. Then, the coefficient of Y^n in $E_n[X \mid Y]$ is

$$\frac{1}{\det \boldsymbol{M}_{Y,n}} \sum_{\substack{\ell \in [n] \\ \ell \text{ odd}}} \mathbb{E}\left[XY^\ell\right]\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n}, \tag{C.6}$$

where $\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n}$ denotes the (ℓ, n)-th entry of $\boldsymbol{M}_{Y,n}^{-1}$. Fix an odd $\ell \in [n]$. Let $T_n^{(\ell,n)}$ denote the set of permutations of $[n]$ that send ℓ to n. We have that

$$\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n} = -\sum_{\pi \in T_n^{(\ell,n)}} \text{sgn}(\pi) \prod_{r \in [n]\setminus\{\ell\}} \mathbb{E}\left[Y^{r+\pi(r)}\right]. \tag{C.7}$$

We have that, for every $\pi \in T_n^{(\ell,n)}$, $\sum_{r \in [n]\setminus\{\ell\}} r + \pi(r) = n(n+1) - \ell - n$, which is odd. Therefore, for at least one $r \in [n] \setminus \{\ell\}$, the integer $r + \pi(r)$ is odd. Hence, $\mathbb{E}[Y^{r+\pi(r)}] = 0$, implying that $\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n} = 0$. As this is true for every odd $\ell \in [n]$, we conclude that the coefficient of Y^n in $E_n[X \mid Y]$

is 0. In other words, we have $E_{2k}[X \mid X+Z] = E_{2k-1}[X \mid X+Z]$, and the proof is complete.

C.1.3 PMMSE Convergence Theorems: Proof of Lemma 4.5

Note that in (i) the sequences $\{X_k Y^j\}_{k\in\mathbb{N}}$, for each fixed $j \in [n]$, are monotone almost surely. Also, X_0 is integrable, as we are assuming that $X_0 \in L^2(\mathcal{F})$. Note also that in (ii) each sequence $\{X_k Y^j\}_{k\in\mathbb{N}}$, for $j \in [n]$, is dominated by $M|Y|^j$, which is integrable since both M and Y^j are square-integrable. Thus, monotone convergence and dominated convergence both hold in $L^1(\mathcal{F})$ for each of the sequences $\{X_k Y^j\}_{k\in\mathbb{N}}$, where $j \in [n]$ is fixed. In addition, the formula

$$\begin{aligned} E_n\left[X_k \mid Y=y\right] &= \mathbb{E}\left[X_k \boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\left(1,y,\cdots,y^n\right)^T \\ &= \sum_{j=0}^{n} c_j \mathbb{E}\left[X_k Y^j\right] \end{aligned} \tag{C.8}$$

expresses $E_n\left[X_k \mid Y=y\right]$ as an $\mathbb{R}$-linear combination of $\{X_k Y^j\}_{j\in[n]}$ (where the c_j do not depend on k). Thus, the convergence theorems in (i) and (ii) also hold.

Remark C.1. A version of Fatou's lemma that holds for a subset of values of y is also derivable. Namely, suppose that there is a random variable $M \in L^1(\mathcal{F})$ such that $X_k Y^j \geq -M$ for every $(k,j) \in \mathbb{N} \times [n]$, and that $\liminf_{k\to\infty} X_k$ is square-integrable. Then, the same argument in the proof of Lemma 4.5 shows that

$$\liminf_{k\to\infty} E_n[X_k \mid Y=y] \geq E_n\left[\liminf_{k\to\infty} X_k \,\middle|\, Y=y\right] \tag{C.9}$$

for every $y \in \mathbb{R}$ such that $\boldsymbol{M}_{Y,n}^{-1}(1,y,\cdots,y^n)^T$ consists of non-negative entries. For example, when $n=1$, Fatou's lemma holds for $y \geq \mathbb{E}[Y]$ if $\mathbb{E}[Y] \leq 0$, and it holds for $y \in [\mathbb{E}[Y], \mathbb{E}[Y^2]/\mathbb{E}[Y]]$ if $\mathbb{E}[Y] > 0$.

C.2 Rationality of the PMMSE (Theorem 4.2): Proofs of Section 4.3.1

C.2.1 Proof of Lemma 4.6

We introduce the following functions. Recall that we denote $\mathcal{X}_k = \mathbb{E}[X^k]$. For $k \in [n]$, we define the function $v_{X,k} : [0,\infty) \to \mathbb{R}$ at each $t \geq 0$ by

$$v_{X,k}(t) := \mathbb{E}\left[X\left(\sqrt{t}X+N\right)^k\right]. \tag{C.10}$$

For example, $v_{X,0}(t) = \mathcal{X}_1$, $v_{X,1}(t) = \sqrt{t}\mathcal{X}_2$, and $v_{X,2}(t) = t\mathcal{X}_3 + \mathcal{X}_1$ if $X \in L^3(P)$. Define the vector-valued function $\boldsymbol{v}_{X,n} : [0,\infty) \to \mathbb{R}^{n+1}$ via

$$\boldsymbol{v}_{X,n} := (v_{X,0}, \cdots, v_{X,n})^T. \tag{C.11}$$

In view of Lemma 4.1, we may represent the PMMSE as

$$\mathrm{pmmse}_n(X,t) = \mathbb{E}\left[X^2\right] - \boldsymbol{v}_{X,n}(t)^T \boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} \boldsymbol{v}_{X,n}(t). \tag{C.12}$$

Therefore, defining $F_{X,n} : [0,\infty) \to [0,\infty)$ by

$$F_{X,n}(t) := \boldsymbol{v}_{X,n}(t)^T \boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} \boldsymbol{v}_{X,n}(t), \tag{C.13}$$

we have the equation

$$\mathrm{pmmse}_n(X,t) = \mathbb{E}\left[X^2\right] - F_{X,n}(t). \tag{C.14}$$

The functions $F_{X,n}$ are non-negative because the matrices $\boldsymbol{M}_{\sqrt{t}X+N,n}$ are positive-definite (see Lemma 4.1). In view of (C.14), PMMSE is fully characterized by $F_{X,n}$, so we focus on this function.

We introduce the following auxiliary polynomials, where R is a random variable independent of N. For ℓ even, we set

$$e_{R,X,\ell}(t) := \mathbb{E}\left[R\left(\sqrt{t}X+N\right)^{\ell}\right], \tag{C.15}$$

and for ℓ odd we set (for $t > 0$)

$$o_{R,X,\ell}(t) := t^{-1/2}\mathbb{E}\left[R\left(\sqrt{t}X+N\right)^{\ell}\right]. \tag{C.16}$$

That $e_{R,X,\ell}$ and $o_{R,X,\ell}$ are polynomials in t can be seen as follows. Recall that $\mathbb{E}[N^r] = 0$ for odd $r \in \mathbb{N}$. If ℓ is even then expanding the right hand side of (C.15) yields

$$e_{R,X,\ell}(t) = \sum_{\substack{k\in[\ell] \\ k \text{ even}}} \binom{\ell}{k} t^{k/2}\mathbb{E}\left[RX^k\right]\mathbb{E}\left[N^{\ell-k}\right], \tag{C.17}$$

whereas if ℓ is odd then the right hand side of (C.16) yields

$$o_{R,X,\ell}(t) = \sum_{\substack{k\in[\ell] \\ k \text{ odd}}} \binom{\ell}{k} t^{(k-1)/2}\mathbb{E}\left[RX^k\right]\mathbb{E}\left[N^{\ell-k}\right]. \tag{C.18}$$

Both expressions on the right hand sides of (C.17) and (C.18) are polynomials of degree at most $\lfloor \ell/2 \rfloor$. Further, the coefficient of $t^{\lfloor \ell/2 \rfloor}$ in either polynomial is $\mathbb{E}\left[RX^{\ell}\right]$.

Let $\mathcal{S}_{[n]}$ denote the symmetric group of permutations on the $n+1$ elements of $[n]$. We utilize the following auxiliary result on the parity of $i + \pi(i)$ for a permutation $\pi \in \mathcal{S}_{[n]}$.

Lemma C.1. *For any permutation $\pi \in \mathcal{S}_{[n]}$, there is an even number of elements $i \in [n]$ such that $i + \pi(i)$ is odd, i.e., the following is an even integer*

$$\delta(\pi) := |\{i \in [n] \ ; \ i + \pi(i) \ \text{ is odd}\}| \,. \tag{C.19}$$

Proof. The integer $i + \pi(i)$ is odd if and only if i and $\pi(i)$ have opposite parities. Thus, the desired result follows from the following more general characterization. For any partition $[n] = A \cup B$, the cardinality of the set

$$I := \{i \in [n] \ ; \ (i, \pi(i)) \in (A \times B) \cup (B \times A)\} \tag{C.20}$$

is even. The desired result follows by letting A and B be even and odd integers, respectively, in $[n]$. Now, we show that the general characterization holds.

Let $A_\pi \subset A$ denote the subset of elements of A that get mapped by π into B, i.e.,

$$A_\pi := \{i \in A \ ; \ \pi(i) \in B\}, \tag{C.21}$$

and define B_π similarly. Then, $I = A_\pi \cup B_\pi$ is a partition. As $|A_\pi| = |B_\pi|$, we get that $|I| = 2|A_\pi|$, and the desired result that $|I|$ is even follows. □

We show first that the function $t \mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t, and show that the coefficient of t^{d_n} in it is $\det \boldsymbol{M}_{X,n}$. By Leibniz's formula,

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi \in \mathcal{S}_{[n]}} \operatorname{sgn}(\pi) \prod_{r \in [n]} \mathbb{E}\left[\left(\sqrt{t}X + N\right)^{r+\pi(r)}\right]. \tag{C.22}$$

With the auxiliary polynomials $e_{1,X,\ell}$ and $o_{1,X,\ell}$ as defined in (C.15) and (C.16) (i.e., with $R = 1$), and δ as defined in (C.19), we may write

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi \in \mathcal{S}_{[n]}} \operatorname{sgn}(\pi) t^{\delta(\pi)/2} \prod_{\substack{i \in [n] \\ i+\pi(i) \text{ odd}}} o_{1,X,i+\pi(i)}(t) \prod_{\substack{j \in [n] \\ j+\pi(j) \text{ even}}} e_{1,X,j+\pi(j)}(t), \tag{C.23}$$

thereby showing that $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t by evenness of each $\delta(\pi)$ (Lemma C.1). Furthermore, for each permutation $\pi \in \mathcal{S}_{[n]}$,

$$\deg\left(t^{\delta(\pi)/2} \prod_{i+\pi(i) \text{ odd}} o_{1,X,i+\pi(i)}(t) \prod_{j+\pi(j) \text{ even}} e_{1,X,j+\pi(j)}(t)\right) \le \frac{\delta(\pi)}{2} + \sum_{i+\pi(i) \text{ odd}} \frac{i+\pi(i)-1}{2} + \sum_{j+\pi(j) \text{ even}} \frac{j+\pi(j)}{2} \tag{C.24}$$

$$= \frac{1}{2} \sum_{k=0}^{n} k + \pi(k) = \frac{n(n+1)}{2} = d_n. \tag{C.25}$$

Therefore, we also have $\deg\left(\det \boldsymbol{M}_{\sqrt{t}X+N,n}\right) \leq d_n$. Finally, taking the terms of highest degrees in $\sqrt{t}$ in (C.22), we obtain that the coefficient of t^{d_n} in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is

$$\sum_{\pi \in \mathcal{S}_{[n]}} \operatorname{sgn}(\pi) \prod_{r \in [n]} \mathcal{X}_{r+\pi(r)}, \tag{C.26}$$

which is equal to $\det \boldsymbol{M}_{X,n}$ by the Leibniz determinant formula. This coefficient is non-negative because $\boldsymbol{M}_{X,n}$ is positive-semidefinite, and it is nonzero if and only if $|\text{supp}(X)| > n$ by Lemma 4.1.

The same approach can be used to show that the mapping $t \mapsto F_{X,n}(t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t and to characterize its leading coefficient. In this case, we utilize $e_{X,X,\ell}$ and $o_{X,X,\ell}$ (i.e., $R = X$).

For each $(i,j) \in [n]^2$ let the subset $T_n^{(i,j)} \subset \mathcal{S}_{[n]}$ denote the collection of permutations sending i to j, i.e.,

$$T_n^{(i,j)} := \left\{\pi \in \mathcal{S}_{[n]} \ ; \ \pi(i) = j\right\}. \tag{C.27}$$

We define, for each $(i,j) \in [n]^2$, the cofactor functions $c_{X,n}^{(i,j)} : [0,\infty) \to \mathbb{R}$ and the products $d_{X,n}^{(i,j)} : [0,\infty) \to \mathbb{R}$ by

$$c_{X,n}^{(i,j)}(t) := \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \prod_{\substack{k \in [n] \\ k \neq i}} \left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{k,\pi(k)}, \tag{C.28}$$

$$d_{X,n}^{(i,j)}(t) := v_{X,i}(t)\, c_{X,n}^{(i,j)}(t)\, v_{X,j}(t). \tag{C.29}$$

Here, $\left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{a,b}$ is the (a,b)-th entry of $\boldsymbol{M}_{\sqrt{t}X+N,n}$, i.e.,

$$\left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{a,b} = \mathbb{E}\left[\left(\sqrt{t}X + N\right)^{a+b}\right]. \tag{C.30}$$

Note that $c_{X,n}^{(i,j)}(t)$ is the (i,j)-th cofactor of $\boldsymbol{M}_{\sqrt{t}X+N,n}$. The cofactor matrix $\boldsymbol{C}_{X,n} : [0,\infty) \to \mathbb{R}^{(n+1)\times(n+1)}$ of $t \mapsto \boldsymbol{M}_{\sqrt{t}X+N,n}$ is given by

$$\boldsymbol{C}_{X,n} := \left(c_{X,n}^{(i,j)}\right)_{(i,j)\in[n]^2}. \tag{C.31}$$

We define the function $D_{X,t} : [0,\infty) \to \mathbb{R}$ by

$$D_{X,n} := \boldsymbol{v}_{X,n}^T \boldsymbol{C}_{X,n} \boldsymbol{v}_{X,n}. \tag{C.32}$$

We have the following two relations. First, $D_{X,n}$ is the sum of the $d_{X,n}^{(i,j)}$

$$D_{X,n}(t) = \sum_{(i,j)\in[n]^2} d_{X,n}^{(i,j)}(t). \tag{C.33}$$

Second, by Cramer's rule, and because symmetry of the matrix $\boldsymbol{M}_{\sqrt{t}X+N,n}$ implies that its cofactor is equal to its adjugate, we have the formula

$$\boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} = \frac{1}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} \, \boldsymbol{C}_{X,n}. \tag{C.34}$$

Therefore, we obtain

$$F_{X,n}(t) = \frac{D_{X,n}(t)}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} = \frac{\sum_{(i,j)\in[n]^2} d_{X,n}^{(i,j)}(t)}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}}. \tag{C.35}$$

Hence, it suffices to study the $d_{X,n}^{(i,j)}$.

We start with a characterization of the cofactors $c_{X,n}^{(i,j)}$. Namely, we show that if $i+j$ is even then $c_{X,n}^{(i,j)}(t)$ is a polynomial in t, and if $i+j$ is odd then $\sqrt{t}c_{X,n}^{(i,j)}(t)$ is a polynomial in t. If $i+j$ is even, then

$$c_{X,n}^{(i,j)}(t) = \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) t^{\delta(\pi)/2} \prod_{\substack{k\in[n] \\ k+\pi(k) \text{ odd}}} o_{1,X,k+\pi(k)}(t) \prod_{\substack{r\in[n],\ r\neq i \\ r+\pi(r) \text{ even}}} e_{1,X,r+\pi(r)}(t), \tag{C.36}$$

whereas if $i+j$ is odd then

$$c_{X,n}^{(i,j)}(t) = \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) t^{\frac{\delta(\pi)-1}{2}} \prod_{\substack{k\in[n],\ k\neq i \\ k+\pi(k) \text{ odd}}} o_{1,X,k+\pi(k)}(t) \prod_{\substack{r\in[n] \\ r+\pi(r) \text{ even}}} e_{1,X,r+\pi(r)}(t). \tag{C.37}$$

Thus, evenness of $\delta(\pi)$ for each $\pi \in \mathcal{S}_{[n]}$ implies that each $c_{X,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is even and that each $\sqrt{t}c_{X,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is odd. Further, the degree of $c_{X,n}^{(i,j)}$ for even $i+j$ is upper bounded by

$$\frac{\delta(\pi)}{2} + \sum_{k+\pi(k) \text{ odd}} \frac{k+\pi(k)-1}{2} + \sum_{r+\pi(r) \text{ even}\,;\, r\neq i} \frac{r+\pi(r)}{2} = \frac{n(n+1)}{2} - \frac{i+j}{2}, \tag{C.38}$$

whereas the degree of $\sqrt{t}c_{X,n}^{(i,j)}$ and for odd $i+j$ is upper bounded by

$$\frac{\delta(\pi)}{2} + \sum_{k+\pi(k) \text{ odd}\,;\, k\neq i} \frac{k+\pi(k)-1}{2} + \sum_{r+\pi(r) \text{ even}} \frac{r+\pi(r)}{2} = \frac{n(n+1)}{2} - \frac{i+j-1}{2}. \tag{C.39}$$

We note that both upper bounds are equal to

$$\frac{n(n+1)}{2} - \left\lfloor \frac{i+j}{2} \right\rfloor. \tag{C.40}$$

Finally, considering the terms of highest order, we see that the term

$$\sum_{\pi \in T_n^{(i,j)}} \mathrm{sgn}(\pi) \prod_{k \in [n] \setminus \{i\}} \mathcal{X}_{k+\pi(k)} \tag{C.41}$$

is the coefficient of $t^{\frac{n(n+1)}{2} - \left\lfloor \frac{i+j}{2} \right\rfloor}$ in $c_{X,n}^{(i,j)}$ when $i+j$ is even and in $\sqrt{t}c_{X,n}^{(i,j)}$ when $i+j$ is odd.

Now, to show that $D_{X,n}$ is a polynomial, it suffices to check that each $d_{X,n}^{(i,j)}$ is. We consider separately the parity of $i+j$ and build upon the characterization of $c_{X,n}^{(i,j)}$. If $i+j$ is even, so i and j have the same parity, then

$$\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^i\right]\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^j\right]$$

is a polynomial in t of degree at most $(i+j)/2$ with the coefficient of $t^{(i+j)/2}$ being $\mathcal{X}_{i+1}\mathcal{X}_{j+1}$. If $i+j$ is odd, so i and j have different parities, then

$$t^{-1/2}\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^i\right]\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^j\right]$$

is a polynomial in t of degree at most $(i+j-1)/2$ with the coefficient of $t^{(i+j-1)/2}$ being $\mathcal{X}_{i+1}\mathcal{X}_{j+1}$.

Thus, from the characterization of $c_{X,n}^{(i,j)}$, regardless of the parity of $i+j$ we obtain that $d_{X,n}^{(i,j)}(t)$ is a polynomial in t of degree at most $n(n+1)/2 = d_n$. Thus, from (C.35), the function $t \mapsto F_{X,n}(t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most d_n. Further, note that $\mathrm{pmmse}_n(X,t) \leq \mathrm{lmmse}(X,t) \to 0$ as $t \to \infty$. Thus, writing

$$\mathrm{pmmse}_n(X,t) = \frac{(\mathcal{X}_2 - F_{X,n}(t)) \det \boldsymbol{M}_{\sqrt{t}X+N,n}}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} \tag{C.42}$$

and recalling that we have shown that $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t of degree at most d_n, we conclude that the numerator $t \mapsto \mathrm{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most $d_n - 1$.

Next, we derive the constant terms. Denote by $a_X^{n,0}$ and $b_X^{n,0}$ the constant terms of the polynomials $t \mapsto \mathrm{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ and $t \mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$, respectively. The formulas for $a_X^{n,0}$ and $b_X^{n,0}$ follow simply by setting $t=0$. Indeed, if $N \sim \mathcal{N}(0,1)$ is independent of X, then

$$F_{X,n}(0) = \mathcal{X}_1^2 \, \mathbb{E}\left[\boldsymbol{N}^{(n)}\right]^T \boldsymbol{M}_{N,n}^{-1} \, \mathbb{E}\left[\boldsymbol{N}^{(n)}\right] = \mathcal{X}_1^2 \tag{C.43}$$

because $\mathbb{E}\left[\boldsymbol{N}^{(n)}\right]$ is the leftmost column of $\boldsymbol{M}_{N,n}$. Therefore,

$$a_X^{n,0} = \sigma_X^2 \det \boldsymbol{M}_{N,n} = \sigma_X^2 b_X^{n,0}. \tag{C.44}$$

Further, by direct computation or using the connection between Hankel matrices and orthogonal polynomials [Sim98, Appendix A] along with the fact that the probabilist's Hermite polynomials q_k satisfy the recurrence $xq_k(x) = q_{k+1}(x) + kq_{k-1}(x)$, it follows that $\det \boldsymbol{M}_{N,n} = \prod_{k=1}^n k! = G(n+2)$ where G is the Barnes G-function.

Next, we show the last statement in the lemma, namely, that each coefficient in either of the two considered polynomials stays unchanged if X is shifted by a constant. This property will allow us to prove the claim that the coefficient of t in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is $\sigma_X^2 G(n+2)d_n$. By what we have shown thus far, we may define constants $a_X^{n,j}$ and $b_X^{n,j}$ by the polynomial identities

$$\text{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n-1]} a_X^{n,j}\ t^j, \tag{C.45}$$

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j}\ t^j. \tag{C.46}$$

Fix $s \in \mathbb{R}$. For any i.i.d. random variables $Z, Z_0, \cdots, Z_n$, we have that (see, e.g., [Sim98, Appendix A])

$$\det \boldsymbol{M}_{Z,n} = \frac{1}{(n+1)!}\ \mathbb{E}\left[\prod_{0\le i<j\le n} (Z_i - Z_j)^2\right]. \tag{C.47}$$

From equation (C.47), since $(Z_i+s)-(Z_j+s) = Z_i - Z_j$, we obtain that

$$\det \boldsymbol{M}_{Z+s,n} = \det \boldsymbol{M}_{Z,n}. \tag{C.48}$$

Let $N \sim \mathcal{N}(0,1)$ be independent of X. Then, for every $t \in [0,\infty)$, considering $Z = \sqrt{t}X + N$ in (C.48), we obtain

$$\det \boldsymbol{M}_{\sqrt{t}(X+s)+N,n} = \det \boldsymbol{M}_{\sqrt{t}X+N,n}. \tag{C.49}$$

As both sides of (C.49) are polynomials in t, we obtain that $b_{X+s}^{n,j} = b_X^{n,j}$ for every $j \in [d_n]$. Since we also have $\text{pmmse}_n(X+s,t) = \text{pmmse}_n(X,t)$, it follows that

$$t \mapsto \sum_{j\in[d_n-1]} a_X^{n,j} t^j = \text{pmmse}_n(X,t) \sum_{j\in[d_n]} b_X^{n,j} t^j \tag{C.50}$$

is also invariant under shifting X, so we also obtain $a_{X+s}^{n,j} = a_X^{n,j}$.

By the shift-invariance of $b_X^{n,1}$, we may assume that $\mathcal{X}_1 = 0$ (so $\mathcal{X}_2 = \sigma_X^2$). Now, as each entry in $\boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in $\sqrt{t}$, we see that we may drop any term of order $(\sqrt{t})^3$ or above for the sake of finding $b_X^{n,1}$ (which is the coefficient of t in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$). In other words,

$$b_X^{n,1} = \det\left(\binom{i+j}{2}\sigma_X^2 \mathbb{E}\left[N^{i+j-2}\right]t + \mathbb{E}\left[N^{i+j}\right]\right)_{(i,j)\in[n]^2}. \tag{C.51}$$

By Leibniz's formula, we conclude

$$b_X^{n,1} = \sigma_X^2 \sum_{\substack{\pi\in\mathcal{S}_{[n]} \\ k\in[n]}} \operatorname{sgn}(\pi)\binom{k+\pi(k)}{2}\mathbb{E}\left[N^{k+\pi(k)-2}\right]\cdot \prod_{i\in[n]\setminus\{k\}}\mathbb{E}\left[N^{i+\pi(i)}\right]. \tag{C.52}$$

But, for any non-negative integer m

$$\binom{m}{2}\mathbb{E}\left[N^{m-2}\right] = \frac{m}{2}\mathbb{E}\left[N^m\right]. \tag{C.53}$$

Therefore, (C.52) simplifies to

$$b_X^{n,1} = \frac{\sigma_X^2}{2}\sum_{\substack{\pi\in\mathcal{S}_{[n]} \\ k\in[n]}} \operatorname{sgn}(\pi)(k+\pi(k))\prod_{i\in[n]}\mathbb{E}\left[N^{i+\pi(i)}\right]. \tag{C.54}$$

Evaluating the summation over k for each fixed π, we obtain that

$$b_X^{n,1} = \binom{n+1}{2}\sigma_X^2 \sum_{\pi\in\mathcal{S}_{[n]}} \operatorname{sgn}(\pi)\prod_{i\in[n]}\mathbb{E}\left[N^{i+\pi(i)}\right]. \tag{C.55}$$

Finally, by Leibniz's formula for $\det \boldsymbol{M}_{N,n}$, we obtain that

$$b_X^{n,1} = \binom{n+1}{2}\sigma_X^2 \det \boldsymbol{M}_{N,n}, \tag{C.56}$$

as desired. This completes the proof of Lemma 4.6.

C.2.2 Expanded Formulas for the Coefficients in (4.48)

As stated in Remark 4.8, we give here fully-expanded formulas for the coefficients $a_X^{n,j}$ and $b_X^{n,j}$, which will yield further restrictions on which moments can appear in any of these coefficients. Recall that we set $\mathcal{X}_k = \mathbb{E}[X^k]$.

We have the expansion (see (C.22))

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi\in\mathcal{S}_{[n]}} \operatorname{sgn}(\pi)\prod_{r\in[n]}\mathbb{E}\left[\left(\sqrt{t}X+N\right)^{r+\pi(r)}\right] \tag{C.57}$$

by the Leibniz formula. In the expressions that follow, we denote the tuple $\boldsymbol{k} = (k_0,\cdots,k_n)$. Expanding the powers inside the expectation and computing the expectation, we get a formula of the form

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k_r\in[r+\pi(r)],\ \forall r\in[n]}} t^{(k_0+\cdots+k_n)/2}\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}\beta_{\pi;\boldsymbol{k}}, \tag{C.58}$$

where the $\beta_{\pi;\boldsymbol{k}}$ are integers given by[1]

$$\beta_{\pi,\boldsymbol{k}} := \operatorname{sgn}(\pi)\prod_{r\in[n]}\binom{r+\pi(r)}{k_r}\mathbb{E}[N^{r+\pi(r)-k_r}]. \tag{C.59}$$

By Lemma 4.6, only the summands for which the integer $k_0+\cdots+k_n$ is even can be non-trivial, because $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t. Thus, we have

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} t^j \sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k_r\in[r+\pi(r)],\ \forall r\in[n]\\ k_0+\cdots+k_n=2j}} \beta_{\pi;\boldsymbol{k}}\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}. \tag{C.60}$$

Because the coefficients $b_X^{n,j}$ were defined by equality of polynomials $\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j}t^j$ (see (4.47)), we obtain that for each $j\in[d_n]$

$$b_X^{n,j} = \sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k_r\in[r+\pi(r)],\ \forall r\in[n]\\ k_0+\cdots+k_n=2j}} \beta_{\pi;\boldsymbol{k}}\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}. \tag{C.61}$$

The coefficient $a_X^{n,j}$ may be expanded similarly to obtain the following formula. Define the integers

$$\gamma_{i,\pi,\boldsymbol{k},w,z} =(-1)^{i+\pi(i)}\operatorname{sgn}(\pi)\binom{i}{w}\binom{\pi(i)}{z}\mathbb{E}[N^{i-w}]\mathbb{E}[N^{\pi(i)-z}]\prod_{r\in[n]\setminus\{i\}}\binom{r+\pi(r)}{k_r}\mathbb{E}[N^{r+\pi(r)-k_r}], \tag{C.62}$$

and the restricted sums

$$s_i(\boldsymbol{k}) = \sum_{r\in[n]\setminus\{i\}} k_r. \tag{C.63}$$

[1] From this formula, one may deduce an alternative proof that $t\mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial. The term $\beta_{\pi;\boldsymbol{k}}$ is nonzero if and only if all the differences $r+\pi(r)-k_r$ are even. Suppose, for the sake of contradiction, that this is true for some fixed permutation $\pi\in\mathcal{S}_{[n]}$ and naturals $k_0,\cdots,k_n$ for which $k_0+\cdots+k_n$ is odd. Then, there is an odd number of odd numbers k_r. But, by Lemma C.1, there is an even number of odd numbers $r+\pi(r)$. Therefore, there is an $r\in[n]$ for which $r+\pi(r)$ and k_r have different parities, contradicting evenness of $r+\pi(r)-k_r$.

Then, we have the formula

$$a_X^{n,j} \quad = \sum_{\substack{\pi\in\mathcal{S}_{[n]} \\ k_r\in[r+\pi(r)],\ \forall r\in[n] \\ k_0+\cdots+k_n=2j}} \beta_{\pi;k_0,\cdots,k_n}\mathcal{X}_2\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n} - \sum_{\substack{(i,\pi)\in[n]\times\mathcal{S}_{[n]} \\ (w,z)\in[i]\times[\pi(i)] \\ k_r\in[r+\pi(r)],\ \forall r\in[n]\setminus\{i\} \\ w+z+s_i(\boldsymbol{k})=2j}} \gamma_{i,\pi,\boldsymbol{k},w,z}\mathcal{X}_{w+1}\mathcal{X}_{z+1}\prod_{r\in[n]\setminus\{i\}}\mathcal{X}_{k_r}. \tag{C.64}$$

From the formulas for $a_X^{n,j}$ and $b_X^{n,j}$ in (C.64) and (C.61), we obtain the following restrictions on how they can contain any of the moments of X. We need to define the following set of homogeneous polynomials in the moments of X. We use the notation $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)^T \in \mathbb{N}^m$.

Definition C.1. For $(\ell, m, k) \in \mathbb{N}^3$, let $\Pi_{\ell,m,k}$ denote the set of unordered partitions of ℓ into at most m parts each of which not exceeding k, i.e., $\Pi_{\ell,m,k} := \left\{\boldsymbol{\lambda} \in \mathbb{N}^m \ ; \ k \geq \lambda_1 \geq \cdots \geq \lambda_m,\ \boldsymbol{\lambda}^T\mathbf{1} = \ell\right\}$. We define the set of homogeneous integer-coefficient polynomials with weighted-degree ℓ and width at most m in the first k moments $\mathcal{X}_1, \cdots, \mathcal{X}_k$ of X as

$$H_{\ell,m,k}(X) := \left\{ \sum_{\boldsymbol{\lambda}\in\Pi_{\ell,m,k}} c_{\boldsymbol{\lambda}} \prod_{i=1}^{m} \mathcal{X}_{\lambda_i} \ ; \ c_{\boldsymbol{\lambda}} \in \mathbb{Z} \right\}. \tag{C.65}$$

If $\Pi_{\ell,m,k} = \varnothing$, we set $H_{\ell,m,k}(X) = \mathbb{Z}$.

Remark C.2. An element $q(X) \in H_{\ell,m,k}(X)$ will be an integer linear combination of terms $\prod_{i=1}^m \mathcal{X}_{\lambda_i}$. Each of these terms is a product of at most m of the moments of X (hence the terminology *width*). The highest moment that can appear is $\mathcal{X}_k$, because $\boldsymbol{\lambda} \in \Pi_{\ell,m,k}$. Suppose $\Pi_{\ell,m,k} \neq \varnothing$. Then, each summand shares the property that $\sum_{i=1}^m \lambda_i = \ell$. Further, looking at each $\mathcal{X}_j$ as an indeterminate of "degree" j, we may view $q(X)$ as a "homogeneous" polynomial in the moments of X of "degree" ℓ. In other words, for any constant c, $q(cX) = c^\ell q(X)$; in fact, this homogeneity holds for each term in the sum defining q, $\prod_{i=1}^m \mathbb{E}\left[(cX)^{\lambda_i}\right] = c^\ell \prod_{i=1}^m \mathbb{E}\left[X^{\lambda_i}\right]$.

Example 6. The partitions of the integer 6 into at most 3 parts each of which not exceeding 4 are given by $\Pi_{6,3,4} = \{(4,2,0),(4,1,1),(3,3,0),(3,2,1),(2,2,2)\}$. Note the resemblance between the elements of $\Pi_{6,3,4}$ and the terms appearing in the expression for $\det \boldsymbol{M}_{X,2}$, namely, (see (4.37))

$$\det \boldsymbol{M}_{X,2} = \mathcal{X}_4\mathcal{X}_2 - \mathcal{X}_4\mathcal{X}_1^2 - \mathcal{X}_3^2 + 2\mathcal{X}_3\mathcal{X}_2\mathcal{X}_1 - \mathcal{X}_2^3. \tag{C.66}$$

A term $\prod_{i=1}^3 \mathcal{X}_{\lambda_i}$ with $\lambda_1 \geq \lambda_2 \geq \lambda_3$ appears in $\det \boldsymbol{M}_{X,2}$ if and only if $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ is in $\Pi_{6,3,4}$. In particular, $\det \boldsymbol{M}_{X,2} \in H_{6,3,4}(X)$. Leibniz's formula for the determinant can be used to show that, in general, $\det \boldsymbol{M}_{X,n} \in H_{n(n+1),n+1,2n}(X)$.

From (C.64) and (C.61), we have that the constants $a_X^{n,j}$ and $b_X^{n,j}$ satisfy

$$a_X^{n,j} \in H_{2j+2,\, \min(n,2j)+2,\, \tau_n(j)}(X) \tag{C.67}$$

$$b_X^{n,j} \in H_{2j,\, \min(n+1,2j),\, 2\min(n,j)}(X), \tag{C.68}$$

with $H_{\ell,m,k}(X)$ as given in Definition C.1 and $\tau_n(j) \leq 2\min(n, j+1)$ is defined by

$$\tau_n(j) = \begin{cases} 2 & \text{if } j = 0, \\ 2j+1 & \text{if } 1 \leq j \leq \frac{n}{2}, \\ 2j & \text{if } \frac{n+1}{2} \leq j \leq n, \\ 2n & \text{if } j > n. \end{cases} \tag{C.69}$$

C.2.3 Proof of Proposition 4.2

We proceed by induction on m. The case $m = 1$ follows because then by assumption on p we have that $p(k) = 0$ for every positive integer k as can be seen by taking $X \sim \mathcal{N}(k, 1)$, but the only polynomial with infinitely many zeros is the zero polynomial. Now, assume that the statement of the proposition holds for every polynomial in $m-1$ variables, where $m \geq 2$.

Fix a polynomial p in m variables, and assume that $p|_{\mathcal{C}^m} = 0$. Regarding p as a polynomial in one of the variables with coefficients being polynomials in the remaining $m-1$ variables, we may write

$$p(u_1, \cdots, u_m) = \sum_{j \in [d]} p_j(u_1, \cdots, u_{m-1}) u_m^j, \tag{C.70}$$

for some polynomials $p_0, \cdots, p_d$ in $m-1$ variables, where d is the total degree of p. We show that $p = 0$ identically by showing that each p_j vanishes on $\mathcal{C}_{m-1}$ and using the induction hypothesis to obtain $p_j = 0$ identically.

Fix $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_{m-1}) \in \mathcal{C}_{m-1}$. Let μ_m be a variable, and set $\ell = \lfloor m/2 \rfloor$. We have that $\ell = (m-1)/2$ if m is odd, and $\ell = m/2$ if m is even. Set $\boldsymbol{H} = (\mu_{i+j})_{(i,j) \in [\ell]^2}$. If m is even, then $\det \boldsymbol{H} = \alpha \mu_m + \beta$ for some constants $\alpha, \beta \in \mathbb{R}$ determined by $\boldsymbol{\mu}$, with $\alpha = \det(\mu_{i+j})_{(i,j) \in [\ell-1]^2} > 0$. In the case m is even, we set $t = -\beta/\alpha$, and in the case m is odd, we set $t = 0$. Then, $\boldsymbol{H}$ is positive definite whenever $\mu_m > t$.

For each integer $k \geq 1$ and real $\varepsilon > 0$, Lemma 4.7 yields a random variable $X_{k,\varepsilon} \in \mathcal{R}_m$ satisfying

$$\delta_{k,\ell}(\varepsilon) := \mathbb{E}[X_{k,\varepsilon}^\ell] - \mu_\ell \in (-\varepsilon, \varepsilon) \tag{C.71}$$

for each $\ell \in \{1, \cdots, m-1\}$ and

$$\delta_{k,m}(\varepsilon) := \mathbb{E}[X_{k,\varepsilon}^m] - (t+k) \in (-\varepsilon, \varepsilon). \tag{C.72}$$

Then, by assumption on p, for every $\varepsilon > 0$ and $k \in \mathbb{N}_{\geq 1}$,

$$\sum_{j \in [d]} p_j \left(\boldsymbol{\mu} + (\delta_{k,\ell}(\varepsilon))_{1 \leq \ell \leq m-1} \right) (t + k + \delta_{k,m}(\varepsilon))^j = 0. \tag{C.73}$$

Taking the limit $\varepsilon \to 0^+$, we deduce that

$$\sum_{j \in [d]} p_j(\mu_1, \cdots, \mu_{m-1})(t+k)^j = 0. \tag{C.74}$$

Considering the left-hand side in (C.74) as a univariate polynomial in k, and noting that the vanishing in (C.74) holds at infinitely many values of k, we deduce that

$$p_j(\mu_1, \cdots, \mu_{m-1}) = 0 \tag{C.75}$$

for every $j \in [d]$. This holds for every $(\mu_1, \cdots, \mu_{m-1}) \in \mathcal{C}_{m-1}$, i.e., the premise of the proposition applies to each p_j (namely, for every $X \in \mathcal{R}_{m-1}$ we have $p_j(\mathbb{E}[X], \cdots, \mathbb{E}[X^{m-1}]) = 0$). By the induction hypothesis, we obtain $p_j = 0$, as polynomials, for every $j \in [d]$. Therefore, $p = 0$, and the proof is complete.

C.3 Convergence of the PMMSE to the MMSE in Gaussian Channels (Theorem 4.3): Proofs of Section 4.3.2

We derive in Appendix C.3.1 the uniform convergence $\sup_{t \geq 0}$ $\text{pmmse}_n(X,t) - \text{mmse}(X,t) \searrow 0$ stated in equation (4.13). Lemma 4.8 regarding Freud weights is derived in Appendix C.3.2, and the bound on the higher-order derivatives of the conditional expectation given in Lemma 4.9 is shown in Appendix C.3.3.

C.3.1 Uniform Convergence of PMMSE to MMSE (4.13)

We start the proof by obtaining from Theorem 4.1 pointwise convergence. Let $N \sim \mathcal{N}(0,1)$ be independent of X. The MGF of $\sqrt{t}X + N$ exists (it is the product of the MGFs of $\sqrt{t}X$ and N) and this implies that $\sqrt{t}X + N$ satisfies Carleman's condition [Sch17, Sec. 4.2]. Hence, by Theorem 4.1, we have $\lim_{n\to\infty} \mathrm{pmmse}_n(X,t) = \mathrm{mmse}(X,t)$ pointwise for each fixed $t \geq 0$. Now, we show that the convergence is in fact uniform in t.

For each $n \in \mathbb{N}$ and $t \in [0,\infty)$, write $g_n(t) := \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)$. We will show that

$$\lim_{n\to\infty} \sup_{t\in[0,\infty)} g_n(t) = 0. \tag{C.76}$$

The limit $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ as $n \to \infty$ says that $g_n(t) \searrow 0$ as $n \to \infty$ for every fixed $t \geq 0$. In addition, the asymptotics given in Corollary 4.1 imply that for each fixed $n \in \mathbb{N}$, $g_n(t) \to 0$ as $t \to \infty$. Note that $\{g_n\}_{n\in\mathbb{N}}$ is a pointwise decreasing sequence of nonnegative functions. We finish the proof via Cantor's intersection theorem.

Fix $\varepsilon > 0$. For each $n \in \mathbb{N}$, let $C_{\varepsilon,n} = g_n^{-1}([\varepsilon,\infty))$, where g_n^{-1} denotes the set-theoretic inverse. As $\{g_n\}_{n\in\mathbb{N}}$ is decreasing, $C_{\varepsilon,1} \supseteq C_{\varepsilon,2} \supseteq \cdots$ is decreasing too. As each g_n is continuous, each $C_{\varepsilon,n}$ is closed. Further, $\lim_{t\to\infty} g_1(t) = 0$ implies that $C_{\varepsilon,1}$ is bounded, hence each $C_{\varepsilon,n}$ is bounded. Thus, each $C_{\varepsilon,n}$ is compact. But, the intersection $\bigcap_{n\in\mathbb{N}} C_{\varepsilon,n}$ is empty, for if t_0 were in the intersection then $\liminf_{n\to\infty} g_n(t_0) \geq \varepsilon$ violating that $\lim_{n\to\infty} g_n(t_0) = 0$. Hence, by Cantor's intersection theorem, it must be that the $C_{\varepsilon,n}$ are eventually empty, so there is an $m \in \mathbb{N}$ such that $\sup_{t\in[0,\infty)} g_n(t) \leq \varepsilon$ for every $n > m$. This is precisely the uniform convergence in (C.76), and the proof is complete.

C.3.2 Proof of Lemma 4.8

Write $Y = X + N$ and $p_Y = e^{-Q}$. To see that p_Y^s is a Freud weight, it suffices to show that p_Y is a Freud weight, since it can be easily seen that the conditions in Definition 4.2 hold for p_Y^s if they hold for p_Y. First, we note that $Q'(y)$ is equal to $\mathbb{E}[N \mid Y = y]$.

Lemma C.2. *Fix a random variable X and let $Y = X + N$ where $N \sim \mathcal{N}(0,1)$ is independent of X. Writing $p_Y = e^{-Q}$, we have that $Q'(y) = \mathbb{E}[N \mid Y = y]$.*

Proof. We have that $p_Y(y) = \mathbb{E}[e^{-(y-X)^2/2}]/\sqrt{2\pi}$. Differentiating this equation, we obtain that $p_Y'(y) = \mathbb{E}[(X-y)e^{-(y-X)^2/2}]/\sqrt{2\pi}$, where the exchange of differentiation and integration is warranted since $t \mapsto te^{-t^2/2}$ is bounded. Now, $Q = -\log p_Y$, so $Q' = -p_Y'/p_Y$, i.e.,

$$Q'(y) = y - \frac{\mathbb{E}[Xe^{-(y-X)^2/2}]}{\mathbb{E}[e^{-(y-X)^2/2}]} = y - \mathbb{E}\left[X \mid Y = y\right]. \tag{C.77}$$

The proof is completed by substituting $X = Y - N$. □

In view of Lemma C.2, that p is even and non-increasing over $[0,\infty) \cap \mathrm{supp}(p)$ imply that Q

satisfies conditions (1)–(4) of Definition 4.2. It remains to show that property (5) holds. To this end, we show that if $\mathrm{supp}(p) \subset [-M, M]$ and $\lambda = M+2$, then for every $y > M+4$ we have that

$$1 < \frac{M^2+5M+8}{2(M+2)} \le \frac{Q'(\lambda y)}{Q'(y)} \le \frac{M^2+7M+8}{4}. \tag{C.78}$$

First, since $Q'(y) = y - \mathbb{E}[X \mid Y = y]$ (see (C.77)), we have the bounds $y - M \le Q'(y) \le y + M$ for every $y \in \mathbb{R}$. Therefore, $y > M$ and $\lambda > 1$ imply that

$$\frac{\lambda y - M}{y+M} \le \frac{Q'(\lambda y)}{Q'(y)} \le \frac{\lambda y + M}{y - M}. \tag{C.79}$$

Further, since $y > M+4$ and $\lambda = M+2$, we have

$$\frac{M^2+5M+8}{2(M+2)} < \lambda - \frac{M(M+3)}{y+M} = \frac{\lambda y - M}{y+M} \tag{C.80}$$

and

$$\frac{\lambda y + M}{y-M} = \lambda + \frac{M(M+3)}{y-M} \le \frac{M^2+7M+8}{4}. \tag{C.81}$$

The fact that $1 < \frac{M^2+5M+8}{2(M+2)}$ follows since the discriminant of M^2+3M+4 is $-7 < 0$. Therefore, p_Y is a Freud weight.

Next, we derive the bound on $a_n(sQ)$ stated in (4.58). By definition of a_n, we have that $a_n(sQ) = a_{n/s}(Q)$. Thus, it suffices to show $a_n(Q) \le (2M+\sqrt{2})\sqrt{n}$. By Lemma C.2,

$$Q'(y) = \mathbb{E}[N \mid Y = y] = y - \mathbb{E}[X \mid Y = y]. \tag{C.82}$$

Therefore $X \le M$ implies that, for any constant $z \ge 0$, we have

$$\int_0^1 \frac{ztQ'(zt)}{\sqrt{1-t^2}}\,dt = \frac{\pi}{4}z^2 - z\int_0^1 \frac{t}{\sqrt{1-t^2}} \frac{\mathbb{E}\left[Xe^{-(X-zt)^2/2}\right]}{\mathbb{E}\left[e^{-(X-zt)^2/2}\right]}dt \tag{C.83}$$

$$\ge \frac{\pi}{4}z^2 - Mz. \tag{C.84}$$

We have $\pi z^2/4 - Mz > n$ for $z = (2M+\sqrt{2})\sqrt{n}$. Since $y \mapsto yQ'(y)$ is strictly increasing over $(0,\infty)$ (condition (3) of Definition 4.2), we conclude that $a_n(Q) \le (2M+\sqrt{2})\sqrt{n}$. This completes the proof of Lemma 4.8.

C.3.3 Proof of Lemma 4.9

We use the formula of the conditional expectation derivative given in Proposition 4.3, with the conditional cumulant being expanded in terms of conditional moments using Bell polynomials, then apply Hölder's inequality to each ensuing summand. We use the following notation. The set of all finite-length tuples of non-negative integers is denoted by $\mathbb{N}^*$. For every integer $r \geq 2$, let Π_r be the set of unordered integer partitions $r = r_1 + \cdots + r_k$ of r into integers $r_j \geq 2$. We encode Π_r via a list of the multiplicities of the parts as

$$\Pi_r := \{(\lambda_2, \cdots, \lambda_\ell) \in \mathbb{N}^* \; ; \; 2\lambda_2 + \cdots + \ell\lambda_\ell = r\}. \tag{C.85}$$

In (C.85), $\ell \geq 2$ is free, and trailing zeros are ignored (i.e., $\lambda_\ell > 0$). For a partition $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$ having $m = \lambda_2 + \cdots + \lambda_\ell$ parts, we denote

$$c_{\boldsymbol{\lambda}} := \frac{1}{m}\binom{m}{\lambda_2, \cdots, \lambda_\ell}\binom{r}{\underbrace{2, \cdots, 2}_{\lambda_2}; \cdots ; \underbrace{\ell, \cdots, \ell}_{\lambda_\ell}} \tag{C.86}$$

and $e_{\boldsymbol{\lambda}} := (-1)^{m-1} c_{\boldsymbol{\lambda}}$. Set $C_r' := \sum_{\boldsymbol{\lambda} \in \Pi_r} c_{\boldsymbol{\lambda}}$. For each $(y, k) \in \mathbb{R} \times \mathbb{N}$, denote $f(y) := \mathbb{E}[X \mid Y = y]$ and

$$g_k(y) := \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^k \mid Y = y\right]. \tag{C.87}$$

For $\ell \geq 2$ and $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \mathbb{N}^{\ell-1}$, denote $\boldsymbol{g}^{\boldsymbol{\lambda}} := \prod_{i=2}^{\ell} g_i^{\lambda_i}$, with the understanding that $g_i^0 = 1$. Using Proposition 4.3, and expanding $\kappa_r(X \mid Y = y)$ in terms of the conditional moments $\mathbb{E}[X^k \mid Y = y]$, we obtain (see [AC21c, Proposition 1])

$$f^{(r-1)} = \sum_{\boldsymbol{\lambda} \in \Pi_r} e_{\boldsymbol{\lambda}} \boldsymbol{g}^{\boldsymbol{\lambda}}. \tag{C.88}$$

Fix $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$. By the generalization of Hölder's inequality stating $\|\psi_1 \cdots \psi_k\|_1 \leq \prod_{i=1}^{k} \|\psi_i\|_k$, we have that

$$\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2^2 = \left\|\prod_{\lambda_i \neq 0} g_i^{2\lambda_i}(Y)\right\|_1 \leq \prod_{\lambda_i \neq 0} \left\|g_i^{2\lambda_i}(Y)\right\|_s \tag{C.89}$$

where s is the number of nonzero entries in $\boldsymbol{\lambda}$. By Jensen's inequality for conditional expectation, for each i such that $\lambda_i \neq 0$, we have that

$$\left\|g_i^{2\lambda_i}(Y)\right\|_s \leq \|X - \mathbb{E}[X \mid Y]\|_{2i\lambda_i s}^{2i\lambda_i}. \tag{C.90}$$

Now, $r = \sum_{i=2}^{\ell} i\lambda_i \geq \sum_{i=2}^{s+1} i = \frac{(s+1)(s+2)}{2} - 1$, so we have that $s^2 + 3s - 2r \leq 0$, i.e., $s \leq q_r$. Further, $i\lambda_i \leq r$ for each i. Hence, monotonicity of norms and inequalities (C.89) and (C.90) imply the uniform (in $\boldsymbol{\lambda}$) bound

$$\left\| g^{\boldsymbol{\lambda}}(Y) \right\|_2 \leq \|X - \mathbb{E}[X \mid Y]\|_{2rq_r}^{r}. \tag{C.91}$$

Observe that $\|X - \mathbb{E}[X \mid Y]\|_k \leq 2 \min\left((k!)^{1/(2k)}, \|X\|_k \right)$ (see [GWSV11]). Therefore, applying the triangle inequality in (C.88) we obtain

$$\left\| f^{(r-1)}(Y) \right\|_2 \leq \sum_{\boldsymbol{\lambda} \in \Pi_r} c_{\boldsymbol{\lambda}} \left\| g^{\boldsymbol{\lambda}}(Y) \right\|_2 \tag{C.92}$$

$$\leq 2^r C_r' \ \min\left(\gamma_r, \|X\|_{2rq_r}^{r} \right), \tag{C.93}$$

where $\gamma_r = (2rq_r)!^{1/(4q_r)}$.

It only remains then to note that $C_r' = C_r$. The integer $c_{\boldsymbol{\lambda}}$ (as defined in (C.86)) can be easily seen to be equal to the number of cyclically-invariant ordered set-partitions of an r-element set into $m = \lambda_2 + \cdots + \lambda_\ell$ subsets where, for each $k \in \{2, \cdots, \ell\}$, exactly λ_k parts have size k. Hence, the integer C_r' equals the total number of cyclically-invariant ordered set-partitions of an r-element set into subsets of sizes at least 2, which is given by sequence A032181 at [OEI]. The formula for C_r' stated in [OEI] coincides with our definition of C_r in (4.60) in the statement of the lemma, from which we obtain $C_r' = C_r$. Finally, since the formula in [OEI] is stated without proof, we provide a proof here for completeness. Using the notation of [BR19], we have that

$$C_r' = \sum_{k=1}^{r} (k-1)! \left\{ {r \atop k} \right\}_{\geq 2} \tag{C.94}$$

where $\left\{ {r \atop k} \right\}_{\geq 2}$ denotes the number of partitions of an r-element set into k subsets each of which contains at least 2 elements (note that there are $(k-1)!$ cyclically-invariant arrangements of k parts). The exponential generating function of the sequence $r \mapsto \left\{ {r \atop k} \right\}_{\geq 2}$ is $(e^x - 1 - x)^k / k!$. Now, we may write

$$(e^x - 1 - x)^k = \sum_{a+b \leq k} \binom{k}{a, b} (-1)^{k-a} x^b \sum_{t \in \mathbb{N}} \frac{(ax)^t}{t!}. \tag{C.95}$$

Therefore, the coefficient of x^r in $(e^x - 1 - x)^k / k!$ is

$$\frac{1}{r!} \left\{ {r \atop k} \right\}_{\geq 2} = \sum_{a+b \leq k} \frac{(-1)^{k-a} a^{r-b}}{a! b! (k-a-b)! (r-b)!} \tag{C.96}$$

$$= \frac{1}{r!} \sum_{b=0}^{k} \binom{r}{b} \sum_{a=0}^{k-b} (-1)^{k-a} \frac{a^{r-b}}{a!(k-a-b)!} \tag{C.97}$$

$$= \frac{1}{r!} \sum_{b=0}^{k} \binom{r}{b} \left\{ \begin{matrix} r-b \\ k-b \end{matrix} \right\} (-1)^b, \tag{C.98}$$

which when combined with (C.94) gives $C'_r = C_r$ in view of (4.60). This completes the proof of the lemma.

Remark C.3. A closer analysis reveals that $i\lambda_i s$ in (C.90) cannot exceed $\beta_r := t_r^2(t_r + 1/2)$ where $t_r := (\sqrt{6r+7} - 1)/3$. For $r \to \infty$, we have $rq_r/\beta_r \sim 3^{3/2}/2 \approx 2.6$. The reduction when, e.g., $r = 7$, is from $rq_r = 14$ to $\beta_r = 10$.

C.4 Generalizations to Arbitrary Bases and Multiple Dimensions

We extend our approximation results for the conditional expectation from the polynomial-basis setting to arbitrary bases, and from conditioning on random variables to conditioning on arbitrary σ-algebras. An extension to the multidimensional case is also presented, which straightforwardly yields an approximation theorem for differential entropy of random vectors. Another byproduct of the multidimensional generalization is the expression for mutual information between two continuous random variables completely in terms of moments, as given in Theorem 4.5.

C.4.1 Arbitrary Bases and σ-Algebras

Up to here, our exposition dealt with the polynomial basis of $L^2(P_Y)$. However, our results can be extended to a more general setup. Recall that we have defined

$$\boldsymbol{M}_{Y,n} = \mathbb{E}\left[\boldsymbol{Y}^{(n)} \left(\boldsymbol{Y}^{(n)} \right)^T \right], \tag{C.99}$$

and derived

$$\mathbb{E}[X \mid Y] = \lim_{n\to\infty} \mathbb{E}\left[X\boldsymbol{Y}^{(n)} \right] \boldsymbol{M}_{Y,n}^{-1} \boldsymbol{Y}^{(n)} \tag{C.100}$$

in Theorem 4.1 under two requirements: Y satisfies Carleman's condition, and $|\text{supp}(Y)| = \infty$. Along similar lines, we derive a generalization where the set of polynomials of Y is replaced with any linearly-independent subset of $L^2(\Sigma)$ having a dense span, where $\Sigma \subset \mathcal{F}$ is any σ-algebra, and $L^2(\Sigma)$ denote the subset of $L^2(P)$ consisting of Σ-measurable random variables. Denseness replaces Carleman's condition, while linear independence replaces the infinite-support requirement.

Theorem C.1. *Fix a σ-algebra $\Sigma \subset \mathcal{F}$ and a set $\{\psi_k\}_{k\in\mathbb{N}} = \mathcal{S} \subset L^2(\Sigma)$. For each $n \in \mathbb{N}$, define the random vector $\boldsymbol{\psi}^{(n)} = (\psi_0, \cdots, \psi_n)^T$ and the matrix of inner products*

$$\boldsymbol{M}_{\mathcal{S},n} := \mathbb{E}\left[\boldsymbol{\psi}^{(n)}\left(\boldsymbol{\psi}^{(n)}\right)^T\right]. \tag{C.101}$$

If $\mathcal{S}$ is linearly independent and $\operatorname{span}(\mathcal{S})$ *is dense in $L^2(\Sigma)$, then*

$$\mathbb{E}[X \mid \Sigma] = \lim_{n\to\infty} \mathbb{E}\left[X\boldsymbol{\psi}^{(n)}\right]^T \boldsymbol{M}_{\mathcal{S},n}^{-1}\,\boldsymbol{\psi}^{(n)} \tag{C.102}$$

in $L^2(\Sigma)$ for any random variable $X \in L^2(P)$.

For the proof of Theorem C.1, we will need the following formula for the closest element in a finite-dimensional subspace of $L^2(P)$ to a random variable $X \in L^2(P)$, which will also be used for the extension of our results to random vectors later in this appendix. The following formula is simply an instantiation of the fact that, in a separable Hilbert space, the orthogonal projection onto a closed subspace is the unique closest element.

Lemma C.3. *For any fixed finite-dimensional subspace $\mathcal{V} \subset L^2(P)$ having a basis $\{V_0, V_1, \cdots, V_n\}$, denoting $\boldsymbol{V} = (V_0, V_1, \cdots, V_n)^T$, we have that for every $X \in L^2(P)$*

$$\mathbb{E}\left[X\boldsymbol{V}\right]^T \mathbb{E}\left[\boldsymbol{V}\boldsymbol{V}^T\right]^{-1}\boldsymbol{V} = \operatorname*{argmin}_{V\in\mathcal{V}} \|X - V\|_2\,. \tag{C.103}$$

In view of Lemma C.3, we introduce the following notation.

Definition C.2. Fix a random variable $X \in L^2(P)$, a σ-algebra $\Sigma \subset \mathcal{F}$, and a linearly-independent set $\{\theta_j\}_{j\in\mathbb{N}} = \Theta \subset L^2(\Sigma)$. Write $\boldsymbol{\theta}^{(n)} = (\theta_0, \cdots, \theta_n)^T$ for each $n \in \mathbb{N}$. We define the n-th approximation of $\mathbb{E}[X \mid \Sigma]$ with respect to Θ by

$$E_{n,\Theta}\left[X \mid \Sigma\right] := \mathbb{E}\left[X\boldsymbol{\theta}^{(n)}\right]\mathbb{E}\left[\boldsymbol{\theta}^{(n)}\left(\boldsymbol{\theta}^{(n)}\right)^T\right]^{-1}\boldsymbol{\theta}^{(n)}. \tag{C.104}$$

Note that $E_{n,\Theta}[X \mid \Sigma]$ belongs to $\operatorname{span}(\{\theta_j\}_{j\in[n]})$. Further, according to Lemma C.3, $E_{n,\Theta}[X \mid \Sigma]$ is the unique closest element in $\operatorname{span}(\{\theta_j\}_{j\in[n]})$ to X,

$$E_{n,\Theta}[X \mid \Sigma] = \operatorname*{argmin}_{V\in\,\operatorname{span}(\{\theta_j\}_{j\in[n]})} \|X - V\|_2\,. \tag{C.105}$$

If $Y \in L^{2n}(P)$, $\Theta = \{Y^j\}_{j\in\mathbb{N}}$, and $\Sigma = \sigma(Y)$, then the estimate reduces to $E_{n,\Theta}[X \mid \Sigma] = E_n[X \mid Y]$.

The central claim in Theorem C.1 is that if $\operatorname{span}(\Theta)$ is dense in $L^2(\Sigma)$ then we have the limit

$$\mathbb{E}[X \mid \Sigma] = \lim_{n\to\infty} E_{n,\Theta}[X \mid \Sigma]. \tag{C.106}$$

The proof of Theorem 4.1 can be adapted *mutatis mutandis* to derive the above limit, so we omit the details.

C.4.2 The Multidimensional PMMSE

We extend our results on the PMMSE of random variables to random vectors. We begin with some notation. The Hilbert space of q-integrable m-dimensional random vectors is denoted by $L^q(\mathbb{R}^m, P)$, with norm also denoted by $\| \cdot \|_q$. The subspace of Σ-measurable random vectors is denoted by $L^q(\mathbb{R}^m, \Sigma)$. We keep the notations $L^q(\mathbb{R}, \Sigma) = L^q(\Sigma)$ and $L^q(\mathbb{R}, P_Y) = L^q(P_Y)$. By a generalization of Hölder's inequality, for any $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T \in L^\beta(\mathbb{R}^m, P)$, we also have that $Y_1^{\alpha_1} \cdots Y_m^{\alpha_m} \in L^1(P)$ for any constants $\alpha_1, \cdots, \alpha_m \geq 0$ such that $\alpha_1 + \cdots + \alpha_m \leq \beta$.

We extend the notation $\boldsymbol{Y}^{(n)}$ to random vectors as follows. For an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$, we let $\boldsymbol{Y}^{(n,m)}$ denote the random vector whose entries are monomials in the Y_j of total degree at most n, ordered first by total degree then reverse-lexicographically in the exponents. For example, if $m = 3$ so $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$, then for $n = 2$

$$\boldsymbol{Y}^{(2,3)} = (1, Y_1, Y_2, Y_3, Y_1^2, Y_1Y_2, Y_1Y_3, Y_2^2, Y_2Y_3, Y_3^2)^T \tag{C.107}$$

because we are taking the totally ordered set of exponents $(\ \{\boldsymbol{v} \in \mathbb{N}^3 \mid \boldsymbol{1}^T\boldsymbol{v} \leq 2\}\ ,\ <\ )$ to have the order[2]

$$\begin{aligned}&(0,0,0) < (1,0,0) < (0,1,0) < (0,0,1) < (2,0,0)\\ &< (1,1,0) < (1,0,1) < (0,2,0) < (0,1,1) < (0,0,2).\end{aligned}$$

A straightforward stars-and-bars counting argument reveals that the length of $\boldsymbol{Y}^{(n,m)}$ is $\binom{n+m}{m}$.

Let $\mathscr{P}_{n,m}$ denote the set of polynomials in m variables with real coefficients of total degree at most n. For a fixed m-dimensional random vector $\boldsymbol{Y}$, denote $\mathscr{P}_{n,m}(\boldsymbol{Y}) := \{p(\boldsymbol{Y})\ ;\ p \in \mathscr{P}_{n,m}\}$. Note that $\mathscr{P}_{n,1} = \mathscr{P}_n$. Also, the notation $\boldsymbol{Y}^{(n,1)}$, while avoided, is disambiguated by interpreting it as $\boldsymbol{Y}^{(n)}$, i.e., $\boldsymbol{Y}^{(n,1)} = (1, Y, \cdots, Y^n)^T$ where the subscript on Y_1 is dropped. We denote the product sets of $\mathscr{P}_{n,m}(\boldsymbol{Y})$ by $\mathscr{P}_{n,m}^\ell(\boldsymbol{Y})$, and consider their elements as vectors rather than tuples. In other words, we denote the set of length-ℓ vectors whose coordinates are multivariate polynomial expressions of an

[2] Note that this ordering is not the same as the degree reverse lexicographical order nor its reverse.

m-dimensional random vector $\boldsymbol{Y}$ with total degree at most n by

$$\mathscr{P}_{n,m}^{\ell}(\boldsymbol{Y}) = \left\{ (p_1(\boldsymbol{Y}), \cdots, p_\ell(\boldsymbol{Y}))^T \, ; \, p_1, \cdots, p_\ell \in \mathscr{P}_{n,m} \right\}. \tag{C.108}$$

The multivariate generalization of the PMMSE is defined as follows.

Definition C.3 (Multivariate Polynomial MMSE)**.** Fix positive integer $\ell, m,$ and n. Fix an ℓ-dimensional random vector $\boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$ and an m-dimensional random vector $\boldsymbol{Y} \in L^{2n}(\mathbb{R}^m, P)$, and set $k = \binom{n+m}{m}$. We define the n-th order PMMSE for estimating $\boldsymbol{X}$ given $\boldsymbol{Y}$ by

$$\mathrm{pmmse}_n(\boldsymbol{X} \mid \boldsymbol{Y}) := \min_{\boldsymbol{C} \in \mathbb{R}^{\ell \times k}} \left\| \boldsymbol{X} - \boldsymbol{C}\boldsymbol{Y}^{(n,m)} \right\|_2^2, \tag{C.109}$$

and the n-th order PMMSE estimate of $\boldsymbol{X}$ given $\boldsymbol{Y}$ by

$$E_n[\boldsymbol{X} \mid \boldsymbol{Y}] := \boldsymbol{C}\boldsymbol{Y}^{(n,m)} \in \mathscr{P}_{n,m}^{\ell}(\boldsymbol{Y}) \tag{C.110}$$

for any minimizing matrix $\boldsymbol{C} \in \mathbb{R}^{\ell \times k}$ in (C.109).

Remark C.4. For any minimizer $\boldsymbol{C}$ in (C.109), the ℓ-dimensional random vector $\boldsymbol{C}\boldsymbol{Y}^{(n,m)}$ is the unique orthogonal projection of $\boldsymbol{X}$ onto $\mathscr{P}_{n,m}^{\ell}(\boldsymbol{Y})$; in particular, $E_n[\boldsymbol{X} \mid \boldsymbol{Y}]$ is well-defined by (C.110).

Denote, for $\boldsymbol{Y} \in L^{2n}(\mathbb{R}^m, P)$,

$$\boldsymbol{M}_{\boldsymbol{Y},n} := \mathbb{E}\left[\boldsymbol{Y}^{(n,m)} \left(\boldsymbol{Y}^{(n,m)} \right)^T \right]. \tag{C.111}$$

For $n \in \mathbb{N}$ and an ℓ-dimensional random vector $(X_1, \cdots, X_\ell)^T = \boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$, if $\boldsymbol{M}_{\boldsymbol{Y},n}$ is invertible, Lemma C.3 yields that

$$\begin{align} E_n[\boldsymbol{X} \mid \boldsymbol{Y}] &= \begin{pmatrix} E_n[X_1 \mid \boldsymbol{Y}] \\ \vdots \\ E_n[X_\ell \mid \boldsymbol{Y}] \end{pmatrix} \tag{C.112} \\ &= \begin{pmatrix} \mathbb{E}\left[X_1 \boldsymbol{Y}^{(n,m)} \right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \\ \vdots \\ \mathbb{E}\left[X_\ell \boldsymbol{Y}^{(n,m)} \right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \end{pmatrix}. \tag{C.113} \end{align}$$

We say that the Y_j do not satisfy a polynomial relation if the monomials $\prod_{j=1}^m Y_j^{\alpha_j}$, for $\alpha_1, \cdots, \alpha_m \in \mathbb{N}$, are linearly independent, i.e., if the mapping

$$\varphi : \bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m} \to \bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y}), \qquad \varphi(p) = p(\boldsymbol{Y}) \tag{C.114}$$

is an isomorphism of vector spaces.

Generalizing our results on random variables to random vectors can be done in view of the following polynomial denseness result.

Theorem C.2 ([Pet82])**.** *For any m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$ and $q > 1$, if we have the denseness $\overline{\bigcup_{n\in\mathbb{N}} \mathscr{P}_n(Y_j)} = L^q(P_{Y_j})$ for each $j \in \{1, \cdots, m\}$, then we have the denseness $\overline{\bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y})} = L^r(P_{\boldsymbol{Y}})$ for every $r \in [1, q)$.*

Since $\|\boldsymbol{Z}\|_r^r = \sum_j \|Z_j\|_r^r$, this inferred denseness in Theorem C.2 over $L^r(P_{\boldsymbol{Y}})$ may be extended to denseness over $L^r(\mathbb{R}^m, P_{\boldsymbol{Y}})$, i.e., we have the following immediate corollary.

Corollary C.1. *Fix an integer $m \geq 1$ and an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$. If each of the random variables $Y_1, \cdots, Y_m$ satisfies Carleman's condition, then the set of vectors of polynomials $\bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m}^m(\boldsymbol{Y})$ is dense in $L^q(\mathbb{R}^m, P_{\boldsymbol{Y}})$ for any $q \geq 1$.*

We deduce the following result on the convergence of the multivariate PMMSE to the MMSE.

Theorem C.3. *Fix an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$ and an ℓ-dimensional random vector $\boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$. If each Y_j satisfies Carleman's condition, and if the Y_j do not satisfy a polynomial relation, then we have the $L^2(\mathbb{R}^\ell, P_{\boldsymbol{Y}})$-limit*

$$\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}] = \lim_{n\to\infty} E_n[\boldsymbol{X} \mid \boldsymbol{Y}]. \tag{C.115}$$

Proof. Since the Y_j do not satisfy a polynomial relation, the matrix $\boldsymbol{M}_{\boldsymbol{Y},n}$ is invertible for each $n \in \mathbb{N}$. Further, the entries of $\boldsymbol{Y}^{(n,m)}$ are linearly independent for each n. Then, by Lemma C.3, equation (C.113) follows, i.e., the PMMSE estimate $E_n[\boldsymbol{X} \mid \boldsymbol{Y}]$ is the ℓ-dimensional random vector whose k-th entry is $\mathbb{E}\left[X_k \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)}$. By Corollary C.1, since each Y_j satisfies Carleman's condition, the set of vectors of polynomials $\bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m}^m(\boldsymbol{Y})$ is dense in $L^2(\mathbb{R}^m, P_{\boldsymbol{Y}})$. In particular, $\bigcup_{n\in\mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y})$ is dense in $L^2(P_{\boldsymbol{Y}})$. By Theorem C.1, we have the $L^2(P_{\boldsymbol{Y}})$ limits

$$\mathbb{E}[X_k \mid \boldsymbol{Y}] = \lim_{n\to\infty} \mathbb{E}\left[X_k \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \tag{C.116}$$

for each $k \in \{1, \cdots, \ell\}$. We conclude that $E_n[\boldsymbol{X} \mid \boldsymbol{Y}] \to \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}]$ in $L^2(\mathbb{R}^\ell, P_{\boldsymbol{Y}})$, as desired. □

The approach for showing the rationality of $t \mapsto \operatorname{pmmse}_n(X,t)$ for a random variable $X \in L^{2n}(P)$ in Theorem 4.2 may be generalized to deduce rationality of $t \mapsto \operatorname{pmmse}_n(\boldsymbol{X},t)$ for an m-dimensional random vector $\boldsymbol{X} \in L^{2n}(\mathbb{R}^m, P)$. Here, we are denoting $\operatorname{pmmse}_n(\boldsymbol{X},t) := \operatorname{pmmse}_n(\boldsymbol{X} \mid \sqrt{t}\boldsymbol{X} + \boldsymbol{N})$, where $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ is independent of $\boldsymbol{X}$. For brevity, we give a blueprint of how this generalization of rationality can be obtained. First, Lemma 4.6 may be generalized to yield that $\det \boldsymbol{M}_{\sqrt{t}\boldsymbol{X}+\boldsymbol{N}}$ is a polynomial in t of degree at most $d_{n,m}$ which is given by

$$\begin{aligned} d_{n,m} &:= \sum_{k\in[n]} k \cdot |\{(\lambda_1, \cdots, \lambda_m) \in \mathbb{N}^m \, ; \, \lambda_1 + \cdots + \lambda_m = k\}| && \text{(C.117)} \\ &= \sum_{k\in[n]} k \binom{k+m-1}{m-1} && \text{(C.118)} \\ &= \sum_{k\in[n]} m \binom{k+m-1}{m} = m\binom{n+m}{m+1}. && \text{(C.119)} \end{aligned}$$

Further, the coefficient of $t^{d_{n,m}}$ in $\det \boldsymbol{M}_{\sqrt{t}\boldsymbol{X}+\boldsymbol{N}}$ is $\det \boldsymbol{M}_{\boldsymbol{X}}$. Note that $d_{n,1} = d_n$. Then, the PMMSE expression in Theorem 4.2 may be generalized to give

$$\begin{aligned} &\operatorname{pmmse}_n(\boldsymbol{X},t) = \\ &\frac{(\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{X}}) \det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + (\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{N}}) \left(\det \boldsymbol{M}_{\boldsymbol{X},n}\right) \, t^{d_{n,m}-1}}{\det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + \left(\det \boldsymbol{M}_{\boldsymbol{X},n}\right) \, t^{d_{n,m}}}. \end{aligned} \tag{C.120}$$

To deduce (C.120), the multidimensional MMSE dimension result in Theorem 4.6 is used, as follows. Note that $\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{N}} = m$ for $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. By Theorem 4.6, we have that $\operatorname{mmse}(\boldsymbol{X},t) \sim m/t$. It is also true that $\operatorname{lmmse}(\boldsymbol{X},t) \sim m/t$. Therefore, $\operatorname{pmmse}_n(\boldsymbol{X},t) \sim m/t$ for every integer $n \geq 1$. Note that $\operatorname{pmmse}_n(\boldsymbol{X},0) = \operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{X}}$. Expression (C.120) follows via the same proof technique for Theorem 4.2.

With the definition of the multivariate PMMSE at hand, we show that the PMMSE estimate satisfies a tower property similar to the conditional expectation.

Proposition C.1 (Tower Property). *Fix $n \in \mathbb{N}$ and three random variables $X \in L^2(P)$ and $Y_1, Y_2 \in L^{2n}(P)$. Suppose that $|\operatorname{supp}(Y_1)|, |\operatorname{supp}(Y_2)| > n$. Then*

$$E_n\left[E_n[X \mid Y_1] \mid Y_1, Y_2\right] = E_n[X \mid Y_1], \tag{C.121}$$

and

$$E_n\left[E_n[X \mid Y_1, Y_2] \mid Y_2\right] = E_n[X \mid Y_2]. \tag{C.122}$$

Proof. Set $\boldsymbol{Y} = (Y_1, Y_2)^T$. Equation (C.121) is straightforward: since $E_n[X \mid Y_1] \in \mathscr{P}_n(Y_1) \subset \mathscr{P}_{n,2}(\boldsymbol{Y})$, the projection of $E_n[X \mid Y_1]$ onto $\mathscr{P}_{n,2}(\boldsymbol{Y})$ is $E_n[X \mid Y_1]$ again. Equation (C.122) also follows by an orthogonal projection argument. There is a unique representation $X = p_{1,2} + p_{1,2}^{\perp}$ for $(p_{1,2}, p_{1,2}^{\perp}) \in \mathscr{P}_{n,2}(\boldsymbol{Y}) \times \mathscr{P}_{n,2}(\boldsymbol{Y})^{\perp}$. There is also a unique representation $p_{1,2} = q_2 + q_2^{\perp}$ for $(q_2, q_2^{\perp}) \in \mathscr{P}_n(Y_2) \times \mathscr{P}_n(Y_2)^{\perp}$. The projection of X onto $\mathscr{P}_{n,2}(\boldsymbol{Y})$ is $p_{1,2}$, whose projection onto $\mathscr{P}_n(Y_2)$ is q_2, i.e.,

$$E_n\left[E_n[X \mid Y_1, Y_2] \mid Y_2\right] = q_2. \tag{C.123}$$

Furthermore, we have the representation $X = q_2 + (q_2^{\perp} + p_{1,2}^{\perp})$, for which $(q_2, q_2^{\perp} + p_{1,2}^{\perp}) \in \mathscr{P}_n(Y_2) \times \mathscr{P}_n(Y_2)^{\perp}$. Hence, the projection of X onto $\mathscr{P}_n(Y_2)$ is q_2 too, i.e.,

$$E_n[X \mid Y_2] = q_2. \tag{C.124}$$

From (C.123) and (C.124) we get (C.122). Equation (C.122) can also be deduced from the formula of $W := \mathbb{E}[X \mid \boldsymbol{Y}]$. Denote $\boldsymbol{Y}_2^{(n)} = (1, Y_2, \cdots, Y_2^n)^T$. We have that

$$W = \mathbb{E}\left[X\boldsymbol{Y}^{(n,2)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,2)} \tag{C.125}$$

and

$$E_n[W \mid Y_2] = \mathbb{E}\left[W\boldsymbol{Y}_2^{(n)}\right]^T \boldsymbol{M}_{Y_2,n}^{-1} \boldsymbol{Y}_2^{(n)}. \tag{C.126}$$

For $k \in [n]$, let $\delta(k) \in \left[\binom{n+2}{2} - 1\right]$ be the index of the entry in $\boldsymbol{Y}^{(n,2)}$ that equals Y_2^k. Then,

$$\mathbb{E}\left[Y_2^k \boldsymbol{Y}^{(n,2)}\right] = \boldsymbol{M}_{\boldsymbol{Y},n} \boldsymbol{e}_{\delta(k)}, \tag{C.127}$$

where $\boldsymbol{e}_0, \cdots, \boldsymbol{e}_{\binom{n+2}{2}-1}$ are the standard basis vectors of $\mathbb{R}^{\binom{n+2}{2}}$. Therefore, plugging (C.125) into (C.126), we obtain

$$E_n[W \mid Y_2] = \mathbb{E}\left[X\boldsymbol{Y}_2^{(n)}\right]^T \boldsymbol{M}_{Y_2,n}^{-1} \boldsymbol{Y}_2^{(n)}, \tag{C.128}$$

which is just $E_n[X \mid Y_2]$, as desired. □

C.5 Information Measures in Terms of Moments: Proofs of Section 4.4

C.5.1 Proof of Lemma 4.10

By finiteness of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, we get that $h(\boldsymbol{X})$ is well defined and less that ∞, but it could be $-\infty$. First, the case that $\det \boldsymbol{\Sigma}_{\boldsymbol{X}} = 0$ follows since both sides of (4.76) would then equal $-\infty$, which can be seen as follows. That $h(\boldsymbol{X}) = -\infty$ follows by a limiting argument starting from $0 \le D_{\mathrm{kl}}\left(P_{\boldsymbol{X}} \| \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}} + \varepsilon \boldsymbol{I}_m)\right)$, and inferring that $h(\boldsymbol{X}) \leq \frac{1}{2}\log\left((2\pi)^m \det\left(\boldsymbol{\Sigma}_{\boldsymbol{X}} + \varepsilon \boldsymbol{I}_m\right)\right) + \frac{1}{2}\text{rank}(\boldsymbol{\Sigma}_{\boldsymbol{X}})$ for all $\varepsilon > 0$, then taking $\varepsilon \to 0^+$. That the right-hand side of (4.76) equals $-\infty$ follows from $\text{mmse}(\boldsymbol{X},t) \leq \text{lmmse}(\boldsymbol{X},t)$ and $\text{lmmse}(\boldsymbol{X},t) \sim \frac{\text{rank}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}{t}$. So, we may assume $\det \boldsymbol{\Sigma}_{\boldsymbol{X}} \neq 0$.

In the same way that (4.74) is derived in [GSV05] (see Lemma 7 and Theorem 14 therein), one may obtain

$$h(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^m \det \boldsymbol{\Sigma}_{\boldsymbol{X}}\right) - \frac{1}{2}\lim_{\gamma\to\infty}\left[\log\left(\det\left(\gamma\boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{I}_m\right)\right) - \int_0^\gamma \text{mmse}(\boldsymbol{X},t)\,dt\right]. \tag{C.129}$$

Building on (C.129), we infer via the monotone convergence theorem that, with the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ denoted by $\lambda_1,\cdots,\lambda_m$,

$$h(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^m \prod_{i=1}^m \lambda_i\right) + \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{X},t) - \sum_{i=1}^m \frac{\lambda_i}{1+\lambda_i t}\,dt. \tag{C.130}$$

This equation yields the desired result $h(\boldsymbol{X}) = \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{X},t) - \frac{m}{2\pi e + t}\,dt$ since $\log\left((2\pi e)^m \prod_{i=1}^m \lambda_i\right) = \int_0^\infty \sum_{i=1}^m \frac{\lambda_i}{1+\lambda_i t} - \frac{m}{2\pi e+t}\,dt$, completing the proof of the lemma.

C.5.2 Proof of Theorem 4.4

We derive the multidimensional version of Theorem 4.4 here. Fix an m-dimensional random vector $\boldsymbol{V}$. We may assume $\det \boldsymbol{\Sigma}_{\boldsymbol{V}} \neq 0$, for otherwise the result follows immediately from $h_n(\boldsymbol{V}) = h(\boldsymbol{V}) = -\infty$ for all n. In view of monotonicity of $\text{pmmse}_n(\boldsymbol{V},t)$ in n, and since $h_1(\boldsymbol{V})$ is finite, it suffices by the monotone convergence theorem and the equation

$$h(\boldsymbol{V}) = \frac{1}{2}\int_0^\infty \text{mmse}(\boldsymbol{V},t) - \frac{m}{2\pi e + t}\,dt \tag{C.131}$$

to show that $\text{pmmse}_n(\boldsymbol{V},t) \to \text{mmse}(\boldsymbol{V},t)$ as $n \to \infty$. Let $\boldsymbol{N} \sim \mathcal{N}(0,\boldsymbol{I}_m)$ be independent of $\boldsymbol{V}$. A simple application of the triangle inequality yields that it suffices to prove the convergence

$$E_n\left[\boldsymbol{V} \mid \sqrt{t}\boldsymbol{V} + \boldsymbol{N}\right] \to \mathbb{E}\left[\boldsymbol{V} \mid \sqrt{t}\boldsymbol{V} + \boldsymbol{N}\right]. \tag{C.132}$$

We deduce (C.132) from Theorem C.3, as follows.

Denote $\boldsymbol{Z}^{(t)} := \sqrt{t}\boldsymbol{V} + \boldsymbol{N}$, and let $Z_j^{(t)}$ be the j-th entry of $\boldsymbol{Z}^{(t)}$. Fix $t \geq 0$. To apply Theorem C.3, we only need to show that the $Z_j^{(t)}$ do not satisfy a nontrivial polynomial relation. We show this by induction on m. The case $m = 1$ follows since $Z_1^{(t)}$ is continuous. Assume that we have shown that $Z_1^{(t)},\cdots,Z_{m-1}^{(t)}$ do not satisfy a nontrivial polynomial relation, and that $m \geq 2$. Suppose, for the sake of contradiction, that q is a polynomial in m variables such that $q(\mathbf{Z}^{(t)}) = 0$. Write $q(u_1, \cdots, u_m) = \sum_{k\in[d]} q_k(u_1, \cdots, u_{m-1})u_m^k$ for some polynomials q_k in $m-1$ variables such that $q_d \neq 0$. Squaring $q(\mathbf{Z}^{(t)}) = 0$ and taking the conditional expectation with respect to N_m we obtain

$$0 = \mathbb{E}\left[q\left(\mathbf{Z}^{(t)}\right)^2 \middle| N_m \right] = \sum_{k\in[2d]} \beta_k N_m^k \tag{C.133}$$

for some real constants β_k with the leading constant $\beta_{2d} := \|q_d(Z_1^{(t)}, \cdots, Z_{m-1}^{(t)})\|_2^2$. Since N_m is continuous, equation (C.133) cannot be a nontrivial polynomial relation for N_m. Thus, we must have $\beta_{2d} = 0$, i.e., $q_d(Z_1^{(t)}, \cdots, Z_{m-1}^{(t)}) = 0$. By the induction hypothesis, $q_d = 0$ identically, a contradiction. Therefore, no nontrivial polynomial relation $q(\mathbf{Z}^{(t)}) = 0$ can hold, and the inductive proof is complete. Finally, applying Theorem C.3, we deduce the limit in (C.132), thereby completing the proof of the theorem.

C.5.3 Proof of Theorem 4.5

Consider the first case, namely, X is discrete with finite support and Y is continuous whose MGF exists and for which $h(Y) > -\infty$. The existence of the MGF of Y implies the existence of the MGFs of $Y^{(x)}$ for each $x \in \text{supp}(X)$. Since $\sigma_Y^2 < \infty$, we have that $h(Y)$ is finite. In addition, for each $x \in \text{supp}(X)$, we infer from $\sigma_{Y^{(x)}}^2 < \infty$ the existence of the differential entropy $h(Y \mid X = x)$ and that $h(Y \mid X = x) < \infty$. If $\min_{x\in\text{supp}(X)} h(Y \mid X = x) > -\infty$, then $I(X;Y) = h(Y) - h(Y \mid X)$; this latter equation also holds if $h(Y \mid X = x) = -\infty$ for some $x \in \text{supp}(X)$. Therefore, Theorem 4.4 implies (4.23).

Now, consider the second case instead, so both X and Y are continuous random variables whose MGFs exist and that satisfy $h(X), h(Y) > -\infty$. We also assume that $I(X;Y) < \infty$ or else (X,Y) is not continuous. From these assumptions, we conclude that both $h(X)$ and $h(Y)$ are finite and $h(X,Y)$ exists. Thus, we obtain $I(X;Y) = h(X) + h(Y) - h(X,Y)$. By Theorem 4.4, we have that $h_n(X) \to h(X)$ and $h_n(Y) \to h(Y)$ as $n \to \infty$. Finally, note that the MGF of (X,Y) exists by the assumption that the MGFs of X and Y exist. Thus, by Theorem 4.4, we have that $h_n(X,Y) \to h(X,Y)$ too. The desired result (4.24) follows.

C.6 Estimator Implementation

We show in this appendix how to implement the proposed estimators numerically. Note that $\mathrm{pmmse}_n(X,t)$ contains roughly n^2 terms, and that numerically integrating this rational function can be done efficiently using built-in quadrature methods. Precomputing the function $t \mapsto \mathrm{pmmse}_{10}(X,t)$ takes a couple of minutes on a commercial laptop, whereas querying this rational function can be done in constant time. However, we need to develop the expressions of our approximations of differential entropy further to avoid possible issues that could arise from numerically computing the improper integral over $[0,\infty)$. To illustrate this issue, consider the expression for $h_2(X)$. For convenience, define the function $\delta_{X,n} : (0,\infty) \to [0,\infty)$ by

$$\delta_{X,n}(t) := \det \boldsymbol{M}_{\sqrt{t}X+N,n} \tag{C.134}$$

for a $2n$-times integrable random variable X. Recall that $\delta_{X,n}$ is the denominator of $\mathrm{pmmse}_n(X,\cdot)$. Recall from (4.10) that a zero-mean unit-variance random variable X satisfies

$$\mathrm{pmmse}_2(X,t) = \frac{2+4t+(\mathcal{X}_4-\mathcal{X}_3^2-1)t^2}{2+6t+(\mathcal{X}_4+3)t^2+(\mathcal{X}_4-\mathcal{X}_3^2-1)t^3}. \tag{C.135}$$

For example, when $X \sim \mathrm{Unif}([-\sqrt{3},\sqrt{3}])$, so

$$(\mathcal{X}_1,\mathcal{X}_2,\mathcal{X}_3,\mathcal{X}_4) = \left(0,1,0,\frac{9}{5}\right), \tag{C.136}$$

we obtain

$$\mathrm{pmmse}_2(X,t) = \frac{5+10t+2t^2}{5+15t+12t^2+2t^3}. \tag{C.137}$$

Now, consider the expression for $h_2(X)$ in (4.79), namely,

$$h_2(X) = \frac{1}{2}\int_0^\infty \frac{5+10t+2t^2}{5+15t+12t^2+2t^3} - \frac{1}{2\pi e+t}\,dt. \tag{C.138}$$

The integral in (C.138) converges, but a numerical computation might not be able to capture this convergence as the expression for the integrand is a difference of non-integrable functions that both decay as $1/t$. To avoid this possible issue, we subtract a $1/t$ term from both of these non-integrable functions. More precisely, denoting differentiation with respect to t by a prime, we write

$$\begin{aligned}\mathrm{pmmse}_2(X,t) &= \frac{5+10t+2t^2-\frac{1}{3}\delta_{X,2}'(t)+\frac{1}{3}\delta_{X,2}'(t)}{\delta_{X,2}(t)} \\ &= \frac{2t}{5+15t+12t^2+2t^3} + \frac{1}{3}\frac{d}{dt}\log\delta_{X,2}(t)\end{aligned}$$

and

$$\frac{1}{2\pi e + t} = \frac{d}{dt} \log(2\pi e + t). \tag{C.139}$$

The integrand $\text{pmmse}_2(X, t) - 1/(2\pi e + t)$ now becomes

$$\frac{2t}{5 + 15t + 12t^2 + 2t^3} + \frac{d}{dt} \log \frac{\delta_{X,2}(t)^{1/3}}{2\pi e + t}. \tag{C.140}$$

The advantage in having the integrand in this form is that the first term is well-behaved (it decays as $1/t^2$), and the second term's integral can be given in closed form

$$\int_0^\infty \left(\log \frac{\delta_{X,2}(t)^{1/3}}{2\pi e + t} \right)' dt = \log \left(2\pi e \left(\frac{2}{5} \right)^{1/3} \right). \tag{C.141}$$

Therefore, equation (C.138) becomes

$$h_2(X) = \frac{1}{2} \log \frac{2\pi e}{(5/2)^{1/3}} + \int_0^\infty \frac{t}{5 + 15t + 12t^2 + 2t^3} \, dt. \tag{C.142}$$

We use equation (C.142) instead of (C.138) for numerical computation. Note that this resolves the same numerical instability issue when estimating from data: if $\mathcal{S} = \{X_j\}_{j=1}^m$ is a multiset of i.i.d. samples distributed according to P_X, and if $U \sim \text{Unif}(\mathcal{S})$, we compute the estimate $\widehat{h}_2(\mathcal{S}) = h_2(U)$ of $h_2(X)$ via an expression analogous to that in (C.142) where X is replaced with U.

The procedure of obtaining expression (C.142) from (C.138) can be carried out for a general X and n such that $\mathbb{E}[X^{2n}] < \infty$ and $|\text{supp}(X)| > n$, as follows. Let $\theta_{X,n} : [0, \infty) \to [0, \infty)$ be the polynomial that is the numerator of $\text{pmmse}_n(X, t)$, i.e., $\theta_{X,n}(t) := \delta_{X,n}(t) \cdot \text{pmmse}_n(X, t)$. Thus, we have that

$$\text{pmmse}_n(X, t) = \frac{\theta_{X,n}(t)}{\delta_{X,n}(t)}. \tag{C.143}$$

We define the function $\rho_{X,n} : [0, \infty) \to \mathbb{R}$ by

$$\rho_{X,n}(t) := \frac{\theta_{X,n}(t) - d_n^{-1} \delta'_{X,n}(t)}{2\delta_{X,n}(t)}, \tag{C.144}$$

where $d_n = \binom{n+1}{2}$. By the analysis of the coefficients in $\text{pmmse}_n(X, t)$ proved in Theorem 4.2, we have that $\rho_{X,n}(0) = 0$ and

$$\rho_{X,n}(t) = O\left(t^{-2}\right) \tag{C.145}$$

as $t \to \infty$. In particular, $\rho_{X,n}$ is integrable over $[0, \infty)$. The following formula for differential entropy directly follows from the definition of h_n in (4.79).

Lemma C.4. *For any random variable X satisfying $\mathbb{E}[X^{2n}] < \infty$ and $|\operatorname{supp}(X)| > n$, we have the formula*

$$h_n(X) = \frac{1}{2} \log \left(2\pi e \left(\frac{\det \boldsymbol{M}_{X,n}}{\det \boldsymbol{M}_{N,n}} \right)^{1/d_n} \right) + \int_0^\infty \rho_{X,n}(t)\, dt, \tag{C.146}$$

where $d_n = \binom{n+1}{2}$, $N \sim \mathcal{N}(0,1)$, and $\rho_{X,n}$ is as defined in (C.144).

A similar conclusion holds for mutual information.

Lemma C.5. *Fix a discrete random variable X with finite support, and a $2n$-times integrable continuous random variable Y. We have that*

$$I_n(X;Y) = \frac{1}{n(n+1)} \log \frac{\det \boldsymbol{M}_{Y,n}}{\prod_{x \in \operatorname{supp}(X)} \left(\det \boldsymbol{M}_{Y^{(x)},n} \right)^{P_X(x)}} + \int_0^\infty \rho_{Y,n}(t) - \mathbb{E}_X \left[\rho_{Y^{(X)},n}(t) \right] \, dt, \tag{C.147}$$

where for each $x \in \operatorname{supp}(X)$ we denote by $Y^{(x)}$ the random variable Y conditioned on $\{X = x\}$.

Note that in Lemmas C.4 and C.5, the determinants $\det \boldsymbol{M}_{A,n}$ and the rational functions $\rho_n(A;t)$, for $A \in \{X, Y\}$ or $A \in \{Y^{(x)}\ ;\ x \in \operatorname{supp}(X)\}$, are completely determined by the first $2n$ moments of A. To obtain the estimates $\widehat{h}_n$ and $\widehat{I}_n$ given samples, the moments of A are replaced with their respective sample moments in formulas (C.146) and (C.147).

C.7 Proofs of Subsection 4.5.1: Consistency

C.7.1 Proof of Theorem 4.9: Consistency of the Differential Entropy Estimator

We use the formula for h_n given in Lemma C.4,

$$h_n(X) = \frac{1}{2} \log \left(2\pi e \left(\frac{\det \boldsymbol{M}_{X,n}}{\det \boldsymbol{M}_{N,n}} \right)^{1/d_n} \right) + \int_0^\infty \rho_{X,n}(t)\, dt, \tag{C.148}$$

where $d_n = \binom{n+1}{2}$ and $N \sim \mathcal{N}(0,1)$. We may assume that N is independent of X and the X_j. For each $m \in \mathbb{N}$, let $\mathcal{S}_m := \{X_j\}_{j \in [m]}$, and consider the sequence $\{U_m \sim \operatorname{Unif}(\mathcal{S}_m)\}_{m \in \mathbb{N}}$. For each $m \in \mathbb{N}$, let $\mathfrak{E}_m$ be the event that $X_0, \cdots, X_m$ are distinct, and let $\mathfrak{E}$ be the event that the X_j, for $j \in \mathbb{N}$, are all distinct. Whenever $m \geq n$ and $\mathfrak{E}_m$ occurs, we have by Definition 4.4 of $\widehat{h}_n$ and formula (C.148) for h_n the following estimate

$$\widehat{h}_n\left(\mathcal{S}_m\right) = \frac{1}{2} \log \left(2\pi e \left(\frac{\det \boldsymbol{M}_{U_m,n}}{\det \boldsymbol{M}_{N,n}} \right)^{1/d_n} \right) + \int_0^\infty \rho_{U_m,n}(t)\, dt. \tag{C.149}$$

Since X is continuous, we have that $P(\mathfrak{E}_m) = 1$ for every $m \in \mathbb{N}$. Further, $\mathfrak{E}_0 \supset \mathfrak{E}_1 \supset \cdots$ and $\mathfrak{E} = \bigcap_{m\in\mathbb{N}} \mathfrak{E}_m$, hence $P(\mathfrak{E}) = 1$. Therefore, for the purpose of proving the almost-sure limit $\widehat{h}_n\left(\mathcal{S}_m\right) \to h_n(X)$, we may assume that $\mathfrak{E}$ occurs. We first treat convergence of the integral part. We show that the integral part is a continuous function of the moments, then the continuous mapping theorem yields that

$$\int_0^\infty \rho_{U_m,n}(t)\, dt \to \int_0^\infty \rho_{X,n}(t)\, dt \tag{C.150}$$

almost surely as $m \to \infty$ because sample moments converge almost surely to the moments. A similar method is then applied to the convergence of the $\log\det \boldsymbol{M}_{X,n}$ part.

We fix $n \in \mathbb{N}_{\geq 1}$, and assume $m \geq n$ throughout the proof. We use the following notation. The $2n$-dimensional random vector $\boldsymbol{\mu}^{(m)}$ consists of the first $2n$ moments of U_m

$$\boldsymbol{\mu}^{(m)} := \left(\frac{\sum_{j=0}^m X_j}{m+1}, \cdots, \frac{\sum_{j=0}^m X_j^{2n}}{m+1} \right)^T. \tag{C.151}$$

Let $\mu_k^{(m)}$ be the k-th coordinate of $\boldsymbol{\mu}^{(m)}$, so $\boldsymbol{\mu}^{(m)} = \left(\mu_1^{(m)}, \cdots, \mu_{2n}^{(m)}\right)^T$. We write $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$ for $k \in \mathbb{N}$, and consider the constant vector

$$\boldsymbol{\mathcal{X}} := (\mathcal{X}_k)_{1\leq k\leq 2n}. \tag{C.152}$$

By the strong law of large numbers, we have the almost-sure convergence $\mu_k^{(m)} \to \mathcal{X}_k$ for each $1 \leq k \leq 2n$. Then, $\boldsymbol{\mu}^{(m)} \to \boldsymbol{\mathcal{X}}$ almost surely as $m \to \infty$. We show next that the function $\boldsymbol{\mathcal{X}} \mapsto \int_0^\infty \rho_{X,n}(t)\, dt$ is continuous.

By definition of $\rho_{X,n}$ (see (C.144)), there are polynomials $A_1, \cdots, A_{d_n-2}$ and $B_1, \cdots, B_{d_n}$ in $2n$ variables such that

$$\rho_{X,n}(t) = \frac{\sum_{j=1}^{d_n-2} A_j(\boldsymbol{\mathcal{X}})\ t^j}{c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}})\ t^j} \tag{C.153}$$

where $c_n := \prod_{k=1}^n k!$ (we are subsuming the $1/2$ factor in (C.144) in the numerator, so we have the equality $\delta_{X,n}(t) = c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}}) t^j$). Being polynomials, each of the A_j and the B_ℓ is continuous over $\mathbb{R}^{2n}$. Then, by the continuous mapping theorem, we have the almost-sure convergences

$$A_j\left(\boldsymbol{\mu}^{(m)}\right) \to A_j(\boldsymbol{\mathcal{X}}) \quad \text{and} \quad B_\ell\left(\boldsymbol{\mu}^{(m)}\right) \to B_\ell(\boldsymbol{\mathcal{X}}) \tag{C.154}$$

as $m \to \infty$ for each $1 \leq j \leq d_n - 2$ and $1 \leq \ell \leq d_n$. Denote

$$\boldsymbol{A}(\boldsymbol{\mathcal{X}}) := \left(A_j(\boldsymbol{\mathcal{X}})\right)_{1\leq j\leq d_n-2}, \tag{C.155}$$

$$\boldsymbol{B}(\boldsymbol{\mathcal{X}}) := \left(B_j(\boldsymbol{\mathcal{X}})\right)_{1\le j\le d_n}. \tag{C.156}$$

We show next that the there is an open set $\mathcal{O} \subset \mathbb{R}^{d_n}$ containing the point $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$ such that the mapping $f : \mathbb{R}^{d_n-2} \times \mathcal{O} \to \mathbb{R}$ defined by

$$f(p_1, \cdots, p_{d_n-2}, q_1, \cdots, q_{d_n}) := \int_0^\infty \frac{\sum_{j=1}^{d_n-2} p_j t^j}{c_n + \sum_{j=1}^{d_n} q_j t^j}\, dt \tag{C.157}$$

is continuous at the point $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$. To this end, we shall show first that the mapping in (C.157) is well-defined on an open neighborhood of $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$. In other words, the denominator of the integrand $t \mapsto c_n + \sum_{j=1}^{d_n} q_j t^j$ cannot have a root $t \in [0, \infty)$ for any $\boldsymbol{q} \in \mathcal{O}$, and the rational function integrand has to be integrable. For integrability, we will restrict the set $\mathcal{O}$ to contain only points having $q_{d_n} > 0$, so showing that the integrand's denominator is strictly positive over $t \in [0, \infty)$ will be enough to deduce integrability in (C.157).

We consider the subset $\mathcal{G} \subset \mathbb{R}^{d_n}$ defined by

$$\mathcal{G} := \left\{ \boldsymbol{g} \in \mathbb{R}^{d_n} ; \, g_{d_n} > 0, \sum_{\ell=1}^{d_n} g_j t^j > -c_n \text{ for all } t \ge 0 \right\} \tag{C.158}$$

where in this definition and the subsequent argument we set $\boldsymbol{g} = (g_1, \cdots, g_{d_n})^T$. Note that $\boldsymbol{B}(\boldsymbol{\mathcal{X}}) \in \mathcal{G}$. Indeed, since X is continuous, $B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X,n} > 0$; similarly, for every $t \in [0, \infty)$, continuity of $\sqrt{t}X + N$ implies that $\det \boldsymbol{M}_{\sqrt{t}X+N} > 0$ (recall that $c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}})t^j = \det \boldsymbol{M}_{\sqrt{t}X+N}$). We show that $\mathcal{G}$ is an open set. Fix $\boldsymbol{g} \in \mathcal{G}$ and $\varepsilon_1 \in (0, g_{d_n})$. We have that the polynomial $\sum_{j=1}^{d_n} (g_j - \varepsilon_1) t^j$ is eventually increasing and approaches infinity as $t \to \infty$. Let $t_0 > 1$ be such that for every $t > t_0$ we have

$$\sum_{\ell=1}^{d_n} (g_j - \varepsilon_1) t^j > -c_n. \tag{C.159}$$

Being continuous, the polynomial $\sum_{j=1}^{d_n} g_j t^j$ attains its minimum over the compact set $[0, t_0]$. Let s denote this minimum, and note that $s > -c_n$. Let $\varepsilon \in (0,1)$ be defined by

$$\varepsilon := \frac{1}{2} \min\left(\varepsilon_1, \frac{(s + c_n)(t_0 - 1)}{t_0(t_0^{d_n} - 1)} \right). \tag{C.160}$$

As $\varepsilon < \varepsilon_1$, inequality (C.159) yields that for every $t > t_0$

$$\sum_{j=1}^{d_n} (g_j - \varepsilon) t^j > -c_n. \tag{C.161}$$

In addition, for any $t \in [0, t_0]$,

$$\begin{aligned}\sum_{j=1}^{d_n}(g_j - \varepsilon)t^j &= \sum_{j=1}^{d_n} g_j t^j - \varepsilon \sum_{j=1}^{d_n} t^j \geq s - \varepsilon \sum_{j=1}^{d_n} t_0^j \\ &> s - \frac{(s + c_n)(t_0 - 1)}{t_0(t_0^{d_n} - 1)} \sum_{j=1}^{d_n} t_0^j \\ &= s - (s + c_n) = -c_n. \end{aligned} \tag{C.162}$$

Thus, combining (C.161) and (C.162) we obtain

$$\sum_{j=1}^{d_n}(g_j - \varepsilon)t^j > -c_n \tag{C.163}$$

for every $t \in [0, \infty)$. Hence, for any $(\delta_j)_{1 \leq j \leq d_n} =: \boldsymbol{\delta} \in \mathbb{R}^{d_n}$ such that $\|\boldsymbol{\delta}\|_2 < \varepsilon$, we have that for all $t \in [0, \infty)$

$$\sum_{j=1}^{d_n}(g_j - \delta_j)t^j \geq \sum_{j=1}^{d_n}(g_j - \|\boldsymbol{\delta}\|_2)t^j \geq \sum_{j=1}^{d_n}(g_j - \varepsilon)t^j > -c_n. \tag{C.164}$$

In other words, the open ball $\{\boldsymbol{q} \in \mathbb{R}^{d_n} ; \|\boldsymbol{q} - \boldsymbol{g}\| < \varepsilon\}$ lies within $\mathcal{G}$. This completes the proof that $\mathcal{G}$ is open. Then, the function f given by (C.157) is well-defined on the open set $\mathbb{R}^{d_n - 2} \times \mathcal{G}$. We will replace $\mathcal{G}$ with an open box $\mathcal{O} \subset \mathcal{G}$ to simplify the notation for the proof of continuity of f.

By openness of $\mathcal{G}$, there is an $\eta_1 \in (0, B_{d_n}(\boldsymbol{\mathcal{X}}))$ such that the open box

$$\mathcal{O}_1 := \prod_{j=1}^{d_n} \left(B_j(\boldsymbol{\mathcal{X}}) - \eta_1, B_j(\boldsymbol{\mathcal{X}}) + \eta_1\right) \subset \mathcal{G} \tag{C.165}$$

contains $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$. Since $\mathcal{O}_1 \subset \mathcal{G}$, we have by the definition of $\mathcal{G}$ in (C.158) that for any $\boldsymbol{g} \in \mathcal{O}_1$ the lower bound

$$c_n + \sum_{\ell=1}^{d_n} g_\ell t^\ell > 0 \tag{C.166}$$

holds for every $t \geq 0$. In particular, with $\eta := \eta_1/2$, the set

$$\mathcal{O} := \prod_{j=1}^{d_n} \left(B_j(\boldsymbol{\mathcal{X}}) - \eta, B_j(\boldsymbol{\mathcal{X}}) + \eta\right) \subset \mathcal{O}_1 \subset \mathcal{G} \tag{C.167}$$

is an open set containing $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$, and the point $(B_j(\boldsymbol{\mathcal{X}}) - \eta)_{1 \leq j \leq 2n}$ lies inside $\mathcal{G}$. Then, the function $f : \mathbb{R}^{d_n - 2} \times \mathcal{O} \to \mathbb{R}$ given by (C.157) is well-defined, and for any $\boldsymbol{g} \in \mathcal{O}$ we have the lower bound (over $t \in [0, \infty)$)

$$c_n + \sum_{\ell=1}^{d_n} g_\ell t^\ell \geq c_n + \sum_{\ell=1}^{d_n}(B_\ell(\boldsymbol{\mathcal{X}}) - \eta)t^\ell > 0. \tag{C.168}$$

From (C.168), Lebesgue's dominated convergence shows continuity of f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$, as follows.

Let $\boldsymbol{w} := (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{d_n - 2} \times \mathcal{O}$ be such that $\|\boldsymbol{w}\|_2 < \eta$. The integrand in f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) - (\boldsymbol{u}, \boldsymbol{v})$ may be bounded as

$$\left| \frac{\sum_{j=1}^{d_n - 2} (A_j(\boldsymbol{\mathcal{X}}) - u_j) t^j}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - v_\ell) t^\ell} \right| = \frac{\left| \sum_{j=1}^{d_n - 2} (A_j(\boldsymbol{\mathcal{X}}) - u_j) t^j \right|}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - v_\ell) t^\ell} \tag{C.169}$$

$$\leq \frac{\sum_{j=1}^{d_n - 2} (|A_j(\boldsymbol{\mathcal{X}})| + \eta) t^j}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - \eta) t^\ell}. \tag{C.170}$$

The bound in (C.170) is uniform in $\boldsymbol{w}$, and the upper bound is integrable over $[0, \infty)$ as the denominator's degree exceeds that of the numerator by at least 2 and the denominator is strictly positive by (C.168). Hence, by Lebesgue's dominated convergence

$$\lim_{\|\boldsymbol{w}\| \to 0} f\left((\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) - \boldsymbol{w} \right) = f\left(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}) \right), \tag{C.171}$$

i.e., f is continuous at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$, as desired. Denote

$$\boldsymbol{A}^{(m)} := \left(A_j(\boldsymbol{\mu}^{(m)}) \right)_{1 \leq j \leq d_n - 2}, \tag{C.172}$$

$$\boldsymbol{B}^{(m)} := \left(B_\ell(\boldsymbol{\mu}^{(m)}) \right)_{1 \leq \ell \leq d_n}. \tag{C.173}$$

We have the formulas

$$f(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) = \int_0^\infty \rho_{U_m, n}(t)\, dt \tag{C.174}$$

and

$$f(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) = \int_0^\infty \rho_{X, n}(t)\, dt. \tag{C.175}$$

Since $(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) \to (\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$ almost surely, continuity of f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$ implies by the continuous mapping theorem that

$$f(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) \to f(\boldsymbol{A}(\boldsymbol{\nu}), \boldsymbol{B}(\boldsymbol{\nu})) \tag{C.176}$$

almost surely as $m \to \infty$, i.e., (C.150) holds.

Now, for the convergence of the logarithmic part, recall that we have the almost sure convergence

$$\det \boldsymbol{M}_{U_m, n} = B_{d_n}(\boldsymbol{\mu}^{(m)}) \to B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X, n} \tag{C.177}$$

as $m \to \infty$. As the mapping $\mathbb{R}_{>0} \to \mathbb{R}$ defined by $q \mapsto \log q$ is continuous, the continuous mapping theorem yields that

$$\log \det \boldsymbol{M}_{U_m, n} \to \log \det \boldsymbol{M}_{X, n} \tag{C.178}$$

almost surely as $m \to \infty$. Combining (C.176) and (C.178), we obtain that

$$\widehat{h}_n\left(\mathcal{S}_m\right) \to h_n(X) \tag{C.179}$$

almost surely as $m \to \infty$. Finally, (4.95) follows from (C.179) by Theorem 4.4.

C.7.2 Proof of Corollary 4.4: Consistency of the Mutual Information Estimator

Denote $\mathcal{S}_m = \{(X_j, Y_j)\}_{j\in[m]}$, and consider the empirical measure

$$\widehat{P}_m(x) := \sum_{j\in[m]} \frac{\delta_x(X_j)}{m+1}. \tag{C.180}$$

Let $\mathfrak{D}_m$ be the event that for each $x \in \operatorname{supp}(X)$ there is a subset of indices $J_x \subset [m]$ of size at least $n+1$ such that: **i)** $X_j = x$ for each $j \in J_x$, and **ii)** the Y_j, for $j \in J_x$, are distinct. If $\mathfrak{D}_m$ occurs, then we may write

$$\widehat{I}_n(\mathcal{S}_m) = \widehat{h}_n(\mathcal{A}_m) - \sum_{x\in\operatorname{supp}(X)} \widehat{P}_m(x)\, \widehat{h}_n(\mathcal{B}_{m,x}), \tag{C.181}$$

where $\mathcal{A}_m := \{Y_j\}_{j\in[m]}$ and $\mathcal{B}_{m,x} := \{Y_j\ ;\ j \in [m], X_j = x\}$. By the assumption of continuity of Y, it holds with probability 1 that the Y_j, for $j \in \mathbb{N}$, are all distinct. In addition, we have that $P_X(x) > 0$ for each $x \in \operatorname{supp}(X)$. Therefore, $P(\mathfrak{D}_m) \to 1$ as $m \to \infty$. Note that $\mathfrak{D}_0 \subset \mathfrak{D}_1 \subset \cdots$.

Let $\mathfrak{C}$ be the event that $\lim_{m\to\infty} \widehat{h}_n(\mathcal{A}_m) = h_n(Y)$ and, for each $x \in \operatorname{supp}(X)$, $\lim_{m\to\infty} \widehat{h}_n(\mathcal{B}_{m,x}) = h_n(Y^{(x)})$. By Theorem 4.9 and finiteness of $\operatorname{supp}(X)$, for each integer $m' \geq (n+1)|\operatorname{supp}(X)|$, we have that $P(\mathfrak{C} \mid \mathfrak{D}_{m'}) = 1$. Let $\mathfrak{F}$ be the event that the empirical measure $\widehat{P}_m$ converges to P_X, i.e., that for each $x \in \operatorname{supp}(X)$ the limit $\widehat{P}_m(x) \to P_X(x)$ holds as $m \to \infty$. By the strong law of large numbers, $P(\mathfrak{F}) = 1$. Therefore,

$$\begin{aligned} P\left(\lim_{m\to\infty} \widehat{I}_n(\mathcal{S}_m) = I_n(X;Y)\right) &\geq P(\mathfrak{C} \cap \mathfrak{F} \cap \mathfrak{D}_{m'}) \\ &\geq P(\mathfrak{F}) + P(\mathfrak{C} \cap \mathfrak{D}_{m'}) - 1 \\ &= P(\mathfrak{D}_{m'}). \end{aligned} \tag{C.182}$$

Taking $m' \to \infty$, we deduce that $\widehat{I}_n(\mathcal{S}_m) \to I_n(X;Y)$ almost surely.

C.8 Proofs of Subsection 4.5.2: Sample Complexity

C.8.1 Proof of Proposition 4.4: Differential Entropy

Suppose $\text{supp}(X) \subset [p,q] \subset (0,\infty)$, and write $\mathcal{S} = \{X_j\}_{j=1}^m$; note that we may assume, without loss of generality, that X is strictly positive because h_n is shift-invariant. We use the same notation in Appendix C.7. In particular, $\mathcal{X}_k = \mathbb{E}[X^k]$, and $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \cdots, \mathcal{X}_{2n})^T$. Let $U \sim \text{Unif}(\mathcal{S})$. Let $\mathfrak{E}_m$ be the event that $X_1, \cdots, X_m$ are distinct. From (C.148)–(C.149), if $m > n$ and $\mathfrak{E}_m$ holds, then we have that

$$\widehat{h}_n(\mathcal{S}) - h_n(X) = \frac{1}{2d_n} \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} + \int_0^\infty \rho_{U,n}(t) \, - \rho_{X,n}(t) \, dt. \tag{C.183}$$

By the assumption of continuity of X, we have that $P(\mathfrak{E}_m) = 1$ for every m. Therefore, for the purpose of proving a sample complexity bound, we may assume that $m > n$ and that $\mathfrak{E}_m$ occurs.

We will consider the determinant part and the integral part in (C.183) separately, but the proof technique will be the same. Let A_j and B_ℓ be the polynomials as defined by (C.153) in Appendix C.7, so

$$\rho_{X,n}(t) = \frac{\sum_{j=1}^{d_n-2} A_j(\boldsymbol{\mathcal{X}}) \ t^j}{c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}}) \ t^j} \tag{C.184}$$

where $c_n := \prod_{k=1}^n j!$. We split each of the polynomials A_j and B_ℓ into a positive part and a negative part. More precisely, we collect the terms in A_j that have positive coefficients into a polynomial $A_j^{(+)}$, and the terms in A_j with negative coefficients into a polynomial $-A_j^{(-)}$ (so $A_j^{(-)}$ has positive coefficients, and $A_j = A_j^{(+)} - A_j^{(-)}$). Define $B_\ell^{(+)}$ and $B_\ell^{(-)}$ from B_ℓ similarly. By positivity of X, each moment $\mathcal{X}_k$ is (strictly) positive. Then, we may write

$$\rho_{X,n}(t) = \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \tag{C.185}$$

with the polynomials in t

$$f_X(t) := \sum_{j=1}^{d_n-2} A_j^{(+)}(\boldsymbol{\mathcal{X}}) t^j \tag{C.186}$$

$$g_X(t) := \sum_{j=1}^{d_n-2} A_j^{(-)}(\boldsymbol{\mathcal{X}}) t^j \tag{C.187}$$

$$u_X(t) := c_n + \sum_{\ell=1}^{d_n} B_\ell^{(+)}(\boldsymbol{\mathcal{X}}) t^\ell \tag{C.188}$$

$$v_X(t) := \sum_{\ell=1}^{d_n} B_\ell^{(-)}(\boldsymbol{\mathcal{X}}) t^\ell, \tag{C.189}$$

having all non-negative coefficients. We note that we have suppressed the dependence on n in the notation used for these polynomials for readability. For $q \in \{f, g, u, v\}$, let q_U be the random variable whose value is what is obtained via q_X when the moments of X are replaced with the sample moments obtained from the samples $\mathcal{S}$, e.g.,

$$f_U(t) := \sum_{j=1}^{d_n-2} A_j^{(+)} \left(\frac{\sum_{i=1}^m X_i}{m}, \cdots, \frac{\sum_{i=1}^m X_i^{2n}}{m} \right) t^j. \tag{C.190}$$

Note that $u_U(t) - v_U(t) = \det \boldsymbol{M}_{\sqrt{t}U+N,n} > 0$, where $N \sim \mathcal{N}(0,1)$ is independent of $X, X_1, \cdots, X_m$. Then the function

$$\rho_{U,n}(t) = \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} \tag{C.191}$$

is well-defined over $t \in [0, \infty)$. By the homogeneity properties proved in Theorem 4.2, we know that the total degree of A_j is at most $2j+2$, and the total degree of B_ℓ is at most 2ℓ. Therefore, for any $\eta \in (0,1)$ and $\boldsymbol{\xi} \in \mathbb{R}_{\geq 0}^{2n}$, we have the inequalities

$$(1-\eta)^{2j+2} A_j^{(\pm)}(\boldsymbol{\xi}) \leq A_j^{(\pm)}((1-\eta)\boldsymbol{\xi}) \tag{C.192}$$

$$A_j^{(\pm)}((1+\eta)\boldsymbol{\xi}) \leq (1+\eta)^{2j+2} A_j^{(\pm)}(\boldsymbol{\xi}) \tag{C.193}$$

$$(1-\eta)^{2\ell} B_\ell^{(\pm)}(\boldsymbol{\xi}) \leq B_\ell^{(\pm)}((1-\eta)\boldsymbol{\xi}) \tag{C.194}$$

$$B_\ell^{(\pm)}((1+\eta)\boldsymbol{\xi}) \leq (1+\eta)^{2\ell} B_\ell^{(\pm)}(\boldsymbol{\xi}) \tag{C.195}$$

for every $1 \leq j \leq d_n - 2$ and $1 \leq \ell \leq d_n$.

For each $\eta \in (0,1)$, we denote the event

$$\mathfrak{A}_{n,\eta}(\mathcal{S}) := \left\{ 1 - \eta \leq \frac{\sum_{i=1}^m X_i^k}{m \mathcal{X}_k} \leq 1 + \eta \text{ for all } k \in [2n] \right\}, \tag{C.196}$$

Hoeffding's inequality yields that, for any $z > 0$ and $1 \leq k \leq 2n$,

$$P\left(\left| \mathcal{X}_k - \frac{1}{m} \sum_{i=1}^m X_i^k \right| \geq z \right) \leq 2e^{-2mz^2/\left(q^k - p^k\right)^2}. \tag{C.197}$$

Setting $z = \eta \mathcal{X}_k \geq \eta p^k > 0$ for $\eta \in (0,1)$ yields that

$$P\left((1-\eta)\mathcal{X}_k < \frac{1}{m} \sum_{i=1}^m X_i^k < (1+\eta)\mathcal{X}_k \right) \geq 1 - 2e^{-2m\eta^2/\left((q/p)^k - 1\right)^2}. \tag{C.198}$$

Therefore, the union bound yields that

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \geq 1 - 4ne^{-2m\eta^2/\left((q/p)^{2n} - 1\right)^2}. \tag{C.199}$$

If $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs, we show a bound on the estimation error that is linear in η

$$\widehat{h}_n(\mathcal{S}) - h_n(X) = O_{X,n}(\eta), \tag{C.200}$$

independent of the number of samples m, for all small enough η. Then, we choose η to be linear in the error ε to conclude the proof.

We may bound $\rho_{U,n}(t)$ (see (C.191)) via the bounds in (C.192)–(C.195) under the assumption that $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs. If $(1-\eta)\mathcal{X}_k \le \frac{1}{m}\sum_{i=1}^m X_i^m \le (1+\eta)\mathcal{X}_k$ holds for every $1 \le k \le 2n$, then by (C.192)–(C.195) we have that for every $t \ge 0$ and $\eta \in (0,1)$

$$\frac{(1-\eta)^2 f_X((1-\eta)^2 t) - (1+\eta)^2 g_X((1+\eta)^2 t)}{u_X((1+\eta)^2 t) - v_X((1-\eta)^2 t)} \le \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} = \rho_{U,n}(t). \tag{C.201}$$

For an analogous upper bound, we first verify the positivity

$$u_X((1-\eta)^2 t) - v_X((1+\eta)^2 t) > 0 \tag{C.202}$$

for every small enough η. Let

$$\mu_X := \sup_{t\in[0,\infty)} \frac{v_X(t)}{u_X(t)}. \tag{C.203}$$

We show that $\mu_X < 1$. We have the limit

$$\xi_X := \lim_{t\to\infty} \frac{v_X(t)}{u_X(t)} = \frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}. \tag{C.204}$$

Recall that $B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}) = B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X,n} > 0$ and both $B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})$ and $B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})$ are non-negative, hence $B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) > 0$. Then, $\xi_X < 1$. Thus, there is a $t_0 \ge 0$ such that $v_X(t)/u_X(t) < (1+\xi_X)/2 < 1$ whenever $t > t_0$. Further, by the extreme value theorem, there is a $t_1 \in [0,t_0]$ such that $v_X(t)/u_X(t) \le v_X(t_1)/u_X(t_1) < 1$ for every $t \in [0,t_0]$. Therefore, $\mu_X \le \max((1+\xi_X)/2, v_X(t_1)/u_X(t_1)) < 1$, as desired. Note that if $\mu_X = 0$ then $v_X \equiv 0$ identically, in which case (C.202) trivially holds by positivity of u_X. So, for the purpose of showing (C.202), it suffices to consider the case $\mu_X \in (0,1)$. Denote

$$\nu := \left(\frac{1+\eta}{1-\eta}\right)^2. \tag{C.205}$$

Now, since v_X is a polynomial of degree at most d_n, we have that $v_X(\alpha\tau) \le \alpha^{d_n} v_X(\tau)$ for every $\alpha \ge 1$ and $\tau \ge 0$. Therefore, for every $1 \le \nu < \mu_X^{-1/d_n}$ and $t \ge 0$, we have that

$$\frac{v_X((1+\eta)^2 t)}{u_X((1-\eta)^2 t)} \le \left(\frac{1+\eta}{1-\eta}\right)^{2d_n} \cdot \frac{v_X((1-\eta)^2 t)}{u_X((1-\eta)^2 t)} \tag{C.206}$$

$$\leq \nu^{d_n}\mu_X < 1, \tag{C.207}$$

i.e., inequality (C.202) holds. Therefore, for every $1 \leq \nu < \mu_X^{-1/d_n}$ (if $\mu_X = 0$, we allow $1 \leq \nu < \infty$), inequalities (C.192)–(C.195) imply the bound

$$\rho_{U,n}(t) = \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} \tag{C.208}$$

$$\leq \frac{(1+\eta)^2 f_X((1+\eta)^2 t) - (1-\eta)^2 g_X((1-\eta)^2 t)}{u_X((1-\eta)^2 t) - v_X((1+\eta)^2 t)}. \tag{C.209}$$

Combining (C.201) and (C.209), then integrating with respect to t over $[0,\infty)$ and performing a change of variables from t to $(1-\eta)^2 t$, we obtain the bounds

$$\int_0^\infty \frac{f_X(t) - \nu g_X(\nu t)}{u_X(\nu t) - v_X(t)}\, dt \leq \int_0^\infty \rho_{U,n}(t)\, dt \tag{C.210}$$

$$\leq \int_0^\infty \frac{\nu f_X(\nu t) - g_X(t)}{u_X(t) - v_X(\nu t)}\, dt. \tag{C.211}$$

Next, we further develop these bounds. For any $s \in (0,1)$, denote

$$\nu_{X,n,s} := \left(\frac{1 - s\mu_X}{1-s}\right)^{1/d_n}. \tag{C.212}$$

Consider the functions

$$\varphi_X(t;\nu) := \frac{u_X(t) - v_X(t)}{u_X(t) - v_X(\nu t)}, \tag{C.213}$$

$$\psi_X(t;\nu) := \frac{u_X(t) - v_X(t)}{u_X(\nu t) - v_X(t)}. \tag{C.214}$$

We show in Appendix C.8.2 that, for any constants $s \in (0, (1-\mu_X)/(1+\mu_X))$ and $1 \leq \nu \leq \nu_{X,n,s}$, the uniform bounds

$$1 - s \leq \psi_X(t;\nu) \leq 1 \leq \varphi_X(t;\nu) \leq 1 + s \tag{C.215}$$

hold over $t \in [0,\infty)$. Fix $s \in (0, (1-\mu_X)/(1+\mu_X))$ and $1 \leq \nu \leq \nu_{X,n,s}$.

Now, the integrand in the upper bound in (C.211) can be rewritten as

$$\frac{\nu f_X(\nu t) - g_X(t)}{u_X(t) - v_X(\nu t)} = \varphi_X(t;\nu)\left(\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + \frac{\nu f_X(\nu t) - f_X(t)}{u_X(t) - v_X(t)}\right). \tag{C.216}$$

The integrand in the lower bound in (C.210) can be rewritten as

$$\frac{f_X(t) - \nu g_X(\nu t)}{u_X(\nu t) - v_X(t)} = \psi_X(t;\nu)\left(\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + \frac{g_X(t) - \nu g_X(\nu t)}{u_X(t) - v_X(t)}\right). \tag{C.217}$$

By the bounds in (C.215), we have that for every $t \geq 0$

$$0 \leq \varphi_X(t;\nu) - 1 \leq s. \tag{C.218}$$

Hence, by non-negativity of f_X and g_X, we deduce

$$(\varphi_X(t;\nu) - 1) \cdot \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \leq s \cdot \frac{f_X(t)}{u_X(t) - v_X(t)}, \tag{C.219}$$

i.e., the inequality

$$\varphi_X(t;\nu) \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \leq \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + s \frac{f_X(t)}{u_X(t) - v_X(t)} \tag{C.220}$$

hold of all $t \geq 0$. In addition, since $f_X(\nu t) \leq \nu^{d_n - 2} f_X(t)$ over $t \in [0, \infty)$, inequality (C.218) implies that

$$\varphi_X(t;\nu) \cdot \frac{\nu f_X(\nu t) - f_X(t)}{u_X(t) - v_X(t)} \leq \frac{(1+s)(\nu^{d_n - 1} - 1) f_X(t)}{u_X(t) - v_X(t)}. \tag{C.221}$$

Therefore, applying inequalities (C.220) and (C.221) in formula (C.216), we deduce in view of the upper bound in (C.211) the inequality

$$\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\, dt \leq \left((1+s)\nu^{d_n - 1} - 1 \right) \int_0^\infty \frac{f_X(t)}{u_X(t) - v_X(t)}\, dt. \tag{C.222}$$

Similarly, we derive a lower bound on (C.217). By (C.215), we have that for every $t \geq 0$

$$s \geq 1 - \psi_X(t;\nu) \geq 0. \tag{C.223}$$

Hence, by non-negativity of f_X and g_X,

$$s \cdot \frac{f_X(t)}{u_X(t) - v_X(t)} \geq (1 - \psi_X(t;\nu)) \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)}, \tag{C.224}$$

i.e., the inequality

$$\psi_X(t;\nu) \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \geq \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} - s \frac{f_X(t)}{u_X(t) - v_X(t)} \tag{C.225}$$

holds for all $t \geq 0$. In addition, from $\psi_X(t;\nu) \leq 1 \leq \nu$ and $g_X(\nu t) \leq \nu^{d_n - 2} g_X(t)$ for $t \geq 0$, we deduce

$$\begin{aligned} \psi_X(t;\nu) \cdot \frac{g_X(t) - \nu g_X(\nu t)}{u_X(t) - v_X(t)} &\geq \psi_X(t;\nu) \cdot \frac{(1 - \nu^{d_n - 1}) g_X(t)}{u_X(t) - v_X(t)} \\ &\geq \left(1 - \nu^{d_n - 1}\right) \frac{g_X(t)}{u_X(t) - v_X(t)}. \end{aligned} \tag{C.226}$$

Therefore, applying inequalities (C.225) and (C.226) in formula (C.217), the lower bound in (C.210) yields the bound

$$\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt \geq -s\int_0^\infty \frac{f_X(t)}{u_X(t)-v_X(t)}\,dt - \left(\nu^{d_n-1}-1\right)\int_0^\infty \frac{g_X(t)}{u_X(t)-v_X(t)}\,dt. \tag{C.227}$$

In particular, (C.227) implies that

$$\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt \geq -\left(\nu^{d_n-1}-(1-s)\right)\int_0^\infty \frac{f_X(t)+g_X(t)}{u_X(t)-v_X(t)}\,dt. \tag{C.228}$$

Now, note that $(1+s)\nu^{d_n-1}-1 \geq \nu^{d_n-1}-(1-s)$. Therefore, combining the upper bound in (C.222) and the lower bound in (C.228), we deduce that

$$\left|\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt\right| \leq \left((1+s)\nu^{d_n-1}-1\right)\int_0^\infty \frac{f_X(t)+g_X(t)}{u_X(t)-v_X(t)}\,dt. \tag{C.229}$$

The upper bound in (C.229) may be made as small as needed by choosing a small s then choosing a small ν.

The second part of the proof, given in Appendix C.8.3, derives the following error bound for estimating $\log\det \boldsymbol{M}_{X,n}$ from samples. If $B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}) > 0$, we denote

$$\tau_{X,n} := \left(\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})/B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})+1}{2}\right)^{1/(n+1)} \in (1,\infty) \tag{C.230}$$

and

$$\eta_{X,n} := \min\left(\frac{1}{2}, \frac{\tau_{X,n}-1}{\tau_{X,n}+1}\right) \in (0,1/2]. \tag{C.231}$$

If $B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}) = 0$, then we set $\tau_{X,n} = \infty$ and $\eta_{X,n} = 1/2$. We show that for all $\eta \in (0,\eta_{X,n})$, if $\mathfrak{A}_{n,\eta}(\mathcal{S})$ holds, then we have the bound

$$\left|\frac{1}{2d_n}\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| \leq \frac{6\eta}{n}\cdot\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})+B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})-B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}. \tag{C.232}$$

To finish the proof, we choose η so that the desired accuracy is achieved with high probability. Recall from (C.199) that

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \geq 1-4ne^{-m\eta^2\alpha_{X,n}} \tag{C.233}$$

where we denote the constant

$$\alpha_{X,n} := 2\cdot\left(\left(\frac{q}{p}\right)^{2n}-1\right)^{-2}. \tag{C.234}$$

In addition, from (C.229) and (C.232), we know that if $s \in (0,(1-\mu_X)/(1+\mu_X))$, $\nu \in [1,\nu_{X,n,s}]$,

$\eta \in (0, \eta_{X,n})$, and $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs, then

$$\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \le \eta \cdot \beta_{X,n} + \left((1+s)\nu^{d_n-1} - 1\right) \cdot \gamma_{X,n} \tag{C.235}$$

where we denote the constants

$$\beta_{X,n} := \frac{6}{n} \cdot \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}, \tag{C.236}$$

$$\gamma_{X,n} := \int_0^\infty \frac{f_X(t) + g_X(t)}{u_X(t) - v_X(t)}\, dt. \tag{C.237}$$

Consider the constant $\varepsilon_{X,n} \in (0, 2\min(\gamma_{X,n}, \beta_{X,n})]$ defined by

$$\varepsilon_{X,n} := \min\left(2\gamma_{X,n} \cdot \frac{1-\mu_X}{1+\mu_X},\ 2\beta_{X,n}\right). \tag{C.238}$$

Fix $\varepsilon \in (0, \varepsilon_{X,n})$, set $s := \varepsilon/(6\gamma_{X,n}) \in (0, 1/3]$, denote

$$\kappa_{X,n} := \min\left(3, \tau_{X,n}, \left(\frac{1-s\mu_X}{1-s}\right)^{\frac{1}{2d_n}}, \frac{1+\varepsilon/(2\beta_{X,n})}{1-\varepsilon/(2\beta_{X,n})}\right), \tag{C.239}$$

and fix $\eta \in (0, (\kappa_{X,n}-1)/(\kappa_{X,n}+1))$. Since $\kappa_{X,n} \le 3$, we obtain $\eta < 1/2$. In addition, $\kappa_{X,n} \le \tau_{X,n}$, hence $\eta < (\kappa_{X,n}-1)/(\kappa_{X,n}+1)$ implies that $\eta < \eta_{X,n}$. Note that, for $a \in (0,1)$ and $b > 1$, the inequality $a \le (b-1)/(b+1)$ is equivalent to $(1+a)/(1-a) \le b$. By definition,

$$\kappa_{X,n} \le \left(\frac{1-s\mu_X}{1-s}\right)^{1/(2d_n)}, \tag{C.240}$$

hence we have

$$(1+s)\nu^{d_n} = (1+s)\left(\frac{1+\eta}{1-\eta}\right)^{2d} < (1+s)\kappa_{X,n}^{2d} \tag{C.241}$$

$$\le (1+s) \cdot \frac{1-s\mu_X}{1-s} \le \frac{1+s}{1-s} \tag{C.242}$$

$$\le \frac{1+s+s(1-3s)}{1-s} = 1+3s. \tag{C.243}$$

In addition, since

$$\kappa_{X,n} \le \frac{1+\varepsilon/(2\beta_{X,n})}{1-\varepsilon/(2\beta_{X,n})}, \tag{C.244}$$

and since we assume $\eta < (\kappa_{X,n}-1)/(\kappa_{X,n}+1)$, we deduce the inequality $\eta < \varepsilon/(2\beta_{X,n})$. Applying the two inequalities $\eta < \varepsilon/(2\beta_{X,n})$ and $(1+s)\nu^{d_n} \le 1+3s$ (see (C.243)) into inequality (C.235), we conclude that

$$\begin{aligned}\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| &\leq \eta \cdot \beta_{X,n} + \Big((1+s)\nu^{d_n-1} - 1\Big) \cdot \gamma_{X,n} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned} \tag{C.245}$$

whenever $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs.

Now, fix $\delta \in (0, 1/(4n))$. Set

$$\eta := \frac{1}{2} \cdot \frac{\kappa_{X,n} - 1}{\kappa_{X,n} + 1}. \tag{C.246}$$

We show that $\eta \geq \varepsilon c_{X,n}$, where we define the constant $c_{X,n}$ by

$$c_{X,n} := \min\left(\frac{1}{8\gamma_{X,n}}, \frac{\tau_{X,n} - 1}{4\gamma_{X,n}(\tau_{X,n}+1)}, \frac{1-\mu_X}{72\gamma_{X,n}d_n}, \frac{1}{4\beta_{X,n}}\right). \tag{C.247}$$

In this definition of $c_{X,n}$, the term involving $\tau_{X,n}$ is removed if $\tau_{X,n} = \infty$. We assume that

$$m \geq \frac{2/(c_{X,n}^2 \alpha_{X,n})}{\varepsilon^2} \log \frac{1}{\delta}. \tag{C.248}$$

From $\eta \geq \varepsilon c_{X,n}$ and (C.248), it follows that the probability that the event $\mathfrak{A}_{n,\eta}(\mathcal{S})$ does not occur is bounded as

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})^c\right) \leq 4ne^{-m\eta^2\alpha_{X,n}} \leq \delta. \tag{C.249}$$

Note that this would conclude the proof, as then we would have that

$$\begin{aligned} P\left(\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \leq \varepsilon\right) &\geq P\left(\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \leq \varepsilon \;\middle|\; \mathfrak{A}_{n,\eta}(\mathcal{S})\right) P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \\ &= P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) > 1 - \delta. \end{aligned} \tag{C.250}$$

The rest of the proof is devoted to showing that $\eta \geq \varepsilon c_{X,n}$ holds.

Let $\rho = (1-\mu_X)/(6d_n)$. We will show that

$$\left(\frac{1 - s\mu_X}{1-s}\right)^{1/(2d_n)} \geq \frac{1+s\rho}{1-s\rho}. \tag{C.251}$$

Inequality (C.251) is equivalent to

$$(1 - s\mu_X)(1-s\rho)^{2d_n} \geq (1+\rho s)^{2d_n}(1-s). \tag{C.252}$$

By Bernoulli's inequality, since $0 \leq s\rho \leq 1$, we have that $(1-s\rho)^{2d_n} \geq 1 - 2d_n\rho s$. In addition, the inequality $1 + 2az \geq e^{az} \geq (1+a)^z$ for $a, z \geq 0$ satisfying $az \leq \log 2$ implies, in view of

$2d_n\rho s \leq 1/9 < \log 2$, that

$$1 + 4d_n\rho s \geq (1+\rho s)^{2d_n}. \tag{C.253}$$

Therefore, to show (C.252), it suffices to show that

$$(1 - s\mu_X)(1 - 2d_n\rho s) \geq (1 + 4d_n\rho s)(1 - s). \tag{C.254}$$

Now, using the definition $\rho = (1-\mu_X)/(6d_n)$, inequality (C.254) follows as

$$\begin{aligned}(1 - s\mu_X)(1 - 2d_n\rho s) &= (1 - s\mu_X)(1 - s(1-\mu_X)/3) \\ &= (1 + 2(1-\mu_X)s/3)(1-s) + s^2(1-\mu_X)(\mu_X+2)/3 \\ &\geq (1 + 2(1-\mu_X)s/3)(1-s) = (1+4d_n\rho s)(1-s).\end{aligned}$$

Since (C.254) holds, we conclude that inequality (C.251) holds.

Now, by the definition of $\kappa_{X,n}$ in (C.239) there are four possible values $\kappa_{X,n}$ can take. First, if $\kappa_{X,n} = 3$, then

$$\eta = \frac{1}{4} = \varepsilon \cdot \frac{1}{4\varepsilon} \geq \varepsilon \cdot \frac{1}{8\gamma_{X,n}} \geq \varepsilon c_{X,n} \tag{C.255}$$

since $\varepsilon < \varepsilon_{X,n} \leq 2\gamma_{X,n}$. Now, if $\kappa_{X,n} = \tau_{X,n}$ (so $B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}) > 0$), then

$$\eta = \frac{1}{2} \cdot \frac{\tau_{X,n} - 1}{\tau_{X,n} + 1} \geq \frac{\varepsilon}{4\gamma_{X,n}} \cdot \frac{\tau_{X,n} - 1}{\tau_{X,n} + 1} \tag{C.256}$$

since $\varepsilon < 2\gamma_{X,n}$. Next, suppose that

$$\kappa_{X,n} = \left(\frac{1 - s\mu_X}{1 - s}\right)^{1/(2d_n)}. \tag{C.257}$$

By (C.251) and (C.257), we deduce that

$$\kappa_{X,n} \geq \frac{1 + s\rho}{1 - s\rho}. \tag{C.258}$$

Recall that, for $0 < a < 1 < b$, the inequalities $(1+a)/(1-a) \leq b$ and $(b-1)/(b+1) \geq a$ are equivalent. Therefore, the definition of η in (C.246) yields from (C.258) that $\eta \geq s\rho/2$. Plugging in the definitions of s and ρ, we conclude that

$$\eta \geq \varepsilon \cdot \frac{1 - \mu_X}{72\gamma_{X,n} d_n} \geq \varepsilon c_{X,n}. \tag{C.259}$$

Finally, when

$$\kappa_{X,n} = \frac{1 + \varepsilon/(2\beta_{X,n})}{1 - \varepsilon/(2\beta_{X,n})}, \tag{C.260}$$

the definition of η implies that $\eta \geq \varepsilon/(4\beta_{X,n}) \geq \varepsilon c_{X,n}$. Combining these four cases, we conclude that we must have $\eta \geq \varepsilon c_{X,n}$ independently of the value of $\kappa_{X,n}$. The proof is thus complete.

C.8.2 Proof of the Uniform Bounds (C.215)

Being polynomials of degree at most d_n with non-negative coefficients, the functions u_X and v_X satisfy $u_X(\nu t) \leq \nu^{d_n} u_X(t)$ and $v_X(\nu t) \leq \nu^{d_n} v_X(t)$ for every $\nu \geq 1$ and $t \geq 0$. Note also that both u_X and v_X are nondecreasing. In addition, we have $v_X(t) < u_X(t)$ for every $t \geq 0$, because $u_X(t) - v_X(t) = \det \boldsymbol{M}_{\sqrt{t}X+N,n} > 0$. We have also shown that $\mu_X < 1$, where μ_X is defined in (C.203) as

$$\mu_X := \sup_{t\in[0,\infty)} \frac{v_X(t)}{u_X(t)}. \tag{C.261}$$

These facts will be enough to deduce the bounds in (C.215).

We show first the bounds on φ_X in (C.215). It suffices to consider the case $\mu_X > 0$, for otherwise v_X vanishes identically and $\varphi_X \equiv 1$ identically. We show that for every $s > 0$ and $1 \leq \nu \leq \nu'_{X,n,s}$, where $\nu'_{X,n,s} := ((1/s + 1/\mu_X)/(1/s+1))^{1/d_n}$, the uniform bound $1 \leq \varphi_X(t;\nu) \leq 1+s$ in (C.215) holds.

Consider the lower bound on φ_X. For every $1 \leq \nu < \mu_X^{-1/d_n}$, we have the uniform bound

$$\frac{v_X(\nu t)}{u_X(t)} \leq \frac{\nu^{d_n} v_X(t)}{u_X(t)} \leq \nu^{d_n}\mu_X < 1 \tag{C.262}$$

over $t \in [0,\infty)$. In particular,

$$u_X(t) - v_X(\nu t) > 0 \tag{C.263}$$

for every $1 \leq \nu < \mu_X^{-1/d_n}$ and $t \geq 0$. Since v_X is nondecreasing, we conclude that $\varphi_X(t;\nu) = (u_X(t) - v_X(t))/(u_X(t) - v_X(\nu t)) \geq 1$ whenever $1 \leq \nu < \mu_X^{-1/d_n}$. Note that $\nu'_{X,n,s} < \mu_X^{-1/d_n}$ for every $s > 0$ since $\mu_X \in (0,1)$.

Next, we show the upper bound on φ_X. Fix $s > 0$ and $\nu \in [1, \nu'_{X,n,s}]$. Since $v_X(t)/\mu_X \leq u_X(t)$, we have for every $t \geq 0$ the bound

$$v_X(\nu t) \leq \nu^{d_n} v_X(t) \leq \frac{1/s + 1/\mu_X}{1/s+1} \cdot v_X(t) \tag{C.264}$$

$$\leq \frac{v_X(t)/s + u_X(t)}{1/s+1} = v_X(t) + \frac{u_X(t) - v_X(t)}{1/s+1}. \tag{C.265}$$

Rearranging (C.265), we obtain the bound

$$\frac{-1}{1/s+1} \le \frac{v_X(t) - v_X(\nu t)}{u_X(t) - v_X(t)}. \tag{C.266}$$

Adding 1 to both sides of (C.266) then inverting, we obtain $\varphi_X(t;\nu) \le 1+s$; for this step, we used the fact that $u_X(t) - v_X(\nu t) > 0$, which follows by (C.263) since $\nu \le \nu'_{X,n,s} < \mu_X^{-1/d_n}$.

Next, we prove the bounds on ψ_X in (C.215). We do not assume $\mu_X > 0$. The upper bound $\psi_X(t;\nu) \le 1$ follows for every $\nu \ge 1$ by monotonicity of u_X. For the lower bound on ψ_X, we show that for every $s \in (0,1)$ and $1 \le \nu \le \nu_{X,n,s}$, where $\nu_{X,n,s} := ((1-s\mu_X)/(1-s))^{1/d_n}$, the uniform bound $\psi_X(t;\nu) \ge 1-s$ holds over $t \in [0,\infty)$. We have, for every $s \in (0,1)$ and $\nu \in [1, \nu_{X,n,s}]$, the bound

$$u_X(\nu t) \le \nu^{d_n} u_X(t) \le \frac{1-s\mu_X}{1-s} \cdot u_X(t) \tag{C.267}$$

$$\le \frac{u_X(t) - s v_X(t)}{1-s} = \frac{u_X(t) - v_X(t)}{1-s} + v_X(t) \tag{C.268}$$

over $t \in [0,\infty)$. Rearranging (C.268), we obtain $\psi_X(t;\nu) \ge 1-s$, as desired.

Finally, note that $\nu_{X,n,s} \le \nu'_{X,n,s}$ is equivalent to $s \le (1-\mu_X)/(1+\mu_X)$. This concludes the proof that, for every $s \in (0,(1-\mu_X)/(1+\mu_X))$ and $\nu \in [1,\nu_{X,n,s}]$, the uniform bounds in (C.215)

$$1 - s \le \psi_X(t;\nu) \le 1 \le \varphi_X(t;\nu) \le 1+s \tag{C.269}$$

hold over $t \in [0,\infty)$.

C.8.3 Proof of Inequality (C.232)

Recall that

$$\det \boldsymbol{M}_{X,n} = B_{d_n}(\boldsymbol{\mathcal{X}}) = B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}). \tag{C.270}$$

We bound the error when estimating $\log\det \boldsymbol{M}_{X,n}$ from the samples $\mathcal{S}$. Denote the random vector $\boldsymbol{\mu} := \left(\frac{\sum_{i=1}^m X_i}{m}, \cdots, \frac{\sum_{i=1}^m X_i^{2n}}{m} \right)$, and note that

$$\det \boldsymbol{M}_{U,n} = B_{d_n}(\boldsymbol{\mu}) = B_{d_n}^{(+)}(\boldsymbol{\mu}) - B_{d_n}^{(-)}(\boldsymbol{\mu}). \tag{C.271}$$

We assume that $m > n$. Let $\eta_{X,n}$ be as defined by (C.230) and (C.231), and fix $\eta \in (0, \eta_{X,n})$. Then we show that under $\mathfrak{A}_{n,\eta}(\mathcal{S})$

$$\left| \frac{1}{2d_n} \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \right| \leq \frac{6\eta}{n} \cdot \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}. \tag{C.272}$$

By (C.68) in the proof of Theorem 4.2, each term in the polynomials $B_{d_n}^{(\pm)}$ is a product of at most $n+1$ monomials. Thus,

$$(1-\eta)^{n+1} B_{d_n}^{(\pm)}(\boldsymbol{\mathcal{X}}) \leq B_{d_n}^{(\pm)}(\boldsymbol{\mu}) \leq (1+\eta)^{n+1} B_{d_n}^{(\pm)}(\boldsymbol{\mathcal{X}}). \tag{C.273}$$

It suffices to consider the case when $B_{d_n}^{(-)}$ is not the zero polynomial, for if $B_{d_n}^{(-)}$ is the zero polynomial then we obtain from (C.273) the bound

$$\left| \frac{1}{2d_n} \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \right| = \frac{1}{2d_n} \left| \log \frac{B_{d_n}^{(+)}(\boldsymbol{\mu})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})} \right| \tag{C.274}$$

$$\leq \frac{\max\left(\pm \log(1 \pm \eta)\right)}{n} \tag{C.275}$$

$$= \frac{-\log(1-\eta)}{n} < \frac{2\eta}{n} \tag{C.276}$$

where the last inequality follow because $-\log(1-z) < 2z$ for $z \in (0, 1/2)$, which can be verified by checking the derivative. Note that the bound $2\eta/n$ in (C.276) is stronger than the bound in (C.272). Assume that $B_{d_n}^{(-)}$ does not vanish identically, so positivity of X yields that $B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}) > 0$.

From (C.273), we have that

$$\log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + (n+1)\log(1-\eta) \leq \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \tag{C.277}$$

and

$$\log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \leq \log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + (n+1)\log(1+\eta) \tag{C.278}$$

where we used our assumption that

$$\nu^{\frac{n+1}{2}} = \left(1 + \frac{2}{1/\eta - 1}\right)^{n+1} < \frac{1}{2}\left(\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + 1\right) \tag{C.279}$$

$$< \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}. \tag{C.280}$$

Now, for every $(w, z, r) \in \mathbb{R}^3$ such that $w > z > 0$ and $w/z > r > 1$, rearranging $r + 1/r > 2$ we have that

$$\frac{w - z/r}{w - z} < \frac{w - z}{w - rz}. \tag{C.281}$$

Setting $(w, z, r) = (B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}), B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}), \nu^{(n+1)/2})$, we obtain that

$$1 < \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \tag{C.282}$$

$$< \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}. \tag{C.283}$$

Therefore,

$$0 < \log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \tag{C.284}$$

$$< \left| \log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \right|. \tag{C.285}$$

Applying (C.284)–(C.285) in (C.278) and combining that with (C.277), we obtain (since $\log(1+\eta) < -\log(1-\eta)$) the bound

$$\left| \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \right| \leq \log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + (n+1) \log \frac{1}{1-\eta}. \tag{C.286}$$

Now, we may write

$$\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} = \left(1 - \frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \left(\nu^{\frac{n+1}{2}} - 1 \right) \right)^{-1}. \tag{C.287}$$

The proof of (C.272) (or, (C.232)) is completed by showing that for $(w, z, r) \in \mathbb{R}^3_{>0}$ such that $(1+z)^r < 1 + \frac{1}{2w}$ we have

$$-\log\left(1 - w\left((1+z)^r - 1\right)\right) \leq (2w+1) rz. \tag{C.288}$$

Before showing that (C.288) holds, we note how it completes the proof. Setting

$$(w, z, r) = \left(\frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}, \frac{2\eta}{1-\eta}, n+1 \right), \tag{C.289}$$

we obtain that

$$\log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \leq \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \cdot (n+1) \cdot \frac{2\eta}{1-\eta} \tag{C.290}$$

since (see (C.279))

$$\nu^{\frac{n+1}{2}} < \frac{1}{2} \left(\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + 1 \right). \tag{C.291}$$

Then $-\log(1-\eta) < 2\eta$ yields from (C.286) and (C.290) that

$$\frac{1}{2d_n} \left| \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \right| \leq \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \cdot \frac{2\eta}{n(1-\eta)} + \frac{2\eta}{n}. \tag{C.292}$$

Then (C.292) yields the desired inequality (C.232) as $\eta \in (0, 1/2)$.

Finally, to see that (C.288) holds, we consider for fixed $w, r > 0$

$$f(z) := (2w+1)rz + \log\left(1 - w\left((1+z)^r - 1\right)\right) \tag{C.293}$$

over $0 \leq z < (1+1/(2w))^{1/r} - 1$. Inequality (C.288) is restated as $f(z) \geq 0$ for every $0 < z < (1+1/(2w))^{1/r} - 1$, which follows since f is continuous, $f(0) = 0$, $f'(0^+) = (w+1)r > 0$, and

$$\begin{aligned} f'(z) &= (2w+1)r - \frac{wr(1+z)^{r-1}}{1 - w((1+z)^r - 1)} && \text{(C.294)} \\ &> (2w+1)r - \frac{wr(1+z)^r}{1 - w((1+z)^r - 1)} && \text{(C.295)} \\ &> (2w+1)r - \frac{wr(1+1/(2w))}{1 - w((1+1/(2w)) - 1)} = 0 && \text{(C.296)} \end{aligned}$$

for every $0 \leq z < (1+1/(2w))^{1/r} - 1$.

C.8.4 Proof of Proposition 4.5: Mutual Information

Let $\{(X_j, Y_j)\}_{j \in \mathbb{N}}$ be i.i.d. samples drawn according to $P_{X,Y}$. Denote $\mathcal{S}_m = \{X_j\}_{j=1}^m$. By continuity of Y, we may assume that all the Y_j, for $j \in \mathbb{N}$, are distinct. For each $x \in \text{supp}(X)$, let $J_x := \{1 \leq j \leq m \,;\, X_j = x\}$. Let $\mathfrak{D}_m$ be the event that, for every $x \in \text{supp}(X)$, we have that $|J_x| > n$. We use Hoeffding's inequality to obtain a lower bound on the probability

$$P(\mathfrak{D}_m) = P\left(\min_{x \in \text{supp}(X)} |J_x| > n \right). \tag{C.297}$$

Let $\widehat{P}_m$ be the empirical measure, i.e., define $\widehat{P}_m(x) := m^{-1} \sum_{j=1}^m \delta_x(X_j)$. Note that $|J_x| = m\widehat{P}_m(x)$.

Let $x_0 \in \mathrm{supp}(X)$ be such that $P_X(x_0)$ is minimal, set $\zeta := P_X(x_0)/2$, and suppose $m \geq \zeta^{-1}n$. Then, the union bound and $\zeta \leq P_X(x) - \zeta$ for each $x \in \mathrm{supp}(X)$ yield that

$$P\left(n \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right) \leq P\left(m\zeta \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right) \tag{C.298}$$

$$\leq \sum_{x\in\mathrm{supp}(X)} P\left(m\zeta \geq |J_x|\right) \tag{C.299}$$

$$\leq \sum_{x\in\mathrm{supp}(X)} P\left(m(P_X(x)-\zeta) \geq |J_x|\right) \tag{C.300}$$

$$= \sum_{x\in\mathrm{supp}(X)} P\left(P_X(x) - \widehat{P}_m(x) \geq \zeta\right). \tag{C.301}$$

Since $\mathbb{E}[\widehat{P}_m(x)] = P_X(x)$ for each $x \in \mathrm{supp}(X)$, Hoeffding's inequality yields that $P\left(P_X(x) - \widehat{P}_m(x) \geq \zeta\right) \leq e^{-2\zeta^2 m}$. Therefore,

$$P\left(n \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right) \leq |\mathrm{supp}(X)| \cdot e^{-2\zeta^2 m}. \tag{C.302}$$

In other words, for every $m \geq 2n/P_X(x_0)$, we have the bound

$$P(\mathfrak{D}_m) \geq 1 - |\mathrm{supp}(X)| \cdot e^{-mP_X(x_0)^2/2}. \tag{C.303}$$

Denote $\pi_X := 4/P_X(x_0)^2$ and

$$\delta_{X,n} := \min\left(\frac{1}{4|\mathrm{supp}(X)|}, e^{-P_X(x_0)n/2}\right). \tag{C.304}$$

We conclude from (C.303) that, for every $\delta \in (0, \delta_{X,n})$, if $m \geq \pi_X \log(1/\delta)$ then $P(\mathfrak{D}_m) > 1 - \delta/4$.

Consider the event $\mathfrak{P}_{m,\varepsilon}$ that the empirical measure $\widehat{P}_m$ is pointwise ε-close to the true measure P_X, i.e.,

$$\mathfrak{P}_{m,\varepsilon} := \left\{ \max_{x\in\mathrm{supp}(X)} \left|\widehat{P}_m(x) - P_X(x)\right| < \varepsilon \right\}. \tag{C.305}$$

By the union bound, we have that

$$P\left(\mathfrak{P}_{m,\varepsilon}^c\right) \leq \sum_{x\in\mathrm{supp}(X)} P\left(\left|\widehat{P}_m(x) - P_X(x)\right| \geq \varepsilon\right). \tag{C.306}$$

By Hoeffding's inequality, for each $x \in \mathrm{supp}(X)$, we have that

$$P\left(\left|\widehat{P}_m(x) - P_X(x)\right| \geq \varepsilon\right) \leq 2e^{-2m\varepsilon^2}. \tag{C.307}$$

Therefore, we obtain the bound

$$P\left(\mathfrak{P}_{m,\varepsilon}\right) > 1 - 2|\mathrm{supp}(X)|e^{-2m\varepsilon^2}. \tag{C.308}$$

In particular, if $\delta \in (0, 1/(4|\mathrm{supp}(X)|))$, then $m \geq (1/\varepsilon^2)\log(1/\delta)$ implies $P(\mathfrak{P}_{m,\varepsilon}) > 1 - \delta/2$.

Recall that, if $\mathfrak{D}_m$ occurs, then we may write

$$\widehat{I}_n(\mathcal{S}_m) = \widehat{h}_n(\mathcal{A}_m) - \sum_{x \in \mathrm{supp}(X)} \widehat{P}_m(x)\, \widehat{h}_n(\mathcal{B}_{m,x}), \tag{C.309}$$

where $\mathcal{A}_m := \{Y_j\}_{j=1}^m$ and $\mathcal{B}_{m,x} := \{Y_j \,;\, 1 \leq j \leq m, X_j = x\}$. Then,

$$\begin{aligned} \left|\widehat{I}_n(\mathcal{S}_m) - I_n(X;Y)\right| \leq{}& \left|\widehat{h}_n(\mathcal{A}_m) - h_n(Y)\right| \\ &+ \sum_{x \in \mathrm{supp}(X)} \widehat{P}_m(x) \left|\widehat{h}_n(\mathcal{B}_{m,x}) - h_n(Y^{(x)})\right| \\ &+ \left(\max_{x \in \mathrm{supp}(X)} \left|\widehat{P}_m(x) - P_X(x)\right|\right) \sum_{x \in \mathrm{supp}(X)} |h_n(Y^{(x)})|. \end{aligned} \tag{C.310}$$

Denote $H_{X,Y,n} := \sum_{x \in \mathrm{supp}(X)} |h_n(Y^{(x)})|$. Consider the events

$$\mathfrak{F}_{x,\varepsilon} := \left\{\left|\widehat{h}_n(\mathcal{B}_{m,x}) - h_n(Y^{(x)})\right| < \frac{\varepsilon}{3}\right\} \tag{C.311}$$

$$\mathfrak{F}'_{\varepsilon} := \left\{\left|\widehat{h}_n(\mathcal{A}_m) - h_n(Y)\right| < \frac{\varepsilon}{3}\right\}. \tag{C.312}$$

Set $\mathfrak{F}_\varepsilon := \bigcap_{x \in \mathrm{supp}(X)} \mathfrak{F}_{x,\varepsilon}$. From Proposition 4.4, we know that there is a constant $C_{X,Y,n}$ such that for every small enough $\varepsilon, \delta > 0$, if $m \geq (C_{X,Y,n}/\varepsilon^2)\log(1/\delta)$ then $P(\mathfrak{F}_{x,\varepsilon} \mid \mathfrak{D}_m) \geq 1 - \delta/(8|\mathrm{supp}(X)|)$ for each $x \in \mathrm{supp}(X)$ and $P(\mathfrak{F}'_\varepsilon \mid \mathfrak{D}_m) > 1 - \delta/8$. Then, $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \mid \mathfrak{D}_m) \geq 1 - \delta/4$. We conclude, possibly after increasing $C_{X,Y,n}$, that $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m) \geq 1 - \delta/2$. Also, $P(\mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}) > 1 - \delta/2$, where we increase $C_{X,Y,n}$, if necessary, to exceed $9H^2_{X,Y,n}$. Then, $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m \cap \mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}) \geq 1 - \delta$. But under the event $\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m \cap \mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}$, we have from (C.310) the bound

$$\left|\widehat{I}_n(\mathcal{S}_m) - I_n(X;Y)\right| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \tag{C.313}$$

and the proof is complete.