Trend Filtering on Graphs: Optimal denoising in k-D TV-classes and the Limitation of Linear Smoothers

Yu-Xiang Wang

Machine Learning Department, Statistics Department
Carnegie Mellon University

Based on joint works with: Ryan Tibshirani, James Sharpnack, Alex Smola, Veeranjaneyulu Sadhanala

Image denoising in the wild

Noisy image

Laplacian smoothing

TV denoising

$$\hat{\theta}^{\mathrm{LS}} = \arg\min_{\theta} \|\theta - y\|^2 + \lambda \|D\theta\|_2^2 = \underbrace{(\lambda D^T D + I)^{-1} y}_{\text{a linear smoother}}$$

$$\hat{\theta}^{\mathrm{TV}} = \arg\min_{\theta} \|\theta - y\|^2 + \lambda \|D\theta\|_1 \text{ — not a linear smoother}$$

- TV-denoising yields a cleaner and sharper denoised image.
- Quantitatively 35% less mean square error (MSE).
- But computationally more expensive.

This talk will be about

Theoretically quantifying the denoising performance

- By connecting it to nonparametric regression.
- How fast does MSE converge to 0 as the image gets finer resolutions?

Information-theoretic limit

- How fast does it get for any method?

Linear vs. Nonlinear estimation

- Could simpler methods perform well/optimally?

Outline

- Locally adaptive nonparametric regression
  1. Univariate trend filtering
  2. Graph trend filtering
- Discrete TV-classes beyond 1D
  3. Minimax rate and the limit of linear smoother

1 Univariate trend filtering
(Tibshirani, 2013, Annals of Statistics)

Classical nonparametric regression

Univariate nonparametric regression: observe independent draws from model

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \dots n$$

Conditional on $X = x_i$, error ϵ_i assumed to have zero mean and constant variance. Want to estimate

$$f_0(x) = \mathbb{E}[Y|X = x]$$

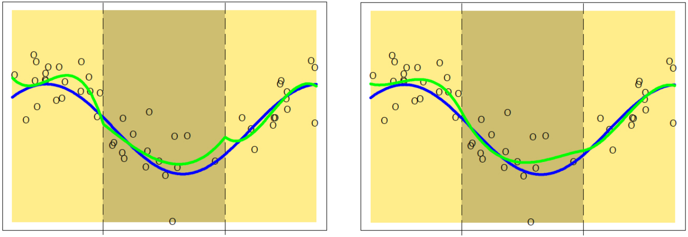
Rich literature, lots of interesting work. E.g., some key words:

- Splines
- Kernels
- Wavelets

Trend filtering: close relative of spline methods; relative newcomer in an old field.

Splines

A kth degree spline is a kth degree piecewise polynomial, with continuous derivatives of orders $0, 1, \ldots k-1$ at its knots



The added (higher-order) continuity constraints make the function smoother; think bias-variance tradeoff, this decreases the variance. Splines play a ubiquitous role in nonparametric modeling ...

Two spline estimators

- Smoothing spline (Schoenberg 1946; Reinsch 1967; Wahba 1990) estimate of order k is defined by

$$\min_f \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \int_0^1 \big(f^{(\frac{k+1}{2})}(t)\big)^2 \, dt$$

  Solution is a natural spline of degree k with knots at each $x_1, \ldots x_n$

- Locally adaptive regression spline (Mammen & van de Geer 1997) estimate of order k is defined by

$$\min_f \frac{1}{2} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \cdot \mathrm{TV}(f^{(k)})$$

  Solution is a spline of degree k whose knots are in $x_1, \ldots x_n$ when $k = 0$ or 1, and are in ??? when $k \geq 2$

Properties comparison

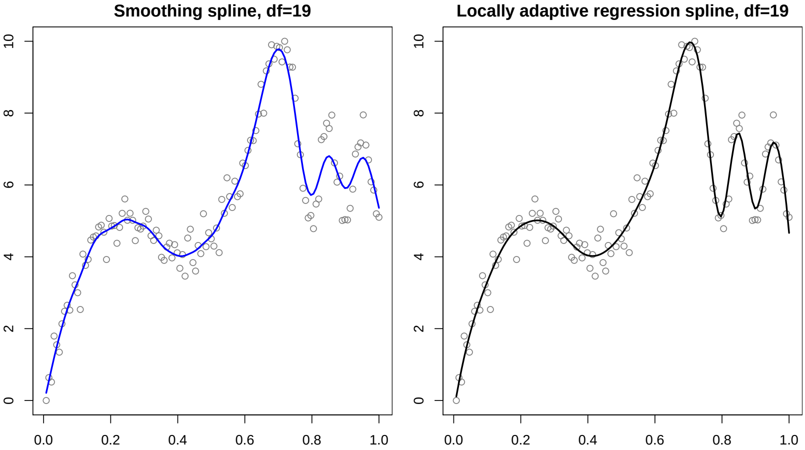
Smoothing splines

- Solution $\hat{f} = \sum_{j=1}^n \hat{\theta}_j \eta_j$, for a natural kth degree spline basis $\eta_1, \ldots \eta_n$
- Computable in $O(n)$ operations
- Coefficients $\hat{\theta}_1, \ldots, \hat{\theta}_n$ are ℓ_2-regularized
- Places knots at all data points $x_1, \ldots x_n$
- Globally smooth, or globally wiggly

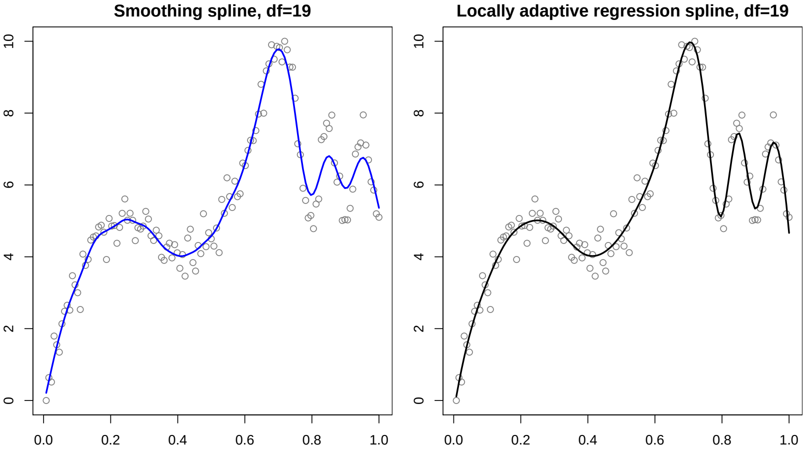
Locally adaptive splines

- Solution (approximate) $\hat{f} = \sum_{j=1}^n \hat{\theta}_j g_j$, for kth degree splines $g_1, \ldots g_n$
- Computable in $\approx O(n^3)$ operations
- Coefficients $\hat{\theta}_1, \ldots, \hat{\theta}_n$ are ℓ_1-regularized
- Selects knots as a subset of $x_1, \ldots x_n$
- Adapts to appropriate local level of smoothness

Example: Heterogeneous smoothness



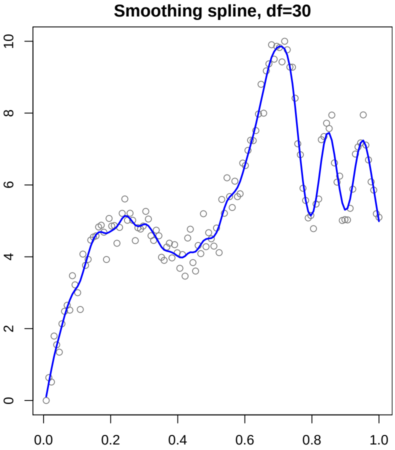
Example: Heterogeneous smoothness



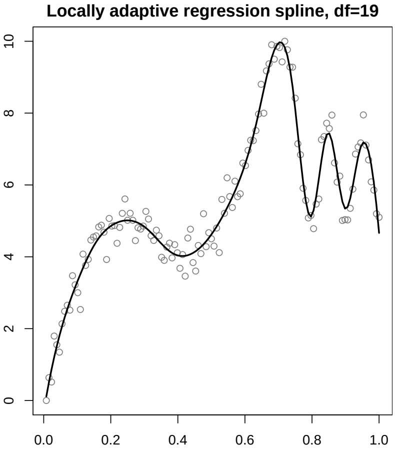
Oversmoothed on right

Adapts throughout

Example: Heterogeneous smoothness

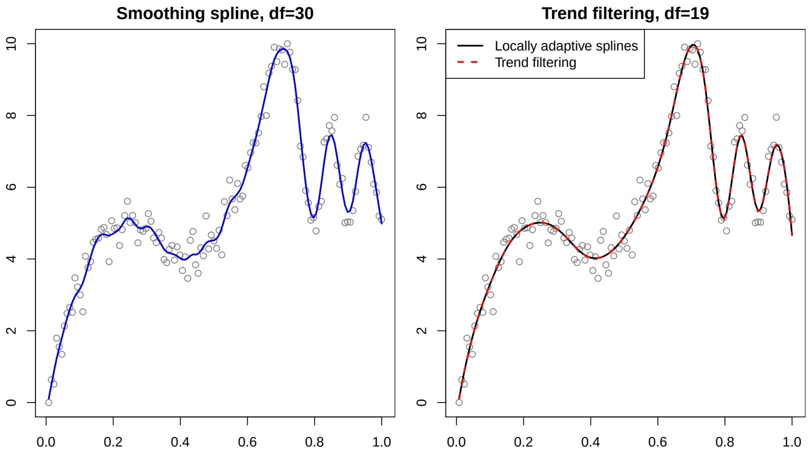


Undersmoothed on left



Adapts throughout

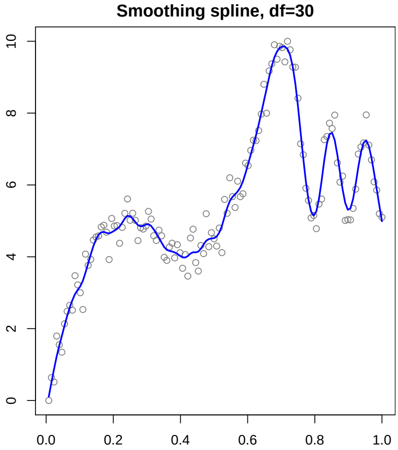
Example: Heterogeneous smoothness

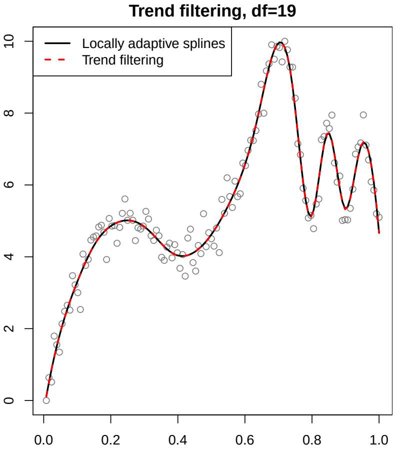


Undersmoothed on left

Adapts throughout

Example: Heterogeneous smoothness





Undersmoothed on left
(any linear smoother)

Adapts throughout
(both)

Trend filtering

Trend filtering (Steidl et al. 2006; Kim et al. 2009; T. 2014) can be seen as a discrete approximation to locally adaptive spline problem

$$\min_f \; \frac{1}{2} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \cdot \mathrm{TV}(f^{(k)})$$
$$\approx \min_{\beta \in \mathbb{R}^n} \; \frac{1}{2} \|y - \beta\|_2^2 + \lambda \|D^{(k+1)} \beta\|_1$$

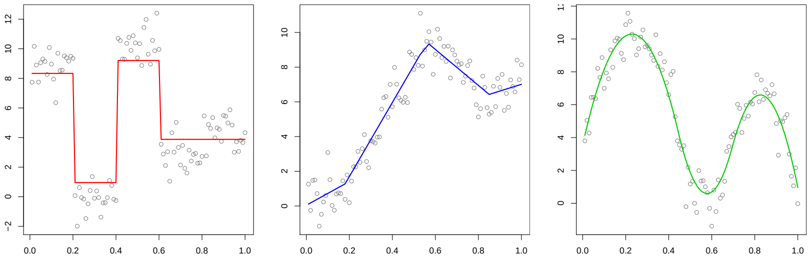
via $\mathrm{TV}(f^{(k)}) \approx \int_0^1 |f^{(k+1)}(t)| \, dt \approx \|D^{(k+1)}\beta\|_1$, where $D^{(k+1)}$ is a discrete derivative operator of order k. Recursive definition:

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ & & & \ldots & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}, \quad \text{and for } k = 1, 2, 3, \ldots,$$

$$D^{(k+1)} = D^{(1)} \, \mathrm{diag}\left(\frac{k}{x_{k+1} - x_1}, \frac{k}{x_{k+2} - x_2}, \ldots \frac{k}{x_n - x_{n-k}} \right) D^{(k)}$$

Trend filtering in continuous space

Intuitively, trend filtering solution $\hat{\theta}$ should exhibit the structure of kth degree piecewise polynomial (since it penalizes changes in kth derivatives across inputs)



Constant, $k = 0$ (Fused lasso) | Linear, $k = 1$ | Quadratic, $k = 2$

This idea can be formalized using falling factorial functions (W., Smola, Tibshirani. 2014)

Convergence theory

Assume observations from the model

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots n$$

for i.i.d. sub-Gaussian errors, and with f_0 in the class, for constant $C > 0$,

$$\mathcal{F}_k = \big\{ f : \mathrm{TV}(f^{(k)}) \leq C \big\}$$

Denote by $\|\cdot\|_n$ the empirical norm, as in $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2$

From Donoho & Johnstone (1998): minimax rate over $\mathcal{F}_k$ is

$$\min_{\hat{f}} \max_{f_0 \in \mathcal{F}_k} \mathbb{E}\|\hat{f} - f_0\|_n^2 = \Theta(n^{-(2k+2)/(2k+3)})$$

Meanwhile, linear smoothers achieve rate at best $n^{-(2k+1)/(2k+2)}$ ... this applies to smoothing splines, kernels, local polynomials, RKHS estimates, etc.!

Note: locally adaptive regression splines achieve the minimax rate, with $\lambda = \Theta(n^{1/(2k+3)})$ (Mammen & van de Geer 1997)

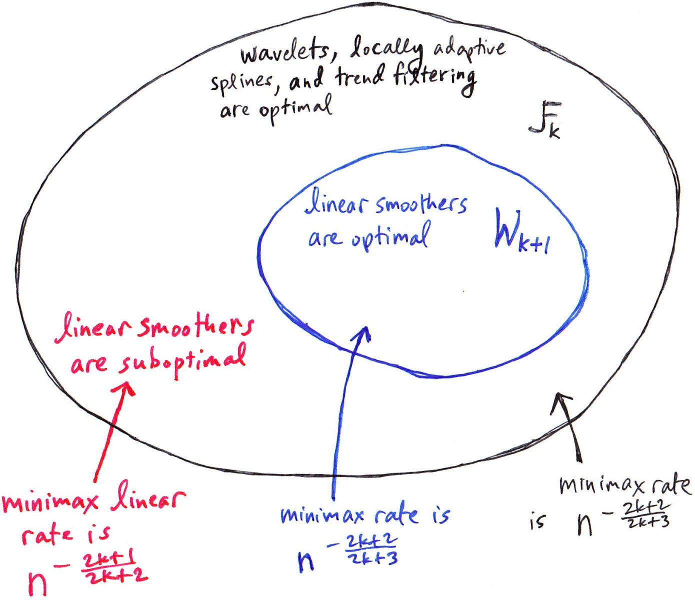
Theorem (Tibshirani, 2014): (informally) Trend filtering with $\lambda = \Theta(n^{1/(2k+3)})$ is "almost" equivalent to the locally adaptive splines, therefore, achieve the minimax rate

$$O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$$

for estimation over $\mathcal{F}_k$.

Same statistical properties, but much faster in computation!

Summary of univariate nonparametric regression



Two interesting points from the picture

- In 1D nonparametric regression/signal denoising, statistically speaking, we get local-adaptivity for free!
- But, we paid a computational price: it cannot be achieved by linear methods.

Question: Does the same picture extend to higher dimension?

2 Trend filtering on Graphs

(W., Sharpnack, Smola, Tibshirani, 2015 AIStats+JMLR)

Nonparametric regression on graphs

Graph smoothing: given a graph $G = (V, E)$, with vertices denoted $V = \{1, \ldots n\}$, we observe

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots n$$

Errors ϵ_i assumed to have zero mean. Want to estimate underlying signal μ, assumed to be smooth with respect to edges E

In comparison to univariate case, a lot less literature. E.g.,

- Eigen-based methods
- Laplacian smoothing
- Wavelets on graphs

Newcomer in this field: graph trend filtering, an extension of the univariate technique with analogous benefits

Graph trend filtering

Graph trend filtering (W., Sharpnack, Smola, Tibshirani, 2015) solves

$$\min_{\theta \in \mathbb{R}^n} \|y - \theta\|_2^2 + \lambda \|\Delta^{(k+1)} \theta\|_1$$

where $\Delta^{(k+1)}$ is a graph difference operator of order $k+1$, over G

Two key properties of univariate trend filtering:

- Computationally fast
- Locally adaptive

With suitably defined difference operators $\Delta^{(k+1)}$, $k = 1, 2, 3, \ldots$, graph trend filtering will share these properties

Discrete differences over graphs

Given graph $G = (V, E)$ with $V = \{1, \ldots n\}$ and $E = \{e_1, \ldots e_m\}$

- Define the first order graph difference operator $\Delta^{(1)}$ to be the edge incidence matrix of G, an $m \times n$ matrix, whose ℓth row is

$$D_\ell = (0, \ldots \underset{\substack{\uparrow \\ i}}{-1}, \ldots \underset{\substack{\uparrow \\ j}}{1}, \ldots 0)$$

  if the ℓth edge is $e_\ell = \{i, j\}$

- For higher orders, use the recursion:

$$\Delta^{(k+1)} = \begin{cases} (\Delta^{(1)})^T \Delta^{(k)} & \text{for } k \text{ odd,} \\ \Delta^{(1)} \Delta^{(k)} & \text{for } k \text{ even} \end{cases}$$

I.e., for D the edge incidence matrix, and $L = D^T D$ the Laplacian:

$$\Delta^{(1)} = D, \;\; \Delta^{(2)} = L, \;\; \Delta^{(3)} = DL, \;\; \Delta^{(4)} = L^2, \;\; \ldots$$

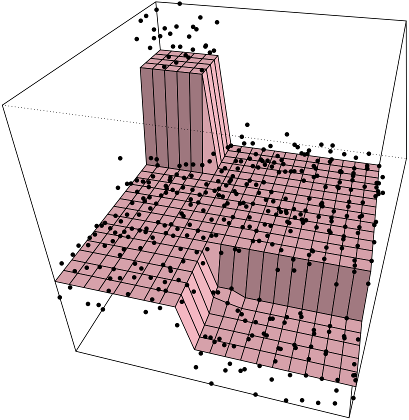
Constant order

The penalty for constant order graph trend filtering:

$$\|\Delta^{(1)}\theta\|_1 = \|D\theta\|_1 = \sum_{\{i,j\}\in E} |\theta_i - \theta_j|$$

Estimate $\hat{\theta}$ is piecewise constant over G

(This is also known as the graph fused lasso or graph TV-denoising)

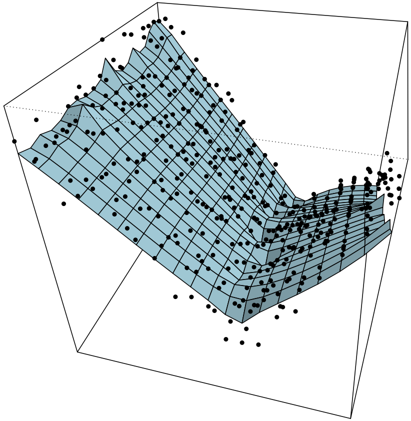


Linear order

The penalty for linear order graph trend filtering:

$$\|\Delta^{(2)}\theta\|_1 = \|L\theta\|_1 = \sum_{i=1}^{n} n_i \left| \theta_i - \frac{1}{n_i} \sum_{\{i,j\} \in E} \theta_i \right|$$

Estimate $\hat{\theta}$ is "piecewise linear" over G

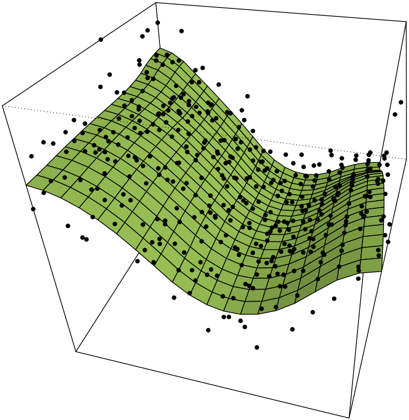


Quadratic order

The penalty for quadratic order graph trend filtering:

$$\|\Delta^{(2)}\theta\|_1 = \|DL\theta\|_1 = \sum_{\{i,j\}\in E} \left| \Big(n_i\theta_i - \sum_{\{i,\ell\}\in E} \theta_\ell\Big) - \Big(n_j\theta_j - \sum_{\{j,\ell\}\in E} \theta_\ell\Big) \right|$$

Estimate $\hat{\theta}$ is "piecewise quadratic" over G



A family of graph differences

What have we done? To recap:

- For odd k, the $(k+1)$st order differences are given by taking first differences of kth differences:

$$\Delta^{(k+1)} = D\Delta^{(k)}$$

- For even k, the $(k+1)$st order differences are given by taking second differences of $(k-1)$st order differences

$$\Delta^{(k+1)} = L\Delta^{(k-1)}$$

In general, $\Delta^{(k+1)}$ is structured enough that we can efficiently solve graph trend filtering problems, even over large graphs

Comparisons and interpretations

Laplacian smoothing (Belkin & Niyogi 2002; Smola & Kondor 2003) estimate is given by

$$\min_{\theta \in \mathbb{R}^n} \|y - \theta\|_2^2 + \lambda \theta^T L \theta$$

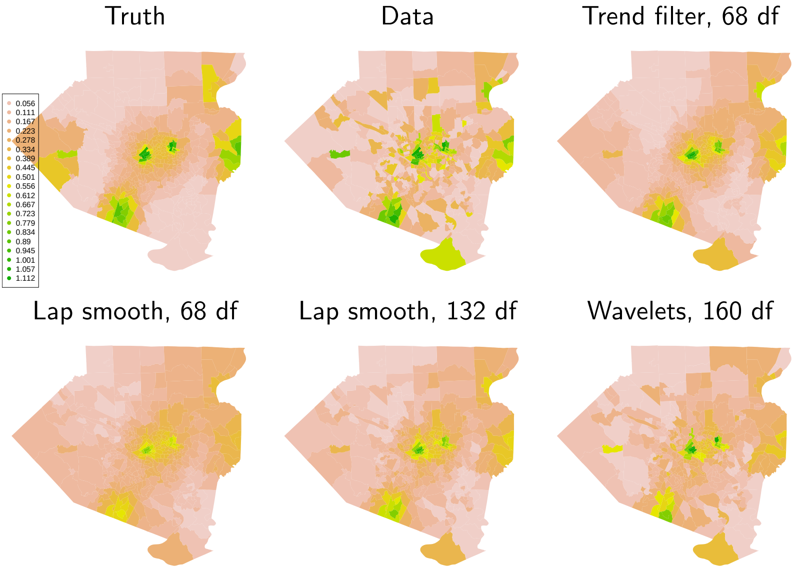
Generalize to higher-orders by replacing L with L^{k+1}, for some k

- Laplacian smoothing: ℓ_2 penalty $\theta^T L^{k+1} \theta = \|(L^{k+1})^{\frac{1}{2}} \theta\|_2^2$
- Graph trend filtering: ℓ_1 penalty $\|\Delta^{(k+1)} \theta\|_1 = \|(L^{k+1})^{\frac{1}{2}} \theta\|_1$
- Just like in univariate case, the latter is better at picking up local level of smoothness

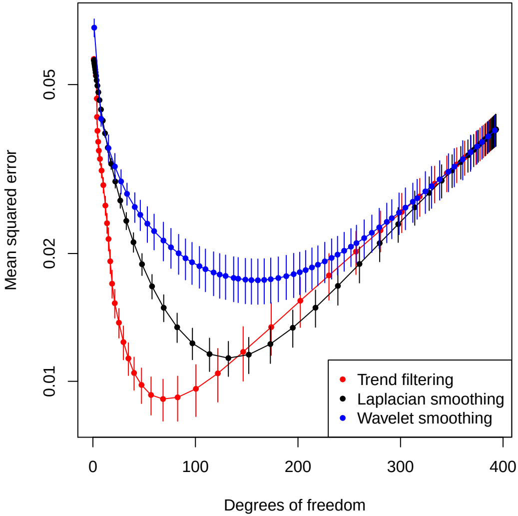
When $k = 0$ and the graph being a grid:

- Graph trend filtering $\equiv$ TV-denoising.
- Laplacian smoothing $\equiv$ an instance of Low-Pass Filtering.

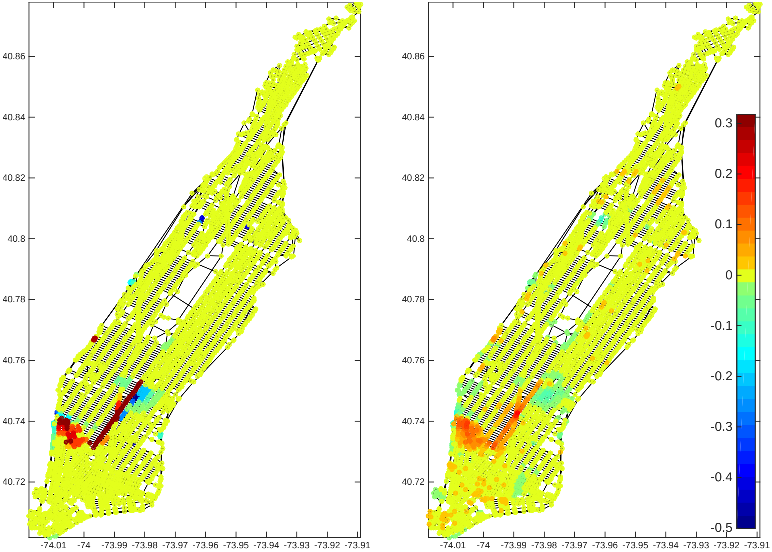
Example: Heteregeneous smoothness



Mean squared errors (averaged over 10 simulations):



Event detection on New York City Taxi counts

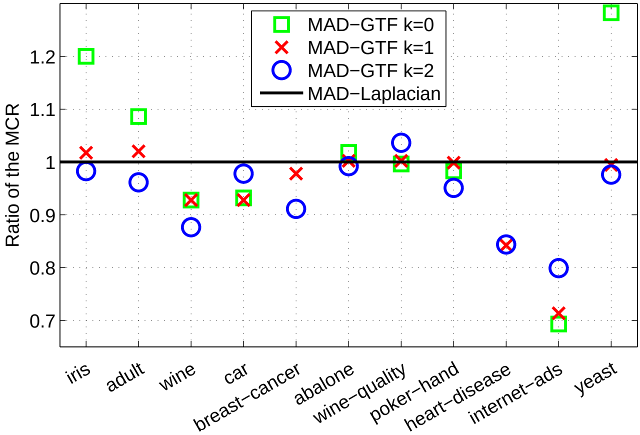


Sparse trend filtering

Sparse Laplacian smoothing

Graph-based Transductive Learning on UCI Datasets

We apply to plain classification problems:



The challenges for a unified theory for GTF

Assume observations from the model

$$y_i = \theta_{0i} + \epsilon_i, \quad i = 1, \dots n$$

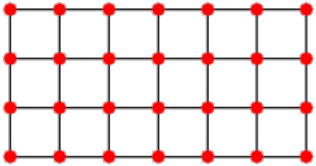
where errors are i.i.d. Gaussian, and $\|\Delta^{(k+1)}\theta_0\|_1$ is well-controlled. This is more challenging to analyze than Euclidean settings

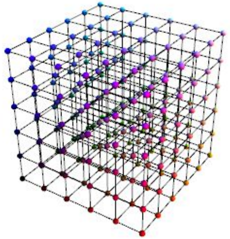
- Inexorable dependence on the underlying graph G; note that $\|\Delta^{(k+1)}\theta_0\|_1$ being small is a statement both about G and θ_0
- Not really any other rates to compare to
- No general notion of optimality (minimax rates)

Theoretical results are separated by the properties assumed about the underlying graph. One such property: graph incoherence

3 Total Variation classes beyond 1D

(Sadhanala, W., and Tibshirani, 2016 to appear in NIPS)





Defining the minimax problem

An **estimator** $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$ that takes in $\theta_0 +$ i.i.d. Gaussian noise and produces an estimator.

Mean square error:

$$\mathrm{MSE}(\hat{\theta}, \theta_0) = \frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2.$$

Minimax risk:

$$R(\mathcal{K}) = \min_{\hat{\theta}} \max_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big].$$

Minimax linear risk:

$$R_L(\mathcal{K}) = \min_{\hat{\theta}\ \text{linear}} \max_{\theta_0 \in \mathcal{K}} \mathbb{E}\big[\mathrm{MSE}(\hat{\theta}, \theta_0)\big],$$

d-dimensional Discrete TV-class

Define "function" classes

$$\text{TV Classes: } \mathcal{F}_d(C_n) = \{\theta : \|D\theta\|_1 \leq C_n\},$$

$$\text{Sobolev Classes: } \mathcal{M}_d(C'_n) = \{\theta : \|D\theta\|_2 \leq C'_n\},$$

Where D is the incidence matrix for the d-dimensional grid graph with a total of n vertices.

Recall that: When $d = 1$, Johnston and Donoho (1998) showed that

$$R(\mathcal{F}_1(C)) \asymp n^{-2/3}.$$

and the minimax linear rate much slower

$$R_L(\mathcal{F}_1(C)) \asymp n^{-1/2}.$$

What happens when $d > 1$ is an open problem!

For (continuous) Sobolev classes, the minimax rates are the standard nonparametric rates

$$n^{-2/(2+d)}.$$

Curse of dimensionality: As d increases the rate gets slower.

We would intuitively expect that the minimax rates on TV-classes should also get slower with increasing d.

A somewhat surprising upper bound for TV-denoising

Theorem (Hütter and Rigollet, 2016): Total variation denoising estimator obeys

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\left(\frac{C_n \log n}{n}\right) \text{ for } d = 2,$$

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{TV}}, \theta_0) = O_{\mathbb{P}}\left(\frac{C_n \sqrt{\log n}}{n}\right) \text{ for } d \geq 3,$$

Isn't this too good to be true?

From 1D to 2D, the rate suddenly becomes parametric rate!

Did we get away from the "curse of dimensionality"?

An even more surprising upper bound for a trivial estimator

Lemma (Sadhanala, W. and Tibshirani, 2016): A trivial estimator $\hat{\theta}^{\mathrm{mean}}$ that outputs $\bar{y}\mathbb{1}$ obeys

$$\sup_{\theta_0 \in \mathcal{F}(C_n)} \mathbb{E}[\mathrm{MSE}(\hat{\theta}^{\mathrm{mean}}, \theta_0)] = O\left(\frac{\sigma^2 + C_n^2 \log n}{n}\right)$$

Note that:

- $\hat{\theta}^{\mathrm{mean}}$ is a linear smoother!
- If C_n is a constant, then the trivial estimator performs as well as TV-denoising!

The only logical explanation: $C_n = O(1)$ is a trivial region! In other word, C_n should increase with n!

Matching lower bounds for the surprising upper bounds

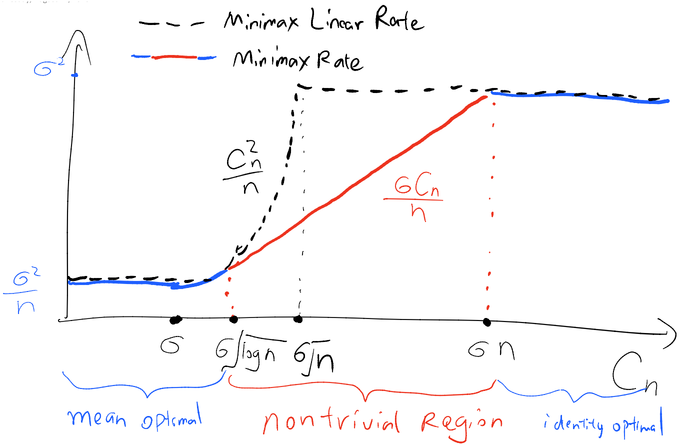
Theorem (Sadhanala, W. and Tibshirani, 2016): For constant d, and nontrivial region of C_n:

$$R(\mathcal{F}_d(C_n)) \asymp \frac{\sigma^2 + \sigma C_n}{n}.$$

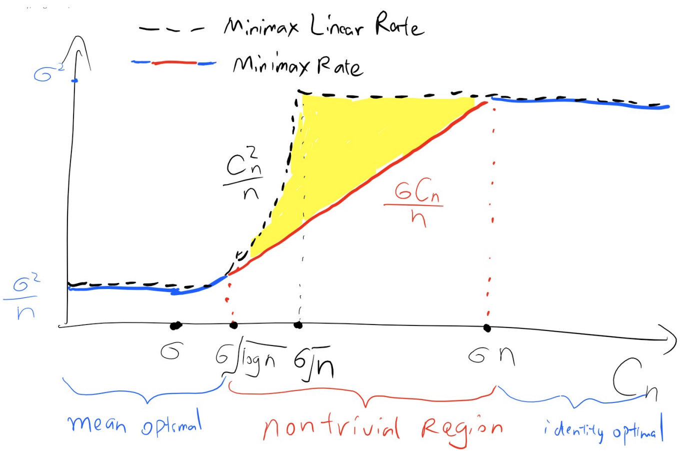
$$R_L(\mathcal{F}_d(C_n)) \asymp \frac{\sigma^2 + C_n^2}{n}.$$

- $\hat{\theta}^{\mathrm{TV}}$ is optimal for the TV-class!
- $\hat{\theta}^{\mathrm{mean}}$ is an optimal linear smoother for the TV-class!
- Spectacular failure: No linear smoother can do better than a trivial linear smoother, in 2-dim and above!

Minimax rate and minimax linear rate



Minimax rate and minimax linear rate



This still does not solve our problem: where did the "Curse-of-dimensionality" go?

A "canonical" scaling

Interpreting the results in the context of continuous space function-classes in $[0, 1]^d$.

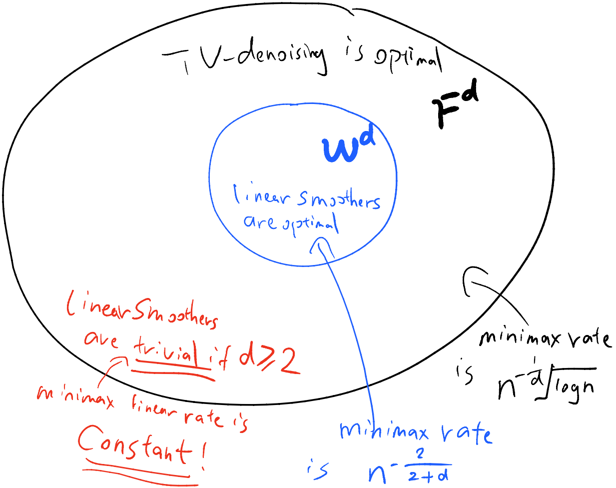
1. The Sobolev class has the canonical nonparametric rate $n^{-\frac{2}{2+d}}$.
2. The TV class is big enough to contain the Sobolev class.

The canonical scaling of C_n is:

TV-class: $\mathcal{F}_d(n^{1-1/d})$

Sobolev-class: $\mathcal{M}_d(n^{1/2-1/d})$

The big picture for d-dim problems



An interesting phase-transition

Function class	Dimension 1	Dimension 2	Dimension $d \geq 3$
TV ball	$n^{-2/3}$	$n^{-1/2}\sqrt{\log n}$	$n^{-1/d}\sqrt{\log n}$
Sobolev ball	$n^{-2/3}$	$n^{-1/2}$	$n^{-\frac{2}{2+d}}$

Table: *Summary of rates for canonically-scaled TV and Sobolev spaces.*

Remarks:

- When $d = 2$, there is a $\sqrt{\log n}$ gap between the minimax rates of TV-class and the Sobolev class contained in it.
- When $d \geq 3$, there is a polynomial gap. We no longer get adaptivity for free.
- Open problem: Is TV-denoising minimax in Sobolev? If not, is there an algorithm that is simultaneously minimax in TV and Sobolev?

A few notes about proof techniques.

For upper bounds:

- A $d-$dim grid's Laplacian matrix can be diagonalized by DCT and inverse DCT.
- Prove that D is constant incoherent.
- Careful calculations of the partial sum of the spectrum.

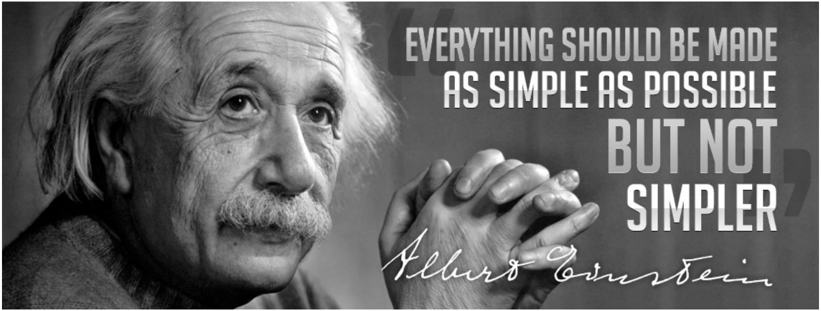
For lower bounds:

- Embedding a big ℓ_1 ball inside the TV-ball.
- Gaussian model selection (Birgé and Massart, 2001) .
- Linear smoother lower bound: use orthosymmetric and quadratically convex set (Donoho, Liu MacGibbon, 1990).

To reiterate the main points

- We derive trend filtering for smoothing heterogeneous signals on graphs.
- Define discrete TV-classes and characterized its minimax rates.
- Show that linear smoothers can fail spectacularly.
- The extra computational cost for solving GTF is often worth it.

The story of trend filtering, linear smoothers and the price of local adaptivity in d-dim TV-classes



Acknowledgments

Ryan Tibshirani

James Sharpnack

Alex Smola

Veeru Sadhanala

Thank you for your attention!