UNDER REVIEW (2021) **00**, 1–35
https://doi.org/DOI HERE

Optimal empirical Bayes estimation for the Poisson model via minimum-distance methods

SOHAM JANA*
Department of ORFE, Princeton University, 98 Charlton Street, Princeton, 08540, NJ, USA

YURY POLYANSKIY
Department of EECS, MIT, 50 Vassar St, Cambridge, 02139, MA, USA

AND

YIHONG WU
Department of Statistics and Data Science, Yale University, 10 Hillhouse Ave., New Haven, 06511, CT, USA
*Corresponding author: soham.jana@princeton.edu

[Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year]

The Robbins estimator is the most iconic and widely used procedure in the empirical Bayes literature for the Poisson model. On one hand, this method has been recently shown to be minimax optimal in terms of the regret (excess risk over the Bayesian oracle that knows the true prior) for various nonparametric classes of priors. On the other hand, it has been long recognized in practice that Robbins estimator lacks the desired smoothness and monotonicity of Bayes estimators and can be easily derailed by those data points that were rarely observed before. Based on the minimum-distance distance method, we propose a suite of empirical Bayes estimators, including the classical nonparametric maximum likelihood, that outperform the Robbins method in a variety of synthetic and real data sets and retain its optimality in terms of minimax regret.

Keywords: Mixture modeling; Robbins method; Poisson mean estimation; Non-parametric estimation; NPMLE.

1. Introduction

Consider the Poisson mean estimation problem. Given observations $Y^n \triangleq (Y_1, \ldots, Y_n)$, independently distributed according to the Poisson distribution with mean parameters $\theta^n \triangleq (\theta_1, \ldots, \theta_n)$, the goal is to learn the parameter vector under squared error loss. It has been well established in the literature that even though the maximum likelihood estimator $\widehat{\theta}_i = Y_i$ is minimax optimal, in practice it is often desirable to use some sort of shrinkage type estimators. For example, consider the scenario when the parameter space is bounded. Then with large sample size n, the inherent variability of the model will likely produce some extreme observations that will derail the maximum likelihood estimates. A class of shrinkage type estimators with nice theoretical properties was proposed in the seminal paper of [Rob51, Rob56], namely the empirical Bayes (EB) methodology. In the regular Bayes setup, which also produces estimators with shrinkage properties, one assumes that the parameter values are and independently distributed according to a prior distribution G. Then the best estimator under the squared error loss (i.e., the Bayes estimator) of θ_j is given by the posterior mean $\widehat{\theta}_G(Y_j) = \mathbb{E}_G[\theta_j | Y_j]$. For the Poisson model this takes the

following simplified form

$$\widehat{\theta}_G(y) = \mathbb{E}_{Y\sim \mathrm{Poi}(\theta),\theta\sim G}[\theta|Y=y] = \frac{\int \theta e^{-\theta}\frac{\theta^y}{y!}G(d\theta)}{\int e^{-\theta}\frac{\theta^y}{y!}G(d\theta)} = (y+1)\frac{f_G(y+1)}{f_G(y)}, \tag{1.1}$$

where $\mathrm{Poi}(\theta)$ denotes the Poisson distribution with mean θ (here and below), and the marginal density of the Y_j-s are given by

$$f_G(y) = \int f_\theta(y)G(d\theta), \quad f_\theta(y) = e^{-\theta}\frac{\theta^y}{y!}, y \in \mathbb{Z}_+ \triangleq \{0,1,\ldots\}. \tag{1.2}$$

The EB theory proposes to bypass the assumed knowledge about G, which is might be unavailable in practice, by approximating the G dependent expressions using the observations. The major achievement of the EB theory is that, when the number of independent observations is large, it is possible to "borrow strength" from these independent (and seemingly unrelated) observations to achieve the asymptotically optimal Bayes risk per coordinate. Since its conception, the theory and methodology of empirical Bayes have been well developed and widely applied in large-scale data analysis in practice cf. e.g. [ETST01, VH96, Bro08, PLLB10]. We refer the reader to the surveys and monographs on the theory and practice of empirical Bayes [Mor83, Cas85, Zha03, Efr14, ML18, Efr21].

In the literature, there are two main avenues to obtain solutions to the EB problem:

- f-modeling: Construct an approximate Bayes estimator by approximating the marginal density. For example, the Robbins estimator [Rob56] is a plug-in estimate of (1.1) replacing the true f_G with the empirical distribution, leading to

$$\widehat{\theta}_j = \widehat{\theta}_{\mathrm{Robbins}}(Y_j|Y_1,\ldots,Y_n) \triangleq (Y_j+1)\frac{N(Y_j+1)}{N(Y_j)}, \quad N(y) \triangleq |\{i\in[n]: Y_i = y\}|. \tag{1.3}$$

- g-modeling: We first obtain an estimate $\widehat{G}$ of the prior G from Y^n and then apply the corresponding Bayes estimator formula $\widehat{\theta}_{\widehat{G}}(Y_j)$. Examples of $\widehat{G}$ include the celebrated nonparametric maximum likelihood estimator (NPMLE) [KW56]

$$\widehat{G} = \operatorname*{argmax}_G \frac{1}{n}\sum_{i=1}^n \log f_G(Y_i) \tag{1.4}$$

  where the maximization is over all priors on $\mathbb{R}_+$ (unconstrained NPMLE). When additional information about the prior is available (e.g., compactly supported), it is convenient to incorporate this as constraints into the above optimization, leading to constrained NPMLE.

In a nutshell, both f-modeling and g-modeling rely on estimate of the population density f_G; the difference is that the former applies improper density estimate such as the empirical distribution or kernel density estimate (see, e.g., [LGL05, BG09, Zha09] for Gaussian models), while the latter applies *proper* density estimate of the form $f_{\widehat{G}}$.

In recent years, there have been significant advances in the theoretical analysis of f-modeling EB estimators for the Poisson model, specifically, the Robbins method. For compactly supported priors, [BGR13] showed that with Poisson sampling (replacing the sample size n by $\mathrm{Poi}(n)$),

the Robbins estimator achieves a $O\left(\frac{(\log n)^2}{n(\log\log n)^2}\right)$ regret for estimating each θ_i. Later [PW21] showed the same regret bound holds with fixed sample size n and established the optimality of the Robbins estimator by proving a matching minimax lower bound. In addition, for the class of subexponential priors, for estimating each θ_i the Robbins estimator also achieves optimal minimax regret $\Theta\left(\frac{(\log n)^3}{n}\right)$.

On the other hand, despite its simplicity and optimality, it has been long recognized that the Robbins method often produces unstable estimates in practice. This occurs particularly for those y which appears few times or none whatsoever, so that $N(y)$ is small or zero. Thus, unless $N(y+1)$ is also small, the formula (1.3) produces exceptionally large value of $\widehat{\theta}_{\text{Robbins}}(y)$. In addition, if $N(y+1)=0$ (e.g., when $y \geq \max\{Y_1,\ldots,Y_n\}$), we have $\widehat{\theta}_{\text{Robbins}}(y)=0$ irrespective of any existing information about y, which is at odds with the fact that the Bayes estimator $\widehat{\theta}_G(y)$ is always monotonically increasing in y for any G [HS83]. These issues of the Robbins estimator have been well-documented and discussed in the literature; see, for example, [Mar68, Section 1] and [ML18, Section 1.9] for a finite-sample study and [EH21, Section 6.1] for the destabilized behavior of Robbins estimator in practice (e.g., in analyzing insurance claims data). To alleviate the shortcomings of the Robbins estimator, a number of modifications have been proposed [Mar68, BGR13] that enforce smoothness or monotonicity; nevertheless, it is unclear they still retain the regret optimality of the Robbins method. This raises the question of whether it is possible to construct a well-behaved EB estimator that is provably optimal in terms of regret.

In this paper, we answer this question in the positive. This is accomplished by a class of g-modeling EB estimators, which are free from the unstable behavior of Robbins estimator, thanks to their Bayesian form which guarantees monotonicity among many other desirable properties. The prior is learned using the *minimum-distance* method, including the NPMLE (1.4) as a special case. Introduced in the pioneering works [Wol53, Wol54, Wol57], the minimum-distance method aims to find the best fit *in class* to the data with respect to a given distance. As such, it is well-suited for the task of estimating the prior and the obtained density estimate is *proper* and of the desired mixture type.

As a concrete example, we consider a simple uniform prior and compare the numerical performance of Robbins and three prototypical examples of minimum-distance estimators of G, with respect to the Kullback-Leibler (KL) divergence (i.e., the NPMLE), the Hellinger distance, and the χ^2-divergence, respectively (see Section 2.1 for the formal definitions). As evident in Fig. 1, the minimum-distance EB estimators provide a much more consistent approximation of the Bayes estimator compared to the Robbins estimator. This advantage is even more pronounced for unbounded priors (cf. Fig. 6 in Section 5.3); see also Fig. 2 for a real-world example where EB methodology is applied to a prediction task with sports data. Notably, in multidimensional settings, such minimum distance based EB methodologies are difficult to implement in practice as they are computationally expensive even in fixed dimensions. However, we propose that the uni-dimensional empirical Bayes methodology can be employed to provide improved analyses in multidimensional setups as well. To demonstrate the above, we considered a supervised classification problem involving the crime data in different London boroughs. We show that the performance of a linear regression algorithm, that uses different crime counts to infer about the boroughs, can be improved significantly when we pre-process each of the integer data columns using minimum-distance EB filters before supplying them to the algorithm.

The superior performance of minimum-distance EB estimators in practice is also justified by theory. In addition to characterizing their structural properties (existence, uniqueness, discreteness)

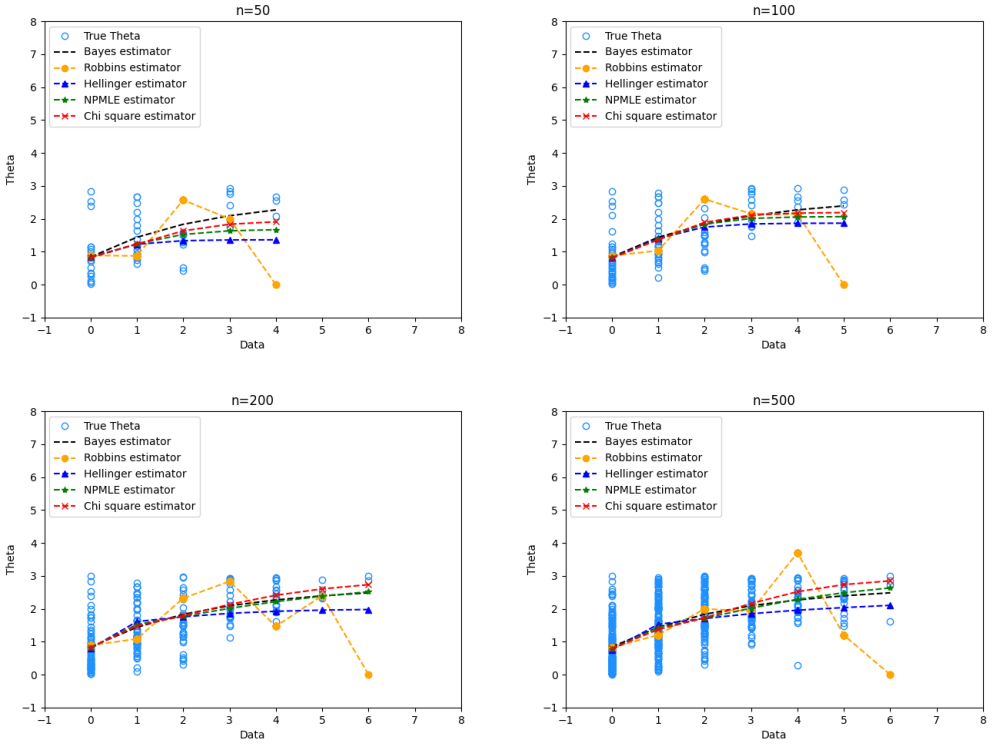

FIG. 1. Comparison of Robbins estimator with different minimum-distance EB estimators. Here the latent $\theta_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,3]$ and the observation $Y_i \overset{\text{ind.}}{\sim} \text{Poi}(\theta_i)$, for $i = 1,\dots,n$. We plot $\widehat{\theta}(Y_i)$ against Y_i for various EB estimators $\widehat{\theta}$. For reference, we also plot the true value θ_i and the Bayes estimator $\widehat{\theta}_G(Y_i)$. The sample sizes are $n = 50, 100, 200, 500$.

in the Poisson model, we show that, under appropriate conditions on the distance functional, their regret is minimax optimal for both compactly supported and subexponential priors. This is accomplished by first proving the optimality of minimum-distance estimate for density estimation in Hellinger distance, then establishing a generic regret upper bound for g-modeling EB estimators in terms of the Hellinger error of the corresponding density estimates. We also extend the theoretical analyses to a multidimensional Poisson models.

1.1. *Related works*

Searching for a stable and smooth alternative to the classical Robbins method for the Poisson EB problem has a long history. [Mar66] was one of the proponents of using g-modeling estimators to resolve this problem. The author considered modeling the prior using the Gamma distribution and estimated the scale and shape parameters using a χ^2-distance minimization; this is a parametric approach as opposed to the nonparametric approach in this paper. Based on the monotonicity of the Bayes estimator, [Mar69] used non-decreasing polynomials to approximate the Bayes oracle. [LK69] proposed an iterative method of estimating of the prior, by first using the empirical distribution of the training sample Y^n and then using corresponding posterior means of the θ_i's to denoise. On a similar vein, [BM72] assumed the existence of density of the prior distribution

and used kernel method to approximate the prior. For a detailed exposition on other smooth EB methods, see [ML18]. However, none of these methods has theoretical guarantees in terms of the regret for nonparametric class of priors considered in the present paper.

Applying NPMLE for estimating the mixture distribution has been well-studied in the literature. [KW56] was one of the preliminary papers to prove the consistency of the NPMLE, which was subsequently extended in [HS84, Jew82, LT84, Pfa88]; for a more recent discussion, see [Che17]. In the present paper we focus on the Poisson mixture model and sharpen these results by obtaining the optimal rate of convergence for the NPMLE. In additional to the aforementioned statistical results, structural understanding of the NPMLE (existence, uniqueness, and discreteness) has been obtained in [Sim76, Jew82, Lin83a, Lin83b, Lin95] for general univariate exponential family. We extend these structural results to a class of minimum-distance estimators for Poisson mixture models following [Sim76]. Finally, we mention the recent work [MKV+21] which explored the application of NPMLE in a related scenario of heterogeneous Poisson mixtures.

Initial work on applying NPMLE for EB estimation was carried out in [Lai82] for the Binomial and the normal location models, and the analysis is primarily numerical. For theoretical results, [GvdV01, Zha09] analyzed the Hellinger risk of NPMLE-based mixture density estimates, which forms the basis of the analysis of NPMLE for EB estimation in [JZ09]. The resulting regret bounds, though state of the art, still differ from the minimax lower bounds in [PW21] by logarithmic factors for both the classes of compactly supported and subgaussian priors. This is because (a) the density estimation analysis in [Zha09] is potentially suboptimal compared to the lower bounds in [Kim14]; (b) the Fourier-analytic reduction from the Hellinger distance for mixture density to regret in [JZ09] is loose. In comparison, in this paper both the density estimation and the regret bounds are optimal with exact logarithmic factors. This can be attributed to the discrete nature of the Poisson model so that for light-tailed priors a simple truncation-based analysis suffices. Additionally, these sharp results are generalized from the NPMLE-based EB estimator to the minimum-distance estimators.

1.2. *Organization*

The rest of the paper is organized as follows. In Section 2 we introduce the class of minimum distance estimators and identify conditions on the distance function that guarantees the existence and uniqueness of the minimizer. The theoretical guarantees of in terms of density estimation and regret are presented in Theorem 2 and Theorem 3 therein. The proof of these theorems are presented in Section 3 and Section 4 respectively. In Section 5 we present an algorithm for computing minimum-distance estimators in the one-dimensional setting and study their numerical performance in empirical Bayes estimation with both simulated and real datasets. In Section 6 we mention . our multidimensional results.

1.3. *Notations*

Denote by $\mathbb{Z}_+$ (resp. $\mathbb{R}_+$) the set of non-negative integers (resp. real numbers). For a Borel measurable subset $\Theta \subset \mathbb{R}$, let $\mathscr{P}(\Theta)$ be the collection of all probability measures on Θ. For any $\theta \in \mathbb{R}_+$ let δ_θ denote the Dirac measure at θ. Denote by $\mathsf{SubE}(s)$ the set of all s-subexponential distributions on $\mathbb{R}_+$: $\mathsf{SubE}(s) = \left\{G : G([t,\infty)]) \le 2e^{-t/s}, \forall t > 0\right\}$. Let $Y_i \sim \mathrm{Poi}(\theta_i)$ for $i = 1,\dots,n$ and $Y \sim \mathrm{Poi}(\theta)$, with $\theta_1,\dots,\theta_n,\theta \overset{\text{i.i.d.}}{\sim} G$. This also implies $Y_1,\dots,Y_n,Y \overset{\text{i.i.d.}}{\sim} f_G$ where f_G is the mixture distribution defined in (1.2). Let $\mathbb{E}_G$ and $\mathbb{P}_G$ respectively denote the expectation and probability where the true mixing distribution is G.

2. Problem formulation and results

2.1. *Minimum-distance estimators*

Denote by $\mathscr{P}(\mathbb{Z}_+)$ the collection of probability distributions (pmfs) on $\mathbb{Z}_+$. We call $\mathsf{dist} : \mathscr{P}(\mathbb{Z}_+) \times \mathscr{P}(\mathbb{Z}_+) \to \mathbb{R}_+$ a *generalized distance* if $\mathsf{dist}(p\|q) \geq 0$ for any $p, q \in \mathscr{P}(\mathbb{Z}_+)$, with equality if and only if $p = q$. Note that any metric or f-divergence [Csi67] qualifies as a generalized distance.

The minimum-distance[1] methodology aims to find the closest fit in the model class to the data. While it is widely used and well-studied in parametric models [Ber77, Ber55, Pol80, Bol77, Mil84], it is also useful in nonparametric settings such as mixture models. Denote by

$$p_n^{\mathsf{emp}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i} \tag{2.1}$$

the empirical distribution of the sample $Y_1, \ldots, Y_n$. The minimum-distance estimator for the mixing distribution with respect to dist, over some target class of distributions $\mathscr{G}$, is

$$\widehat{G} \in \operatorname*{argmin}_{Q \in \mathscr{G}} \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q). \tag{2.2}$$

Primary examples of minimum-distance estimators considered in this paper include the following

- Maximum likelihood: $\mathsf{dist}(p\|q) = \mathrm{KL}(p\|q) \triangleq \sum_{y\geq 0} p(y) \log \frac{p(y)}{q(y)}$ is the KL divergence. In this case, one can verify that the minimum-KL estimator coincides with the NPMLE (1.4).
- Minimum-Hellinger estimator: $\mathsf{dist}(p\|q) = H^2(p,q) \triangleq \sum_{y\geq 0} \left(\sqrt{p(y)} - \sqrt{q(y)}\right)^2$ is the squared Hellinger distance.
- Minimum-χ^2 estimator: $\mathsf{dist}(p\|q) = \chi^2(p\|q) \triangleq \sum_{y\geq 0} \frac{(p(y)-q(y))^2}{q(y)}$ is the χ^2-divergence.

We note that there are other minimum-distance estimators previously studied for Gaussian mixture models such as those respect to L_p-distance of the CDFs, aiming at estimation of the mixing distribution [DK68, Che95, HK18, Ede88]. These are outside the scope of the theory developed in this paper.

In general, the solution to (2.2) need not be unique; nevertheless, for the Poisson mixture model, the uniqueness is guaranteed provided that the generalized distance dist admits the following decomposition:

Assumption 1 *There exist maps $t : \mathscr{P}(\mathbb{Z}_+) \to \mathbb{R}$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ such that for any two distributions $q_1, q_2 \in \mathscr{P}(\mathbb{Z}_+)$*

$$\mathsf{dist}\,(q_1\|q_2) = t(q_1) + \sum_{y\geq 0} \ell(q_1(y), q_2(y)),$$

where $b \mapsto \ell(a,b)$ is strictly decreasing and strictly convex for each $a > 0$ and $\ell(0,b) = 0$ for each $b \geq 0$.

The following theorem guarantees the existence, uniqueness, and discreteness of both unconstrained and support-constrained minimum-distance estimators. For the special case of

[1] We adopt this conventional terminology even when dist need not be a distance.

unconstrained NPMLE this result was previously shown by [Sim76] and later extended to all one-dimensional exponential family [Lin95].

Theorem 1 *Let* dist *satisfy Assumption 1. Let p be a probability distribution on* $\mathbb{Z}_+$ *with support size m. Then for any* $h > 0$*, the constrained solution* $\operatorname{argmin}_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q)$ *exist uniquely and is a discrete distribution with support size at most m. Furthermore, the same conclusion also applies to the unconstrained solution* $\operatorname{argmin}_{Q\in\mathscr{P}(\mathbb{R}_+)} \mathsf{dist}(p\|f_Q)$*, which in addition is supported on* $[\min_{i=1,\dots,m} y_i, \max_{i=1,\dots,m} y_i]$*, where* $\{y_1,\dots,y_m\}$ *is the support of p.*

To analyze the statistical performance of minimum-distance estimators, we impose the following regulatory condition on the generalized distance dist:

Assumption 2 *There exist absolute constants* $c_1, c_2 > 0$ *such that*

$$c_1 H^2(q_1,q_2) \le \mathsf{dist}(q_1\|q_2) \le c_2\chi^2(q_1\|q_2) \tag{2.3}$$

for pmfs q_1, q_2 *on* $\mathbb{Z}_+$.

Major examples of generalized distance satisfying Assumptions 1 and 2 include the KL divergence, squared Hellinger distance, and χ^2-divergence. This follows from noting that $2H^2 \le \mathrm{KL} \le \chi^2$ and each of them satisfies the decomposition Assumption 1: for squared Hellinger $t \equiv 2, \ell(a,b) = -2\sqrt{ab}$, for KL divergence $t \equiv 0, \ell(a,b) = a\log\frac{a}{b}$, for χ^2-divergence $t \equiv -1, \ell(a,b) = \frac{a^2}{b}$. On the other hand, total variation (TV) satisfies neither Assumption 1 nor 2 so the theory in the present paper does not apply to the minimum-TV estimator.

2.2. *Main results*

In this section we state the statistical guarantee for the minimum-distance estimator $\widehat{G}$ defined in the previous section. Our main results are two-fold (both minimax optimal):

1. Density estimation, in terms of the Hellinger distance $f_{\widehat{G}}$ and the true mixture f_G;
2. Empirical Bayes, in terms of the regret of the Bayes estimator with the learned prior $\widehat{G}$.

As mentioned in Section 1, the regret analysis in fact relies on bounding the density estimation error. We start with the result for density estimation. Recall from Section 1.3 $\mathscr{P}([0,h])$ and $\mathsf{SubE}(s)$ denote the class of compactly supported and subexponential priors respectively.

Theorem 2 (Density estimation) *Let* dist *satisfy Assumptions 1 and 2. Given any* $h, s > 0$*, there exist constants* $C_1 = C_1(h)$ *and* $C_2 = C_2(s)$ *such that the following are satisfied.*

1. *Let* $G \in \mathscr{P}([0,h])$ *and* $\widehat{G} = \operatorname{argmin}_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q)$*, then* $\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{C_1}{n}\frac{\log n}{\log\log n}$ *for any* $n \ge 3$.
2. *Let* $G \in \mathsf{SubE}(s)$ *and* $\widehat{G} = \operatorname{argmin}_Q \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q)$*, then* $\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{C_2}{n}\log n$ *for any* $n \ge 2$.

Remark 1 *It has been shown recently in [PW21, Theorem 21] that for any fixed h, s, the minimax squared Hellinger density estimation errors are at least* $\Omega\left(\frac{\log n}{n\log\log n}\right)$ *and* $\Omega\left(\frac{\log n}{n}\right)$ *for priors in the class* $\mathscr{P}([0,h])$ *and* $\mathsf{SubE}(s)$*, respectively. This establishes the minimax optimality of our minimum-distance density estimates.*

Next we turn to the problem of estimating $\theta_1,\dots,\theta_n$ from $Y_1,\dots,Y_n$, under the squared error loss, using the empirical Bayes methodology. In this work we study the estimation guarantees of the g-modeling type estimators. Notably, to produce an estimator $\widehat{\theta}_j$ of θ_j, we use the observations $Y^{-j} = (Y_1,\dots,Y_{j-1},Y_{j+1},\dots,Y_n)$ to approximate G and then plug it in the formula of the Bayes estimator $\widehat{\theta}_{\widehat{G}}(Y_j)$ in (1.1). Given any class of distributions $\mathscr{G}$ and any distribution estimator strategy characterized by $\widehat{G}$, define the total regret as its worst case excess risk over the Bayes error:

$$\begin{aligned}\mathsf{TotRegret}_n(\widehat{G};\mathscr{G}) &\triangleq \sup_{G\in\mathscr{G}}\left\{\mathbb{E}_G\left[\|\widehat{\theta}^n(Y^n)-\theta^n\|^2 - n\cdot\mathrm{mmse}(G)\right]\right\},\\ \widehat{\theta}_j &= \widehat{\theta}_{\widehat{G}(Y^{-j})}(Y_j),\quad j=1,\dots,n.\end{aligned} \tag{2.4}$$

where $\mathrm{mmse}(G)$ denotes the minimum mean squared error of estimating $\theta\sim G$ based on a single observation $Y\sim f_\theta$, i.e., the Bayes risk

$$\mathrm{mmse}(G) \triangleq \inf_{\widehat{\theta}} \mathbb{E}_G\left[\left(\widehat{\theta}(Y)-\theta\right)^2\right] = \mathbb{E}_G\left[(\widehat{\theta}_G(Y)-\theta)^2\right]. \tag{2.5}$$

In addition, one can define the problem of quantifying the individual regret for the estimator $\widehat{G}$

$$\mathsf{Regret}_n(\widehat{G};\mathscr{G}) \triangleq \sup_{G\in\mathscr{G}}\left\{\mathbb{E}_G\left[(\widehat{\theta}_n(Y^n)-\theta_n)^2\right] - \mathrm{mmse}(G)\right\},\quad \widehat{\theta}_n(Y^n) = \widehat{\theta}_{\widehat{G}(Y^{n-1})}(Y_n). \tag{2.6}$$

Here $Y_1,\dots,Y_{n-1}$ can be viewed as training data which is used to learn the estimator and then we apply it on a fresh (unseen) data point Y_n to predict θ_n. Getting back to the loss function under consideration, it is not difficult to see that the total regret with sample size n can be bounded from above using n times the individual regret with training sample size $n-1$

$$\mathsf{TotRegret}_n(\widehat{G};\mathscr{G}) \le n\cdot\mathsf{Regret}_n(\widehat{G};\mathscr{G}). \tag{2.7}$$

In view of the above inequality of the total and individual regret functions, we limit ourselves to studying the individual regret only, as this will suffice to achieve the desired optimal rates.

Now we are in a position to describe the main results for empirical Bayes estimation. For an ease of notations, suppose that given a fresh sample $Y\sim\mathrm{Poi}(\theta)$, where θ is generated from an unknown prior distribution G, we want to predict the value of θ in the squared error loss and training sample to construct the estimator $\widehat{G}$ is given by $Y_1,\dots,Y_n$. Given any estimator $\widehat{G}$ of G we define the regret of the empirical Bayes estimate $\widehat{\theta}_{\widehat{G}}$ as

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) &= \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y)-\theta\right)^2\right] - \mathrm{mmse}(G)\\ &\overset{\text{(a)}}{=} \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_G(Y)\right)^2\right] = \mathbb{E}_G\left[\sum_{y\ge0}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_G(y))^2 f_G(y)\right],\end{aligned} \tag{2.8}$$

where the identity (a) followed by using the orthogonality principle: the average risk of any estimator $\widehat{\theta}$ can be decomposed as

$$\mathbb{E}_G[(\widehat{\theta}-\theta)^2] = \mathrm{mmse}(G) + \mathbb{E}_G[(\widehat{\theta}-\widehat{\theta}_G)^2]. \tag{2.9}$$

Similarly we define the maximum regret of $\widehat{G}$ over the class of model distributions $\mathscr{G}$

$$\mathsf{Regret}(\widehat{G};\mathscr{G}) = \sup_{G\in\mathscr{G}} \mathsf{Regret}(\widehat{G};G). \tag{2.10}$$

Then we have the following estimation guarantees.

Theorem 3 (Empirical Bayes) *Let* dist *satisfy Assumptions 1 and 2. Given any $h,s>0$, there exist constants $C_1 = C_1(h)$ and $C_2 = C_2(s)$ such that the following are satisfied.*

1. *If $\widehat{G} = \operatorname{argmin}_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q)$, then for any $n\geq 3$,*

$$\mathsf{Regret}(\widehat{G};\mathscr{P}([0,h])) \leq \frac{C_1}{n}\left(\frac{\log n}{\log\log n}\right)^2. \tag{2.11}$$

2. *If $\widehat{G} = \operatorname{argmin}_{Q} \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q)$, then for any $n\geq 2$,*

$$\mathsf{Regret}(\widehat{G};\mathsf{SubE}(s)) \leq \frac{C_2}{n}(\log n)^3. \tag{2.12}$$

Remark 2 1. *As mentioned in Section 1, for fixed h and s, both (2.11) and (2.12) match the minimax lower bounds recently shown in [PW21, Theorem 1]. This establishes the regret optimality of minimum-distance EB estimators, which was only known for the f-modeling-based Robbins estimator before.*

2. *When* dist *is the KL divergence, the minimum-distance estimator $\widehat{G} = \operatorname{argmin}_Q \mathsf{KL}(p\|q)$ is the NPMLE. This follows from the expansion*

$$KL(p_n^{\mathsf{emp}}\|f_Q) = \sum_{y\geq 0} p_n^{\mathsf{emp}}(y)\log\frac{p_n^{\mathsf{emp}}(y)}{f_Q(y)} = \sum_{y\geq 0} p_n^{\mathsf{emp}}(y)\log p_n^{\mathsf{emp}}(y) - \frac{1}{n}\sum_{i=1}^n \log f_Q(Y_i).$$

3. *Theorem 3 holds for approximate solutions. Consider the following approximate minimum-distance estimators $\widehat{G}$, over some target class of distributions $\mathscr{G}$, that satisfies*

$$\mathsf{dist}(p_n^{\mathsf{emp}}\|f_{\widehat{G}}) \leq \inf_{Q\in\mathscr{G}} \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q) + \delta. \tag{2.13}$$

for some $\delta>0$. Then (2.11) (resp. (2.12)) continues to hold if $\delta \lesssim \frac{\log n}{n\log\log n}$ (resp. $\frac{\log n}{n}$). Note that $\widehat{G}$ is the NPMLE over $\mathscr{G}$ if $\delta = 0$ and dist *is given by KL divergence. In case of NPMLE, (2.13) translates to an approximate likelihood maximizer $\widehat{G}$ such that*

$$\frac{1}{n}\sum_{i=1}^n \log f_{\widehat{G}}(Y_i) \geq \operatorname*{argmax}_{G\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^n \log f_G(Y_i) - \delta.$$

This type of results is well-known in the literature, see, for example, [JZ09, Zha09] for the normal location-mixture model.

3. Proof for density estimation

The proof of Theorem 2 is based on a simple truncation idea. It is straightforward to show that the density estimation error for any minimum dist-distance estimator can be bounded from above, within a constant factor, by the expected squared Hellinger distance between the empirical distribution p_n^{emp} and the data-generating distribution f_G, which is further bounded by the expected χ^2-distance. The major contribution to $\chi^2(p_n^{\mathsf{emp}}\|f_G)$ comes from the "effective support" of f_G, outside of which the total probability is $o(\frac{1}{n})$. For the the prior classes $\mathscr{P}([0,h])$ and $\mathsf{SubE}(s)$, the Poisson mixture f_G is effectively supported on $\{0,\dots,O(\frac{\log n}{\log\log n})\}$ and $\{0,\dots,O(\log n)\}$. Each point in the effective support contributes $\frac{1}{n}$ to $\chi^2(p_n^{\mathsf{emp}}\|f_G)$ from which our results follow.

Proof of Theorem 2 For any $K \geq 1$ and distribution G denote the tail probabilities of the Poisson mixture as

$$\varepsilon_K(G) \triangleq \mathbb{P}[Y \geq K] = \sum_{y=K}^{\infty} f_G(y) \tag{3.1}$$

Note that dist satisfies Assumption 2, namely (2.3). We first prove the following general inequality

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \leq \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1} + 2n\right)\varepsilon_K(G). \tag{3.2}$$

Using the triangle inequality, the elementary fact $(a+b)^2 \leq 2(a^2+b^2)$, and the minimizing property of $\widehat{G}$, we get

$$\begin{aligned} H^2(f_G, f_{\widehat{G}}) &\leq \left(H(p_n^{\mathsf{emp}}, f_{\widehat{G}}) + H(p_n^{\mathsf{emp}}, f_G)\right)^2 \\ &\leq 2[H^2(p_n^{\mathsf{emp}}, f_{\widehat{G}}) + H^2(p_n^{\mathsf{emp}}, f_G)] \\ &\leq \frac{2}{c_1}(\mathsf{dist}(p_n^{\mathsf{emp}}\|f_{\widehat{G}}) + \mathsf{dist}(p_n^{\mathsf{emp}}\|f_G)) \leq \frac{4}{c_1}\mathsf{dist}(p_n^{\mathsf{emp}}\|f_G). \end{aligned} \tag{3.3}$$

Define

$$Y_{\max} \triangleq \max_{i=1}^{n} Y_i. \tag{3.4}$$

as before. Then, bounding $\frac{1}{c_2}d$ by χ^2 we get the following chain

$$\begin{aligned}
\frac{1}{c_2}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] &\le \mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] \\
&= \sum_{y\ge 0}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y)-f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)} \\
&\overset{\text{(a)}}{=} \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y)-f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)} + \sum_{y\ge K} f_G(y)\mathbb{P}[Y_{\max}<K] \\
&= \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y)-f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)} + \varepsilon_K(G)(1-\varepsilon_K(G))^n \\
&\le \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y)-f_G(y))^2\right]}{f_G(y)} + \varepsilon_K(G)(1-\varepsilon_K(G))^n \\
&\overset{\text{(b)}}{=} \frac{1}{n}\sum_{y<K}(1-f_G(y)) + \varepsilon_K(G)(1-\varepsilon_K(G))^n \le \frac{K}{n}+\varepsilon_K(G). \quad (3.5)
\end{aligned}$$

where (a) follows from the fact that under $\{Y_{\max}<K\}$ we have $p_n^{\mathsf{emp}}(y)=0$ for any $y\ge K$; and (b) follows from $\mathbb{E}[p_n^{\mathsf{emp}}(y)]=f_G(y)$ and, thus, $\mathbb{E}[(p_n^{\mathsf{emp}}(y)-f_G(y))^2]=\mathrm{Var}(p_n^{\mathsf{emp}}(y))=\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathbf{1}_{\{Y_i=y\}})=\frac{f_G(y)(1-f_G(y))}{n}$.

Using the union bound and the fact $H^2\le 2$ we have

$$\mathbb{E}\left[H^2(f_G,f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}\ge K\}}\right] \le 2\mathbb{P}[Y_{\max}\ge K] \le 2n\varepsilon_K(G).$$

Combining this with (3.3) and (3.5) yields

$$\begin{aligned}
\mathbb{E}\left[H^2(f_G,f_{\widehat{G}})\right] &\le \mathbb{E}\left[H^2(f_G,f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}<K\}}\right] + \mathbb{E}\left[H^2(f_G,f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}\ge K\}}\right] \\
&\le \frac{4}{c_1}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] + 2n\varepsilon_K(G) \le \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1}+2n\right)\varepsilon_K(G),
\end{aligned}$$

which completes the proof of (3.2).

To complete the proof of the theorem we need to estimate the value of K such that $\varepsilon_K(G)\lesssim\frac{1}{n^2}$. This is done slightly differently for each of the two different classes of priors:

1. Let $G\in\mathscr{P}([0,h])$. Note that for $y>0$ the function $e^{-\theta}\theta^y$ is increasing in $\theta\in[0,y]$. So for any $K>2h$,

$$\varepsilon_K(G)=\sum_{y=K}^{\infty}\int_0^h \frac{e^{-\theta}\theta^y}{y!}G(d\theta) \le \sum_{y=K}^{\infty}\frac{e^{-h}h^y}{y!} \le \frac{e^{-h}h^K}{K!}\sum_{y-K=0}^{\infty}\left(\frac{h}{K}\right)^{y-K} \le \frac{2e^{-h}h^K}{K!}. \quad (3.6)$$

We choose $K = \left\lceil \frac{2(2+he)\log n}{\log\log n} \right\rceil$. Using $K! \geq \left(\frac{K}{e}\right)^K$ from the Stirling's formula and the fact $\log\log\log n < \frac{\log\log n}{2}$ for all $n \geq 3$ we continue the last display to get

$$\varepsilon_K(G) \leq 2\left(\frac{he}{K}\right)^K \leq 2\left(\frac{\log\log n}{2\log n}\right)^{\frac{2(2+he)\log n}{\log\log n}} \leq 2e^{-(\log\log n - \log\log\log n)\frac{2(2+he)\log n}{\log\log n}} \leq 2e^{-2\log n} \leq \frac{2}{n^2} \tag{3.7}$$

as required.

2. Let $G \in \mathsf{SubE}(s)$. Choose $K = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}$. Then (C.1) in Appendix C implies that $\varepsilon_K(G) \leq \frac{3}{2n^2}$. Plugging this in (3.2) completes the proof.

□

4. Proof of regret upper bound

4.1. *General regret upper bound via density estimation*

The proof of Theorem 3 relies on relating the regret in EB estimation to estimating the mixture density in the Hellinger distance. This idea has been previously noted in [JZ09, Theorem 3] for the Gaussian location models using Fourier analysis and an ingenious induction argument. Here the analysis turns out to be much simpler thanks in part to the discreteness of the Poisson model and the light tail of the prior, leading to the following deterministic result which is crucial for proving the regret optimality of minimum-distance EB estimators.

Lemma 4 *Let G be a distribution such that $\mathbb{E}_G[\theta^4] \leq M$ for some constant M. Then for any distribution $\widehat{G}$ supported on $[0,\widehat{h}]$, any $h > 0$ with $G([0,h]) > \frac{1}{2}$ and any $K \geq 1$,*

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) \leq &\left\{12(h^2+\widehat{h}^2)+48(h+\widehat{h})K\right\}(H^2(f_G,f_{\widehat{G}})+2G((h,\infty))) \\ &+2(h+\widehat{h})^2\varepsilon_K(G)+2(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}\end{aligned}$$

where $\mathsf{Regret}(\widehat{G};G)$ *and* $\varepsilon_K(G)$ *were defined in* (3.1) *and* (2.8) *respectively.*

We provide a sketch of the proof here (see Appendix B for the full proof.) It is relatively easy to bound the regret if the corresponding Bayes estimator is also bounded, which is the case if the prior G is compactly supported. Otherwise, one can consider its restriction G_h on $[0,h]$ defined by $G_h(\cdot) = \frac{G(\cdot\cap[0,h])}{G([0,h])}$. The truncation error can be controlled using properties of the mmse as follows:

$$\mathsf{Regret}(\widehat{G};G) \leq \mathsf{Regret}(\widehat{G};G_h) + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}.$$

Then we use the structure of the Bayes estimator (1.1) in the Poisson model to relate $\mathsf{Regret}(\widehat{G};G_h)$ to the squared Hellinger distance between f_{G_h} and $f_{\widehat{G}}$

$$\mathsf{Regret}(\widehat{G};G_h) \leq \left\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right\}H^2(f_{G_h},f_{\widehat{G}})+(h+\widehat{h})^2\varepsilon_K(G_h), \tag{4.1}$$

for any $K \geq 0$. We then show that $\varepsilon_K(G_h)$ and $H^2(f_{G_h}, f_{\widehat{G}})$ can be related to those for the original prior G as

$$\varepsilon_K(G_h) \leq 2\varepsilon_K(G), \quad H^2(f_{\widehat{G}}, f_{G_h}) \leq 2H^2(f_G, f_{\widehat{G}}) + 4G((h,\infty)).$$

Replacing these bounds in (4.1) we get the desired result.

4.2. *Proof of Theorem 3*

For rest of the section, let $C_1, C_2, \ldots$ denote constants depending on h, s, d as required.

For Part (a), recall that $\widehat{G} = \operatorname{argmin}_{Q \in \mathscr{P}([0,h])} \mathsf{dist}\left(p_n^{\mathsf{emp}} \| f_Q\right)$ is the support-constrained NPMLE. To apply Lemma 4, set $\widehat{h} = h, M = h^4$, and $K = \left\lceil \frac{2(2+he)\log n}{\log\log n} \right\rceil$. For any $G \in \mathscr{P}([0,h])$ we have from the proof of Theorem 2(a)

$$\varepsilon_K(G) \leq \frac{2}{n^2}, \quad G((h,\infty)) = 0, \quad \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \leq C_1 \left(\frac{1}{n} \frac{\log n}{\log\log n}\right).$$

Then Lemma 4 yields

$$\begin{aligned} \mathsf{Regret}(\widehat{G}; G) &= \mathbb{E}\left[\sum_{y=0}^{\infty} (\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_G(y))^2 f_G(y)\right] \\ &\leq \left\{12h^2 + 48hK\right\} \mathbb{E}\left[H^2\left(f_G, f_{\widehat{G}}\right)\right] + \frac{C_2}{n} \leq \frac{C_3}{n} \left(\frac{\log n}{\log\log n}\right)^2, \end{aligned}$$

as required.

For Part (b), $\widehat{G} = \operatorname{argmin}_Q \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q)$ is the unconstrained minimum-distance estimator. Choose

$$h = 4s\log n, \quad K = \frac{2\log n}{\log\left(1 + \frac{1}{2s}\right)}, \quad M = 30s^4. \tag{4.2}$$

Since G is s-subexponential, we have (see Appendix C for details)

$$\mathbb{E}_G[\theta^4] \leq M, \quad G((h,\infty)) \leq \frac{2}{n^4}, \quad \varepsilon_K(G) \leq \frac{3}{2n^2}. \tag{4.3}$$

In view of Lemma 8 we get that $\widehat{G}$ is supported on $[0, \widehat{h}]$ where $\widehat{h} = Y_{\max}$ as defined in (3.4). Then using Lemma 4 and $(\mathbb{E}_G[Y_{\max}^2])^2 \leq \mathbb{E}_G[Y_{\max}^4] \leq C_4(\log n)^4)$ (see Appendix C for a proof) we get

$$\mathsf{Regret}(\widehat{G}; G) \leq \mathbb{E}\left[\left\{6(h^2 + Y_{\max}^2) + 24K(h + Y_{\max})\right\} H^2(f_G, f_{\widehat{G}})\right] + \frac{C_5}{n}. \tag{4.4}$$

Next we bound the expectation in the last display. Using the fact that $H^2 \leq 2$, we get

$$\begin{aligned} &\mathbb{E}\left[\left\{(h^2 + Y_{\max}^2) + 4K(h + Y_{\max})\right\} H^2(f_G, f_{\widehat{G}})\right] \\ &\leq (h^2 + 4Kh + 12K^2)\mathbb{E}\left[H^2\left(f_G, f_{\widehat{G}}\right)\right] + 2\mathbb{E}\left[\left\{(h^2 + Y_{\max}^2) + 4K(h + Y_{\max})\right\} \mathbf{1}_{\{Y_{\max} \geq 2K\}}\right] \end{aligned} \tag{4.5}$$

Using Theorem 2(b) we get $\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{C_6 \log n}{n}$. For the second term in (4.5) we use Cauchy-Schwarz inequality and union bound to get

$$\begin{aligned}
&\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}\mathbf{1}_{\{Y_{\max}\ge 2K\}}\right] \\
&\le \sqrt{\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}^2\right]\mathbb{P}_G\left[Y_{\max}\ge 2K\right]} \\
&\le \sqrt{\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}^2\right] n\varepsilon_{2K}(G)} \\
&\le \frac{6}{n}\sqrt{\mathbb{E}\left[\left\{4(h^4+Y_{\max}^4)+16K^2(h^2+Y_{\max}^2)\right\}\right]} \le \frac{12}{n}\sqrt{h^4+(\log n)^4+4K^2(h^2+C_7(\log n)^2))}.
\end{aligned}$$

Plugging the bounds back in (4.5) and in view of (4.4), we complete the proof.

5. Numerical experiments

In this section we analyze the performances of the empirical Bayes estimators based on the minimum-H^2, the minimum-χ^2, and the minimum-KL divergence estimator (i.e., the NPMLE). We compare them against the Robbins estimator and also draw comparisons among their individual performances. Unlike the Robbins estimator, the minimum-distance based estimators do not admit a closed form solution. Our algorithm to compute the solution is closely related to the vertex direction method (VDM) algorithms for finding NPMLE [Lin83a, Lin95], specialized for the Poisson family and modified to work with the generalized distance that we considered. In case of the NPMLE, the convergence of the VDM method to the unique optimizer is well-known [Fed72, Wyn70], and the algorithms for finding the other minimum dist-distance estimators are expected to show similar convergence guarantees as well. Additionally, thanks to the form of the Poisson density, the first-order optimality condition takes on a polynomial form, which allow us to use existing root-finding algorithms for polynomial to update the support points of the solution. See [Sim76] for a similar VDM-type algorithm for Poisson mixtures and [KM14, KG17] for discretization-based algorithms.

5.1. *First-order optimality condition and algorithm*

In the numerical experiments we focus on the unconstrained minimum-distance estimator $\widehat{G} = \operatorname{argmin}_Q \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q)$, which is a discrete distribution (Theorem 1). For any $\theta \in \mathbb{R}_+$ let δ_θ denote the Dirac measure at θ. Suppose that the support of p_n^{emp} be $\{y_1, \ldots, y_m\}$. The optimality of $\widehat{G}$ implies that for all $\theta, \varepsilon \in [0,1]$

$$\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{\widehat{G}}) \le \mathsf{dist}(p_n^{\mathsf{emp}} \| f_{(1-\varepsilon)\widehat{G}+\varepsilon\delta_\theta}), \tag{5.1}$$

leading to the first-order optimality condition $\frac{d}{d\varepsilon}\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{(1-\varepsilon)\widehat{G}+\varepsilon\delta_\theta})\Big|_{\varepsilon=0} \ge 0$, namely

$$D_{\widehat{G}}(\theta) \triangleq \sum_{i=1}^{m} \frac{d}{df}\ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_{\widehat{G}}(y_i)} (f_\theta(y_i) - f_{\widehat{G}}(y_i)) \ge 0. \tag{5.2}$$

Averaging the left hand side over $\theta \sim \widehat{G}$, we get $\int D_{\widehat{G}}(\theta) d\widehat{G}(\theta) = 0$. This implies that each θ in the support of $\widehat{G}$ satisfies $D_{\widehat{G}}(\theta) = 0$. Simplifying the above equation we get that the atoms of $\widehat{G}$

satisfies the following polynomial equation in θ

$$\sum_{i=1}^{m} w_i(\widehat{G})\left(y_i\theta^{y_i-1}-\theta^{y_i}\right)=0, \quad w_i(\widehat{G})=\frac{\frac{d}{df}\ell(p_n^{\mathsf{emp}}(y_i),f)\Big|_{f=f_{\widehat{G}}(y_i)}}{y_i!}.$$

Iterating the above conditions leads to following algorithm to compute the minimum-distance estimators.

Algorithm 1 Computing the minimum dist-distance estimators

Input: Data points $Y_1,\ldots,Y_n$. Target distribution $G_{\boldsymbol{\theta},\boldsymbol{\mu}}=\sum_j \mu_j\delta_{\theta_j}$. Divergence dist with $t-\ell$ decomposition $\mathsf{dist}(q_1\|q_2)=t(q_1)+\sum_{y\geq 0}\ell(q_1(y),q_2(y))$. Initialization of $(\boldsymbol{\theta},\boldsymbol{\mu})$. Tolerance $\varepsilon,\eta_1,\eta_2$.

Steps:

1: Calculate empirical distribution p_n^{emp}. Obtain the set of distinct sample entries $\{y_1,\ldots,y_m\}$.
2: **while** $\mathsf{dist}(p_n^{\mathsf{emp}}\|f_{G_{\boldsymbol{\theta},\boldsymbol{\mu}}})$ decreases by less than ε in the new update, **do**
3: $\quad$ *newroots* $=\left\{\theta:\theta\geq 0,\sum_{i=1}^m w_i(G_{\boldsymbol{\theta},\boldsymbol{\mu}})\left(y_i\theta^{y_i-1}-\theta^{y_i}\right)=0\right\}$.
4: $\quad$ Combine $\boldsymbol{\theta}$ and *newroots* and denote the new vector as $\boldsymbol{\theta}'$.
5: $\quad$ Merge entries of $\boldsymbol{\theta}'$ that are within η_1 distance of each other.
6: $\quad$ Find $\mathrm{argmin}_{\widetilde{\boldsymbol{\mu}}}\sum_{i=1}^m \ell(p_n^{\mathsf{emp}}(y_i),f_{G_{\boldsymbol{\theta}',\widetilde{\boldsymbol{\mu}}}}(y_i))$, via gradient descent with initialization at $\widetilde{\boldsymbol{\mu}}=\boldsymbol{\mu}$.
7: $\quad$ Remove entries of $\boldsymbol{\theta}'$ and $\boldsymbol{\mu}'$ corresponding to the entries of $\boldsymbol{\mu}'$ that are less than η_2 and re-normalize $\boldsymbol{\mu}'$.
8: $\quad$ $(\boldsymbol{\theta},\boldsymbol{\mu})\leftarrow(\boldsymbol{\theta}',\boldsymbol{\mu}')$.
9: **end while**

Output: $(\boldsymbol{\theta},\boldsymbol{\mu})$.

We apply this algorithm for finding the minimum-distance estimators in the following examples. In all our experiments we used $\eta_1=0.01,\eta_2=0.001$. Instead of a pre-specified tolerance ε we set the maximum number of iteration to be 15. We choose the initialization for $\boldsymbol{\theta}$ as the uniform grid of size 1000 over the interval $[0,Y_{\max}]$, with initial probability assignment $\boldsymbol{\mu}$ being uniform as well.

5.2. *Real-data analysis*

5.2.1. Prediction of hockey goals

We study the data on total number of goals scored in the National Hockey League for the seasons 2017-18 and 2018-19 (the data is available at https://www.hockey-reference.com/). We consider the statistics of $n=745$ players, for whom the data were collected for both the seasons. Let Y_i be the total number of goal scored by the i^{th} player in the season 2017-18. We model Y_i as independently distributed $\mathrm{Poi}(\theta_i)$, where θ_i's are independently distributed according to some prior G on $\mathbb{R}_+$. Based on the observations we intend to predict the goal scored by each player in the season 2018-19. Specifically, for the i^{th} player our prediction is $\widehat{\theta}(Y_i)$, where $\widehat{\theta}$ is an empirical Bayes estimator driven by the 2017-18 data, either through f-modeling (e.g. $\widehat{\theta}=\widehat{\theta}_{\text{Robbins}}$) or g-modeling (e.g. $\widehat{\theta}=\widehat{\theta}_{\widehat{G}}$, where $\widehat{G}$ is learned by minimum-distance methods.) In Fig. 2 we plot the EB estimators based

TABLE 1 *Robbins vs. minimum-distance: Prediction error comparison.*

Methods	Robbins	minimum-H^2	NPMLE	minimum-χ^2
RMSE	15.59	6.02	6.04	6.05
MAD	6.64	4.37	4.38	4.39

on the Robbins method, the minimum H^2, the minimum-χ^2 distance estimator and the NPMLE against the 2017-18 data (denoted as "Past" on the x-axis) and compare their estimates against the real values of goals in 2018-19 (denoted by "Future" on the y-axis).

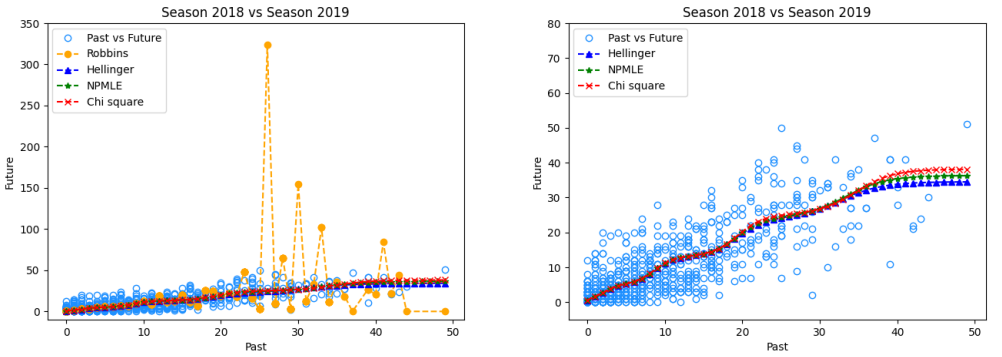

FIG. 2. Prediction of hockey goals with empirical Bayes, comparing Robbins and minimum-distance estimators. On the right panel, the Robbins estimator is omitted.

The left panel shows that there exists a large number of individuals for whom the Robbins estimator produces unstable prediction that is significantly worse than all minimum-distance methods. This difference is significant for values of scored goals which have lower sample representations. Thus on the right panel we omit the Robbins estimator and provide a more detailed comparison for the three minimum-distance estimators, which shows that their behavior are mostly comparable except near the tail end of the data-points. Interestingly, all three estimators seem to do shrinkage towards several fixed values. There could be several explanations for this multi-modality. One is that different clusters correspond to different player positions (defense, winger, center). The other is that clusters correspond to the line of the player (different lines get different amount of ice time). To test this hypothesis we also redid on Fig. 3 the estimation for each position separately. Since the multi-modality is retained, we conclude that the second option is more likely to be the real explanation. We also tabulate performance of the estimators in terms of the root mean squared error (RMSE) and the Mean absolute deviation (MAD) with respect to the true goal values in 2018-19:[2]

In addition we also compared the four goal-prediction methods based on different EB estimators across the possible playing positions: defender, center, winger. Similar as before, we

[2] Given data points $Y_1, \ldots, Y_n$ and their predictions $\widehat{Y}_1, \ldots, \widehat{Y}_n$ the RMSE is defined as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$ and the MAD is defined as $\frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$.

used the Poisson model and tried to predict the goal scoring for the year 2019 using the goal scoring data from the year 2018 for players in each playing position separately. As expected, the minimum distance methodology provides more stable and accurate estimates that the estimates based on the Robbins method. The plots showing closeness of the predictions to the true number of goals for the different EB methods are provided below.

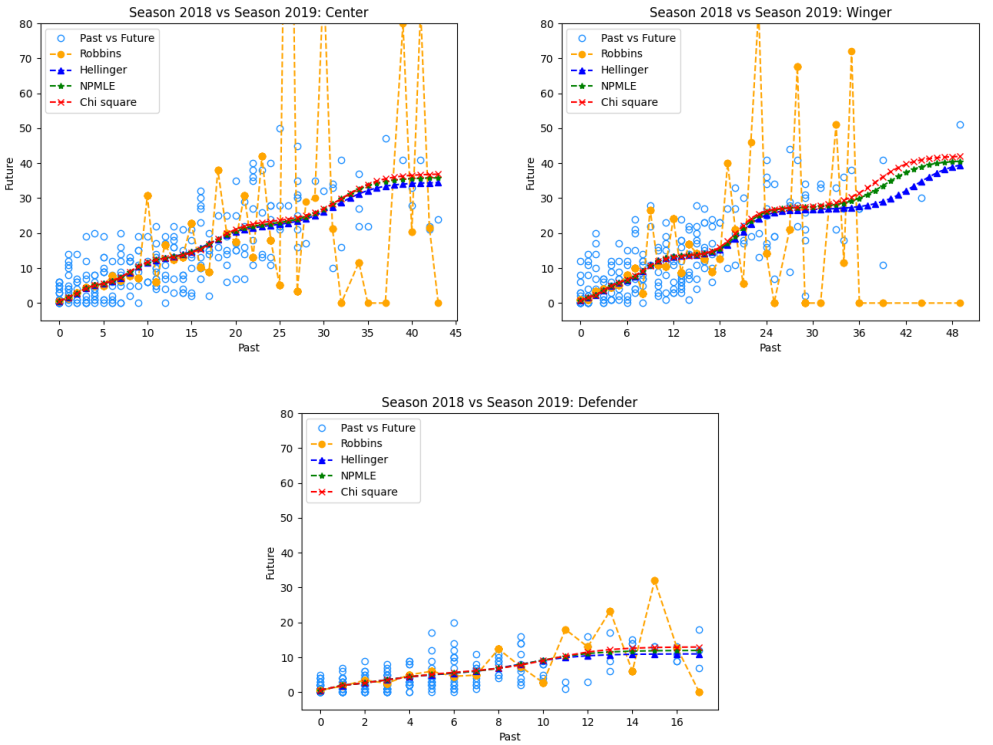

FIG. 3. Prediction of hockey goals at different playing positions.

5.2.2. Classification of London boroughs using crime data

In this section we demonstrate an application of the EB methodology for data cleaning. We propose to show that given a standard statistical methodology, incorporating an EB-based filter on the data before feeding it to the algorithm, can significantly improve the existing performance guarantees. For our analysis we used the London crime dataset which includes crime count over different geographic levels of London (borough, ward, LSOA) per month, according to crime type. The dataset is available publicly at https://data.london.gov.uk/dataset/recorded_crime_summary. Using this data, we study the task of accurately classifying the boroughs using data for different minor crime types. We work with the subset of the data corresponding to the first six boroughs according to the alphabetical ordering. We use the crime data corresponding to the smaller geographical subdivisions called the Lower Super Output Area (LSOA) as independent representatives of the crime patterns in different boroughs. Next we explain our statistical analysis. The base level methodology we intend to improve upon is the multiple Logistic regression performed by the

Scikit machine learning library for the Python programming language. The software library can be obtained freely on Python 3.10 or higher versions. We run the Logistic Regression on the borough names using the counts for different minor crime types (those with at least 50 incidents over the corresponding month) as the independent variables. We train the model using a fixed percentage of randomly choosen entries in the data set. We measure the accuracy (on a scale from 0 to 1) of the fit using the rest of the data points as the test set. Next we implement the EB filtering on the data rows separately for the different minor crime types before performing the same modeling as follows. For each crime type, we fit a Poisson mixture model similar to our setup and then replace the data vector with the NPMLE-based empirical Bayes estimates given by $\widehat{\theta}_{\widehat{G}}$, where $\widehat{G}$ is the estimated prior. Then use the a similar train-test split as before to measure the accuracy of classification for all the boroughs.

We first demonstrate the effect of the test-train split on the classification process. We consider the data corresponding to January, 2023. Varying the proportion of test set in the set $\left\{\frac{i}{6} : i = 1,\ldots,5\right\}$ we repeat the experiment 1000 time in each scenario and measure the average accuracy for the regression processes, with and without the EB filtering. As it can be seen in Fig. 4, the regression performed post the EB-filtering produces 10%-12% more accurate classification. Note that the plots include 95% confidence intervals as well.

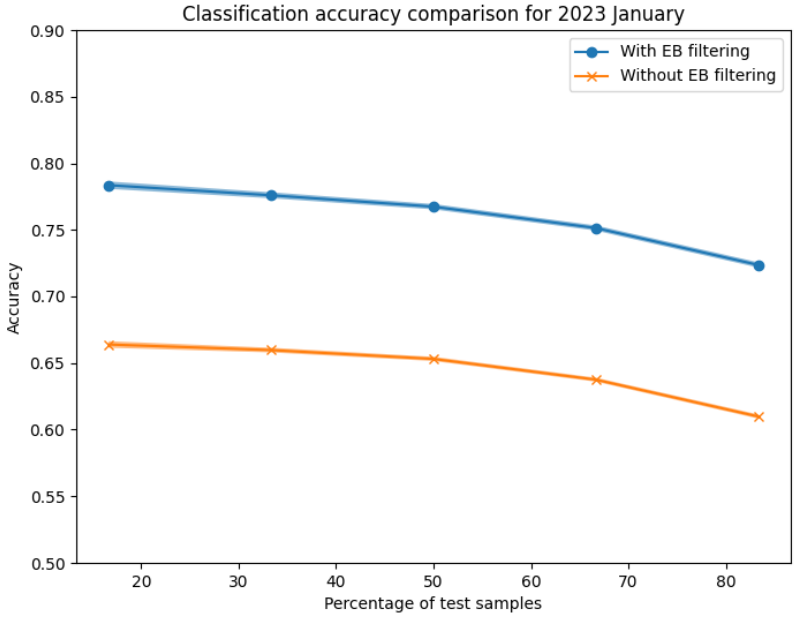

FIG. 4. Accuracy comparison for different proportion of test samples.

Next we check whether the improvement can be observed consistently over different time periods. For this purpose we consider the first five months in the same year, 2023, and two test

train split at 1:2 and 2:1 ratio respectively. We repeat the experiment 1000 time in each scenario and measure the average accuracy for the regression processes, with and without the EB filtering. Fig. 5 presents a comparison bar plot for the two methods. As it can be seen below, the regression performed post the EB-filtering maintains similar increased accuracy for all the five months.

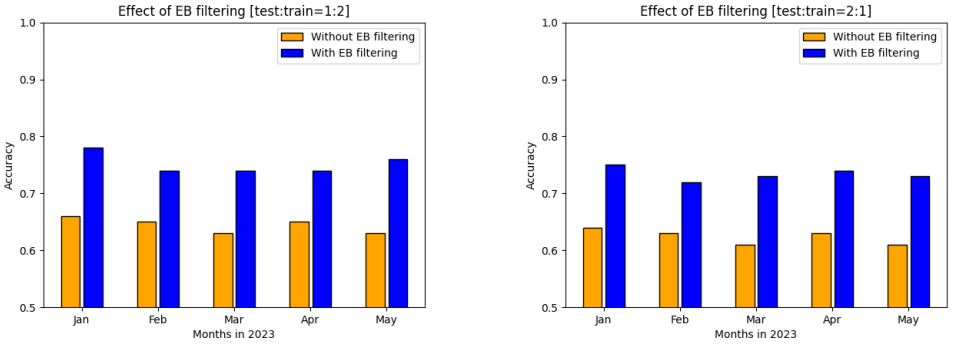

FIG. 5. Accuracy of borough classification over different months.

5.3. *More simulation studies*

In this subsection, we test more priors in addition to the uniform prior in Fig. 1, including discrete priors and priors with unbounded support. In Section 5.2.1 we see that the three minimum-distance estimators performed similarly. However, question arises whether the best choice among the minimum-distance EB methods can be argued when some information about the prior is available. With the specific goal of differentiating the three minimum-distance estimators among themselves, we carry out simulation studies in the end of this section using different priors.

For comparing the EB methods in the discrete setup we choose the prior G to be $0.2\text{Poi}(1) + 0.3\text{Poi}(2) + 0.5\text{Poi}(8)$ and for the continuous unbounded setup we choose the prior G to be the Gamma distribution with scale parameter 2 and shape parameter 4, i.e. with prior density $f(x) = \frac{1}{96}x^3e^{-\frac{x}{2}}$. In both of the cases we simulate $\{\theta_i\}_{i=1}^{600}$ independently from the prior distribution and correspondingly generate data $Y_i \sim \text{Poi}(\theta_i)$. For each of the priors we calculate the Bayes estimator numerically (denoted by the black dashed line in the plots). Then from the generated datasets we compute the Robbins estimator, the NPMLE based EB estimator, the H^2-distance based EB estimator and the χ^2-distance based EB estimator. All the estimators are then plotted against θ and the data (Fig. 6). As expected, the Robbins estimator shows high deviation from the true θ values in many instances whereas the minimum-distance based estimators are much more stable.

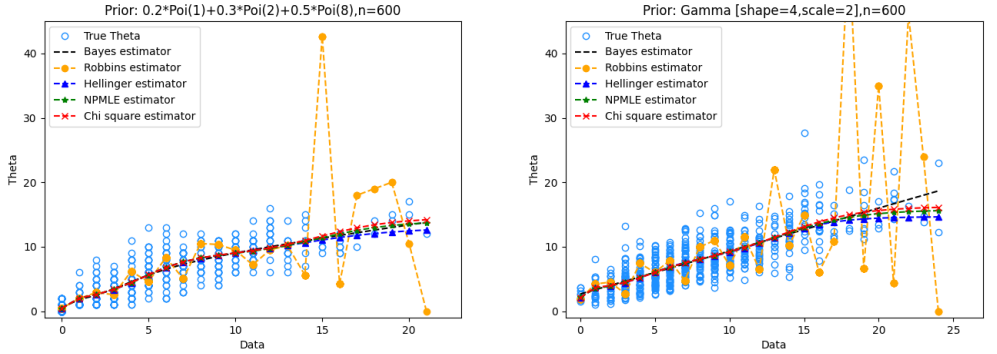

FIG. 6. Robbins vs. minimum-distance estimators: Unbounded priors

To differentiate the different minimum-distance based EB methods we analyze the effect of the tail properties of the prior in the simulations below. Consider the exponential distribution parameterized by scale (α) and with density $g_\alpha(x) = \frac{1}{\alpha}e^{-x/\alpha}$. Note that the higher values of α generate distributions with heavier tails. We consider three values of α: 0.3,1.05 and 2. For each α we estimate the training regret for sample sizes n in the range $[50, 300]$. Given sample $Y_1, \ldots, Y_n$ from the mixture distribution with prior G we define the training regret for any estimator $\widehat{G}$ of G as $\mathbb{E}_G[\frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_G(Y_i) - \widehat{\theta}_{\widehat{G}}(Y_i))^2]$. We compute the Bayes estimator $\widehat{\theta}_G(y)$ numerically for each y. For every pair (α, n) we replicate the following experiment independently 10,000 times for each minimum-distance method:

- Generate $\{\theta_i\}_{i=1}^n$ and $Y_i \sim \text{Poi}(\theta_i)$,
- Calculate $\widehat{G}$ using minimum-distance method,
- Calculate prediction error $\mathsf{E}(Y^n) = \frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_G(Y_i) - \widehat{\theta}_{\widehat{G}}(Y_i))^2$.

Then we take the average of $\mathsf{E}(Y^n)$ values from all the 10,000 replications to estimate the training error. For each α and each minimum distance method, at every n we also estimate the 95% confidence interval as $[\overline{\mathsf{E}(Y^n)} \mp 0.0196 * \mathsf{sd}(\mathsf{E}(Y^n))]$ where $\overline{\mathsf{E}(Y^n)}$ and $\mathsf{sd}(\mathsf{E}(Y^n))$ define respectively the sample mean and the sample standard deviation of the $\mathsf{E}(Y^n)$ values over the 10,000 independent runs. Below we plot the training regrets and their 95% confidence bands against the training sample sizes (Fig. 7).

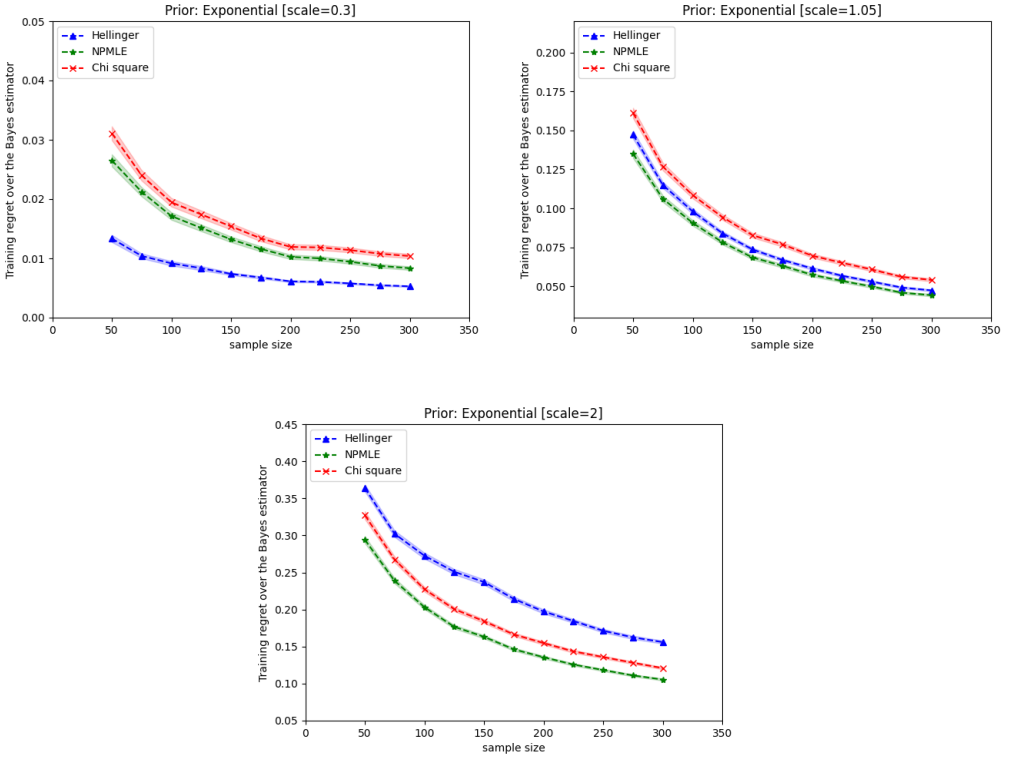

FIG. 7. Comparison of minimum-distance estimators

We observe that that minimum-H^2 based estimator outperforms the other estimators when the scale of the exponential distribution is small. As the tails of the prior distributions become heavier, the performance of the minimum-H^2 based estimator gets worse and the NPMLE based estimator comes out as a better choice.

6. Results in multiple dimensions

The minimum distance estimator (2.2) can be easily extended to the d-dimension Poisson model. For clarity, we use the bold fonts to denote a vector, e.g., $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d), \boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id}), \boldsymbol{Y} = (Y_1, \ldots, Y_d), \boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id}), \boldsymbol{y} = (y_1, \ldots, y_d)$, etc. Let G be a prior distribution on $\mathbb{R}_+^d$. Consider the following data-generating process

$$\boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} G \qquad Y_{ij} \overset{\text{ind.}}{\sim} \text{Poi}(\theta_{ij}). \tag{6.1}$$

Note that the marginal distribution of the multidimensional Poisson mixture is given by

$$f_G(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \prod_{i=1}^{d} e^{-\theta_i} \frac{\theta_i^{y_i}}{y_i!} dG(\boldsymbol{\theta}), \quad \boldsymbol{y} \in \mathbb{Z}_+^d.$$

To construct the multidimensional minimum distance estimator we use the same minimization as in (2.2), the only difference being that now the prior class used for the minimization resides in $\mathbb{R}_+^d$. Next we construct the empirical Bayes estimator. Denote by $\widehat{\boldsymbol{\theta}}_G$ the Bayes estimator, whose i-th coordinate $\widehat{\theta}_{G,i}$ is given by

$$\widehat{\theta}_{G,j}(\boldsymbol{y}) = \mathbb{E}_G[\theta_j|\boldsymbol{y}] = \frac{\int_{\boldsymbol{\theta}} \theta_j \prod_{j=1}^d e^{-\theta_j} \frac{\theta_j^{y_j}}{y_j!} dG(\boldsymbol{\theta})}{f_G(\boldsymbol{y})} = (y_j+1)\frac{f_G(\boldsymbol{y}+\boldsymbol{e}_j)}{f_G(\boldsymbol{y})}, \quad j = 1,\ldots,d,$$

where $\boldsymbol{e}_i$ denote the i-th coordinate vector. Suppose that $\widehat{G}$ gives us an estimate of the prior distribution G and consider the corresponding empirical Bayes estimator

$$\widehat{\boldsymbol{\theta}}_{\widehat{G}} = (\widehat{\theta}_{\widehat{G},1},\ldots,\widehat{\theta}_{\widehat{G},d}).$$

Similar to (2.8), let us define the regret of any plug-in estimator given estimator $\widehat{G}$ as

$$\mathsf{Regret}(\widehat{G},G) = \mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_G(\boldsymbol{Y})\|^2\right] = \mathbb{E}_G\left[\sum_{\boldsymbol{y}\in\mathbb{Z}_+^d} \|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{y}) - \widehat{\boldsymbol{\theta}}_G(\boldsymbol{y})\|^2 f_G(\boldsymbol{y})\right], \tag{6.2}$$

where $\boldsymbol{Y} \sim f_G$ is a test point independent from the training sample $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n \overset{\text{i.i.d.}}{\sim} f_G$. We will prove regret bounds for the minimum distance estimator of the form (2.2) where the

Assumption 3 (Loss function in multi-dimension) *There exist maps $\boldsymbol{t} : \mathscr{P}(\mathbb{Z}_+^d) \to \mathbb{R}$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ such that for any two distributions $q_1, q_2 \in \mathscr{P}(\mathbb{Z}_+^d)$*

$$\mathsf{dist}\,(q_1\|q_2) = \boldsymbol{t}(q_1) + \sum_{\boldsymbol{y}\in\mathbb{Z}_+^d} \ell(q_1(\boldsymbol{y}), q_2(\boldsymbol{y})),$$

where $b \mapsto \ell(a,b)$ is strictly decreasing and strictly convex for each $a > 0$ and $\ell(0,b) = 0$ for each $b \geq 0$.

The above assumption is identical to Assumption 1 except for the fact that now the summation is defined over the set of all d-tuple of non-negative integers. Nonetheless, this is just a matter of notations and loss functions such as the Kullback-Leibler divergence, squared Hellinger distance, Chi-squared divergence still satisfy the above assumption.

Theorem 5 *Suppose that $\widehat{G}$ is a minimum distance estimator satisfying (2.2), with the distance functional* dist *obeying Assumption 2 and Assumption 3. Then the corresponding minimum distance EB estimator attains the following regret bounds:*

1. *If G is supported on $[0,h]^d$, then $\mathbb{E}\left[H^2(f_{\widehat{G}}, f_G)\right] \leq O(\frac{1}{n}\max\{c_1, c_2 h\}^d (\frac{\log(n)}{\log\log(n)})^d)$;*
2. *If all marginals of G belong to the $\mathsf{SubE}(s)$ class of distributions for some $s > 0$, then $\mathbb{E}\left[H^2(f_{\widehat{G}}, f_G)\right] \leq O(\frac{1}{n}(\max\{c_3, c_4 s\}\log(n))^d)$,*

where $c_1, c_2, c_3, c_4 > 0$ are absolute constants.

Theorem 6 *Suppose that $\widehat{G}$ is a minimum distance estimator satisfying (2.2), with the distance functional* dist *obeying Assumption 2 and Assumption 3. Then the corresponding minimum distance EB estimator attains the following regret bounds whenever $n \geq d$:*

1. *If G is supported on $[0,h]^d$ and $\widehat{G}$ is chosen to be the solution over $[0,h]^d$, then* $\mathsf{Regret}(\widehat{G};G) \leq O(\frac{d}{n}\max\{c_1, c_2 h\}^{d+2}(\frac{\log(n)}{\log\log(n)})^{d+1})$ *;*
2. *If all marginals of G belong to the* $\mathsf{SubE}(s)$ *class of distributions for some $s > 0$, then* $\mathsf{Regret}(\widehat{G};G) \leq O(\frac{d}{n}(\max\{c_3, c_4 s\}\log(n))^{d+2})$,

where $c_1, c_2, c_3, c_4 > 0$ are absolute constants.

The proof of the above results are identical as in the one dimensional case, and they are provided in Appendix D below. We conjecture these regret bounds in Theorem 6 are nearly optimal and factors like $(\log n)^d$ are necessary. Indeed, for the Gaussian model in d dimensions, the minimax squared Hellinger risk for density estimation is shown to be at least $O((\log n)^d/n)$ for subgaussian mixing distributions and the minimax regret is typically even larger. A rigorous proof of matching lower bound for Theorem 6 will likely involve extending the regret lower bound based on Bessel kernels in [PW21] to multiple dimensions; this is left for future work.

REFERENCES

Ber55. Joseph Berkson. Maximum likelihood and minimum χ^2 estimates of the logistic function. *Journal of the American statistical association*, 50(269):130–162, 1955.

Ber77. Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463, 1977.

BG09. Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

BGR13. Lawrence D Brown, Eitan Greenshtein, and Ya'acov Ritov. The poisson compound decision problem revisited. *Journal of the American Statistical Association*, 108(502):741–749, 2013.

BM72. G Kemble Bennett and HF Martz. A continuous empirical Bayes smoothing technique. *Biometrika*, 59(2):361–368, 1972.

Bol77. E Bolthausen. Convergence in distribution of minimum-distance estimators. *Metrika*, 24(1):215–227, 1977.

Bro08. Lawrence D Brown. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 2(1):113–152, 2008.

Cas85. George Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

Che95. Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23:221–233, 1995.

Che17. Jiahua Chen. Consistency of the MLE under mixture models. *Statistical Science*, 32(1):47–63, 2017.

Csi67. I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

DK68. JJ Deely and RL Kruse. Construction of sequences estimating the mixing distribution. *The Annals of Mathematical Statistics*, 39(1):286–288, 1968.

Ede88. David Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622, 1988.

Efr14. Bradley Efron. Two modeling strategies for empirical Bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.

Efr21. Bradley Efron. Empirical Bayes: Concepts and Methods. 2021. http://statweb.stanford.edu/~ckirby/brad/papers/2021EB-concepts-methods.pdf.

EH21. Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press, 2021.

ETST01. Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

Fed72. Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

GvdV01. S. Ghosal and A.W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, 2001.

HK18. Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

HS83. JC van Houwelingen and Th Stijnen. Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Statistica Neerlandica*, 37(1):29–43, 1983.

HS84. James Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984.

Jew82. Nicholas P Jewell. Mixtures of exponential distributions. *The annals of statistics*, pages 479–484, 1982.

JPTW23. Soham Jana, Yury Polyanskiy, Anzo Teh, and Yihong Wu. Empirical bayes via erm and rademacher complexities: the poisson model. 2023.

JZ09. Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

KG17. Roger Koenker and Jiaying Gu. Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82:1–26, 2017.

Kim14. Arlene KH Kim. Minimax bounds for estimation of Normal mixtures. *bernoulli*, 20(4):1802–1818, 2014.

KM14. Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

KW56. Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

Lai82. Nan M Laird. Empirical Bayes estimates using the nonparametric maximum likelihood estimate for the prior. *Journal of Statistical Computation and Simulation*, 15(2-3):211–220, 1982.

LGL05. Jianjun Li, Shanti S Gupta, and Friedrich Liese. Convergence rates of empirical Bayes estimation in exponential family. *Journal of statistical planning and inference*, 131(1):101–115, 2005.

Lin83a. Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pages 86–94, 1983.

Lin83b. Bruce G Lindsay. The geometry of mixture likelihoods, part II: the Exponential family. *The Annals of Statistics*, 11(3):783–792, 1983.

Lin95. Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

LK69. Glen H Lemon and Richard G Krutchkoff. An empirical Bayes smoothing technique. *Biometrika*, 56(2):361–365, 1969.

LT84. Diane Lambert and Luke Tierney. Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *The Annals of Statistics*, pages 1388–1399, 1984.

Mar66. JS Maritz. Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika*, 53(3-4):417–429, 1966.

Mar68. JS Maritz. On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. *Biometrika*, 55(1):83–100, 1968.

Mar69. JS Maritz. Empirical bayes estimation for the Poisson distribution. *Biometrika*, 56(2):349–359, 1969.

Mil84. PW Millar. A general approach to the optimality of minimum distance estimators. *Transactions of the American Mathematical Society*, 286(1):377–418, 1984.

MKV+21. Zhen Miao, Weihao Kong, Ramya Korlakai Vinayak, Wei Sun, and Fang Han. Fisher-pitman permutation tests based on nonparametric Poisson mixtures with application to single cell genomics. *arXiv preprint arXiv:2106.03022*, 2021.

ML18. Johannes S Maritz and T Lwin. *Empirical Bayes methods*. Chapman and Hall/CRC, 2018.

Mor83. Carl N Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55, 1983.

Pfa88. Johann Pfanzagl. Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference*, 19(2):137–158, 1988.

PLLB10. Bhagwant Persaud, Bo Lan, Craig Lyon, and Ravi Bhim. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention*, 42(1):38–43, 2010.

Pol80. David Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980.

PS98. G Pólya and G Szegö. *Problems and Theorems in Analysis II,(reprint ed.)*. Springer, Heidelburg, 1998.

PW21. Yury Polyanskiy and Yihong Wu. Sharp regret bounds for empirical Bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*, 2021.

Rob51. Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pages 131–149. University of California Press, 1951.

Rob56. Herbert Robbins. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.

Sim76. Leopold Simar. Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, 4(6):1200–1209, 1976.

VH96. Jay M Ver Hoef. Parametric empirical Bayes methods for ecological applications. *Ecological Applications*, 6(4):1047–1055, 1996.

Wol53. J Wolfowitz. Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, 5(1):9–23, 1953.

Wol54. J Wolfowitz. Estimation by the minimum distance method in nonparametric stochastic difference equations. *The Annals of Mathematical Statistics*, 25(2):203–217, 1954.

Wol57. Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

WV10. Yihong Wu and Sergio Verdú. Functional properties of MMSE. In *2010 IEEE International Symposium on Information Theory*, pages 1453–1457. IEEE, 2010.

Wyn70. Henry P Wynn. The sequential generation of d-optimum experimental designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664, 1970.

Zha03. Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *Annals of Statistics*, pages 379–390, 2003.

Zha09. Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

A. Proof of Theorem 1

We first prove the result for the constrained solution $\operatorname{argmin}_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q)$. As mentioned towards the end of the proof, this also implies the desired result for the unconstrained solution. Suppose that p is supported on $\{y_1,\dots,y_m\}\subset\mathbb{Z}_+$. Define

$$S \triangleq \{(f_Q(y_1),\dots,f_Q(y_m)) : Q\in\mathscr{P}([0,h])\}, \tag{A.1}$$

where $f_Q(y)=\mathbb{E}_{\theta\sim Q}[f_\theta(y)]$ is the probability mass function of the Poisson mixture (1.2), and $f_\theta(y)=e^{-\theta}\theta^y/y!$. We claim that S is convex and compact.[3] The convexity follows from definition. For compactness, note that S is bounded since $\sup_{\theta\geq 0} f_\theta(y)=e^{-y}y^y/y!$, so it suffices to check S is closed. Let $(f'_1,\dots,f'_m)\in\mathbb{R}_+^m$ be the limiting point of $(f_{Q_k}(y_1),\dots,f_{Q_k}(y_m))$ for some sequence $\{Q_k\}$ in $\mathscr{P}([0,h])$. By Prokhorov's theorem, there is a subsequence $\{Q_{k_\ell}\}$ that converges weakly to some $Q'\in\mathscr{P}([0,h])$. Since $\theta\mapsto f_\theta(y)$ is continuous and bounded, we have $f'_j=f_{Q'}(y_j)$ for all j. In other words, S is closed.

Next, define $v: S\to\mathbb{R}$ by $v(f_1,\dots,f_m)=\sum_{i=1}^m \ell(p(y_i),f_i)$. By Assumption 1, the value of the min-distance optimization can be written as

$$\min_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q) = t(p) + \min_{(f_1,\dots,f_m)\in S} v(f_1,\dots,f_m). \tag{A.2}$$

Furthermore, by assumption $\ell(0,b)\equiv 0$ and $b\mapsto\ell(a,b)$ is strictly convex for $a>0$. Thus v is strictly convex. Therefore, there exists a unique point $(f_1^*,\dots,f_m^*)\in S$ that achieves the minimum on the right side of (A.2). Thus, the left side has a minimizer $\widehat{G}\in\mathscr{P}([0,h])$ that satisfies

$$f_{\widehat{G}}(y_j)=f_j^*, j=1,\dots,m \tag{A.3}$$

It remains to show that the above representation is unique at the special point $(f_1^*,\dots,f_m^*)$; this argument relies on the specific form of the Poisson density. Let $\widehat{G}$ be one such minimizer. By the first-order optimality condition (see (5.2) in Section 5.1),

$$\begin{aligned} &D_{\widehat{G}}(\theta)=\sum_{i=1}^m a_i(f_\theta(y_i)-f_i^*)\geq 0, \quad \forall 0\leq\theta\leq h;\\ &D_{\widehat{G}}(\theta)=0, \quad \text{for } \widehat{G}\text{-almost every } \theta, \end{aligned} \tag{A.4}$$

where $a_i\triangleq\frac{d}{df}\ell(p(y_i),f)|_{f=f_i^*}<0$, since ℓ is strictly decreasing in the second coordinate and $f_i^*>0$. Define

$$b_i=\frac{a_i}{\sum_{i=1}^m a_i f_i^*}>0.$$

[3] In this case, S is in fact the closed convex hull of the set $\{(f_\theta(y_1),\dots,f_\theta(y_m)) : \theta\in[0,h]\}$.

As ℓ is strictly decreasing in second coordinate, $\frac{d}{df}\ell(p(y_i),f) < 0$ for all $f \in \mathbb{R}_+, i = 1,\ldots,m$. Using this, we rearrange (A.4) to get

$$\begin{aligned}
&\sum_{i=1}^{m} \frac{b_i}{y_i!}\theta^{y_i} \le e^{\theta}, \forall \theta \in [0,h],\\
&\sum_{i=1}^{m} \frac{b_i}{y_i!}\theta^{y_i} = e^{\theta} \text{ for each } \theta \text{ in the support of } \widehat{G}.
\end{aligned} \tag{A.5}$$

Then the following lemma shows that the support of $\widehat{G}$ has at most m points.

Lemma 7 *Suppose that $\sum_{i=1}^{m}\beta_i\theta^{y_i} \le e^{\theta}$ for all $\theta \in [0,h]$ where $\beta_i \in \mathbb{R}$ and $h > 0$. Then the number of solutions to $\sum_{i=1}^{m}\beta_i\theta^{y_i} = e^{\theta}$ in $\theta \in [0,h]$ is at most m.*

Proof The proof is a modification of [Sim76, Lemma 3.1(2)], which deals with the specific case $h = \infty$. Recall the following version of Descartes' rule of signs [PS98, Part V, Problem 38 and 40]: Consider an entire function (i.e., a power series whose radius of convergence is infinity) $\phi(x) = a_0 + a_1x + a_2x^2 + \ldots$ with real coefficients. Let r be the number of strictly positive zeros of ϕ counted with their multiplicities and let s be the number of sign changes[4] in the sequence $a_0, a_1, \ldots$. Then $r \le s$. We apply this fact to the function

$$\phi(\theta) = \sum_{i=1}^{m}\beta_i\theta^{y_i} - e^{\theta} = \sum_{j=0}^{\infty} a_j\theta^j,$$

where

$$a_j = \begin{cases} \beta_i - \frac{1}{y_i!} & j = y_i, i = 1,\ldots,m \\ -\frac{1}{j!} < 0 & \text{else} \end{cases}$$

Case 1: Suppose that 0 is a root of $\phi(\cdot)$. Then $a_0 = 0$. As there are at most $m-1$ positive coefficients in $a_0, a_1, \ldots$, there can be at most $2(m-1)$ sign changes, which implies at most $2(m-1)$ positive roots of s counting multiplicities. Note that, as $\phi(\theta)\mathbf{1}_{\{\theta\in(0,h)\}} \le 0$ and s is an entire function, each root of s inside $(0,h)$ has multiplicity at least 2. Suppose that m_h is the multiplicity of h as a root of $\phi(\cdot)$, which we define to be 0 when h is not a root. This means that the total number of distinct roots in $(0,h)$ is at most the largest integer before $(2(m-1)-m_h)/2$. If h is not a root then the number distinct roots in $(0,h)$ is at most $m-1$. If h is a root, then its multiplicity is at least 1, and hence, the number of distinct roots in $(0,h)$ is at most $m-2$. Hence, there are at most m many distinct roots in $[0,h]$.

Case 2: Suppose that 0 is not a root of $\phi(\cdot)$. As there are at most m positive coefficients in $a_0, a_1, \ldots$, there can be at most $2m$ sign changes, which implies at most $2m$ positive roots counting multiplicities. By a similar argument as in the previous case, the total number of distinct roots in $(0,h)$ is at most than the largest integer before $(2m-m_h)/2$. If h is not a root then the number

[4] The number of sign changes is the number of pairs $0 \le i < j$ such that $a_ia_j < 0$ and either $j = i+1$ or $a_k = 0$ for all $i < k < j$.

distinct roots in $(0,h)$ is at most m. If h is a root, then the number of distinct roots in $(0,h)$ is at most $m-1$. Hence, in total at most m distinct roots in $[0,h]$. $\square$

Suppose that there are $r(\leq m)$ different θ_i's (denote them by $\theta_1 \ldots, \theta_r$) for which (A.5) holds. This implies given any optimizer $\widehat{G}$ its atoms form a subset of $\{\theta_1 \ldots, \theta_r\}$. Let w_j be the weight $\widehat{G}$ puts on θ_j. Then in view of (A.3) we get that

$$\sum_{j=1}^{r} w_j e^{-\theta_j} \theta_j^{y_i} = f_i^* y_i!, \quad i = 1, \ldots, r.$$

The matrix $\{\theta_j^{y_i} : j = 1, \ldots, r, i = 1, \ldots, m\}$ has full column rank, and hence the vector $(w_1, \ldots, w_r)$ can be solve uniquely. This implies uniqueness of the optimizer $\widehat{G}$ as well. This finishes the proof for the constrained solution.

Next we argue for the unconstrained minimizer $\operatorname{argmin}_Q \mathsf{dist}(p\|f_Q)$. In view of Lemma 8 below, we get that the unconstrained minimum-distance estimator is supported on $[0,h]$ with $h = \max_{i=1,\ldots,m} y_i$. Then from the above proof for $\operatorname{argmin}_{Q \in \mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q)$ the existence and uniqueness of the unconstrained estimator follow.

Lemma 8 *Let* dist *satisfy Assumption 1 and let p be a probability distribution on* $\mathbb{Z}_+$ *with support* $\{y_1, \ldots, y_m\}$. *Then the minimizer* $\operatorname{argmin}_Q \mathsf{dist}(p\|f_Q)$ *is supported on the interval* $[y_{\min}, y_{\max}]$, *where* $y_{\min} = \min_{i=1,\ldots,m} y_i, y_{\max} = \max_{i=1,\ldots,m} y_i$.

Proof Let Q be a distribution with $Q([0,y_{\min})) + Q((y_{\max},\infty)) > 0$. Define another distribution $\widetilde{Q}$ by

$$\widetilde{Q}(\cdot) = Q([0,y_{\min}))\delta_{y_{\min}}(\cdot) + Q(\cdot \cap [y_{\min}, y_{\max}]) + Q((y_{\max},\infty))\delta_{y_{\max}}(\cdot).$$

In other words, $\widetilde{Q}$ moves the masses of Q on the intervals $[0,y_{\min})$ (resp. $(y_{\max},\infty)$) to the point $y_{\min}$ (resp. $y_{\max}$). As $f_\theta(y)$ is strictly increasing in $\theta \in [0,y)$ and strictly decreasing in $\theta \in (y,\infty)$ we get for each $i = 1, \ldots, n$

$$\begin{aligned}
f_Q(y_i) &= \int f_\theta(y_i) dQ(\theta) \\
&= \int_{0 \leq \theta < y_{\min}} f_\theta(y_i) dQ(\theta) + \int_{y_{\min} \leq \theta \leq y_{\max}} f_\theta(y_i) dQ(\theta) + \int_{y_{\max} > \theta} f_\theta(y_i) dQ(\theta) \\
&< Q([0,y_{\min}))f_{y_{\min}}(y_i) + \int_{y_{\min} \leq \theta \leq y_{\max}} f_\theta(y_i) dQ(\theta) + Q((y_{\max},\infty))f_{y_{\max}}(y_i) \\
&= \int f_\theta(y_i) d\widetilde{Q}(\theta) = f_{\widetilde{Q}}(y_i).
\end{aligned}$$

Hence, by Assumption 1, we get

$$\begin{aligned}\mathsf{dist}(p\|f_Q) &= t(p) + \sum_{y:p(y)>0} \ell(p(y), f_Q(y)) \\ &\overset{\text{(a)}}{=} t(p) + \sum_{y:p(y)>0} \ell(p(y), f_Q(y)) \overset{\text{(b)}}{>} t(p) + \sum_{y:p(y)>0} \ell\left(p(y), f_{\widetilde{Q}}(y)\right) = \mathsf{dist}(p\|f_{\widetilde{Q}}),\end{aligned} \tag{A.6}$$

where (a) follows from $\ell(0,\cdot) = 0$; (b) follows as the function $b \mapsto \ell(a,b)$ is strictly decreasing. In other words, given any Q with $Q([0,y_{\min})) + Q((y_{\max},\infty)) > 0$ we can produce $\widetilde{Q}$ supported on $[y_{\min}, y_{\max}]$ such that $\mathsf{dist}(p\|f_{\widetilde{Q}}) < \mathsf{dist}(p\|f_Q)$. Hence, the claim follows. □

B. Proof of Lemma 4

Let $\theta \sim G, Y|\theta \sim f_\theta$. Then for any $\widehat{G}$ independent of Y, we can write $\mathsf{Regret}(\widehat{G};G) = \sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_G(y))^2 f_G(y) = \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_G(Y)\right)^2\right]$; cf. (2.8). Fix $h > 0$ and note the following

- $\mathrm{mmse}(G_h) \le \frac{\mathrm{mmse}(G)}{G([0,h])}$ [WV10, Lemma 2],
- $\mathrm{mmse}(G) \le \sqrt{\mathbb{E}_G[\theta^4]} \le \sqrt{M}$, and
- For any fixed distribution $\widehat{G}$

$$\begin{aligned}\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] &\le \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \mathbf{1}_{\{\theta \le h\}}\right] + \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \mathbf{1}_{\{\theta > h\}}\right] \\ &\overset{\text{(a)}}{\le} \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \middle| \theta \le h\right] + \sqrt{\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^4\right] \mathbb{E}_G\left[\mathbf{1}_{\{\theta > h\}}\right]} \\ &\overset{\text{(b)}}{\le} \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] + \sqrt{8(\widehat{h}^4 + \mathbb{E}_G[\theta^4])G((h,\infty))} \\ &= \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] + \sqrt{8(\widehat{h}^4 + M)G((h,\infty))}.\end{aligned}$$

where step (a) followed by Cauchy-Schwarz inequality and step (b) followed as $(x+y)^4 \le 8(x^4+y^4)$ for any $x, y \in \mathbb{R}$.

Using these we get

$$\begin{aligned}
\mathsf{Regret}(\widehat{G};G) &= \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right] - \mathrm{mmse}(G) \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right] - \mathrm{mmse}(G_h) + \mathrm{mmse}(G_h) - \mathrm{mmse}(G) + \sqrt{8(\widehat{h}^4+M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right] + \left(\frac{1}{G([0,h])}-1\right)\mathrm{mmse}(G) + \sqrt{8(\widehat{h}^4+M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right] + \frac{G((h,\infty))}{G([0,h])}\sqrt{M} + \sqrt{8(\widehat{h}^4+M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right] + \frac{(1+2\sqrt{2})\sqrt{(\widehat{h}^4+M)G((h,\infty))}}{G([0,h])}. \qquad \text{(B.1)}
\end{aligned}$$

Next we bound the first term. Fix $K \ge 1$. Using $\widehat{\theta}_{G_h}(y) \le h, \widehat{\theta}_{\widehat{G}}(y) \le \widehat{h}$ we have

$$\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\mathbf{1}_{\{Y\le K-1\}}\right] \\
&= \sum_{y=0}^{K-1}(y+1)^2 f_{G_h}(y)\left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2 \\
&\overset{(a)}{\le} \sum_{y=0}^{K-1}(y+1)^2 f_{G_h}(y)\left\{3\left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{2f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2 + 3\left(\frac{f_{G_h}(y+1)}{f_{G_h}(y)} - \frac{2f_{G_h}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2\right. \\
&\quad \left. +3\left(\frac{2f_{G_h}(y+1)-2f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2\right\} \\
&\le 3\sum_{y=0}^{K-1}\left\{\left(\frac{(y+1)f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)}\right)^2\frac{(f_{G_h}(y)-f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + \left(\frac{(y+1)f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2\frac{(f_{G_h}(y)-f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right. \\
&\quad \left. +4(y+1)^2\frac{(f_{G_h}(y+1)-f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right\} \\
&= 3(\{\widehat{\theta}_{G_h}(y)\}^2+\{\widehat{\theta}_{\widehat{G}}(y)\}^2)\sum_{y=0}^{K-1}\frac{(f_{G_h}(y)-f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1}(y+1)^2\frac{(f_{G_h}(y+1)-f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} \\
&\le 3(h^2+\widehat{h}^2)\sum_{y=0}^{K-1}\frac{(f_{G_h}(y)-f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1}(y+1)^2\frac{(f_{G_h}(y+1)-f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)}
\end{aligned}$$

where (a) followed from $(x+y+z)^2 \le 3(x^2+y^2+z^2)$ for any $x,y,z \in \mathbb{R}$. Using $(\sqrt{f_{G_h}(x)} + \sqrt{f_{\widehat{G}}(x)})^2 \le 2(f_{G_h}(x)+f_{\widehat{G}}(x))$ for $x = y, y+1$ we continue the last display to get

$$\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \le K-1\}}\right] \\
&\le 6(h^2+\widehat{h}^2)\sum_{y=0}^{K-1}(\sqrt{f_{G_h}(y)} - \sqrt{f_{\widehat{G}}(y)})^2 \\
&\quad + 24K \max_{y=0}^{K-1} \frac{(y+1)f_{G_h}(y+1) + (y+1)f_{\widehat{G}}(y+1)}{f_{G_h}(y) + f_{\widehat{G}}(y)} \sum_{y=0}^{K-1}(\sqrt{f_{G_h}(y+1)} - \sqrt{f_{\widehat{G}}(y+1)})^2 \\
&\le \left(6(h^2+\widehat{h}^2) + 24(h+\widehat{h})K\right) H^2(f_{\widehat{G}}, f_{G_h}).
\end{aligned}$$

Again using $\widehat{\theta}_{G_h}(y) \le h, \widehat{\theta}_{\widehat{G}}(y) \le \widehat{h}$ we bound $\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \ge K\}}\right]$ by $(h+\widehat{h})^2 \varepsilon_K(G_h)$. Combining this with the last display we get

$$\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2\right] \le \left\{6(h^2+\widehat{h}^2) + 24(h+\widehat{h})K\right\} H^2(f_{\widehat{G}}, f_{G_h}) + (h+\widehat{h})^2 \varepsilon_K(G_h).$$

In view of above continuing (B.1) we have

$$\begin{aligned}
\mathsf{Regret}(\widehat{G}; G) \le &\left\{6(h^2+\widehat{h}^2) + 24(h+\widehat{h})K\right\} H^2(f_{\widehat{G}}, f_{G_h}) \\
&+ (h+\widehat{h})^2 \varepsilon_K(G_h) + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}.
\end{aligned} \tag{B.2}$$

Using triangle inequality and $(x+y)^2 \le 2(x^2+y^2)$ we get

$$H^2(f_{\widehat{G}}, f_{G_h}) \le 2\left\{H^2(f_G, f_{\widehat{G}}) + H^2(f_{G_h}, f_G)\right\}. \tag{B.3}$$

Note that

$$H^2(f_{G_h}, f_G) \le 2\mathrm{TV}(f_{G_h}, f_G) \le 2\mathrm{TV}(G_h, G) = 2G((h,\infty)).$$

where TV denotes the total variation and the middle inequality applies the data-processing inequality [Csi67].

Then, combining (B.2), (B.3) and (3.7) we get

$$\begin{aligned}
\mathsf{Regret}(\widehat{G}; G) \le &\left\{12(h^2+\widehat{h}^2) + 48(h+\widehat{h})K\right\}\left(H^2(f_{\widehat{G}}, f_G) + 2G((h,\infty))\right) \\
&+ (h+\widehat{h})^2 \frac{\varepsilon_K(G)}{G([0,h])} + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}
\end{aligned}$$

This finishes the proof.

C. Auxiliary results

Given any $s > 0$ and $G \in \mathsf{SubE}(s)$, the following are satisfied.

1. If $\theta \sim G$, then $\mathbb{E}[\theta^4] \le 30s^4$.
2. If $\{Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} f_G$, then

$$\mathbb{P}[Y_1 \ge K] \le \frac{3}{2} e^{-K \log\left(1+\frac{1}{2s}\right)}, \quad \mathbb{E}\left[Y_{\max}^4\right] \le \frac{64(\log n)^4 + 45}{\left(\log\left(1+\frac{1}{2s}\right)\right)^4}. \tag{C.1}$$

To prove (i) we note that $z^3 \le 15 e^{\frac{z}{2}}, z \in \mathbb{R}$. Then we have

$$\mathbb{E}[\theta^4] = 4\int y^3 \mathbb{P}[\theta > y]\, dy \le 2\int y^3 e^{-\frac{y}{s}} dy \le 30 s^4 \int e^{-\frac{z}{2}} dz \le 30 s^4.$$

The proof of the property (ii) is as follows. Using $\mathbb{E}_{Z\sim\mathrm{Poi}(\theta)}\left[e^{Zt}\right] = e^{\theta(e^t-1)}, t > 0$ and denoting $c(s) = \log\frac{1+2s}{2s}$ we have

$$\begin{aligned}
&\mathbb{E}\left[e^{Y_1 c(s)}\right] = \mathbb{E}_{\theta\sim G}\left[\mathbb{E}_{Y_1\sim\mathrm{Poi}(\theta)}\left[e^{Y_1 c(s)}\,\middle|\,\theta\right]\right] = \mathbb{E}_G\left[e^{\frac{\theta}{2s}}\right] = \int e^{\theta/2s} G(d\theta) \\
&= \int_\theta \int_{x<\theta} \frac{e^{x/2s}}{2s} dx G(d\theta) = \int_x \frac{e^{x/2s}}{2s} G([x,\infty))dx \overset{\text{(a)}}{\le} \int_{x<0} \frac{e^{x/2s}}{2s} dx + \int_{x>0} \frac{e^{-x/2s}}{s} dx \le \frac{3}{2}
\end{aligned}$$

where (a) followed by using tail bound for $\mathsf{SubE}(s)$ distribution G. In view of Markov inequality

$$\mathbb{P}[Y_1 \ge K] \le \mathbb{E}\left[e^{Y_1 c(s)}\right] e^{-c(s)K} \le \frac{3}{2} e^{-K\log\left(1+\frac{1}{2s}\right)}. \tag{C.2}$$

For the expectation term we have for any $L > 0$

$$\begin{aligned}
\mathbb{E}\left[(Y_{\max})^4\right] &= 4\int y^3 \mathbb{P}[Y_{\max} > y] \\
&\le 4L^4 + n\int_{y>L} y^3 \mathbb{P}[Y_1 > y]\, dy \\
&\le 4L^4 + \frac{3n}{2}\int_{y>L} y^3 e^{-y\log\left(1+\frac{1}{2s}\right)} dy \\
&\le 4L^4 + \frac{3n}{2\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4} \int_{z>L\log\left(1+\frac{1}{2s}\right)} z^3 e^{-z} dz \\
&\le 4L^4 + \frac{45n}{2\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4} \int_{z>L\log\left(1+\frac{1}{2s}\right)} e^{-z/2} dz \le 4L^4 + \frac{45 n e^{-\frac{L}{2}\log\left(1+\frac{1}{2s}\right)}}{\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4}.
\end{aligned}$$

Choosing $L = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}$ we get the desired result.

D. Proofs of the multidimensional results

D.1. *Density estimation in multiple dimensions*

The proof of Theorem 5 is based on a similar truncation idea as in the proof of Theorem 2. At this end, we note the following result.

Lemma 9 *There exist absolute constants $\widetilde{c}_1, \widetilde{c}_2$ such that the following holds.*

1. *If $G \in \mathscr{P}([0,h]^d)$ and $\boldsymbol{Y} \sim f_G$ then $\mathbb{P}\left[\boldsymbol{Y} \notin [0, \widetilde{c}_1 h \frac{\log n}{\log\log n}]^d\right] \le \frac{d}{n^{10}}$,*
2. *If all the marginals of G lie in in $\mathsf{SubE}(s)$ and $\boldsymbol{Y} \sim f_G$ then $\mathbb{P}\left[\boldsymbol{Y} \notin [0, \widetilde{c}_2 s \log n]^d\right] \le \frac{d}{n^{10}}$.*

Proof From the proof of Theorem 2 we get that there exists constants c_1, c_2 such that with probability at least $1 - \frac{1}{n^{10}}$ all the coordinates of the random variable $\boldsymbol{Y}$ lie within $[0, c_1 h \frac{\log n}{\log\log n}]$ if G is supported on $[0,h]$, and lie within $[0, c_2 \log n]$ if the marginals of G are $\mathsf{SubE}(s)$. Then using a union bound over all the coordinates we achieve the desired result. □

Proof of Theorem 5 Suppose that the dist function, for which we compute the minimum distance estimator, satisfy Assumption 2, namely (2.3). Then recall the general result (3.3)

$$H^2(f_G, f_{\widehat{G}}) \le \frac{2}{c_1}(\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{\widehat{G}}) + \mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)) \le \frac{4}{c_1}\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G), \tag{D.1}$$

where $c_1 > 0$ is an absolute constant. Then, bounding $\frac{1}{c_2}\mathsf{dist}$ by χ^2 we get the following chain for any $K = \widetilde{c}_1 h \frac{\log n}{\log\log n}$ for G supported on $[0,h]^d$ and $K = \widetilde{c}_2 s \log n$ for Ghaving all the marginals in $\mathsf{SubE}(s)$

$$\begin{aligned}
&\frac{1}{c_2}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i=1,\ldots,n\}}\right] \\
&\le \mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i=1,\ldots,n\}}\right] \\
&= \sum_{\boldsymbol{y}} \frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y}) - f_G(\boldsymbol{y}))^2 \mathbf{1}_{\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i=1,\ldots,n\}}\right]}{f_G(\boldsymbol{y})} \\
&\overset{(a)}{=} \sum_{\boldsymbol{y} \in [0,K]^d} \frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y}) - f_G(\boldsymbol{y}))^2 \mathbf{1}_{\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i=1,\ldots,n\}}\right]}{f_G(\boldsymbol{y})} + \sum_{y \notin [0,K]^d} f_G(\boldsymbol{y}) \mathbb{P}[\boldsymbol{Y}_i \in [0,K]^d \ \forall i = 1,\ldots,n] \\
&\le \sum_{\boldsymbol{y} \in [0,K]^d} \frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y}) - f_G(\boldsymbol{y}))^2\right]}{f_G(\boldsymbol{y})} + \mathbb{P}_{\boldsymbol{Y} \sim f_G}[\boldsymbol{Y} \notin [0,K]^d] \\
&\overset{(b)}{\le} \frac{1}{n} \sum_{\boldsymbol{y} \in [0,K]^d} (1 - f_G(\boldsymbol{y})) + \frac{d}{n^{10}} \le \frac{2(K+1)^d}{n}.
\end{aligned} \tag{D.2}$$

where (a) follows from the fact that under $\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i = 1,\dots,n\}$ we have $p_n^{\mathsf{emp}}(\boldsymbol{y}) = 0$ for any $y \notin [0,K]^d$; and (b) follows from $\mathbb{E}[p_n^{\mathsf{emp}}(\boldsymbol{y})] = f_G(\boldsymbol{y})$ and, thus, $\mathbb{E}[(p_n^{\mathsf{emp}}(\boldsymbol{y}) - f_G(\boldsymbol{y}))^2] = \mathrm{Var}(p_n^{\mathsf{emp}}(\boldsymbol{y})) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathbf{1}_{\{\boldsymbol{Y}_i=\boldsymbol{y}\}}) = \frac{f_G(\boldsymbol{y})(1-f_G(\boldsymbol{y}))}{n}$, and due to choices of K.

Using the union bound and the fact $H^2 \le 2$ we have

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\left\{\boldsymbol{Y}_i \notin [0,K]^d \text{ for some } i\in\{1,\dots,n\}\right\}}\right] \le \frac{2d}{n^9}.$$

Combining this with (D.2) yields

$$\begin{aligned}\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] &\le \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\left\{\boldsymbol{Y}_i \in [0,K]^d \ \forall i=1,\dots,n\right\}}\right] + \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\left\{\boldsymbol{Y}_i \notin [0,K]^d \text{ for some } i\in\{1,\dots,n\}\right\}}\right] \\ &\le \frac{4(K+1)^d}{n},\end{aligned}$$

which completes the proof. □

D.2. *Regret bounds in multiple dimensions*

We first note that it suffices to only prove the case where the data generating distribution G satisfies $G \in \mathscr{P}([0,h]^d)$. To prove the case where the marginals of G belong to the $\mathsf{SubE}(s)$ class, it suffices to choose $h = \widetilde{c}s\log n$ where $\widetilde{c} > 0$ is a sufficiently large constant. This is because of the following. Using the property of the Poisson mixture and the the result on the support of $\widehat{G}$ for the one dimensional case in Lemma 8 we get

1. $\widehat{G}$ is supported on $[0, \max_{j=1}^d \max_{i=1}^n Y_{ij} + 1]^n$, which itself is a subset of $[0, \widetilde{c}s\log n]^d$ with probability at least $1 - \frac{d}{n^9}$ for large enough constant $\widetilde{c} > 0$.
2. As a result of the above, with probability at least $1 - \frac{d}{n^9}$, each coordinate of $\widehat{\boldsymbol{\theta}}_{\widehat{G}}$ lies in the interval $[0, \widetilde{c}s\log n]$.

Hence, using arguments similar to [JPTW23, Theorem 2] in the multidimensional case and the one dimensional case in (B.1) we can argue that given any estimate $\widehat{G}$, we can use the above choice of h to guarantee

$$\mathsf{Regret}(\widehat{G}, G) \le \mathsf{Regret}(\widehat{G}, G_h) + O(\frac{d^2 s^2}{n^4}).$$

where given any G supported on $\mathbb{R}_+^d$, G_h denote its restriction on the hypercube $[0,h]^d$, i.e., $G_h[\boldsymbol{Y} \in \cdot] = G[\boldsymbol{Y} \in \cdot | \boldsymbol{Y} \in [0,h]^d]$. As a result, it suffices to bound $\mathsf{Regret}(\widehat{G}, G_h)$ to get the desired regret upper bound.

To bound $\mathsf{Regret}(\widehat{G}, G_h) = \mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\right]$ we use the following decomposition that is similar to the decomposition in the proof of the one dimensional case. First we restrict the expectation on the event $\{\boldsymbol{Y}_i \in [0,K]^d,\ i \in \{1,\dots,n\}\}$, where for some absolute constants $\widetilde{c}_1, \widetilde{c}_2$ to be chosen later we pick

$$K = \widetilde{c}_1 h \frac{\log n}{\log\log n}\mathbf{1}_{\{G \in \mathscr{P}([0,h]^d\}} + \widetilde{c}_2 s\log n \mathbf{1}_{\{\text{marginals of } G \text{ are } \mathsf{SubE}(s)\}}.$$

We also pick $\widehat{h}$ as either h when $G \in \mathscr{P}([0,h]^d$ or $\max_{j=1}^d \max_{i=1}^n Y_{ij} + 1$ when the marginals of G are $\mathsf{SubE}(s)$. Note that in the later case, as we argued above, $\widehat{h}$ is bounded with probability $1 - \frac{d}{n^9}$

by $\widetilde{cs}\log n$. This implies that $\widehat{G}$ is supported on $[0,\widehat{h}]^d$. Then we have

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\mathbf{1}_{\{\boldsymbol{Y}\in[0,K]^d\}}\right]\\
&=\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}(y_j+1)^2 f_{G_h}(\boldsymbol{y})\left(\frac{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}-\frac{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}\right)^2\\
&\overset{(a)}{\leq}\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}(y_j+1)^2 f_{G_h}(\boldsymbol{y})\left\{3\left(\frac{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}-\frac{2f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2+3\left(\frac{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}-\frac{2f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2\right.\\
&\quad\left.+3\left(\frac{2f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-2f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2\right\}\\
&\leq 3\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}\left\{\left(\frac{(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}\right)^2\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}+\left(\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}\right)^2\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right.\\
&\quad\left.+4(y_j+1)^2\frac{(f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right\}\\
&\leq 3(h^2+\widehat{h}^2)\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}+12\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}(y_j+1)^2\frac{(f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}
\end{aligned}
$$

where (a) followed from $(x+y+z)^2\leq 3(x^2+y^2+z^2)$ for any $x,y,z\in\mathbb{R}$. Using $(\sqrt{f_{G_h}(\boldsymbol{x})}+\sqrt{f_{\widehat{G}}(\boldsymbol{x})})^2\leq 2(f_{G_h}(\boldsymbol{x})+f_{\widehat{G}}(\boldsymbol{x}))$ for $\boldsymbol{x}=\boldsymbol{y},\boldsymbol{y}+\boldsymbol{e}_j$ we continue the last display to get

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\mathbf{1}_{\{\boldsymbol{Y}\in[0,K]^d\}}\right]\\
&\leq 6(h^2+\widehat{h}^2)\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}(\sqrt{f_{G_h}(\boldsymbol{y})}-\sqrt{f_{\widehat{G}}(\boldsymbol{y})})^2\\
&\quad+24K\max_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)+(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\sum_{\boldsymbol{y}\in[0,K]^d}\sum_{j=1}^{d}(\sqrt{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}-\sqrt{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)})^2\\
&\leq d\left(6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right)H^2(f_{\widehat{G}},f_{G_h}).
\end{aligned}
$$

Again using $\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{y})\leq h,\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{y})\leq\widehat{h}$ we bound $\mathbb{E}_{G_h}\left[(\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y}))^2\mathbf{1}_{\{\boldsymbol{Y}\notin[0,K]^d\}}\right]$ by $(h+\widehat{h})^2\varepsilon_K(G_h)$, where $\varepsilon_K(G_h)=\mathbb{P}_{G_h}\left[\boldsymbol{Y}\notin[0,K]^d\right]$. Combining this with the last display we get

$$
\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\right]\leq d\left\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right\}H^2(f_{\widehat{G}},f_{G_h})+(h+\widehat{h})^2\varepsilon_K(G_h).
$$

Finally we take expectation on both sides with respect to the training sample, $\widehat{G}$, and $\widehat{h}$. Using the high probability bound on $\widehat{h}$ and the bound on $\mathbb{E}\left[H^2(f_{\widehat{G}},f_{G_h})\right]$ as in Theorem 5, and the probabilistic bound on $\varepsilon_K(G_h)$ as in Lemma 9 we get the result. This finishes the proof.