





Exponential Family Trend Filtering on Lattices

Veeranjaneyulu Sadhanala¹ , Robert Bassett² ,
James Sharpnack³ , and Daniel J. McDonald⁴ 

¹Google Research, e-mail: veeranjaneyulus@gmail.com

²Operations Research Department, Naval Postgraduate School, e-mail: robert.bassett@nps.edu

³Duolingo, e-mail: jsharpna@gmail.com

⁴Department of Statistics, University of British Columbia, e-mail: daniel@stat.ubc.ca

Abstract: Trend filtering is a modern approach to nonparametric regression that is more adaptive to local smoothness than splines or similar basis procedures. Existing analyses of trend filtering focus on estimating a function corrupted by homoskedastic Gaussian noise, but our work extends this technique to general exponential family distributions. This extension is motivated by the need to study massive, gridded climate data derived from polar-orbiting satellites. We present algorithms tailored to large problems, theoretical results for general exponential family likelihoods, and principled methods for tuning parameter selection without excess computation.

MSC2020 subject classifications: Primary 62G08; secondary 62G20.

Keywords and phrases: denoising, Kronecker, local adaptivity, nonparametric, total variation.

Contents

1	Introduction	2
1.1	Estimators	5
1.2	Properties of exponential families	6
1.3	Summary of our contributions	8
1.4	Overview of theoretical contributions	9
1.5	Related work	10
2	Penalized MLE	12
2.1	Additional penalty on the null space component of θ	13
2.2	Error bounds for penalized MLE	15
2.3	Penalized MLE in special cases	17
3	Error bounds for the Mean Trend Filter	18
3.1	Error bounds with canonical scaling	19
3.2	Lower bounds for mean trend filtering	20
4	Algorithmic implementation	21
5	Degrees of freedom and tuning parameter selection	23

arXiv: [2209.09175](https://arxiv.org/abs/2209.09175)

6	Empirical results	25
6.1	Simulation study	25
6.2	Example applications	26
6.2.1	UC COVID-19 hospitalization data	27
6.2.2	Temperature variability	28
7	Discussion	30
A	Proofs for preliminary results	31
A.1	Proof of Theorem 1.2	31
A.2	Sub-exponential parameters for some standard distributions	32
A.3	Some properties of Sub-exponentials	32
B	Proofs of upper bounds	35
B.1	Proof of Theorem 2.2	35
B.2	Proofs of Corollaries to Theorem 2.2	36
B.3	Uniform risk bound with null space penalty	38
B.4	Proof of Theorem 3.1	38
B.5	Proof of Theorem 3.2	39
B.6	Proofs of Corollaries to Theorem 3.2	40
B.7	Error rates assuming that the estimate is bounded	44
B.8	Empirical process bound	46
B.9	Eigenvalue bounds	48
C	Proofs for lower bounds	52
C.1	Proof of Theorem 3.6	52
C.2	Proof of Theorem 3.7	58
D	Algorithmic details	59
E	Degrees of freedom and tuning parameter selection	61
	Acknowledgments	62
	Funding	63
	References	63

1. Introduction

Modeling data using exponential family distributions on the vertices of a graph is a standard task in statistics and artificial intelligence. Examples include satellite images or photographs, traffic or mobility patterns, communications networks, spatiotemporal data, and many others. Suppose we observe $y_i \in \mathbb{R}$ for $i = 1, \dots, n$ on the nodes of a graph and assume that they independently follow a natural exponential family with density of the form

$$p(y_i \mid \theta_i^*) = h(y_i) \exp \{y_i \theta_i^* - \psi(\theta_i^*)\}, \quad (1)$$

for functions $h : \mathbb{R} \rightarrow [0, \infty)$ and $\psi : \Theta \rightarrow \mathbb{R}$ and natural parameter $\theta_i^* \in \Theta$. The maximum likelihood estimator for θ^* is easily shown to be $\psi'^{-1}(y)$ where we apply the function component wise. Unfortunately, this estimator fails to respect the known graphical structure (relationships between “neighbouring” observations y_i), and therefore has high estimation risk. For example, $E\|\psi'^{-1}(y) - \theta^*\|_2^2 \propto n$ for the Gaussian family. In this paper, we imagine that

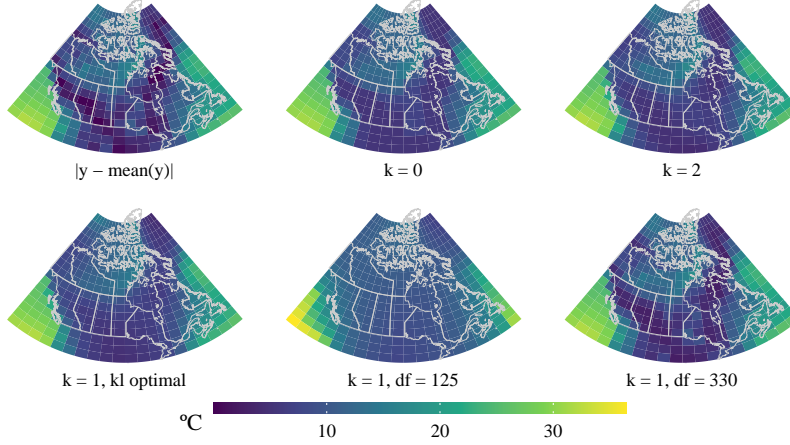


FIG 1. Estimates of the instantaneous temperature variance for 1 January 2010 over Canada. The top row shows the absolute centered data, 0th-order trend filter, and 2nd-order trend filter, in the latter 2 cases, with reasonable values of the tuning parameter. The bottom row shows the 1st-order trend filter for different tuning parameters, with the left most map, labeled “optimal”, corresponding to the estimate when the degrees-of-freedom is chosen by minimizing an unbiased risk estimate.

the natural parameter vector $\theta^* \in \Theta^n \subseteq \mathbb{R}^n$ is smooth on the graph in a total variation sense described below. We study methods to filter (estimate) the true parameter vector θ^* , given observations $y \in \mathbb{R}^n$ subject to this structure.

As an example, Figure 1 shows estimates for the instantaneous variance (that is, imagining $y_i \sim N(0, \sigma_i^2)$) of the temperature for New Year’s Day 2010 over a grid for Canada using maximum likelihood and a few configurations of the main family of estimators we investigate. The smoothness imposed by the grid of neighbouring locations leads to predictable patterns in the estimate that follow topographical features like mountain ranges and bodies of water. We will revisit this example in more detail in Section 6. Before describing our methodology more carefully, we define notation.

Notation. Throughout this paper, we will focus on lattice graphs in d dimensions, though we note that our main theoretical results can be extended to arbitrary graphs with appropriate conditions on the graph-Laplacian. We define a *graph difference operator* D that is crucial for defining our estimators. In one dimension, on a chain graph, the difference operator $D_{n,1}^{(1)}$ is defined by

$$(D_{n,1}^{(1)}\theta)_i = \theta_{i+1} - \theta_i \text{ for all } i \in [n-1], \theta \in \mathbb{R}^n,$$

where $n > 1$. We use the notation $[m]$ to denote the set $\{1, 2, \dots, m\}$ for positive integers m . The $(k+1)^{\text{th}}$ order (forward) difference matrix $D_{n,1}^{(k+1)} \in$

$\mathbb{R}^{(n-k-1) \times n}$ is defined with the recurrence relation

$$D_{n,1}^{(k+1)} = D_{n-k,1}^{(1)} D_{n,1}^{(k)} \text{ for } k > 0, n > k.$$

For example, the 3rd-order differences look like: $(D_{n,1}^{(3)}\theta)_i = \theta_{i+3} - 3\theta_{i+2} + 3\theta_{i+1} - \theta_i$. For a general graph, let $D^{(1)}$ denote its incidence matrix. That is, if edge e connects vertices i and j , then $D_{ei}^{(1)} = 1$ and $D_{ej}^{(1)} = -1$. In $d > 1$ dimensions, we focus on lattice graphs with a length of N on each side and with a total number of vertices $n = N^d$. In our estimators, unless otherwise specified, we penalize the variation of signals only along axis-parallel directions. To make this concrete, consider two dimensional grid graphs. We apply a univariate difference operator $D_{N,1}^{(k_1+1)}$ on all rows and another univariate operator $D_{N,1}^{(k_2+1)}$ on all columns. Concretely, the two dimensional difference operator of order $(k_1 + 1, k_2 + 1)$ is defined using Kronecker products of identity and univariate difference operators $D_{N,1}^{(k_1+1)}, D_{N,1}^{(k_2+1)}$. For an $N \times N$ grid, with $k_1, k_2 \geq 0, N \geq k_1, k_2$, and $y, \theta \in \mathbb{R}^{N \times N}$ stacked column-wise as $\text{vec}(\theta)$,

$$D_{n,2}^{(k_1+1,k_2+1)} = \begin{bmatrix} D_{N,1}^{(k_1+1)} \otimes I_N \\ I_N \otimes D_{N,1}^{(k_2+1)} \end{bmatrix}.$$

More generally, For d -dimensional grids, let $(k+1)$ denote the d -vector $(k_1 + 1, \dots, k_d + 1)$, and define

$$D_{n,d}^{(k+1)} = \begin{bmatrix} D_{N,1}^{(k_1+1)} \otimes I_N \otimes \dots \otimes I_N \\ I_N \otimes D_{N,1}^{(k_2+1)} \otimes \dots \otimes I_N \\ \vdots \\ I_N \otimes I_N \otimes \dots \otimes D_{N,1}^{(k_d+1)} \end{bmatrix},$$

where the Kronecker products consist of d terms each, one term for each dimension. In such cases, we adopt the convention that y and $\theta \in \mathbb{R}^{N \times \dots \times N}$ are stacked such that they vary fastest along the first dimension and slowest along the d^{th} . In the following, we suppress the subscripts and superscripts of the difference operator when they are clear from the context.

Define $\|\cdot\|_2$ to be the usual Euclidean norm and $\|\cdot\|_n = n^{-1/2}\|\cdot\|_2$ to be the empirical norm. We will similarly denote other ℓ_p norms with an appropriate subscript, where, to be explicit, we write $\|x\|_{\inf} = \max_j |x_j|$ and $\|x\|_0 = |\{j : x_j \neq 0\}|$. When there is no chance of confusion, we will assume that the ψ function in (1) applies component-wise. We use $a \otimes b$ to denote the Kronecker product of vectors a and b , $a \odot b$ to denote the elementwise product, and $\langle a, b \rangle = a^\top b$ to be the dot product. When clear, we will use f', f'' to denote componentwise first and second derivatives of the function f . We use \vee/\wedge for maximum/minimum respectively and $(x)_+ = x \vee 0$; while $\mathbf{1}\{A\}$ is the indicator of the event A , taking the value one if true and zero otherwise. We use $a_n \lesssim b_n$ to mean $a_n \leq cb_n$ eventually for some constant $c > 0$, $a_n = \Omega(b_n)$ to mean that $a_n \geq cb_n$ eventually, and $Y_n = O_{\mathbb{P}}(a_n)$ to mean that the sequence of random variables Y_n/a_n is

bounded in probability eventually. We will also use $Y_n = \tilde{O}_{\mathbb{P}}(a_n)$ to mean that $\exists c > 0$ such that $Y_n = O_{\mathbb{P}}(\log^c(n) a_n)$. Finally, for the graph difference operator, we will write the singular value decomposition (SVD) of $D = U\Sigma V^{\top} \in \mathbb{R}^{m \times n}$ where $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$ and $V \in \mathbb{R}^{n \times n}$, and we write the null-space of D as $\mathcal{N} = \mathcal{N}(D)$.

1.1. Estimators

We consider two canonical estimators. The first filters the natural parameter θ^* based on maximizing the likelihood while the second filters the mean $\beta^* := \psi'(\theta^*)$ directly. This distinction is important with respect to the nature of the expected smoothness. If we were to consider the data without regard for the graphical structure, then there is a direct correspondence between these two: the MLE for β^* is given by applying ψ' to the MLE for θ^* . Furthermore, this equivalence holds trivially for estimating the mean of a Gaussian because $\beta^* = \theta^*$. However, any requirement for smoothness over the graph destroys this relation for general exponential families.

Penalized MLE. We minimize negative log-likelihood with a smoothness imposing penalty:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n -y_i \theta_i + \psi(\theta_i) + \lambda \|D\theta\|_1. \quad (2)$$

Here λ is a parameter for balancing fidelity to any anticipated smoothness over the graph, as encoded by D , with fit to the data y . Taking $\lambda \rightarrow 0$ will result in the minimum occurring at $\hat{\theta} = \psi'^{-1}(y)$ while letting $\lambda \rightarrow \infty$ gives the Kullback–Leibler projection of y on to $\mathcal{N}(D)$ (Bauschke and Borwein, 1997):

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathcal{N}(D) \cap \operatorname{dom}(\psi)} \frac{1}{n} \sum_{i=1}^n -y_i \theta_i + \psi(\theta_i).$$

By the likelihood principle, $\hat{\theta}$ is the natural estimator to use when we expect that θ^* is smooth with respect to the graph. However, as we will demonstrate, this estimator can have high excess estimation risk when $\psi''(\theta^*)$ approaches 0. In Section 2.1 we will argue that this issue can be addressed by adding a penalty on the null-space component of θ . Specifically, the MLE with TF and null space penalty is

$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n -y_i \theta_i + \psi(\theta_i) + \lambda_1 \|D\theta\|_1 + \lambda_2 \|P_{\mathcal{N}}\theta\|_2 \quad (3)$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters and $P_{\mathcal{N}}$ is the projection operator on to $\mathcal{N}(D)$.

Mean Trend Filter. When the expected smoothness is in the mean rather than the natural parameter, it may be more appropriate to penalize the roughness in mean directly. For such a scenario, we consider the trend filtering estimator:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|_2^2 + \lambda \|D\beta\|_1. \quad (4)$$

As before, λ balances data fidelity with smoothness, but here, the interpretation as $\lambda \rightarrow \infty$ is more straightforward. In this case, the minimum occurs at the orthogonal projection onto the null space of D : $\hat{\beta} = (I - D^\top(DD^\top)^{-1}D)y$. This estimator was proposed in Steidl, Didas and Neumann (2006), Kim et al. (2009) and statistically analyzed in Tibshirani (2014), Wang et al. (2016) and others. We provide a thorough overview of previous work on mean trend filtering in a later section.

To understand the nature of the penalty in the above formulations, it is clearly important to understand its null space. Sadhanala et al. (2017) showed that the null space of D consists of Kronecker products of polynomials. We give a generalized version of their Lemma 1 here.

Lemma 1.1. *A basis for the null space of D is given by the family of polynomials*

$$\{p(x) = x_1^{a_1} \otimes x_2^{a_2} \otimes \cdots \otimes x_d^{a_d} : a_j \in \{0, \dots, k_j\}\}$$

where $x_j = (1, \dots, N)^\top$, $\forall j$. Note that $p(x) \in \mathbb{R}^n$, and \otimes is the Kronecker product. The dimension of the null space is $\operatorname{nullity}(D) = \prod_{j=1}^d (k_j + 1)$.

Therefore, writing P as the matrix formed by the evaluations of this collection of polynomials over the grid, we can also write the Euclidean projection onto the null space of D as $P_N := P(P^\top P)^{-1}P^\top$.

Characterizing the null space defines the vectors θ^* that have $\|D\theta^*\|_1 = 0$, but it does not say anything about vectors with bounded trend filtering penalty. Consider the ℓ_0 penalty instead, $\|D\theta\|_0$, for $k_1 = \cdots = k_d = k$. This is small when there are few changepoints, which are the indices j_1, \dots, j_M at which the k^{th} derivative is non-zero, $(D\theta^*)_{j_1, \dots, j_M} \neq 0$. Because the ℓ_1 penalty tends to produce sparse vectors with small $\|D\hat{\theta}\|_0$, the reconstructed signals are piecewise polynomials with a few changepoints that are automatically selected. The result is that trend filtering produces estimators that are locally adaptive, which means that the reconstructed signal is not oversmooth in regions of high signal variability (in θ^*) and not undersmooth in regions of low variability. In short the filter does not have one fixed resolution or bandwidth, but adapts the resolution to the observed signal. For a more complete explanation of this phenomenon, see Wang et al. (2016); Bassett and Sharpnack (2019). To simplify the theoretical exposition below, we will assume that $k_1 = \cdots = k_d = k$, but our results are easily modified for other situations.

1.2. Properties of exponential families

In this section, we review properties of exponential families, many of which will play a key role in our theoretical development. Considering the univariate

random variable Y with density of the form in (1), we define the domain $\Theta = \{\theta \in \mathbb{R} : \psi(\theta) < \infty\}$ and assume that Θ has a non-empty interior. Recall that the mean and variance of the distributions $p(\cdot | \theta)$ are $\psi'(\theta)$ and $\psi''(\theta)$ respectively, for natural parameter $\theta \in \Theta$. Therefore, $\epsilon := Y - \psi'(\theta^*)$ has mean zero and a simple expression for its moment generating function (MGF)

$$E[e^{s\epsilon}] = \exp\{\psi(\theta^* + s) - \psi(\theta^*) - s\psi'(\theta^*)\}$$

for s in a neighborhood of 0. Furthermore, ψ is convex and all its derivatives exist for all $\theta \in \Theta$ (see Brown 1986).

We say that a random variable X with mean 0 is *sub-exponential* if there are non-negative parameters ν, b such that

$$E[\exp\{tX\}] \leq \exp\{\nu^2 t^2/2\} \quad \text{for all } |t| < 1/b.$$

For shorthand, we also say X is $\text{SE}(\nu^2, b)$. We can show that random variables following exponential family distributions are sub-exponential in this sense.

Lemma 1.2. *Fix θ^* in $\text{interior}(\Theta)$, and let Y be from a univariate exponential family with parameter θ^* . Then $Y - \psi'(\theta^*)$ is sub-exponential with parameters ν and b . Specifically, for any $\delta > 0$, there exist parameters $\nu(\delta)$ and $b(\delta)$ depending on θ^* such that for all $|s| < 1/b$ the Taylor remainder of the second-order expansion $\psi(s + \theta^*)$ around θ^* is bounded by δ . In this case, ν satisfies $\nu^2 = \psi''(\theta^*) + \delta$.*

Table 1 gives the log-partition function $\psi(\theta)$ and sub-exponential parameters for Gaussian, Poisson, exponential, and χ -squared families. These calculations and the proof of Theorem 1.2 are in Appendix A. In each of the examples in Table 1, ν^2 is selected to be a multiple of the variance, but these are not the only choices of (ν, b) that would constitute valid sub-exponential parameters. Theorem 1.2 is not surprising given the form of the MGF, but seems not to be well-known. Related results can be seen in Brown (1986) or Kakade et al. (2010). Note that many exponential families have tails which decay faster (e.g., Gaussian or Binomial distributions), but *all* exponential families have at least sub-exponential tails.

Finally, we note that in three of these examples (Poisson, exponential, χ -square) the variance, and hence the curvature of $\psi(\theta^*)$ depends on θ^* , resulting in heteroskedasticity. This is one of the main complications of the exponential family setting that we consider in this paper. Along with the heavy-tailed residuals, this setting is a major departure from the sub-Gaussian homoskedastic setting of most prior work.

KL divergence. The Kullback–Leibler (KL) divergence between exponential distributions of the same family has a simple algebraic form in terms of ψ ; see Wainwright and Jordan (2008). The KL divergence with parameter vectors θ_0 and $\theta_1 \in \mathbb{R}^n$ is

$$\text{KL}(p(y | \theta_0) \| p(y | \theta_1)) = \text{KL}(\theta_0 \| \theta_1) := \int p(y | \theta_0) \log \frac{p(y | \theta_0)}{p(y | \theta_1)} dy. \quad (5)$$

TABLE 1
Sub-exponential parameters for some exponential family distributions

Distribution	$\psi(\theta)$	ν^2, b
Gaussian (common variance σ^2)	$\frac{1}{2}\theta^2$	$\sigma^2, 0$
Poisson (mean = μ)	e^θ	$2\mu, 0.55$
Exponential (mean = μ)	$-\log(-\theta)$	$4\mu^2 \log \frac{4}{e}, 2\mu$
χ_k^2 (mean = k)	$\log(\Gamma(\theta + 1)2^{\theta+1})$	$4k, 4$

In the asymptotic setting with $n \rightarrow \infty$, it makes more sense to examine the average divergence per coordinate. Thus we define $\overline{\text{KL}}(\theta_0 \parallel \theta_1) := \frac{1}{n} \text{KL}(\theta_0 \parallel \theta_1)$. For an exponential family as in (1), the KL divergence is the Bregman divergence of ψ ,

$$\text{KL}(\theta_0 \parallel \theta_1) = \psi(\theta_1) - \psi(\theta_0) - (\theta_1 - \theta_0)^\top \psi'(\theta_0).$$

1.3. Summary of our contributions

Most of the existing work on trend filtering referenced above assumes sub-Gaussian noise, that is,

$$y_i = \beta_i + \epsilon_i,$$

for $i \in [n]$ where ϵ_i is mean-zero and sub-Gaussian with common variance σ^2 . For general exponential families of the form in (1), $y_i - Ey_i$ has heavier than sub-Gaussian tails. Furthermore, for general exponential families, the variance, as well as higher moments, are tied to the mean parameter. Therefore, consideration of heteroskedasticity is a necessary and fundamental component of our analysis.

Direct analysis for specific exponential families, such as Poisson (Bassett and Sharpnack, 2019) are rare. Van de Geer (2020) analyses a penalized MLE for the logistic family. However, the logistic family has sub-Gaussian tails and uniformly bounded variance which allows key parts of the analysis to work: most importantly for the theoretical results the Dudley entropy integral bound holds. In other words, the theoretical approach there cannot generalize to arbitrary exponential families.

Our results here apply to the entire exponential family. However, due to this generality, the results are necessarily weaker than could potentially be achieved under additional, more stringent conditions (such as by assuming Gaussian or logistic distributions, or requiring additional bounds on higher moments).

A key ingredient for previous analyses in the sub-Gaussian setting is that the Bregman divergence $\psi(\hat{\theta}) - \psi(\theta^*) - (\hat{\theta} - \theta^*)^\top \psi'(\theta^*)$, can be lower bounded by a multiple of $\|\hat{\theta} - \theta^*\|_2^2$, because ψ is strongly convex. However, for general exponential families, ψ is not strongly convex, even if $\|D\theta^*\|_1$ is well-controlled, unless θ^* satisfies additional conditions. Without such assumptions, $\psi''(\theta^*)$ can be arbitrarily small. If we make the (rather implausible) assumption that both the estimate $\hat{\theta}$ and the parameter θ^* are bounded, then we recover this strong convexity in the relevant region where $\hat{\theta}$ and θ^* lie. In this case, we can apply

the same techniques used to analyze the sub-Gaussian case. We derive these bounds in [Appendix B.7](#). However, without such an assumption, analysis requires entirely different techniques, and we show these results in [Section 2.2](#).

Our main contributions are the following.

1. We derive error bounds on excess KL-risk for the penalized maximum likelihood estimator for general exponential families with sub-exponential noise ([Section 2](#)). We argue that there is a need to constrain the component of the natural parameter vector that lies in the null space of D as in equation (3).
2. We delineate two types of heteroskedasticity that are relevant under general assumptions: strong heteroskedasticity and mild heteroskedasticity. We show how our general KL-bounds behave under these regimes and how the heteroskedasticity interacts with the smoothness constraints and the dimensionality of the problem.
3. For $k = 0$, we show that the mean trend filter and the MLE with penalty are equivalent estimators, and hence, results for the mean trend filter apply immediately in this special case (though under different smoothness assumptions; [Section 3](#)).
4. We show that the mean trend filter nearly achieves the minimax optimal rate under squared error loss for mildly heteroskedastic data and all smoothness levels k and lattice dimensions d ([Section 3](#)). This result in fact holds for general sub-exponential noise ϵ , not just for the exponential families we consider in the paper. We incur an additional $\log n$ factor in the error bound for sub-exponential noise. It is specific to distributions where the mean parameter has bounded trend filtering penalty.
5. We give an algorithm for solving all of these cases for arbitrary likelihood, smoothness levels, and dimension, with the goal of operating on large data ([Section 4](#)).
6. We give a simple estimator for the out-of-sample prediction risk (at the original grid locations) to enable tuning parameter selection without requiring complicated forms of cross validation or other re-estimation procedures ([Section 5](#)).

It is important to note that the results for MLE trend filtering and mean trend filtering are not directly comparable because they make different assumptions. The former constrains the natural parameter, while the latter constrains the mean parameter. These only coincide in the Gaussian case. We present empirical results demonstrating our methods on synthetic and real datasets in [Section 6](#). We conclude with a discussion of the results. The remainder of this section gives a concise overview of our theoretical contributions and a thorough discussion of related work.

1.4. Overview of theoretical contributions

We provide a theoretical characterization of two different estimators (the penalized MLE and the mean trend filter) across all single parameter exponential

family distributions, smoothness levels k , and dimensions d . For the mean trend filter, we will measure the estimation risk with the mean squared error of the mean parameter. This is natural as it matches both the empirical loss, and standard results in the Gaussian setting. However, for the penalized MLE, we prefer to use $\overline{\text{KL}}(\theta^* \parallel \theta)$. This preference is for two main reasons. First, KL matches the log-likelihood loss function used in the optimization problem. Second, it captures the curved geometry of the probability spaces accurately, collapsing to MSE in the Gaussian (constant variance) case. And third, as in Equation (5), the KL between the distributions in the same exponential family simplifies to a convenient functional form depending only on the natural parameters.

To better fix the context for our results, we provide here a concise description of these in the simplest cases (more precise statements are in Sections 2 and 3). Define $\alpha = (k + 1)/d$, and define the “canonical scaling” as $\|D\theta^*\|_1 \lesssim n^{1-\alpha}$. The canonical scaling is called such because it holds for evaluations of Hölder functions—functions where the k th order partial derivatives are Lipschitz continuous—at the grid locations. Under the canonical scaling, it is shown (Sadhanala et al., 2021) that for Gaussian data and ℓ_2 loss, the minimax rate over this class is given by

$$\text{MSE}(\Theta) = \begin{cases} \Omega(n^{-\alpha}) & \alpha \leq 1/2, \\ \Omega(n^{-\frac{2\alpha}{2\alpha+1}}) & \alpha > 1/2, \end{cases}$$

where $\text{MSE}(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \frac{1}{n} E \|\hat{\theta} - \theta\|_2^2$. Furthermore, the mean trend filter is rate optimal up to logarithmic factors in the Gaussian case.

Because, for Gaussian data, $\text{KL}(\theta_0 \parallel \theta_1) \propto \|\theta_0 - \theta_1\|_2^2$, the above immediately provides a lower bound for $\overline{\text{KL}}(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \frac{1}{n} \text{KL}(\theta \parallel \hat{\theta})$ where the supremum is taken over all exponential families. We show that, under additional boundedness conditions on Θ and similar constraints on the estimator, the MLE trend filter in Equation (2) achieves this rate up to additional logarithmic factors, and is thus minimax rate optimal. The case of the MLE trend filter without the artificial boundedness constraint described above is more complicated (Section 2.2). With the additional penalty on the null space in Equation (3), an addition we prove necessary for consistency, we can achieve the minimax rate for $\alpha \leq 1/2$. For $\alpha > 1/2$, the upper bound is weaker than for Gaussian noise: we can show only that $\overline{\text{KL}}(\theta^* \parallel \hat{\theta}) = O_{\mathbb{P}}(n^{-1/2})$. While we are able to show consistency in this setting, we suspect that this bound is loose.

We also show that, under homoskedastic sub-exponential noise, the mean trend filter achieves the minimax rate up to additional logarithmic factors. The homoskedasticity condition can be relaxed, and this is examined in Section 3. We consolidate these results in Tables 2 and 3.

1.5. Related work

Much is known about trend filtering in one dimension (1d). The trend filtering method in (4) was proposed in Steidl, Didas and Neumann (2006); Kim et al.

TABLE 2

Overview of theoretical results for the Penalized MLE under canonical scaling. Logarithmic factors are ignored with \tilde{O} notation, and additional details are described in [Section 2](#).

Conditions	Regime	Lower bound	Upper bound	Literature
Gaussian	$\alpha \leq 1/2$	$\Omega(n^{-\alpha})$	$\tilde{O}_{\mathbb{P}}(n^{-\alpha})$	Sadhanala et al. (2021)
	$\alpha > 1/2$	$\Omega(n^{-\frac{2\alpha}{2\alpha+1}})$	$\tilde{O}_{\mathbb{P}}(n^{-\frac{2\alpha}{2\alpha+1}})$	
Exponential family (bounded)	$\alpha \leq 1/2$	$\Omega(n^{-\alpha})$	$\tilde{O}_{\mathbb{P}}(n^{-\alpha})$	Theorem B.3
	$\alpha > 1/2$	$\Omega(n^{-\frac{2\alpha}{2\alpha+1}})$	$\tilde{O}_{\mathbb{P}}(n^{-\frac{2\alpha}{2\alpha+1}})$	
Exponential family (null-space penalty)	$\alpha \leq 1/2$	$\Omega(n^{-\alpha})$	$\tilde{O}_{\mathbb{P}}(n^{-\alpha})$	Theorem 2.3
	$\alpha > 1/2$	$\Omega(n^{-\frac{2\alpha}{2\alpha+1}})$	$\tilde{O}_{\mathbb{P}}(n^{-1/2})$	

TABLE 3

Overview of theoretical results for the Mean Trend Filter under canonical scaling. Logarithmic factors are ignored with \tilde{O} notation, and additional details are described in [Section 3](#).

Conditions	Regime	Lower bound	Upper bound	Literature
Gaussian	$\alpha \leq 1/2$	$\Omega(n^{-\alpha})$	$\tilde{O}_{\mathbb{P}}(n^{-\alpha})$	Sadhanala et al. (2021)
	$\alpha > 1/2$	$\Omega(n^{-\frac{2\alpha}{2\alpha+1}})$	$\tilde{O}_{\mathbb{P}}(n^{-\frac{2\alpha}{2\alpha+1}})$	
sub-exponential noise (mild heteroskedasticity)	$\alpha \leq 1/2$	$\Omega(n^{-\alpha})$	$\tilde{O}_{\mathbb{P}}(n^{-\alpha})$	Theorem 3.3
	$\alpha > 1/2$	$\Omega(n^{-\frac{2\alpha}{2\alpha+1}})$	$\tilde{O}_{\mathbb{P}}(n^{-\frac{2\alpha}{2\alpha+1}})$	
sub-exponential noise	$d = 2, k = 0$	$\Omega(1)$	not consistent	Theorem 3.7

(2009) for 1d problems. [Tibshirani \(2014\)](#) connected trend filtering to locally adaptive regression splines, proposed in [Mammen and van de Geer \(1997\)](#), and analyzed its statistical properties. [Tibshirani \(2022\)](#) gives an in-depth background of the key ideas that make trend filtering and related methods work. [Johnson \(2013\)](#); [Kim et al. \(2009\)](#); [Ramdas and Tibshirani \(2016\)](#) propose methods to solve the convex optimization problem in 1d trend filtering. Trend filtering with $k = 0$, or total variation (TV) regularization, is an important technique for denoising images (two dimensions). TV methodology and computation was studied in [Rudin, Osher and Fatemi \(1992\)](#); [Tibshirani et al. \(2005\)](#); [Condat \(2013\)](#); [Barbero and Sra \(2018\)](#). Trend filtering over general graphs was first proposed in [Wang et al. \(2016\)](#), and subsequently, other variants of trend filtering have been studied, for example depth-first search TV regularization ([Madrid Padilla et al., 2018](#)), kNN TV denoising ([Madrid Padilla et al., 2020](#)), quantile trend filtering ([Madrid Padilla and Chatterjee, 2021](#)), and sequential TV denoising ([Baby and Wang, 2021](#)). These methods use squared error loss, with the exception of [Madrid Padilla and Chatterjee \(2021\)](#), and so are not necessarily suitable for general exponential families. [Zhang and Chatterjee \(2023\)](#) addresses the specific case of Fused Lasso, though under convex loss, allowing for an extension to heavy-tailed noise.

General exponential family distributions have a long history in statistics. [Brown \(1986\)](#) is a definitive treatment for studying the properties of exponential families while [McCullagh and Nelder \(1989\)](#) covers the details of general-

ized linear models. Direct analysis of trend filtering in this setting is more rare than for Gaussian loss. [Van de Geer \(2020\)](#) derived error bounds for estimating Bernoulli family parameters with bounded variation in 1d. In contrast to most other results, the theory applies without assuming boundedness of the estimated natural parameter. [Khodadadi and McDonald \(2019\)](#) examine computational approaches for variance estimation on spatiotemporal grids. [Kakade et al. \(2010\)](#) discuss strong convexity of general exponential families and use the results to analyze ℓ_1 -penalized maximum likelihood. [Vaiter et al. \(2017\)](#) examine the geometry of penalized generalized linear models and derive important results for general regularizers that we use for specialized risk estimation in [Section 4](#). [Bassett and Sharpnack \(2019\)](#) provides a bound on the Hellinger error for total variation denoising for the estimation of densities over edge segments in a general graph. Our results here are the first to analyze trend filtering over lattice graphs for general exponential families.

An important distinction exists between two varieties of theoretical results for trend filtering examined in the literature: (1) nearly parametric rates under sparsity assumptions with $\|D\theta^*\|_0$ bounded; and (2) non-parametric rates for signals with bounded trend filtering norm $\|D\theta^*\|_1$. In general, these bounds are difficult to compare because they hold under different conditions, and either bound can be tighter for specific signals. [Rinaldo \(2009\)](#); [Harchaoui and Levy-Leduc \(2010\)](#); [Lin et al. \(2017\)](#); [Guntuboyina et al. \(2020\)](#); [Ortelli and van de Geer \(2021\)](#) give more general and tighter error bounds when the true signal is sparse (bounded L_0 norm). Throughout this work, we will focus on establishing non-parametric rates with trend filtering norm bounds.

[Mammen and van de Geer \(1997\)](#) provide one of the earliest theoretical results on 1d trend filtering. In higher dimensions and on general graphs, researchers have typically confined their theory to special cases—e.g., specific dimensions, graph structure, and trend filtering order. [Hütter and Rigollet \(2016\)](#); [Sadhanala, Wang and Tibshirani \(2016\)](#) derive error bounds for total variation denoising (trend filtering with $k = 0$) on lattice graphs. [Chatterjee and Goswami \(2021\)](#); [Ortelli and van de Geer \(2020\)](#) show stronger error bounds when the signal has axis-parallel patches. [Sadhanala et al. \(2017, 2021\)](#), extend the analysis to higher-order trend filtering on lattice graphs of arbitrary dimension. All of the aforementioned works study squared error loss with sub-Gaussian noise. [Wang et al. \(2016\)](#) analyze error bounds for graph trend filtering for specific cases (lattice graphs with a specific trend filtering order). In that work, the “eigenvector incoherence” technique is developed as a tool to analyze the mean squared error of any graph trend filtering problem. In this work, we adapt this technique to work with general exponential families.

2. Penalized MLE

In this section, we provide general results for trend filtering on d -dimensional lattice graphs with exponential family observations. As mentioned above, general exponential families have two interesting features. First, the distributions can

be more heavy tailed than Gaussians, and as we have seen, they are generally sub-exponential. This is reflected in rates that are typically worse than in the Gaussian case. Second, the variance (as well as the sub-exponential parameters ν , b) is a function of the natural parameter, which results in heteroskedasticity. We find that our bounds rely heavily on the “level” of this heteroskedasticity. However, this reliance is most salient with respect to two asymptotic regimes.

We say *mild heteroskedasticity* occurs when both sub-exponential parameters, ν , b , are bounded as n increases. Henceforth, let ν , b denote the vectors $(\nu_i), (b_i)$ for $i \in [n]$ where these are the sub-exponential parameters of centered Y . That is, if there exists an ω such that $\|\nu\|_\infty, \|b\|_\infty \leq \omega$ for all n , we say that the problem is only mildly heteroskedastic. Analysis in this case turns out to be largely similar to the standard homoskedastic setting. We say that *strong heteroskedasticity* occurs whenever it is not *mild*, however, typically we can measure the strength via $\|\nu\|_\infty / \|\nu\|_n$. When this is close to 1, there is little variation of ν across coordinates. However, when $\|\nu\|_\infty / \|\nu\|_n$ is close to \sqrt{n} , only a few coordinates dominate. Importantly, smoothness of θ^* (such as a bound on $\|D\theta^*\|_1$) does not generally have any implications for the level of heteroskedasticity, and furthermore, it is not generally possible to determine the level from data. Thus, considering both situations is necessary for a complete understanding.

Much of the difficulty for both estimation and theoretical analysis in the exponential family setting is that the negative log-likelihood is not strongly convex in general. If we assume that $\psi''(\theta_i^*) > 1/K$ for all i , then we can add this constraint to Equation (2) which will ensure strong convexity. We provide an analysis of this approach in [Appendix B.7](#), which is tight (minimax rate optimal) up to logarithmic factors, see, for example, [Sadhanala et al. \(2021\)](#). Similar results were already derived in the literature, for example, in [Prasad et al. \(2020\)](#). As we will see, however, bounding the curvature in this way excludes important cases, and cannot be verified from data. Nonetheless, this assumption has a long history in statistics. For example, the standard approach to proving estimation consistency in low-dimensional generalized linear models is much the same ([McCullagh and Nelder, 1989](#)).

An alternative method to address the lack of strong convexity would be to add an ℓ_2 (or squared ℓ_2) penalty on θ . Such a modification is similar to the motivation for Tikhonov regularization as a stabilization for ill-posed problems: inhibiting $\theta \rightarrow \pm\infty$ enforces strong convexity. However, note that $\theta = (I - P_N)\theta + P_N\theta$ and $\|D\theta\|_1$ already prevents divergence due to the first part. Thus, it is only necessary to control divergence of θ in the null-space of D . Our main theoretical results in this section demonstrate that this penalty is sufficient for estimation consistency.

2.1. Additional penalty on the null space component of θ

The boundedness constraint discussed above is not desirable for at least two reasons. The first is that it is difficult to calibrate the constraint using data. The second is that strong convexity is an indirect way to get control of the

nullspace of D , which is what we actually need. We now argue why this is the case.

Let the empirical and population risks at a parameter θ be

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \psi(\theta_i) - y_i \theta_i, \quad \text{and} \quad R(\theta) = \frac{1}{n} \sum_{i=1}^n \psi(\theta_i) - E[Y_i] \theta_i,$$

respectively, and note that $\overline{\text{KL}}(\theta_0 \parallel \theta_1) = R(\theta_1) - R(\theta_0)$. For Gaussian data, minimization of the empirical risk, the $\|D\theta\|_1$ constraint, and the fact that the likelihood is strongly convex, together control the discrepancy between the empirical risk and the population risk. The reason is that strong convexity controls behaviour of $\hat{\theta}$ in the nullspace of D . But outside this setting, we no longer have strong convexity, and unfortunately, the penalty alone does not give sufficient control to guarantee the existence of a solution. The result is that, for non-Gaussian data, $\sup_{\theta \in \Theta} |R_n(\theta) - R(\theta)|$ can become arbitrarily large with high probability, even in simple settings, despite bounds on $\|D\theta\|_1$. Suppose $\Theta = \{\theta : \|D\theta\|_1 \leq 1\}$ where $D = D_{n,1}^{(1)}$.

Remark (Degenerate Poisson example). Consider the Poisson family, let $n \geq 1$, and define the true parameter to be $\theta^* \in \mathbb{R}^n$ such that $\theta_j^* = -2 \log n$. This θ_j^* is constant, so that $(D\theta^*)_j = 0, \forall j$, that is $\theta^* \in \mathcal{N}(D)$. Then, the probability that all y_i 's are 0 is $e^{-1/n} \rightarrow 1$. On this event (where $y_i = 0 \forall i$), for any $\lambda \geq 0$,

$$\inf_{\theta} \sum_{i=1}^n e^{\theta_i} - y_i \theta_i + \lambda \|D\theta\|_1 = \inf_{\theta} \sum_{i=1}^n e^{\theta_i} + \lambda \|D\theta\|_1 = 0,$$

because $\lim_{c \rightarrow -\infty} \sum_{i=1}^n e^c + \lambda \|Dc\mathbf{1}\|_1 = 0$. Furthermore, observe that as $c \rightarrow -\infty$, $R(c\mathbf{1}) \rightarrow \infty$ even though $R_n(c\mathbf{1}) \rightarrow 0$. Notice that, here, $\psi''(\theta_i^*) = n^{-2}$, so the strong convexity bound is diminishing with n .

In this example, as $n \rightarrow \infty$, y will be identically 0 with increasingly high probability, in which case the MLE will be $\hat{\theta}_j = -\infty$, and the penalized MLE will result in the same outcome. However $\theta_j^* \rightarrow -\infty$ relatively slowly, so our estimates become increasingly inaccurate. The trend filtering penalty, $\|D\theta\|_1$, does fix this issue. Therefore, we must impose additional constraints on the null space component $P_{\mathcal{N}}\theta$ to guarantee a (finite) solution to the optimization problem. The trend filtering penalty ignores any part of $\theta \in \mathcal{N}(D)$, so, as long as that component is asymptotically bounded, the negative loglikelihood will be bounded as well. But if the null-space component is unbounded, then this estimator will fail, because the MLE will be infinite, and the penalty doesn't compensate.

One can observe similar behaviour for the logistic family whenever θ^* approaches the boundary of Θ . Consider the same θ^* and verify that all y_i 's are 0 with probability $(1 + n^{-2})^{-n} \approx e^{-1/n}$. The MLE with only the $\|D\theta\|_1$ penalty behaves similarly to the Poisson example described above.

In both cases, signals θ^* which have pathological behaviour in the null space of D are unaffected by the trend filter penalty. Furthermore, this phenomenon

is not restricted to the high-dimensional setting. Rather it is a consequence of minimizing the negative log-likelihood and can be easily demonstrated in the univariate setting. Consider a single Poisson random variable, X with intensity e^θ and natural parameter θ . Then the MLE for θ is $\hat{\theta} := \log X$, which is $-\infty$ with a non-zero probability and, hence, has no mean or variance. This unfortunate behaviour can be mitigated by regularizing the problem with an additional ridge penalty ($-\ell(\theta; X) + \lambda\theta^2$ where $\lambda > 0$). The solution to this problem satisfies $e^{\hat{\theta}} = X - 2\lambda\hat{\theta}$ which is finite for all X , guaranteeing the existence of an estimator with finite risk. Unfortunately, the trend filter penalty alone can not avoid this pathology.

While artificially imposing strong convexity addresses this issue, it is both more direct and results in a more tractable estimator to constrain the component of θ in the null space of D . With this additional constraint, we can show the following risk bound. The proof is in [Appendix B.3](#).

Proposition 2.1. *Let $\Theta = \{\theta \in \mathbb{R}^n : \|P_{\mathcal{N}}\theta\|_n \leq a_n, \|D\theta\|_1 \leq c_n n^{1-\alpha}\}$ where $\mathcal{N} = \text{null}(D)$ and $\alpha = (k+1)/d$. Suppose ϵ_i is zero mean sub-exponential with parameters (ν_i^2, b_i) for $i \in [n]$. Assume $\|\nu\|_\infty, \|b\|_\infty \leq c$ where c is a constant. Then*

$$\sup_{\theta \in \Theta} |R_n(\theta) - R(\theta)| = O_{\mathbb{P}} \left(\frac{a_n \log n}{\sqrt{n}} + \frac{c_n \gamma \log n}{n^{\alpha \wedge 1/2}} \right)$$

where $\gamma = \log^{1/2} n$ if $2\alpha = 1$ and 1 otherwise.

For the above example of degenerate Poisson, we can set $a_n = 2 \log n$, $c_n = 0$ to see that the right hand side converges to 0 as $n \rightarrow \infty$. This motivates us to penalize the null space component of θ in the MLE and use the estimator defined in (3) rather than that in (2). In the following, we call this estimator (3), the MLE and define $\alpha = (k+1)/d$. The minimizer in the optimization problem is unique because ψ is strictly convex.

2.2. Error bounds for penalized MLE

Generally, there are three degrees of freedom when stating results: (1) the trend filtering order k , (2) the dimension d , and (3) the exponential family and resulting sub-exponential parameters (ν, b) . There is a natural trade-off between generality and interpretability of the results presented here, so we will prefer to present specific interpretable results as corollaries.

We introduce some additional notation to state our results. Let ρ_ℓ , $\ell \in [N]$ be the eigenvalues of $D_1^\top D_1$ where $D_1 = D_{N,1}^{(k+1)}$, $N = n^{1/d}$, ordered in from largest to smallest. Abbreviate $D = D_{n,d}^{(k+1)}$ and let $\xi_i^2 : i = (i_1, \dots, i_d) \in [N]^d$ be the eigenvalues of $D^\top D$ ordered from smallest to largest. Due to the Kronecker-sum structure of $D^\top D$, we have $\xi_i^2 = \sum_{j=1}^d \rho_{i_j}$. Let $\kappa = (k+1)^d$ denote the nullity of $D^\top D$. A nonzero vector $x \in \mathbb{R}^n$ is said to be *incoherent* with a constant $\mu \geq 1$ if $\|x\|_\infty / \|x\|_n \leq \mu$. Note that, for arbitrary nonzero $x \in \mathbb{R}^n$, $\|x\|_\infty / \|x\|_n \in$

$[1, \sqrt{n}]$. For $J \subset [N]^d$ containing $[k+1]^d$, define

$$L_{J,p} = \left(\frac{\mu^2}{n} \sum_{i \in [N]^d \setminus J} \frac{1}{\xi_i^p} \right)^{1/p} \quad (6)$$

where μ is the constant with which the left singular vectors of D are incoherent. We can derive the following error bound on the excess risk of the estimator in (3).

Theorem 2.2. *Let $y_i = \beta_i^* + \epsilon_i$ where ϵ_i are independent, have zero mean, and are sub-exponential with parameters (ν_i^2, b_i) for $i \in [n]$. Let L be as defined in (6). For $t \geq 1$, abbreviate $A_n = 2t\mu\sqrt{\kappa/n}(\|\nu\|_2 \vee \|b\|_\infty)$, $B_n = 2t(\min\{\|\nu\|_\infty L_{\kappa,2}, \|\nu\|_2 L_{\kappa,1}\} \vee \|b\|_\infty L_{\kappa,1})$ where $L_{\kappa,p} = L_{[k+1]^d,p}$ for $p \geq 1$. Let $\hat{\theta}$ be our estimate in (3) with parameters $\lambda_2 = 2A_n/n$ and $\lambda_1 = 2B_n/n$. Then, with probability at least $1 - 4nde^{-t}$,*

$$\begin{aligned} \overline{\text{KL}}(\theta^* \parallel \hat{\theta}) &\leq \frac{3}{n}(A_n \|P_{\mathcal{N}}\theta^*\|_2 + B_n \|D\theta^*\|_1), \quad \text{and} \\ A_n \|P_{\mathcal{N}}\hat{\theta}\|_2 + B_n \|D\hat{\theta}\|_1 &\leq 3(A_n \|P_{\mathcal{N}}\theta^*\|_2 + B_n \|D\theta^*\|_1). \end{aligned}$$

See the proof in [Appendix B.1](#).

Remark. [Theorem 2.2](#) is general. As mentioned in [Section 1](#), it applies not only when Y_i is (single parameter) exponential family, but also to the case that it comes from any distribution with sub-exponential tails under arbitrary heteroskedasticity. Furthermore, replacing D with $\Delta^{(k+1)}$, the k^{th} order graph-difference operator as defined by [Wang et al. \(2016, Equation 6\)](#), results in the same bound. However, controlling $L_{\kappa,p}$ for other graphs is outside the scope of this paper, and we do not pursue it here. Along the same lines, contrary to the Gaussian case, conditions for good behaviour of $\|\nu\|_2$, $\|\nu\|_\infty$, and $\|b\|_\infty$ requires specialized treatment for the specific distribution of Y . The corollaries that follow are all special cases of this main theorem, treating specific cases where we can provide precise and interpretable rates of convergence.

For regular grids, [Theorem B.6](#) (in [Appendix B.9](#)) controls the magnitude of $L_{\kappa,1}, L_{\kappa,2}$ and hence the bounds in [Theorem 2.2](#). Applying the lemma to the expression for B_n in [Theorem 2.2](#), we get the following corollary for regular grids.

Corollary 2.3. *Assume canonical scaling $\|D\theta^*\|_1 \lesssim n^{1-\alpha}$, and $\|P_{\mathcal{N}}\theta^*\|_n \lesssim 1$. Then for $t \geq 1$,*

$$\overline{\text{KL}}(\theta^* \parallel \hat{\theta}) = O_{\mathbb{P}}(r_n \log n),$$

with

$$r_n = \begin{cases} (\|\nu\|_\infty + \|b\|_\infty)n^{-\alpha} & \alpha < 1/2 \\ \|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-\frac{1}{2}\gamma_2}, \|\nu\|_2 n^{-\alpha} \gamma_1\} & \alpha \in [1/2, 1] \\ \frac{1}{n}(\|\nu\|_2 + \|b\|_\infty) & \alpha > 1 \end{cases}$$

and $\gamma_p = (\log n)^{1/p} \mathbf{1}(p\alpha = 1)$ for $p \geq 1$.

As for [Theorem 2.2](#), this result applies to data with sub-exponential tails, and therefore also the subclass of (single-parameter) exponential family distributions. For the special case of Gaussian errors with $\nu = \sigma \mathbf{1}$, $b = 0$, we recover minimax optimal rates where $\alpha \leq 1/2$ up to logarithmic factors (see for example [Sadhanala et al. 2021](#)). However, we get suboptimal rates when $\alpha > 1/2$.

2.3. Penalized MLE in special cases

We now illustrate [Theorem 2.3](#) in a few special cases to provide intuition. As above, we focus on grid graphs with Poisson and Exponential distributions, and we assume that these are all of width N and dimension d , so that $n = N^d$. Recall that for natural parameter θ^* , the Poisson distribution has mean $\beta^* = \exp(\theta^*)$, while the Exponential distribution has mean $\beta^* = -1/\theta^*$. For the Poisson distribution, an additive change in θ^* results in a multiplicative change in the mean, and $\nu^2 = 2\beta^*$, which can easily result in strong heteroskedasticity. Only in special cases does a constraint on $\|D\theta^*\|_1$ result in a bound on ν^2 , and generally, $\|\nu\|_\infty$ will depend on the signal in question. The specific choices of k below allow us to vary d and illustrate the effects via $\alpha = (k+1)/d$.

The first result is an example of weak heteroskedasticity, where the natural parameter is uniformly bounded.

Corollary 2.4. *Consider the Poisson distribution where the natural parameter vector θ^* satisfies $\|\theta^*\|_\infty = O(1)$. Let $k = 1$ and assume that θ^* satisfies the canonical scaling, $\|D\theta^*\|_1 = O(n^{1-2/d})$. Then, we have the following rate bound for penalized MLE trend filtering.*

$$\overline{\text{KL}}(\theta^* \parallel \hat{\theta}) = O_{\mathbb{P}}(r_n \log n), \quad \text{where} \quad r_n = \begin{cases} n^{-1/2}, & d = 1 \\ n^{-1/2} \log n, & d = 2 \\ n^{-1/2}, & d = 3 \\ n^{-1/2} \log^{1/2} n, & d = 4 \\ n^{-2/d}, & d > 4. \end{cases}$$

A simple example of such a signal is $\theta_i^* = \frac{2}{N} \sum_{j=1}^d |i_j - N/2|$, where $i = (i_1, \dots, i_d) \in [N]^d$. For a proof, see [Appendix B.2](#).

The next example demonstrates [Theorem 2.3](#) under strong heteroskedasticity.

Corollary 2.5. *Consider any exponential family on a d -dimensional grid ($d > 1$) with a natural parameter that satisfies $\|\nu\|_\infty, \|b\|_\infty = O(n^c)$ and $\|\nu\|_2 = O(n^c)$ for some $c > 0$. Under the canonical scaling for $k = 0$, $\|D\theta^*\|_1 = O(n^{1-1/d})$, then*

$$\overline{\text{KL}}(\theta^* \parallel \hat{\theta}) = O_{\mathbb{P}}(r_n \log n), \quad \text{where} \quad r_n = \begin{cases} n^{c-1/2}, & d = 1 \\ n^{c-1/2} \log^{1/2} n, & d = 2 \\ n^{c-1/d}, & d > 2. \end{cases}$$

An example of a signal satisfying these conditions is the Exponential distribution with $\theta_i^* = -n^{-c}\mathbf{1}\{i = 0\} - n^{1-1/d}\mathbf{1}\{i \neq 0\}$. The proof is in [Appendix B.2](#).

In this case, $\|\nu\|_\infty$ is diverging, and so we have strong heteroskedasticity. The level of heteroskedasticity, parameterized by c , determines the rate of convergence and for $c > 1/d$ we cannot guarantee convergence.

3. Error bounds for the Mean Trend Filter

When $k = 0$, remarkably, it turns out that the penalized MLE in (2) is equivalent to the mean trend filtering estimator (4). In fact, this equivalence between the two estimators holds over arbitrary graphs, not just grids.

Theorem 3.1. *Suppose $k = 0$ and let D be the graph incidence matrix. Then, the penalized MLE $\hat{\theta}$ in (2) and the least squares estimator $\hat{\beta}$ in (4) satisfy $\hat{\beta} = \psi'(\hat{\theta})$.*

The proof is elegant and short, relating the Karush–Kuhn–Tucker conditions for the two optimization problems, and is deferred to [Section B.4](#). Therefore, in the case $k = 0$, the penalized MLE can be solved quickly by solving the equivalent mean trend filter problem. This fact was also noted by [Vu \(2018\)](#) and [Dümbgen and Kovac \(2009\)](#) though the proofs and goals are quite different.

For $k \geq 1$, equivalence between the two estimators need not hold in general, with the exception of the mean parameterized Gaussian family, where it holds trivially. The remainder of this section will focus on the general case. For the estimator in (4), we derive the following error bound.

Theorem 3.2. *Let $y_i = \beta_i^* + \epsilon_i$ where ϵ_i is zero mean sub-exponential with parameters (ν_i^2, b_i) for $i \in [n]$. Let $J \subset [N]^d$ and L be as defined in (6). For $t \geq 1$, abbreviate $A_n = 2t\mu\sqrt{|J|}/n(\|\nu\|_2 \vee \|b\|_\infty)$, $B_n = 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1})$. For any $J \subset [N]^d$ containing $[k+1]^d$, the estimator (4) with $\lambda = B_n/n$, satisfies*

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{4A_n^2}{n} + \frac{8B_n}{n} \|D\beta^*\|_1$$

with probability at least $1 - 4nde^{-t}$ for $t \geq 1$.

The set of indices J can be chosen to minimize the bound. The proof is in [Appendix B.5](#) and follows an approach similar to that in [Wang et al. \(2016\)](#). Tail bounds on sums of sub-Gaussian variables in their results are replaced with those on sums of sub-exponential variables. This results in additional $\log n$ factors in the error bound compared to the sub-Gaussian setting.

The proof technique for [Theorem 3.2](#) relies on the properties of D . A potential alternative route to get error rates is via bounding the empirical process $\frac{1}{n}\epsilon^\top(\hat{\theta} - \theta^*)$ with the Dudley entropy integral. However, the empirical process in our case is not sub-Gaussian and we could only derive a trivial upper bound in this way. This should not be entirely surprising however, because the entropy method was also used in [Wang et al. \(2016\)](#) in the sub-Gaussian noise setting, and it also

failed to give a tight characterization in that context. We note, as above, that this theorem is general, applying equally to the case of graphs by replacing D with $\Delta^{(k+1)}$. We do not pursue this generality further.

3.1. Error bounds with canonical scaling

We simplify this bound in some special cases. Assuming that ν , b are uniformly bounded, we get the following result for regular grids. Denote $\gamma_p = \log^{1/p}(n)$ if $p\alpha = 1$ and 1 otherwise.

Corollary 3.3. *Assume $\|\nu\|_\infty \leq \omega$, $\|b\|_\infty \leq \omega$. Let $\alpha = (k+1)/d$. For d -dimensional grids, assume that $\|D\beta^*\|_1 \asymp n^{1-\alpha}$. Then there is a choice of λ such that for $\alpha \leq 1/2$,*

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\frac{\omega^2 \log^2 n}{n} + \frac{\omega \gamma_2 \log n}{n^\alpha} \right)$$

and for $\alpha > 1/2$ and $n^{-\alpha} \leq \omega \log n \lesssim \sqrt{n}$

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\left(\frac{\omega^2 \log^2 n}{n} \right)^{\frac{2\alpha}{2\alpha+1}} + \frac{\omega \gamma_1 \log n}{n^\alpha} \right).$$

The proof is in [Appendix B.6](#). This corollary does not discuss the case where $\alpha > 1/2$ and $\omega \log n$ is outside of $[n^{-\alpha}, \sqrt{n}]$. In that case, when the noise is high ($\omega \log n \gtrsim \sqrt{n}$), the polynomial projection estimator $\hat{\beta} = P_{\mathcal{N}} y$ gives the tightest bound, and, when the noise is low ($\omega \log n < n^{-\alpha}$), the identity estimator gives the tightest bound.

The following corollary examines this result for some special cases.

Corollary 3.4. *Consider the both Poisson and Exponential distributions on a d -dimensional grid ($d > 1$) where the mean parameter is constrained. Specifically, suppose that $\|\beta^*\|_\infty = O(1)$ such that the canonical scaling holds with $k = 1$. Then mean trend filter satisfies*

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}}(r_n) \text{ where } r_n = \begin{cases} (n/\log^2 n)^{-4/(4+d)}, & d = 1, 2, 3 \\ n^{-1/2} \log^{3/2} n, & d = 4 \\ n^{-2/d} \log n, & d > 4. \end{cases}$$

This result matches with rates in the homoskedastic Gaussian case up to logarithmic factors, shown for example in [Sadhanala et al. \(2021\)](#). An example of a signal satisfying the conditions is a grid graph with width N and dimension d , so that $n = N^d$ and $\beta_i^* = \frac{d}{N} + \frac{2}{N} \sum_{j=1}^d |i_j - \frac{N}{2}|$, where $i = (i_1, \dots, i_d) \in [N]^d$. The proof is given in [Appendix B.6](#).

While the previous result treated the (effectively) homoskedastic case by controlling the largest components of ν , b , the following corollary specializes [Theorem 3.2](#) to canonical scaling under strongly heteroskedastic noise.

Corollary 3.5. *Let $\sigma = (\|\nu\|_2 \vee \|b\|_\infty)/\sqrt{n}$, $\sigma_\infty = \|\nu\|_\infty \vee \|b\|_\infty$. Suppose $\|D\beta^*\|_1 \lesssim n^{1-\alpha}$, and assume $\sigma^2 \lesssim n/\log^2 n$ and $\sigma_\infty \lesssim n^\alpha/(\gamma_1\gamma_2 \log n)$. Then, the estimator $\hat{\beta}$ in [Theorem 3.2](#) satisfies*

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = \begin{cases} O_{\mathbb{P}} \left(\frac{\sigma^2 \log^2 n}{n} + \frac{\sigma_\infty \gamma_2 \log n}{n^\alpha} \right), & \alpha \leq 1/2 \\ O_{\mathbb{P}} \left(\frac{\sigma^2 \log^2 n}{n} + \frac{\min\{\sigma_\infty, \sigma \gamma_1 n^{1-\alpha}\} \log n}{n^{1/2}} \right), & 1/2 < \alpha \leq 1, \\ O_{\mathbb{P}} \left(\left(\frac{\sigma^2 \log^2 n}{n} \right)^{\frac{2\alpha-1}{2\alpha}} + \frac{\sigma_\infty \log n}{n^\alpha} \right), & \alpha > 1. \end{cases}$$

This result is most useful under strong heteroskedasticity where $\sigma_\infty/\sigma \propto \sqrt{n}$, and slightly stronger rates with weaker heteroskedasticity can be obtained in the $1/2 < \alpha \leq 1$ case (see [Theorem B.2](#) in [Appendix B.6](#)). Suppose ϵ_i in [Theorem 3.2](#) is mean-zero Laplace noise with standard deviation parameter τ_i and that $\|D\beta^*\|_1$ satisfies canonical scaling. For this case, $\nu_i = b_i = c\tau_i$ for a constant c independent of β_i^* , while $\sigma = c\|\tau\|_n$ and $\sigma_\infty = c\|\tau\|_\infty$ with the natural constraint that $\sigma_\infty/\sqrt{n} \leq \sigma \leq \sigma_\infty$. For $\alpha < 1$, the scaling requirement on σ_∞ is stronger, meaning that the estimator can only tolerate heteroskedasticity on the order of $\sigma_\infty/\sigma \propto n^\alpha < \sqrt{n}$. On the other hand, for $\alpha > 1/2$, the constraint on σ is stronger, meaning that we can tolerate $\sigma_\infty/\sigma \propto \sqrt{n}$. The associated rates of convergence will necessarily be much slower than in the homoskedastic sub-Gaussian case.

Importantly, [Theorem 3.5](#) illustrates that without control of the amount of heteroskedasticity, we cannot guarantee convergence of the estimator. In other words, while the estimator can tolerate strong heteroskedasticity as we have defined it here, it cannot tolerate arbitrary heteroskedasticity. Simply controlling $\|D\beta^*\|_1$ is not generally enough to guarantee estimation consistency. In the next section, we make this precise, illustrating that in certain settings, there is no estimator that can achieve consistency without additional constraints.

3.2. Lower bounds for mean trend filtering

We now show that the upper bound in [Theorem 3.3](#) is minimax optimal up to logarithmic factors. Consider the observation model

$$y_i = \beta_i + \epsilon_i, \quad i \in [n] \tag{7}$$

where $\beta \in \mathbb{R}^n$ is the true signal and ϵ_i , $i \in [n]$ are mean-zero noise terms. For a set $S \subset \mathbb{R}^n$ denote its minimax risk

$$\text{MSE}(S) = \inf_{\hat{\beta}} \sup_{\beta \in S} E \left[\|\hat{\beta} - \beta\|_n^2 \right]$$

where $\hat{\beta}$ is measurable in the observations $y \in \mathbb{R}^n$. Consider the Kronecker total variation (KTV) set

$$T_{n,d}^k(C_n) = \left\{ \beta : \|D_{n,d}^{(k+1)} \beta\|_1 \leq C_n \right\},$$

for integers $k \geq 0$, $d \geq 1$, $n \geq (k+1)^d$ and $C_n \geq 0$. Let $\text{Lap}(\mu, \sigma)$ denote the Laplace distribution centered at $\mu \in \mathbb{R}$ and with scale parameter $\sigma > 0$ with density $p(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}$ over \mathbb{R} .

Proposition 3.6. *Consider the observation model in (7) where ϵ_i , $i \in [n]$ are i.i.d. $\text{Lap}(0, \sigma)$ for a parameter $\sigma > 0$. Then,*

$$\text{MSE}(T_{n,d}^k(C_n)) = \Omega \left(\frac{\sigma^2}{n} + \frac{\sigma C_n}{n} \log \left(\frac{\sigma n}{C_n} \right) + \left(\frac{C_n}{n} \right)^{\frac{2}{2\alpha+1}} \left(\sigma^{\frac{4\alpha}{2\alpha+1}} \wedge \sigma^2 \right) \right)$$

where the Ω notation absorbs constants depending only on k , d .

The first term in the bound is due to the null space of D . To derive the second term, we embed an ℓ_1 ball in $T_{n,d}^k(C_n)$ and adapt arguments from Birge and Massart (2001). The final term is obtained similarly to Sadhanala et al. (2017, Theorem 4), by embedding a Hölder ball of appropriate size in $T_{n,d}^k(C_n)$. The proof is in Appendix C.1.

Let us compare the lower bound in Theorem 3.6 with the upper bound in Theorem 3.3. The Laplace distribution with scale parameter σ is sub-exponential with parameters $\nu = c\sigma$, $b = c\sigma$ for some constant $c > 0$. Plugging in $C_n = n^{1-\alpha}$ in the lower bound, and $\omega = c\sigma$ in the upper bound stated in Theorem 3.3, we can verify that the bounds match up to logarithmic factors.

The lower bound in Theorem 3.6 is for homoskedastic noise. When the noise is heteroskedastic, the estimation can be harder, in the sense that the minimax risk can be larger. Specifically, we can show the following lower bound on a TV class of signals for the Exponential family.

Proposition 3.7. *Assume $C_n > 1$. Consider the class of signals over a 2d grid*

$$\Theta(C_n) = \left\{ \beta \in \mathbb{R}^n : \|D_{n,2}^{(1)} \beta\|_1 \leq C_n, \|\beta\|_\infty \leq 2C_n \right\}$$

and the observation model $y_i \sim \text{Exp}(\text{mean} = \beta_i)$ for $i \in [n]$. Then

$$\text{MSE}(\Theta(C_n)) \geq \frac{3}{256} \frac{C_n^2}{n}.$$

The proof is in Appendix C.2. With canonical scaling $C_n \asymp n^{1-\alpha} = \sqrt{n}$, this means a lower bound of $\Omega(1)$. In other words, there is no consistent estimator for the class of signals $\Theta(\sqrt{n})$. This result also hints at the difficulty of handling various regimes of noise parameters ν , b .

4. Algorithmic implementation

In this section, we discuss our algorithmic implementation, focusing on the multivariate setting for the MLE trend filter for which there are not currently generic procedures. For the Mean Trend Filter, there are many standard approaches that can apply immediately since this is a quadratic program. In the one dimensional

case with $k = 0$, [Kim et al. \(2009\)](#) use a Primal-Dual Interior-Point method. [Ramdas and Tibshirani \(2016\)](#) examine a fast ADMM algorithm for $k > 0$. [Wang et al. \(2016\)](#) develop ADMM and Newton methods for general graphs and arbitrary k . We follow the approach of [Khodadadi and McDonald \(2019\)](#) for the MLE trend filter (2) and use an algorithm called linearized ADMM. We focus on (2) rather than (3) for simplicity and to provide intuition. A more complete description is given in [Appendix D](#). Note that the necessary changes to solve (3) are minor.

First, rewrite Equation (2) (substituting x for θ) as

$$\min_{Dx=z} \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1.$$

This is equivalent to (2) but with additional variables. The scaled form of the augmented Lagrangian for this problem is

$$L_\rho(x, z, u) = \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1 + \frac{\rho}{2} \|Dx - z + u\|_2^2 - \frac{\rho}{2} \|u\|_2^2.$$

The scaled ADMM algorithm iteratively solves this problem by minimizing over x , then z and then updating u with gradient ascent. However the x solution involves a matrix inversion due to the quadratic in Dx which is best avoided when n is large. So we linearize $L_\rho(x, z, u)$ around the current value x^o resulting in the following update for x

$$x \leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \rho (D^\top Dx^o - D^\top z + D^\top u)^\top x + \frac{\mu}{2} \|x - x^o\|_2^2, \quad (8)$$

where μ is chosen as the largest eigenvalue of $D^\top D$. To include the null space penalty, the changes only impact the x update, and (8) is adjusted accordingly:

$$\begin{aligned} x \leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \rho (D^\top Dx^o - D^\top z + D^\top u)^\top x \\ + \lambda_2 (g(x^o))^\top x + \frac{\mu}{2} \|x - x^o\|_2^2, \end{aligned}$$

where $g(v)$ is a subgradient of the function $v \mapsto \|P_{\mathcal{N}}v\|_2$ given by $g(v) = P_{\mathcal{N}}v/\|P_{\mathcal{N}}v\|_2$ when $P_{\mathcal{N}}v \neq 0$ and $g(v) = 0$ when $P_{\mathcal{N}}v = 0$.

The solution for the z -update is easily shown to be given by elementwise soft-thresholding, and the u -update is simply vector addition. Solving the x -update is potentially more challenging. Note that the form of (8) is the same for each i , so we can solve n one-dimensional problems. The KKT stationarity condition requires

$$0 = (\psi'(x_i) - y_i) + \rho (D^\top (Dx^o - z + u))_i + \mu(x_i - x_i^o).$$

Therefore, for any negative loglikelihood as given by ψ , we want to solve $\psi'(x_i) + \mu x_i = b_i$, for each $i \in [n]$. For many functions ψ , the solution has a closed form. The binomial distribution with $\psi(x) = \log(1 + e^x)$ is an exception, though

Algorithm 1 Linearized ADMM for the MLE trend filter

```

1: Input:  $y, \phi, D, \lambda_1 > 0, \lambda_2 \geq 0$ 
2: Set:  $x^o = \phi'^{-1}(y), \rho = \lambda_1, z = u = 0, \mu = \lambda_{\max}(D^\top D)$ 
3: while Not converged do
4:   Set  $b = y - \rho D^\top (Dx^o - z + u) + \mu x^o + \lambda_2 P_{\mathcal{N}} x^o / \|P_{\mathcal{N}} x^o\|_2$ 
5:   Update  $x$  by solving  $\psi'(x_i) + \mu x_i = b_i$  for  $i \in [n]$ .
6:   Update  $z \leftarrow \text{Soft}_{\lambda/\rho}(Dx + u)$  with  $\text{Soft}_a(v) = \text{sign}(v)(|v| - a)_+$ .
7:   Update  $u \leftarrow u + Dx - z$ 
8: end while
9: return  $z$ 

```

standard root finding methods have no difficulties. To include the nullspace penalty, the x update changes slightly, but the logic is the same. This procedure is shown in [Algorithm 1](#). In practice, we have found the algorithm to converge quickly when initialized from a small value of λ_1 (because the solution will be close to the MLE) and then calculated for an increasing sequence with the solution at smaller λ_1 used as a warm start. This is the opposite of most coordinate descent-style procedures which use a decreasing sequence of λ_1 . In terms of computational complexity, this ADMM algorithm is linear in n for each iteration. Careful implementations, customized to both a particular lattice and a specific exponential family, could yield PDIP or other alternatives with linear per-iteration complexity in n as well (due to the structure of D). Our implementation is easily generalizable across these conditions, though it may be empirically slower in individual cases. In practice, we have found that our algorithm requires more iterations, though each iteration is much faster than alternatives.

5. Degrees of freedom and tuning parameter selection

We describe an unbiased estimator for the KL divergence between the estimate and the truth for the purposes of tuning parameter selection. Additional justification and description of its derivation are given in [Appendix E](#). If $Y \sim \mathcal{N}(\theta^*, \sigma^2)$, a now common method of risk estimation makes use of Stein's Lemma. The utility of this result comes from examining the decomposition of the mean squared error of $\hat{\theta}(Y)$ as an estimator of θ^* .

$$\begin{aligned}
E \left[\|\theta^* - \hat{\theta}(Y)\|_2^2 \right] &= E \left[\|Y - \hat{\theta}(Y)\|_2^2 \right] - n\sigma^2 + 2 \text{tr Cov}(Y, \hat{\theta}(Y)) \\
&= E \left[\|Y - \hat{\theta}(Y)\|_2^2 \right] - n\sigma^2 + 2\sigma^2 E \left[\text{tr } J\hat{\theta}(z) \Big|_Y \right],
\end{aligned}$$

where J denotes the Jacobian. This characterization motivates the definition of degrees-of-freedom for linear predictors: $\text{df} := \frac{1}{\sigma^2} \text{tr } J\hat{\theta}(z) \Big|_y$ ([Efron, 1986](#)), where $\hat{\theta}(y) = Hy$. Using Stein's Lemma, assuming σ^2 is known, we have Stein's Unbiased Risk Estimator

$$\text{SURE}(\hat{\theta}) = \|y - \hat{\theta}\|_2^2 - n\sigma^2 + 2\sigma^2 \text{tr} \left(J\hat{\theta}(z) \Big|_y \right),$$

which satisfies $E[\text{SURE}(\hat{\theta})] = E\|\theta^* - \hat{\theta}(Y)\|_2^2$. Note that this is the risk for estimating the n -dimensional parameter θ^* . This estimator is appropriate for the mean trend filter, but, for the MLE trend filter, we prefer “Stein’s Unbiased KL” estimator due to Deledalle (2017) that applies to continuous exponential families.

Lemma 5.1 (simplified from Theorem 4.1 in Deledalle 2017). *For the MLE trend filter, whenever h is weakly differentiable,*

$$\text{SUKL}(\hat{\theta}) = \left\langle \hat{\theta} + \frac{\nabla h(y)}{h(y)}, \hat{\beta} \right\rangle + \text{tr} \left(J\hat{\beta}(z)|_y \right) - \psi(\hat{\theta})$$

is unbiased for $E[\text{KL}(\hat{\theta}(Y) \parallel \theta^)] - \psi(\theta^*)$.*

Because $\psi(\theta^*)$ does not depend on $\hat{\theta}$, we can ignore it for the purposes of choosing λ_1, λ_2 in the MLE trend filter. Note that for any continuous exponential family, h is weakly differentiable. To evaluate $\text{SUKL}(\hat{\theta})$ we need an expression for $J\hat{\beta}(y)$. This is given in the following result (the proof is deferred to Appendix E).

Theorem 5.2. *For the MLE trend filter, the divergence of $\hat{\beta}(y)$, defined to be the trace of the Jacobian of $y \mapsto \hat{\beta}(y)$, written as $\text{tr} \left(J\hat{\beta}(y) \right)$, is given by*

$$\begin{aligned} & \text{tr} \left(J\hat{\beta}(y) \right) \\ &= \text{tr} \left(\text{diag} \left(\psi''(\hat{\theta}) \right) P_{\mathcal{N}(\check{D})} \left(P_{\mathcal{N}(\check{D})} \text{diag} \left(\psi''(\hat{\theta}) \right) P_{\mathcal{N}(\check{D})} + \lambda_2 P_{\mathcal{N}} \right)^\dagger P_{\mathcal{N}(\check{D})} \right), \end{aligned}$$

where $P_{\mathcal{N}(\check{D})}$ is the projection onto the null-space of \check{D} , and \check{D} contains the rows of D such that $D\hat{\theta} = 0$.

Unfortunately, estimating the risk in this manner is not known to be possible for general discrete exponential families, though a few specific cases are possible. One such is the Poisson distribution. The following result more closely resembles an empirical derivative of $\hat{\beta}$ rather than the theoretical expression for $J\hat{\beta}(y)$ used in the previous results.

Lemma 5.3 (Theorem 4.2 in Deledalle 2017). *Assume Y is Poisson with independent coordinates conditional on θ . Then*

$$\text{PUKL}(\hat{\theta}) = \|\hat{\beta}\|_1 - \langle y, \log \hat{\beta}_\downarrow(y) \rangle,$$

is unbiased for $E[\text{KL}(\theta^ \parallel \hat{\theta}(Y))] - z(\theta^*)$ where $[\hat{\beta}_\downarrow(y)]_i = [\hat{\beta}(y - e_i)]_i$, where e_i is the i^{th} standard basis vector, and z is a known function of the true parameter.*

With these expressions in hand, we can select the tuning parameters λ_1, λ_2 with minimal additional computations by minimizing $\text{SUKL}(\hat{\theta})$ or $\text{PUKL}(\hat{\theta})$ as appropriate.

6. Empirical results

We demonstrate the performance of both the MLE and the Mean trend filter estimators in a small scale simulation designed to compare the two in challenging settings. We also examine two applications: modeling hospital admissions by age due to COVID-19 in Davis, California; and describing changes in temperature measurements for the Northern hemisphere. Code to reproduce all results is available on [GitHub](#).

6.1. Simulation study

We briefly investigate the relative performance of the Mean Trend Filter and the MLE Trend Filter on a few synthetic examples. Our intention is to push the limits of both, thereby illustrating that the user should choose between the two based on whether smoothness is desired in the mean or in the natural parameter. We focus on one dimension for ease of visualization and $k = 1$. We examine both the exponential distribution and the Poisson distribution.

To create the true signal, we begin with a v-shaped function on the unit interval:

$$f_n(x) = \frac{1}{n} + \left(1 - \frac{2}{n}\right) \left|x - \frac{1}{2}\right|$$

Evaluating this at n equally-spaced points for any n gives a signal such that $\|Df_n(x)\|_1$ has the canonical scaling $1/n$.

For the exponential distribution, we set either θ^* or β^* equal to $f_n(x)$ and evaluate both the Mean Trend Filter and the MLE Trend Filter on sample data. When θ^* is controlled, the mean at $x = 0.5$ approaches infinity as n grows, making estimation very challenging. The reverse occurs if β^* is controlled. For the Poisson, because the mapping from natural parameter to mean is exponential, controlling one does not particularly challenge the opposite procedure with the above f_n . To increase the discrepancy, we use $g_n(x) = 0.5 - f_n(x) + \log(n)$. The signal should create more discrepancy between the estimators as n grows, but results are less dramatic than those in the exponential case.

[Figure 2](#) shows estimation accuracy for both trend filters across four different scenarios. In all cases, we generated data using the signals described above for 20 values of n ranging from 20 to 1000. The values are evenly spaced on the logarithmic scale. For each n , we repeated the experiment 10 times. The left column (panel A) shows results for both distributions when the mean is smooth (mean is given by the smooth functions above) and error is measured using the mean-squared error between the estimate and the truth. In the exponential case, the mean trend filter is slightly more accurate for larger n , but the overall error also decreases with n since the problem is becoming easier. In the Poisson case, the estimates (and therefore their errors) are nearly the same. The right column (panel B) shows results when the natural parameter is smooth. Here, for both distributions, the MLE trend filter performs better (as measured by KL divergence), but the difference is again more pronounced for the exponential

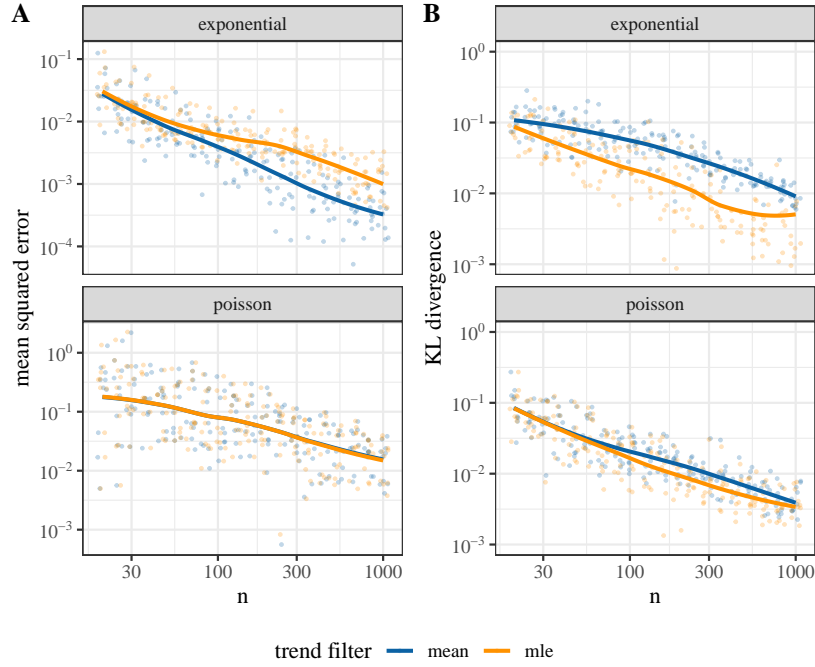


FIG 2. Estimation accuracy for both types of trend filters. The left column (panel A) compares the estimators when the mean is smooth. The right column (panel B) compares the estimators when the natural parameter is smooth. Solid lines show the average error across replications while the points show the error for each replication.

distribution. Figure 3 shows all the estimates for all four scenarios when $n = 104$. In the left two panels, for the exponential distribution, it is clear that whether the mean or natural parameter is smooth makes a substantial difference for the accuracy of the estimator. For the Poisson case (right two panels), there is much less discrepancy. In the case that the mean is smooth, both estimators appear relatively poor, though the MSE remains small in both cases. The reason is that the mean and the variance are the same, and both nearly constant. The difficulty is further exacerbated due to the discreteness of the data and only a small handful of values with non-negligible probability. Therefore, this setting is actually quite challenging. For context, on the typical dataset, the average absolute difference between observations at neighbouring points is about 2.5 compared with a 0.01 change in the signal.

6.2. Example applications

We apply the Penalized MLE to two real-world datasets for illustrative purposes. The first examines Poisson trend filtering for estimating the age-time hospitalization rates due to COVID-19 in the University of California system.

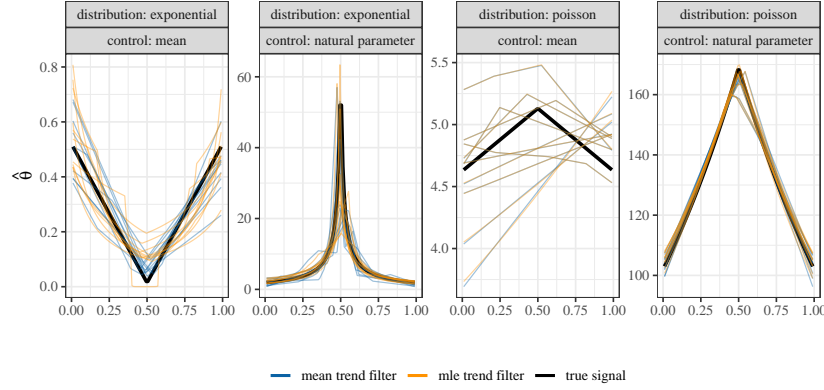


FIG 3. Estimates from both trend filters for the 4 scenarios when $n = 104$.

The second estimates the instantaneous temperature variability over the Northern hemisphere from publicly available observations.

6.2.1. UC COVID-19 hospitalization data

We analyzed the COVID-19 hospitalization rate within five hospitals in the University of California system: UC Davis, UC Los Angeles, UC Irvine, UC San Diego, and UC San Francisco. The data is based on 4,730 patients, all 18 years old or greater, that were admitted between February 12, 2020 to January 6, 2021. We aggregate the hospitalization counts at the weekly level—there are 48 weeks in total—and by age (in 15 bins of 5 years each). This results in noisy and sparse hospitalization counts at the week-by-age level with an average count-per-bin of 6.57. The data was obtained from the authors of [Nuño et al. \(2021\)](#), where they perform a more comprehensive analysis. It is used under a data use agreement and has not been made available to the public due to privacy concerns.

We apply $k = 1$ trend filtering with the Poisson exponential family in 2 dimensions to COVID-19 hospitalizations. We set $\lambda_2 = 0$ and tune λ_1 by minimizing $\text{PUKL}(\hat{\theta})$. One can see the results in [Figure 4](#), where the smoothed version is on the left. Due to the low average count per bin, trends in hospitalization rate are much more clearly visible after applying trend filtering. Using $k = 1$ makes it easy to identify local extrema in the estimated signal, and we have marked the local maxima in the smoothed signal which produces only 4 points—this would not have been possible in the raw data.

Some broad trends are clearly visible from [Figure 4](#). First, we can see two distinct waves for COVID-19 hospitalizations in summer 2020 and winter 2020–2021. Moreover, we can see that the highest hospitalization rates within the summer 2020 wave are among those aged 50–65, while in the winter 2020–2021 wave the highest rates are both within the 50–65 age range but also the 80+ age

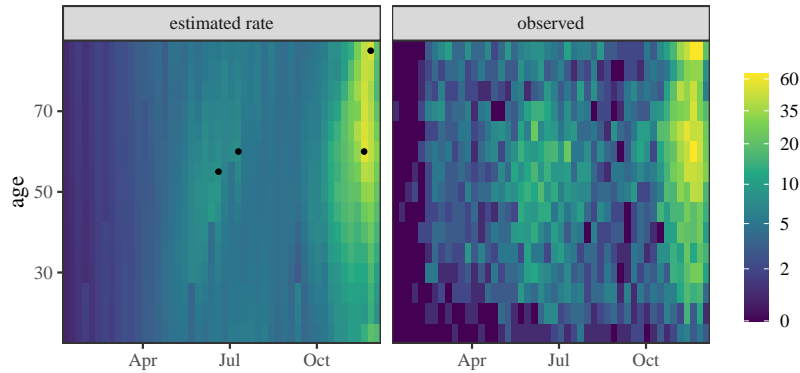


FIG 4. Estimated daily hospitalization rate due to COVID-19 by 5 year age group and week in five UC hospitals. We apply the Penalized MLE with $k = 1$ the Poisson exponential family (left) to the raw count data (right).

range. This suggests that the age distribution is not stationary, and changes with successive waves. This may be due to a number of factors, such as behavioral shifts and holiday effects.

6.2.2. Temperature variability

Trends in temperature variability (rather than in mean) have direct implications for plant and animal life (Huntingford et al., 2013), because changes in variability also impact the probability of extreme weather events (Vasseur et al., 2014). Hansen, Sato and Ruedy (2012) and Huntingford et al. (2013) suggest that adaptation to extremes is more difficult than to gradual increases in the mean temperature. Nevertheless, research examining trends in the volatility of spatio-temporal climate data is relatively scarce. Hansen, Sato and Ruedy (2012) studied changes in the standard deviation (SD) of surface temperatures at each spatial location relative to that location’s SD over a base period and showed that these estimates are increasing. Huntingford et al. (2013) took a similar approach for a different data set. They argued that, while there is an increase in the SDs from 1958-1970 to 1991-2001, it is much smaller than found by Hansen, Sato and Ruedy (2012). Huntingford et al. (2013) also computed the time-evolving global SD from the detrended time-series at each position and argued that the global SD has been stable.

The first row in Figure 5 shows the change in mean temperature averaged over the winter and summer months separately in the 1960s relative to the 2000s using the ERA 20C dataset (Poli et al., 2016). It shows strong increases in average temperatures in both periods over the majority of the hemisphere. The second row shows the estimated standard deviations from the Penalized MLE over the same period. We use $k = 1$ in the temporal dimension and $k = 2$ spa-

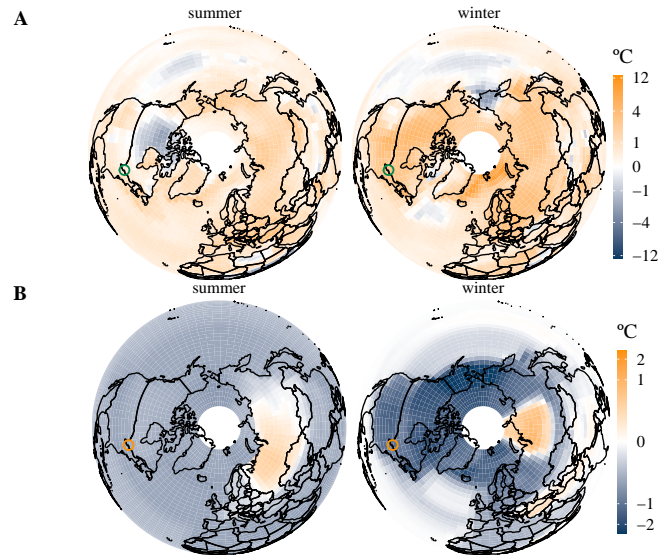


FIG 5. Panel A shows the change in average temperatures observed in the northern hemisphere from the 1960s relative to the 2000s in degrees Celsius. Panel B shows the change in estimated standard deviation (using the Penalized MLE with $k = 1$ in the temporal dimension and $k = 2$ over space) from the 1960s relative to the 2000s. Standard deviations were estimated at each spatio-temporal grid location before being averaged separately over winter / summer months over the appropriate decade.

tially. These estimated SDs are then averaged over the two periods for summer and winter separately and we plot the difference. There is a slight decrease in the SD during the summer and a more pronounced pole-ward decrease during the winter with the exception of Siberia which shows a dramatic increase over both periods.

7. Discussion

We studied estimation error bounds for two estimators with a trend filtering penalty on grid graphs. One estimator minimizes squared distance from the mean and the other maximizes log likelihood. The bounds are more involved, compared to, say, the homoskedastic sub-Gaussian noise case. Such cumbersome bounds are due to the fact there are many more parameters that influence the estimation error. We illustrated the bounds in several interesting regimes of signals with heteroskedastic and homoskedastic noise. We analyzed two datasets with our models showing the applicability of our methodology to real world problems. We showed that both estimators achieve minimax optimal error rates in some scenarios, though unfortunately, addressing all cases remains for future work.

Because our analysis examines the entire class of observations corrupted by sub-exponential noise, the result is a general bound on the error for all exponential families. But, this is a large class, and far from the only way to study the estimation error. More specific analysis in specific cases will likely result in sharper bounds. For example, [van de Geer \(2020\)](#) gets sharper rates for the Bernoulli family and [Brown, Cai and Zhou \(2010\)](#) examines a set of 6 families where the variance can be written as a quadratic function of the mean. However, those analyses are much less comprehensive than ours.

While it is not settled how to define piecewise polynomials in multiple dimensions, one way is to consider signals with $\|D\theta\|_0 \leq s_0$ and $\|\theta\|_2 \leq s_2$. Then this naturally leads to a bound on $\|D\theta\|_1$. However, the bound that one would obtain by embedding the ℓ_0 class to the ℓ_1 class may not be optimal. Hence, while this work provides some results and intuition for the behavior under piecewise polynomial signals, these are likely far from optimal for the piecewise polynomial class.

It may be of interest to examine the risk of the penalized MLE under ℓ_p loss, rather than KL, as undertaken here. One way to relate our results to the ℓ_p case is to note that, up to two terms, the Taylor expansion of the KL loss is the Fisher Information Metric: ℓ_2 weighted by the Fisher Information Matrix. From this perspective, if we think of Θ as a curved manifold determined by the probability measure, then KL is natural captures this curvature correctly. Using the Fisher Metric has similar behavior, due to the weighting, while ℓ_p will be flat, and ignore the manifold. Whether one metric is more appropriate than another is not clear, but we suspect that in order to give bounds in ℓ_p (or even just ℓ_2 , without the Fisher weighting) would require stronger conditions on Θ , particularly for heavy tailed distributions. We leave this for future work.

Other possible extensions are “mixed” loss and penalties. One could try to penalize the mean parameter combined with likelihood loss or the opposite. Preliminary investigations into the first case revealed similar issues as with the penalty on the natural parameter, namely an inability to control the error in the null space of D . Another natural avenue for future work would note that all of these (the estimators examined here and the mixed versions) have connections to state space models in time series. So the relationship between trend filtering and Kalman-type filters may yield new theoretical insights and computational algorithms.

Appendix A: Proofs for preliminary results

A.1. Proof of [Theorem 1.2](#)

Proof of [Theorem 1.2](#). Without loss of generality assume Y has mean zero. We have

$$\begin{aligned} E[e^{sY}] &= \int e^{sy} h(y) e^{y\theta^* - \psi(\theta^*)} dy \\ &= \int h(y) e^{y(s+\theta^*) - \psi(s+\theta^*) + \psi(s+\theta^*) - \psi(\theta^*)} dy \\ &= e^{\psi(s+\theta^*) - \psi(\theta^*)} \int h(y) e^{y(s+\theta^*) - \psi(s+\theta^*)} dy \end{aligned}$$

Since $\theta^* \in \text{interior}(\Theta)$, there is b such that $|s| < \frac{1}{b}$ gives $h(y) e^{y(s+\theta^*) - \psi(s+\theta^*)}$ is a member of the exponential family and hence integrates to 1. Therefore the above display equals $e^{\psi(s+\theta^*) - \psi(\theta^*)}$. A Taylor expansion of $\psi(s + \theta^*) - \psi(\theta^*)$ is possible because θ is infinitely differentiable ([Brown, 1986](#)):

$$\psi(s + \theta^*) - \psi(\theta^*) = \nabla \psi(\theta^*) s + \frac{1}{2} \nabla^2 \psi(\theta^*) s^2 + \frac{1}{2} R(\theta^*, s) s^2$$

where $\lim_{s \rightarrow 0} R(\theta^*, s) \rightarrow 0$. Combined with the fact that $E[Y] = \nabla \psi(\theta^*) = 0$, we have that

$$\psi(s + \theta^*) - \psi(\theta^*) = \frac{1}{2} (\nabla^2 \psi(\theta^*) + R(\theta^*, s)) s^2$$

Fixing $\delta > 0$, we can then choose a b , which depends on δ , such that $\sup_{|s| < \frac{1}{b}} |R(\theta^*, s)| < \delta$. We conclude that there exists a b (where we increase b from our previous choice guaranteeing $s + \theta^* \in \Theta$ as necessary) such that for all $|s| < \frac{1}{b}$

$$\frac{1}{2} (\nabla^2 \psi(\theta^*) - \delta) s^2 \leq \psi(s + \theta^*) - \psi(\theta^*) \leq \frac{1}{2} (\nabla^2 \psi(\theta^*) + \delta) s^2.$$

This gives the second claim of the lemma. Taking $\nu^2 = \nabla^2 \psi(\theta^*) + \delta$ gives $E[e^{sY}] \leq e^{\frac{s^2 \nu^2}{2}}$ and proves the result. \square

TABLE 4
Sub-exponential parameters for some exponential family distributions

Distribution	$\psi(\theta)$	ν^2, b
Poisson (mean = μ)	e^θ	$2\mu, 0.55$
Exponential (mean = μ)	$-\log(-\theta)$	$4\mu^2 \log \frac{4}{e}, 2\mu$
χ_k^2 (mean = k)	$\log(\Gamma(\theta + 1)2^{\theta+1})$	$4k, 4$

A.2. Sub-exponential parameters for some standard distributions

For a Poisson random variable X with mean μ , note that for $s \in \mathbb{R}$,

$$Ee^{s(X-\mu)} = e^{\mu(e^s - s - 1)}.$$

Therefore $Ee^{s(X-\mu)} \leq e^{\mu s^2}$ for s satisfying $e^s - 1 - s \leq s^2$. Let s^* be the non-zero solution to $e^x = 1 + x + x^2$. Then $s^* \approx 1.793$. From this, we can show that

$$X - \mu \text{ is } \text{SE}(\nu^2, b) \text{ with } \nu^2 = 2\mu, b = 1/s^* \leq 0.55.$$

For exponential distribution, we can do a similar calculation to get the results in Table 4. For an exponential variable X with mean μ , for $s \in \mathbb{R}$,

$$Ee^{s(X-\mu)} = \frac{e^{-\mu s}}{1 - \mu s}.$$

We can verify that $X - \mu$ is sub-exponential with parameters (ν^2, b) given in Table 4. To arrive at these parameters, we set $b = 2\mu$ and find the ν^2 of the form $c\mu^2$ for a constant c such that $Ee^{s(X-\mu)} \leq e^{\nu^2 s^2/2}$ for $|s| \leq 1/b$. In a similar fashion, one can also verify the sub-exponential parameters for the χ^2 distribution specified in the bottom row of the table.

A.3. Some properties of Sub-exponentials

Tail bounds on linear combinations of sub-exponentials

We use the following tail bound for sums of sub-exponential variables at multiple places in our proofs.

Lemma A.1. *Let ν and c be vectors such that ϵ_i is sub-exponential with parameters (ν_i, c_i) . Given a matrix $A \in \mathbb{R}^{n \times r}$, assume we have K and H such that $\sup_{i=1, \dots, r} \|\nu \odot A_i\|_2 \leq K$ and $\sup_{i=1, \dots, r} \|c \odot A_i\|_\infty \leq H$, where A_1, \dots, A_r are the columns of A . Then*

$$P(\|A^\top \epsilon\|_\infty \geq t) \leq \begin{cases} 2r \exp\left(-\frac{t^2}{2K^2}\right) & t < \frac{K^2}{H} \\ 2r \exp\left(-\frac{t}{H} + \frac{K^2}{2H^2}\right) & t \geq \frac{K^2}{H} \end{cases}$$

The proof is similar to that of Bernstein inequality from Theorem 2.8.1 in (Ver-shynin, 2018).

Proof of Theorem A.1. We have

$$\begin{aligned} \log E [\exp(s \|A^\top \epsilon\|_\infty)] &= \log E [\exp(s \max\{|A_1^\top \epsilon|, \dots, |A_r^\top \epsilon|\})] \\ &\leq \log E \left[\exp \left(s \sum_{i=1}^r |A_i^\top \epsilon| \right) \right] = \log E \left[\exp \left(s \sum_i \left| \sum_{j=1}^n a_{ij} \epsilon_j \right| \right) \right]. \end{aligned}$$

Note that $A_i^\top \epsilon$ is mean zero with parameters $(\|\nu \otimes A_i\|_2, \|c \otimes A_i\|_\infty)$. This is because

$$\log E [\exp(s A_i^\top \epsilon)] = \log E \left[\exp \left(s \sum_j a_{ij} \epsilon_j \right) \right] = \sum_j \log E [\exp(s a_{ij} \epsilon_j)],$$

by independence of ϵ_j . When $|s| < \frac{1}{a_{ij} c_j}$ for all j , which is satisfied when $|s| < \frac{1}{\|c \otimes A_i\|_\infty}$,

$$\sum_j \log E [\exp(s a_{ij} \epsilon_j)] \leq \sum_j \frac{\nu_j^2 (s a_{ij})^2}{2} = \frac{\|\nu \otimes A_i\|_2^2 s^2}{2}.$$

Therefore, for $|s| < \frac{1}{\sup_{i=1, \dots, r} \|c \otimes A_i\|_\infty}$,

$$\begin{aligned} \log E [\exp(s \|A^\top \epsilon\|_\infty)] &= \log E [\exp(s \max\{|A_1^\top \epsilon|, \dots, |A_r^\top \epsilon|\})] \\ &\leq \log \sum_{i=1}^r E [\exp(s |A_i^\top \epsilon|)] \\ &\leq \log \sum_i E [\exp(s A_i^\top \epsilon) + \exp(-s A_i^\top \epsilon)] \\ &\leq \log \left(2 \sum_{i=1}^r \exp \left(\frac{\|\nu \otimes A_i\|_2^2 s^2}{2} \right) \right). \end{aligned}$$

Therefore, using the Chernoff bound, we have

$$P(\|A^\top \epsilon\|_\infty > t) \leq \exp(-ts) \left(2 \sum_{i=1}^r \exp \left(\frac{\|\nu \otimes A_i\|_2^2 s^2}{2} \right) \right)$$

for $|s| < \frac{1}{\sup_{i=1, \dots, r} \|v_i \otimes c\|_\infty}$, which we minimize in s to get our bound. This is intractable, so we require $\|\nu \otimes A_i\|_2 \leq K$ and $\|c \otimes A_i\|_\infty \leq H$ for all $i \in [r]$. We then have

$$P(\|A^\top \epsilon\|_\infty > t) \leq 2 \sum_{i=1}^r \exp \left(-ts + \frac{\|\nu \otimes A_i\|_2^2 s^2}{2} \right) \leq 2r \exp \left(-ts + \frac{K^2 s^2}{2} \right).$$

Minimizing in s , for $|s| < \frac{1}{H}$, we have $s = t/K^2$ or $1/H$ depending on which is smaller. Therefore,

$$P(\|A^\top \epsilon\|_\infty \geq t) \leq \begin{cases} 2r \exp \left(-\frac{t^2}{2K^2} \right) & t < \frac{K^2}{H} \\ 2r \exp \left(-\frac{t}{H} + \frac{K^2}{2H^2} \right) & t \geq \frac{K^2}{H} \end{cases}$$

□

We state a few convenient ways of using Bernstein's tail bound inequality on linear combinations of sub-exponential random variables. Denote the sub-exponential tail bound function

$$\phi(t; \nu^2, b) = 2 \exp \left(- \frac{1}{2} \min \left\{ \frac{t^2}{\nu^2}, \frac{t}{b} \right\} \right) \quad (9)$$

for $t \geq 0$ with parameters $\nu > 0, b > 0$. Note that if $\mathbb{P}(|X| > t) \leq \phi(t; \nu^2, b)$ for all $t \geq 0$, then

$$|X| \leq 2(\nu \vee b)u \quad (10)$$

with probability at least $1 - 2e^{-u}$ for $u \geq 1$.

Lemma A.2. *Let ϵ_i be independent, mean-zero, sub-exponential variates with parameters (ν_i^2, b_i) for $i \in [n]$. Let $a \in \mathbb{R}^n$ be a fixed vector. Let ϕ be the sub-exponential tail bound function defined in (9). Then for $t \geq 0$,*

$$\mathbb{P}(|a^\top \epsilon| > t) \leq \phi\left(t; \sum_{i=1}^n a_i^2 \nu_i^2, \max_{i \in [n]} |a_i| b_i\right) \quad (11)$$

$$\leq \phi\left(t; \|a\|_2^2 \|\nu\|_\infty^2, \|a\|_\infty \|b\|_\infty\right) \quad (12)$$

Also,

$$\mathbb{P}(|a^\top \epsilon| > t) \leq \phi\left(t; \|a\|_\infty^2 \|\nu\|_2^2, \|a\|_\infty \|b\|_\infty\right). \quad (13)$$

Further, if $\nu = b$, then for $t \geq 1$, with probability at least $1 - 2e^{-t}$, both the following hold:

$$|a^\top \epsilon| \leq 2\|a\|_2 \|b\|_\infty t \quad (14)$$

$$|a^\top \epsilon| \leq 2\|a\|_\infty \|b\|_2 t \quad (15)$$

Proof of Theorem A.2. (11) follows by applying Bernstein's inequality from Theorem 2.8.1 in (Vershynin, 2018). The inequalities (12), (13) follow from (11) by applying Hölder's inequality to the first parameter in different ways.

From (10), observe that for $t \geq 1$,

$$|a^\top \epsilon| \leq 2(\|a \odot b\|_2 \vee \|a \odot b\|_\infty) t \leq 2\|a \odot b\|_2 t$$

holds with probability at least $1 - 2e^{-t}$, where $a \odot b \in \mathbb{R}^n$ with $(a \odot b)_i = a_i b_i$, $i \in [n]$. By applying Hölder's inequality in two different ways we get the high probability bounds (14) and (15). □

Tail bound on maximum of sub-exponentials

Lemma A.3. *Suppose X_i are sub-exponential with parameters (ω^2, ω) for $i \in [m]$. Then for $t \geq 1$*

$$\mathbb{P} \left(\max_{i \in [m]} |X_i| \leq 2\omega(\log 2m + t) \right) \geq 1 - 2e^{-t}.$$

Proof of Theorem A.3. Denote $X_{m+j} = -X_j$ for $j \in [m]$. By union bound, for $u > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_{j \in [m]} |X_j| > u\right) &= \mathbb{P}\left(\max_{j \in [2m]} X_j > u\right) \\ &\leq \sum_{j=1}^{2m} \mathbb{P}(X_j > u) \\ &\leq 4m \exp\left(-\left\{\frac{u^2}{2\omega^2} \wedge \frac{u}{2\omega}\right\}\right). \end{aligned}$$

Set $u = 2\omega(\log 2m + t)$ to get the desired bound. \square

Appendix B: Proofs of upper bounds

B.1. Proof of Theorem 2.2

We first state a basic inequality.

Lemma B.1 (Basic inequality). *Let R be as defined in Section 2.1 and let $\hat{\theta}$ be the estimate in (3). Then,*

$$R(\hat{\theta}) - R(\theta^*) + \lambda_2 \|P_{\mathcal{N}} \hat{\theta}\|_2 + \lambda_1 \|D\hat{\theta}\|_1 \leq \frac{1}{n} \epsilon^\top (\hat{\theta} - \theta^*) + \lambda_2 \|P_{\mathcal{N}} \theta^*\|_2 + \lambda_1 \|D\theta^*\|_1.$$

Further, this inequality is true if we replace $\hat{\theta}$ with $\hat{\theta}_t = t\hat{\theta} + (1-t)\theta^*$ for any $t \in [0, 1]$.

Proof of Theorem B.1. Optimality of $\hat{\theta}$ and the equality $R_n(\theta) - R(\theta) = -\frac{1}{n} \epsilon^\top \theta$ gives

$$\begin{aligned} &R_n(\hat{\theta}) + \lambda_2 \|P_{\mathcal{N}} \hat{\theta}\|_2 + \lambda_1 \|D\hat{\theta}\|_1 \\ &\leq R_n(\theta^*) + \lambda_2 \|P_{\mathcal{N}} \theta^*\|_2 + \lambda_1 \|D\theta^*\|_1 \\ \Leftrightarrow &R(\hat{\theta}) - \frac{1}{n} \epsilon^\top \hat{\theta} + \lambda_2 \|P_{\mathcal{N}} \hat{\theta}\|_2 + \lambda_1 \|D\hat{\theta}\|_1 \\ &\leq R(\theta^*) - \frac{1}{n} \epsilon^\top \theta^* + \lambda_2 \|P_{\mathcal{N}} \theta^*\|_2 + \lambda_1 \|D\theta^*\|_1 \end{aligned}$$

This is equivalent to the main statement in the lemma. The inequality for $\hat{\theta}_t$ follows from the fact that $\theta \mapsto R_n(\theta) + \lambda_1 \|D\theta\|_1 + \lambda_2 \|P_{\mathcal{N}} \theta\|_2$ is convex. \square

Proof of Theorem 2.2. For brevity, define the shorthand

$$\tau(\theta, \lambda_1, \lambda_2) = \lambda_1 \|D\theta\|_1 + \lambda_2 \|P_{\mathcal{N}} \theta\|_2$$

for $\theta \in \mathbb{R}^n, \lambda_1, \lambda_2 \geq 0$. From the basic inequality in Theorem B.1,

$$R(\hat{\theta}) - R(\theta^*) + \tau(\hat{\theta}, \lambda_1, \lambda_2) \leq \frac{1}{n} \epsilon^\top (\hat{\theta} - \theta^*) + \tau(\theta^*, \lambda_1, \lambda_2).$$

Applying [Theorem B.4](#) with $J = [k + 1]^d$,

$$\begin{aligned} \frac{1}{n} \epsilon^\top (\hat{\theta} - \theta^*) &\leq \frac{A}{n} \|P_{\mathcal{N}}(\hat{\theta} - \theta^*)\|_2 + \frac{B}{n} \|D(\hat{\theta} - \theta^*)\|_1 \\ &= \tau(\hat{\theta} - \theta^*, B/n, A/n) \end{aligned}$$

where

$$\begin{aligned} A &= 2t\mu\sqrt{\frac{\kappa}{n}}(\|\nu\|_2 \vee \|b\|_\infty), \\ B &= 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}), \end{aligned}$$

for $t \geq 1$, on an event $\Omega(t)$ with probability at least $1 - 2(m + \kappa)e^{-t}$. Here m is the number of rows of D and $\kappa = (k + 1)^d$. Therefore, on the event $\Omega(t)$,

$$\begin{aligned} R(\hat{\theta}) - R(\theta^*) + \tau(\hat{\theta}, \lambda_1, \lambda_2) &\leq \tau(\hat{\theta} - \theta^*, B/n, A/n) + \tau(\theta^*, \lambda_1, \lambda_2) \\ &\leq \tau(\hat{\theta}, B/n, A/n) + \tau(\theta^*, B/n, A/n) + \tau(\theta^*, \lambda_1, \lambda_2) \end{aligned}$$

where we used triangle inequality in the second line. If we choose $\lambda_1 \geq 2B/n$, $\lambda_2 \geq 2A/n$, by linearity of τ in regularization parameters, we have

$$\tau(\theta, B/n, A/n) \leq \tau(\theta, \lambda_1/2, \lambda_2/2) = \frac{1}{2}\tau(\theta, \lambda_1, \lambda_2),$$

for any $\theta \in \mathbb{R}^n$. Therefore

$$R(\hat{\theta}) - R(\theta^*) + \frac{1}{2}\tau(\hat{\theta}, \lambda_1, \lambda_2) \leq \frac{3}{2}\tau(\theta^*, \lambda_1, \lambda_2)$$

As θ^* minimizes R , we should have $R(\hat{\theta}) \geq R(\theta^*)$. That means, both the terms $R(\hat{\theta}) - R(\theta^*)$ and $\frac{1}{2}\tau(\hat{\theta}, \lambda_1, \lambda_2)$ are non-negative. Therefore,

$$\begin{aligned} R(\hat{\theta}) - R(\theta^*) &\leq \frac{3}{2}\tau(\theta^*, \lambda_1, \lambda_2) \quad \text{and} \\ \frac{1}{2}\tau(\hat{\theta}, \lambda_1, \lambda_2) &\leq \frac{3}{2}\tau(\theta^*, \lambda_1, \lambda_2). \end{aligned}$$

This completes the proof as these inequalities hold with probability $\mathbb{P}(\Omega(t)) \geq 1 - 2(m + \kappa)e^{-t}$. \square

B.2. Proofs of Corollaries to [Theorem 2.2](#)

Proof of [Theorem 2.4](#). We have the following bounds,

$$\begin{aligned} \|\nu\|_2 &= O(\sqrt{n}) \\ \|\nu\|_\infty, \|b\|_\infty &= O(1) \end{aligned}$$

When $d = 1$ then $\alpha = 2$, and so

$$\frac{1}{n}(\|\nu\|_2 + \|b\|_\infty) = O(n^{-1/2}).$$

When $d = 2$, $\alpha = 1$ and $\gamma_1 = \log n$, $\gamma_2 = 1$, thus

$$\|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-1/2} \gamma_2, \|\nu\|_2 n^{-\alpha} \gamma_1\} = O(n^{-1/2} \log n).$$

When $d = 3$ then $\alpha = 2/3$ and $\gamma_1 = \gamma_2 = 1$,

$$\|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-1/2} \gamma_2, \|\nu\|_2 n^{-\alpha} \gamma_1\} = O(n^{-1/2}).$$

When $d = 4$ then $\alpha = 1/2$ and $\gamma_1 = 1$, $\gamma_2 = \log^{1/2} n$ and

$$\|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-1/2} \gamma_2, \|\nu\|_2 n^{-\alpha} \gamma_1\} = O(n^{-1/2} \cdot \log^{1/2} n).$$

Finally, when $d > 4$ then $\alpha = 2/d < 1/2$ and

$$(\|\nu\|_\infty + \|b\|_\infty) n^{-\alpha} = O(n^{-2/d}).$$

Next we show that the example signal satisfies the necessary conditions.

Consider the Poisson distribution where the natural parameter vector θ^* is constrained. For $i = (i_1, \dots, i_d) \in [N]^d$, let

$$\theta_i^* = \frac{2}{N} \sum_{j=1}^d |i_j - N/2|.$$

Then the mean vector is

$$\beta_i^* = \prod_{j=1}^d \exp\left(\frac{2}{N} |i_j - N/2|\right).$$

Because the distribution is Poisson, we have $\|b\|_\infty$ is constant while $\nu_i^2 = 2\beta_i^*$ (see [Table 1](#)). Thus, $\|\nu\|_\infty = \sqrt{2}e^{d/2}$ which is achieved at $i = (0, \dots, 0)$. The canonical scaling holds for $\|D\theta^*\|_1 \lesssim n^{1-\alpha}$ with $k = 1$ because there are on the order of N^{d-1} points at which the Laplacian is non-zero and they are on the order of $1/N$. \square

Proof of [Theorem 2.5](#). For $d = 1$ we have that $\alpha = 1$, and $\gamma_1 = \log n$, $\gamma_2 = 1$, thus

$$\|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-1/2} \gamma_2, \|\nu\|_2 n^{-\alpha} \gamma_1\} = O(n^{c-1/2}).$$

For $d = 2$ we have that $\alpha = 1/2$ and $\gamma_1 = 1$, $\gamma_2 = \log^{1/2} n$ and

$$\|b\|_\infty n^{-\alpha} \gamma_1 + \min\{\|\nu\|_\infty n^{-1/2} \gamma_2, \|\nu\|_2 n^{-\alpha} \gamma_1\} = O(n^{-1/2} \cdot \log^{1/2} n).$$

For $d > 2$ we have that $\alpha = 1/d < 1/2$ and $\gamma_1 = \gamma_2 = 1$, thus

$$(\|\nu\|_\infty + \|b\|_\infty) n^{-\alpha} = O(n^{c-1/d}).$$

To show that the specified signal satisfies the necessary properties, let $d > 1$, $k = 0$, and $c > 0$. Consider the Exponential distribution with natural parameter

$$\theta_i^* = -n^{-c} \mathbf{1}\{i = 0\} - n^{1-1/d} \mathbf{1}\{i \neq 0\}.$$

where i indexes the lattice. We have that for $k = 0$, $\|D\theta^*\|_1 \leq d(n^{1-1/d} - n^{-c}) \asymp n^{1-\alpha}$, so the canonical scaling holds. We apply MLE trend filtering with $k = 0$. From Table 1, we have that $\|\nu\|_\infty, \|b\|_\infty \leq 2n^c$ and $\|\nu\|_2^2 \leq 2(n^{2c} + n^{1/d-1})$. \square

B.3. Uniform risk bound with null space penalty

Proof of Theorem 2.1. From the definitions of R, R_n ,

$$|R(\theta) - R_n(\theta)| = \frac{1}{n} |\epsilon^\top \theta|.$$

Applying Theorem B.4 with $J = [k+1]^d$, we get

$$|\epsilon^\top \theta| \leq A_n \|P_N \theta\|_2 + B_n \|D\theta\|_1$$

where $A_n = 2t\mu\sqrt{\frac{\kappa}{n}}(\|\nu\|_2 \vee \|b\|_\infty)$, $B_n = 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1})$, with probability at least $1 - 4nde^{-t}$, for $t \geq 1$. Here $\kappa = (k+1)^d$ and we used the fact that $m < dn$. By definition of Θ , θ should satisfy $\|D\theta\|_1 \leq c_n n^{1-\alpha}$ and $\|P_N \theta\|_n \leq a_n$. Therefore,

$$|\epsilon^\top \theta| \leq A_n a_n \sqrt{n} + B_n c_n n^{1-\alpha}$$

From the assumptions $\|\nu\|_\infty, \|b\|_\infty \leq c$, we can write $A \leq 2t\mu c \sqrt{\kappa}$. From Theorem B.6, for $p \geq 1$, $L_{\ell,p}^p \leq c_1 n^{(p\alpha-1)+} (\log n)^{\mathbf{1}\{p\alpha=1\}}$. This yields the following bound on B_n :

$$B_n \leq 2tc_1 c \gamma n^{(\alpha-\frac{1}{2})+}.$$

Therefore, with probability at least $1 - 4nde^{-t}$,

$$\begin{aligned} \frac{1}{n} |\epsilon^\top \theta| &\leq c_2 tc (a_n n^{-\frac{1}{2}} + c_n \gamma n^{-\alpha} n^{(\alpha-\frac{1}{2})+}) \\ &= c_2 tc (a_n n^{-\frac{1}{2}} + c_n \gamma n^{-\min\{\alpha, \frac{1}{2}\}}) \end{aligned}$$

for a constant c_2 depending only on k, d . This is sufficient to show the desired bound. \square

B.4. Proof of Theorem 3.1

Proof of Theorem 3.1. Writing the KKT conditions, $\hat{\theta}$ and $\hat{\beta}$ are solutions to (2) and (4) iff

$$\begin{aligned} \psi'(\hat{\theta}) - y + n\lambda D^\top S(D\hat{\theta}) &\ni 0 \\ \hat{\beta} - y + n\lambda D^\top S(D\hat{\beta}) &\ni 0 \end{aligned} \tag{16}$$

where $S(u)$ is the set of subgradients of $x \mapsto \|x\|_1$. $S(u)$ depends only $\text{sgn}(u)$. As ψ' is a strictly increasing function, for any $a, b \in \mathbb{R}$, $\text{sgn}(\psi'(a) - \psi'(b)) = \text{sgn}(a - b)$. Therefore

$$\text{sgn}(D\psi'(\hat{\theta})) = \text{sgn}(D\hat{\theta}),$$

and hence the subgradients $S(D\psi'(\hat{\theta})) = S(D\hat{\theta})$. Plugging this in (16), we see that the KKT conditions for the least squares problem are satisfied by $\psi'(\hat{\theta})$ and therefore it is a solution to the least squares problem (4). The solution to the least squares optimization problem (4) is unique because the objective is strictly convex. Therefore, by definition of $\hat{\beta}$, $\hat{\beta} = \psi'(\hat{\theta})$. \square

B.5. Proof of Theorem 3.2

Proof of Theorem 3.2. The proof follows the strategy in Theorem 6 in Wang et al. (2016).

Abbreviate $\hat{\delta} = \hat{\beta} - \beta^*$. From the optimality in the definition of $\hat{\beta}$,

$$\frac{1}{2n} \|y - \hat{\beta}\|_2^2 + \lambda \|D\hat{\beta}\|_1 \leq \frac{1}{2n} \|y - \beta^*\|_2^2 + \lambda \|D\beta^*\|_1$$

Rearranging and substituting $y = \beta^* + \epsilon$,

$$\frac{1}{2n} \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{1}{n} \epsilon^\top (\hat{\beta} - \beta^*) + \lambda \|D\beta^*\|_1 - \lambda \|D\hat{\beta}\|_1.$$

Bound the empirical process term on the right hand side using Theorem B.4. By Theorem B.4, for $t \geq 1$ and $J \subset [N]^d$, the following holds with probability at least $1 - 2(m + |J|)e^{-t}$:

$$\begin{aligned} & \frac{1}{2n} \|\hat{\beta} - \beta^*\|_2^2 \\ & \leq \frac{A}{n} \|P_J(\hat{\beta} - \beta^*)\|_2 + \frac{B}{n} \|D(\hat{\beta} - \beta^*)\|_1 + \lambda \|D\beta^*\|_1 - \lambda \|D\hat{\beta}\|_1 \end{aligned}$$

where

$$\begin{aligned} A &= 2t\mu \sqrt{\frac{|J|}{n}} (\|\nu\|_2 \vee \|b\|_\infty), \\ B &= 2t (\min \{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}). \end{aligned}$$

Applying Young's inequality on the first term and setting $\lambda \geq \frac{B}{n}$,

$$\begin{aligned} \frac{1}{2n} \|\hat{\beta} - \beta^*\|_2^2 &\leq \frac{1}{4n} \|\hat{\beta} - \beta^*\|_2^2 + \frac{A^2}{n} + \lambda \|D(\hat{\beta} - \beta^*)\|_1 + \lambda \|D\beta^*\|_1 - \lambda \|D\hat{\beta}\|_1 \\ &\leq \frac{1}{4n} \|\hat{\beta} - \beta^*\|_2^2 + \frac{A^2}{n} + 2\lambda \|D\beta^*\|_1 \end{aligned}$$

We used triangle inequality on the penalty terms to get the second line. Canceling terms,

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{4A^2}{n} + 8\lambda \|D\beta^*\|_1.$$

This bound holds with probability at least $1 - 2(m + |J|)e^{-t} \geq 1 - 4nde^{-t}$, and so the proof is complete. \square

B.6. Proofs of Corollaries to Theorem 3.2

Denote $\sigma^2 = \frac{1}{n}(\|\nu\|_2^2 \vee \|b\|_\infty)$. From Theorem 3.2, for any $J \subset [N]^d$ containing $[k+1]^d$, assuming the scaling $\|D\beta^*\|_1 = O(n^{1-\alpha})$,

$$\frac{1}{n}\|\widehat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}}\left(\frac{|J|t^2\sigma^2}{n} + \frac{tB_n}{n^\alpha}\right) \quad (17)$$

where $t = \log n$,

$$B_n = 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}). \quad (18)$$

Compared to the bound in Theorem 3.2, additional $\log n$ factors are incurred when translating from the high-probability statement to $O_{\mathbb{P}}$ notation. B_n can be bound more explicitly by writing down bounds for $L_{J,1}, L_{J,2}$ using Theorem B.6. For $r \in [1, N\sqrt{d}]$, we can write

$$L_{J,2}^2 \leq \begin{cases} c\mu^2\gamma_2^2 & \alpha \leq 1/2, J = [k+1]^d \\ c\mu^2(n/r^d)^{2\alpha-1} & \alpha > 1/2, J = \{i \in [N]^d : \|(i-k-2)_+\|_2 < r\} \end{cases} \quad (19)$$

and

$$L_{J,1} \leq \begin{cases} c\mu^2\gamma_1 & \alpha \leq 1, J = [k+1]^d \\ c\mu^2(n/r^d)^{\alpha-1} & \alpha > 1, J = \{i \in [N]^d : \|(i-k-2)_+\|_2 < r\}. \end{cases} \quad (20)$$

where $\gamma_p = \log^{1/p}(n)$ if $p\alpha = 1$ and 1 otherwise.

Proof of Theorem 3.3. Case $\alpha \leq 1/2$: Set $\alpha \leq 1/2, J = [k+1]^d$ in (20), (19), plugin the resulting bounds for $L_{J,1}, L_{J,2}$ in equation (18) :

$$B_n = O(\min\{\|\nu\|_\infty\gamma_2, \|\nu\|_2\gamma_1\} \vee \|b\|_\infty\gamma_1)t. \quad (21)$$

Then use the assumptions $\|\nu\|_\infty, \|b\|_\infty \leq \omega$, to write $B_n = O(t\omega\gamma_2)$ where $t = \log n$. Plug this expression for B_n in (17), again use the assumption that $\|\nu\|_\infty, \|b\|_\infty \leq \omega$, to write

$$\frac{1}{n}\|\widehat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}}\left(\frac{t^2\omega^2}{n} + \frac{t\omega\gamma_2}{n^\alpha}\right).$$

Case $\alpha > 1/2$: We can write

$$\begin{aligned} B_n &= 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}) \\ &\leq 2t(\|\nu\|_\infty L_{J,2} + \|b\|_\infty L_{J,1}) \\ &\leq 2t\omega(L_{J,2} + L_{J,1}). \end{aligned}$$

Let $J = \{i \in [N]^d : \|(i-k-2)_+\|_2 < r\}$ for an r to be chosen later from $[1, \sqrt{d}N]$. Plugging in the bounds for $L_{J,1}, L_{J,2}$ from (20), (19) with $\alpha > 1/2$,

and then using (17),

$$\begin{aligned} & \frac{1}{n} \|\widehat{\beta} - \beta^*\|_2^2 \\ &= O_{\mathbb{P}} \left(\frac{(r+k+2)^d t^2}{n} \omega^2 + \frac{t}{n^\alpha} \left(\omega(n/r^d)^{\alpha-1/2} \gamma_2 + \omega(n/r^d)^{(\alpha-1)+} \gamma_1 \right) \right) \end{aligned} \quad (22)$$

where $t = \log n$. Select r such that

$$\frac{r^d t^2}{n} \omega^2 \asymp \frac{t \omega}{n^\alpha} (n/r^d)^{\alpha-1/2}.$$

Then the following is sufficient,

$$r^d = \left\lfloor n(n^\alpha t \omega)^{-2/(2\alpha+1)} \right\rfloor.$$

and the following condition ensures that this choice of r is in $[1, \sqrt{d}N]$:

$$n^{-\alpha} \leq t \omega \leq \sqrt{n}.$$

Plugging this choice of r , the first two terms in (22) are bounded by

$$c_1 \frac{r^d t^2}{n} \omega^2 = c_2 (t \omega)^2 (n^\alpha t \omega)^{-2/(2\alpha+1)} \leq c_2 \left(\frac{t^2 \omega^2}{n} \right)^{2\alpha/(2\alpha+1)}$$

where c_1, c_2 are universal constants. The remaining term is bounded by

$$\frac{t}{n^\alpha} \omega \gamma_1 \text{ if } \alpha \leq 1 \quad \text{and} \quad n^{-\frac{3\alpha}{2\alpha+1}} (\omega t)^{\frac{4\alpha-1}{2\alpha+1}} \text{ if } \alpha > 1.$$

When $t \omega \geq n^{-\alpha}$, $n^{-\frac{3\alpha}{2\alpha+1}} (\omega t)^{\frac{4\alpha-1}{2\alpha+1}} \leq (t^2 \omega^2 / n)^{\frac{2\alpha}{2\alpha+1}}$ and so the desired bound holds. \square

Proof of Theorem 3.4. In both the Poisson and Exponential cases $\|\nu\|_\infty, \|b\|_\infty = O(1)$. For $d = 1, 2, 3$ we have that $\alpha > 1/2$ and

$$\left(\frac{\omega^2 \log^2 n}{n} \right)^{\frac{2\alpha}{2\alpha+1}} + \frac{\omega \gamma_1 \log n}{n^\alpha} = O \left(\left(\frac{\log^2 n}{n} \right)^{\frac{2\alpha}{2\alpha+d}} \right).$$

For $d = 4$, $\alpha = 1/2$,

$$\frac{\omega^2 \log^2 n}{n} + \frac{\omega \gamma_2 \log n}{n^\alpha} = O \left(\frac{\log^{3/2} n}{n^\alpha} \right).$$

For $d \geq 5$, $\alpha < 1/2$,

$$\frac{\omega^2 \log^2 n}{n} + \frac{\omega \gamma_2 \log n}{n^\alpha} = O \left(\frac{\log n}{n^\alpha} \right).$$

To show that the example signal satisfies the conditions, consider the Poisson and Exponential families where the mean parameter is constrained. Consider a grid graph with width N and dimension d , so that $n = N^d$. For $i = (i_1, \dots, i_d) \in [N]^d$, let

$$\beta_i^* = \frac{d}{N} + \frac{2}{N} \sum_{j=1}^d |i_j - N/2|.$$

For the Poisson distribution $\nu_i^2 \asymp \beta_i^*$ hence $\|\nu\|_\infty = O(1)$. Similarly, for the Exponential distribution $\|\nu\|_\infty, \|b\|_\infty = O(1)$. \square

Corollary B.2. *Let $\sigma = \max\{\|\nu\|_2, \|b\|_\infty\}/\sqrt{n}$, and $\sigma_\infty = \max\{\|\nu\|_\infty, \|b\|_\infty\}$. Suppose $\|D\beta^*\|_1 \lesssim n^{1-\alpha}$. If $\alpha \leq 1/2$, then the estimator $\hat{\beta}$ in [Theorem 3.2](#) satisfies*

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\frac{\sigma^2 \log^2 n}{n} + \frac{\sigma_\infty \gamma_2 \log n}{n^\alpha} \right).$$

If $\alpha > 1/2$ and $\sigma^2/\sigma_\infty \lesssim \sqrt{n}/\log n$, then

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\left[\frac{\sigma^2 \log^2 n}{n} \right]^{\frac{2\alpha}{2\alpha+1}} \left[\frac{\sigma_\infty}{\sigma} \right]^{\frac{2}{2\alpha+1}} + \frac{\sigma_\infty \gamma_1 \log n}{n^\alpha} \right). \quad (23)$$

Simultaneously, if $\alpha > 1/2$,

$$\begin{aligned} & \frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 \\ &= \begin{cases} O_{\mathbb{P}} \left(\frac{\sigma^2 \log^2 n}{n} + n^{-1/2} (\sigma_\infty \wedge \sigma \gamma_1 n^{1-\alpha}) \log n \right) & \text{if } \alpha \leq 1 \\ O_{\mathbb{P}} \left(\left[\frac{\sigma^2 \log^2 n}{n} \right]^{1-\frac{1}{2\alpha}} + \frac{\sigma_\infty \log n}{n^\alpha} \right) & \text{if } \alpha > 1, \sigma^2 \lesssim n/\log^2 n. \end{cases} \end{aligned} \quad (24)$$

In some situations we can get improved results using (24), particularly in situations when $\sigma \lesssim \sigma_\infty$. This can happen for the Poisson family when the signal β^* is dominated by a few components.

Proof of Theorem B.2. Throughout let $t = \log n$. Start from the bound (17):

$$\frac{1}{n} \|\hat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\frac{|J| t^2 \sigma^2}{n} + \frac{t B_n}{n^\alpha} \right)$$

In the case $\alpha \leq 1/2$, set $J = [k+1]^d$ and recall the bound (21) for B_n . This gives the desired result in this case. In the other case of $\alpha > \frac{1}{2}$, we prove the bounds (23) and (24) now. Set $J = \{i : \|(i - k - 2)_+\|_2 < r\}$ for an r that we choose later.

Bound (23). Recall from (18) that

$$\begin{aligned} B_n &= 2t (\min \{ \|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1} \} \vee \|b\|_\infty L_{J,1}) \\ &\leq 2t (\|\nu\|_\infty L_{J,2} \vee \|b\|_\infty L_{J,1}) \end{aligned}$$

where we get the inequality by taking only the first term of the inner minimum. Plug in the bounds for L terms from (19), (20) to write

$$B_n = O \left(\|\nu\|_\infty \left(\frac{n}{r^d} \right)^{\alpha - \frac{1}{2}} + \left(\frac{n}{r^d} \right)^{(\alpha-1)_+} \gamma_1 \right) t.$$

Plug this back in (17) to get

$$\begin{aligned} & \frac{1}{n} \|\widehat{\beta} - \beta^*\|_2^2 \\ &= O_{\mathbb{P}} \left(\frac{(r+k+2)^d t^2 \sigma^2}{n} + \frac{t}{n^\alpha} \left\{ \|\nu\|_\infty \left(\frac{n}{r^d} \right)^{\alpha - \frac{1}{2}} \vee \|b\|_\infty \left(\frac{n}{r^d} \right)^{(\alpha-1)_+} \gamma_1 \right\} \right) \end{aligned} \quad (25)$$

For $\alpha \neq 1$, when possible we will choose $r \in [1, N\sqrt{d}]$ such that

$$\frac{r^d t^2 \sigma^2}{n} \asymp \frac{t}{n^\alpha} \sigma_\infty \left(\frac{n}{r^d} \right)^{\alpha - \frac{1}{2}}.$$

which is equivalent to

$$r^d \asymp \left(\frac{\sqrt{n} \sigma_\infty}{t \sigma^2} \right)^{\frac{2}{2\alpha+1}}.$$

Selecting this r when possible gives the bound in (23) and the assumption $\frac{\sqrt{n} \sigma_\infty}{t \sigma^2} \gtrsim 1$ ensures that we are not choosing an impossibly small r . When $\alpha = 1$, we can retrace the argument with the additional γ_1 factor in (25) to get the bound.

Bound (24). When $\alpha \leq 1$, set $J = [k+1]^d$ to get the stated bound. Now consider $\alpha > 1$. Simplify (18) by taking only the second term of the minimum, plug the bound for B_n in (17) to get

$$\frac{1}{n} \|\widehat{\beta} - \beta^*\|_2^2 = O_{\mathbb{P}} \left(\frac{(r+k+2)^d t^2 \sigma^2}{n} + t n^{-\alpha + \frac{1}{2}} \sigma \left(\frac{n}{r^d} \right)^{\alpha-1} \right)$$

When possible we will choose $r \in [1, N\sqrt{d}]$ to balance the two terms above, that is,

$$\frac{r^d t^2 \sigma^2}{n} \asymp t n^{-\alpha + \frac{1}{2}} \sigma \left(\frac{n}{r^d} \right)^{\alpha-1}$$

which means,

$$r^d \asymp \left(\frac{n}{\sigma^2 t^2} \right)^{\frac{1}{2\alpha}}.$$

This choice of r gives the desired bound. Our assumption that $\frac{n}{\sigma^2 t^2} \gtrsim 1$ makes sure that this choice of r is not impossibly small. This completes the proof. \square

Proof of Theorem 3.5. This is a direct result of Theorem B.2, simplifying the cases. \square

B.7. Error rates assuming that the estimate is bounded

Consider the penalized maximum likelihood estimator (MLE)

$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n (\psi(\theta_i) - y_i \theta_i) + \lambda \|D\theta\|_1. \quad (26)$$

The minimum may not be achieved at an interior point of the domain. In that case, we set $\hat{\theta}$ to a limit point of a sequence on which the objective converges to the infimum.

If we assume that $\hat{\theta}$ in (26) is constrained in such a way that $\psi''(\hat{\theta})$ is bounded away from 0, then the error bounding analysis essentially reduces to that in the Gaussian family case. Consider the constrained estimator

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta(K)^n} \sum_{i=1}^n -y_i \theta_i + \psi(\theta_i) + \lambda \|D\theta\|_1 \quad (27)$$

where $\Theta(K) = \{\theta \in \mathbb{R} : \psi''(\theta) \geq \frac{1}{K}\}$ for some $K > 0$. Assume that $\Theta(K)$ is a convex set for any $K > 0$. This can be verified for Poisson, exponential and logistic families. Suppose

$$\tilde{\theta} = \operatorname{argmin}_{\theta \in \Theta(K)^n} \sum_{i=1}^n -E[Y_i] \theta_i + \psi(\theta_i)$$

is the best approximation of θ^* within $\Theta(K)^n$. Also define $\tilde{\beta} = \nabla \psi(\tilde{\theta})$. Then the constrained estimator in (27) satisfies the following error bound.

Proposition B.3. *Let $y_i = \beta_i^* + \epsilon_i$ where ϵ_i is zero mean sub-exponential with parameters (ν_i^2, b_i) for $i \in [n]$. Let $L_{J,p}$ be as defined in (6) for $J \subset [N]^d, p \geq 1$.*

Abbreviate $A_n = \mu \sqrt{\frac{|J|}{n}} (\|\nu\|_2 \vee \|b\|_\infty) \log n$, and

$$B_n = (\min \{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}) \log n.$$

Then the estimator (27) with $\lambda = \frac{B_n}{n}$, satisfies

$$\overline{\text{KL}}(\tilde{\theta} \parallel \hat{\theta}) = \frac{1}{n} O_{\mathbb{P}}(K A_n^2 + B_n \|D\tilde{\theta}\|_1 + K \|\tilde{\beta} - \beta^*\|_2^2).$$

The proof is below. We choose $J \subset [N]^d$ to minimize the bound. If we set $K = 1/v_{\min}$ where $v_{\min} = \min_{i \in [n]} \psi''(\theta_i^*)$, then $\tilde{\theta} = \theta^*, \tilde{\beta} = \beta^*$ and the above bound reads

$$\overline{\text{KL}}(\theta^* \parallel \hat{\theta}) = \frac{1}{n} O_{\mathbb{P}}\left(\frac{A_n^2}{v_{\min}} + B_n \|D\theta^*\|_1\right).$$

Proof of Theorem B.3. Similar to the argument in Theorem 3.2, from the optimality of $\hat{\theta}$, we have the basic inequality,

$$R(\hat{\theta}) - R(\tilde{\theta}) \leq \frac{1}{n} \epsilon^\top (\hat{\theta} - \tilde{\theta}) + \lambda \|D\tilde{\theta}\|_1 - \lambda \|D\hat{\theta}\|_1 \quad (28)$$

To lower bound the left hand side, we see that

$$\begin{aligned}
nR(\widehat{\theta}) - nR(\widetilde{\theta}) &= \mathbf{1}^\top \psi(\widehat{\theta}) - \beta^* \widehat{\theta} - \mathbf{1}^\top \psi(\widetilde{\theta}) + \beta^* \widetilde{\theta} \\
&= \mathbf{1}^\top \psi(\widehat{\theta}) - \mathbf{1}^\top \psi(\widetilde{\theta}) - \widetilde{\beta}(\widehat{\theta} - \widetilde{\theta}) + (\widetilde{\beta} - \beta^*)^\top (\widehat{\theta} - \widetilde{\theta}) \\
&\geq \frac{1}{2K} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 + (\widetilde{\beta} - \beta^*)^\top (\widehat{\theta} - \widetilde{\theta}) \\
&\geq \frac{1}{2K} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 - K \|\widetilde{\beta} - \beta^*\|_2^2 - \frac{1}{4K} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 \\
&= \frac{1}{4K} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 - K \|\widetilde{\beta} - \beta^*\|_2^2
\end{aligned}$$

In the above display, the first inequality holds because both $\widehat{\theta}, \widetilde{\theta} \in \Theta(K)^n$ and $\Theta(K)^n$ is convex. (For $i \in [n]$, write $\psi(\widehat{\theta}_i) - \psi(\widetilde{\theta}_i) - \widetilde{\beta}_i(\widehat{\theta}_i - \widetilde{\theta}_i) = \psi''(u_i)(\widehat{\theta}_i - \widetilde{\theta}_i)^2$ for some u_i between $\widehat{\theta}_i$ and $\widetilde{\theta}_i$. As $\Theta(K)$ is convex and u_i lies between $\widehat{\theta}_i$ and $\widetilde{\theta}_i$, we should have $u_i \in \Theta(K)$ and so $\psi''(u_i)$ should be at least $1/K$.) The second inequality follows from the fact that $2ab \geq -ca^2 - \frac{1}{c}b^2$, for any $a, b, c \in \mathbb{R}$ with $c > 0$. Applying this to half of the left hand side of (28),

$$\frac{1}{2}(R(\widehat{\theta}) - R(\widetilde{\theta})) + \frac{1}{8nK} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 - \frac{K}{2n} \|\widetilde{\beta} - \beta^*\|_2^2 \leq \frac{1}{n} \epsilon^\top (\widehat{\theta} - \widetilde{\theta}) + \lambda \|D\widetilde{\theta}\|_1 - \lambda \|D\widehat{\theta}\|_1$$

Rearranging,

$$\frac{1}{2}(R(\widehat{\theta}) - R(\widetilde{\theta})) - \frac{K}{2n} \|\widetilde{\beta} - \beta^*\|_2^2 \leq -\frac{1}{8nK} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 + \frac{1}{n} \epsilon^\top (\widehat{\theta} - \widetilde{\theta}) + \lambda \|D\widetilde{\theta}\|_1 - \lambda \|D\widehat{\theta}\|_1$$

By Theorem B.4, for $t \geq 1$ and $J \subset [N]^d$, the following holds with probability at least $1 - 2(m + |J|)e^{-t}$,

$$\begin{aligned}
&\frac{1}{2}(R(\widehat{\theta}) - R(\widetilde{\theta})) - \frac{K}{2n} \|\widetilde{\beta} - \beta^*\|_2^2 \\
&\leq -\frac{1}{8nK} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 + \frac{A}{n} \|P_J(\widehat{\theta} - \widetilde{\theta})\|_2 \\
&\quad + \frac{B}{n} \|D(\widehat{\theta} - \widetilde{\theta})\|_1 + \lambda \|D\widetilde{\theta}\|_1 - \lambda \|D\widehat{\theta}\|_1
\end{aligned}$$

where

$$\begin{aligned}
A_n &= 2t\mu \sqrt{\frac{|J|}{n}} (\|\nu\|_2 \vee \|b\|_\infty), \\
B_n &= 2t (\min \{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}).
\end{aligned}$$

The sum of the first two terms on the right hand side can be bound by completing squares:

$$\begin{aligned}
&-\frac{1}{8nK} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 + \frac{A}{n} \|P_{[J]}(\widehat{\theta} - \widetilde{\theta})\|_2 \\
&\leq -\frac{1}{8nK} \|\widehat{\theta} - \widetilde{\theta}\|_2^2 + \frac{A}{n} \|\widehat{\theta} - \widetilde{\theta}\|_2 \\
&\leq \frac{2KA^2}{n}.
\end{aligned}$$

Plug this into the bound in the previous display to get

$$\frac{1}{2}(R(\hat{\theta}) - R(\tilde{\theta})) - \frac{K}{2n}\|\tilde{\beta} - \beta^*\|_2^2 \leq \frac{2KA^2}{n} + \frac{B}{n}\|D(\hat{\theta} - \tilde{\theta})\|_1 + \lambda\|D\tilde{\theta}\|_1 - \lambda\|D\hat{\theta}\|_1$$

The argument from here is similar to that in the proof of [Theorem 3.2](#). \square

B.8. Empirical process bound

Let $D = D_{n,d}^{(k+1)} = U\Sigma V^\top$ be the full singular value decomposition of D . For $j \in [N]^d$, let V_j denote $\tilde{V}_{j_1} \otimes \cdots \otimes \tilde{V}_{j_d}$ where \tilde{V}_ℓ is the eigenvector of $(D_{N,1}^{(k+1)})^\top D_{N,1}^{(k+1)}$ corresponding to its ℓ th smallest eigenvalue. For $J \subset [N]^d$, let V_J denote a $n \times |J|$ matrix formed by picking the columns of V corresponding to J . Let $P_J = V_J V_J^\top$ be the projection matrix onto those columns.

Lemma B.4. *Let $y_i = \beta_i^* + \epsilon_i$ where ϵ_i is zero mean sub-exponential with parameters (ν_i^2, b_i) for $i \in [n]$. Let $J \subset [N]^d$ and L be as defined in [\(6\)](#). Let m be the number of rows in D . For any $J \subset [N]^d$ containing $[k+1]^d$, and $t \geq 1$, with probability at least $1 - 2(m + |J|)e^{-t}$, the following holds uniformly for all $\theta \in \mathbb{R}^n$:*

$$|\epsilon^\top \theta| \leq A\|P_J \theta\|_2 + B\|D\theta\|_1$$

where

$$\begin{aligned} A &= 2t\mu\sqrt{\frac{|J|}{n}}(\|\nu\|_2 \vee \|b\|_\infty), \\ B &= 2t(\min\{\|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1}\} \vee \|b\|_\infty L_{J,1}). \end{aligned}$$

Proof of Theorem B.4. Decompose

$$\begin{aligned} |\epsilon^\top \theta| &= |\epsilon^\top P_J \theta + \epsilon^\top (I - P_J) \theta| \\ &= |\epsilon^\top P_J \theta + \epsilon^\top (I - P_J) D^\dagger D \theta| \\ &\leq \|P_J \epsilon\|_2 \|P_J \theta\|_2 + \|(D^\dagger)^\top (I - P_J) \epsilon\|_\infty \|D \theta\|_1 \end{aligned}$$

where we applied Hölder's inequality on each of the two terms separately. We give high probability bounds for $\|P_J \epsilon\|_2$ and $\|(D^\dagger)^\top (I - P_J) \epsilon\|_\infty$ separately. A union bound will yield the stated result.

Bounding $\|P_J \epsilon\|_2$. For $j \in J$, $V_j^\top \epsilon$ is $\text{SE}(\|\nu \odot V_j\|_2^2, \|b \odot V_j\|_\infty)$. Therefore, from [\(10\)](#),

$$|V_j^\top \epsilon| \leq 2t(\|\nu \odot V_j\|_2 \vee \|b \odot V_j\|_\infty)$$

should hold with probability at least $1 - 2e^{-t}$ for any $t \geq 1$. From the incoherence property ($\|V_j\|_\infty \leq \frac{\mu}{\sqrt{n}}$), we get $\|\nu \odot V_j\|_2 \leq \frac{\mu}{\sqrt{n}}\|\nu\|_2$ and $\|b \odot V_j\|_\infty \leq \frac{\mu}{\sqrt{n}}\|b\|_\infty$. Therefore,

$$|V_j^\top \epsilon| \leq 2t\frac{\mu}{\sqrt{n}}(\|\nu\|_2 \vee \|b\|_\infty).$$

By union bound over $j \in J$, for any $t \geq 1$,

$$\|P_J \epsilon\|_2^2 = \sum_{j \in J} (V_j^\top \epsilon)^2 \leq |J| \left(2t \frac{\mu}{\sqrt{n}} (\|\nu\|_2 \vee \|b\|_\infty) \right)^2$$

should hold with probability at least $1 - 2|J|e^{-t}$.

Bounding $\|(D^\dagger)^\top (I - P_J) \epsilon\|_\infty$. Rewrite this term as

$$\|(D^\dagger)^\top (I - P_J) \epsilon\|_\infty = \max_{j \in [m]} |g_j^\top \epsilon|$$

where $g_j = (I - P_J)D^\dagger e_j$ for $j \in [m]$ and where m is the number of rows in D . From [Theorem A.2](#), one can deduce that

$$\max_{j \in [m]} |g_j^\top \epsilon| \leq 2t \left(\max_{j \in [m]} \|\nu \odot g_j\|_2 \vee \|b \odot g_j\|_\infty \right).$$

holds with probability at least $1 - 2me^{-t}$ for $t \geq 1$. Observe that $\|b \odot g_j\|_\infty \leq \|b\|_\infty \|g_j\|_\infty$ and

$$\|\nu \odot g_j\|_2 \leq \min \{ \|\nu\|_\infty \|g_j\|_2, \|\nu\|_2 \|g_j\|_\infty \}.$$

Therefore, substituting the bounds on $\|g_j\|_2, \|g_j\|_\infty$ from [Theorem B.5](#), we get

$$\max_{j \in [m]} |g_j^\top \epsilon| \leq 2t \left(\min \{ \|\nu\|_\infty L_{J,2}, \|\nu\|_2 L_{J,1} \} \vee \|b\|_\infty L_{J,1} \right).$$

with probability at least $1 - 2me^{-t}$. \square

Lemma B.5. *Let J be a subset of $[N]^d$ containing $[k+1]^d$. Define $g_j = (I - P_J)D^\dagger e_j$ for $j \in [m]$ where m is the number of rows in D . Then for all $j \in [m]$,*

$$\begin{aligned} \|g_j\|_2 &\leq L_{J,2}, \\ \|g_j\|_\infty &\leq L_{J,1}. \end{aligned}$$

Proof of Theorem B.5. Let $\tilde{\Sigma} \in \mathbb{R}^{m \times n}$ denote the diagonal matrix such that $\tilde{\Sigma}_{i,i} = \xi_i$ for $i \in J$ and 0 otherwise. Here we abuse the subscript notation: i is a multi-index in $[N]^d$ but we use it as an integer mapped to $\{1, 2, \dots, N^d\}$ by row-major order. Let $\dot{\Sigma} = \Sigma - \tilde{\Sigma}$, which is also diagonal $m \times n$. Then

$$g_j = V \dot{\Sigma}^\dagger U^\top e_j.$$

Therefore, we can write

$$\|g_j\|_2^2 = \|V \dot{\Sigma}^\dagger U^\top e_j\|_2^2 = \|\dot{\Sigma}^\dagger U^\top e_j\|_2^2 = \sum_{i \in [N]^d \setminus J} U_{ij}^2 \frac{1}{\xi_i^2} \leq \frac{\mu^2}{n} \sum_{i \in [N]^d \setminus J} \frac{1}{\xi_i^2} = L_{J,2}^2.$$

Here also, we abuse the subscript notation for U by mapping $i \in [N]^d$ to $\{1, 2, \dots, N^d\}$ with row-major order. The sole inequality in the above display

follows from the incoherence property of U . This shows the upper bound on the ℓ_2 norms of $g_j, j \in [m]$.

For the ℓ_∞ -norm bound, we write,

$$\|g_j\|_\infty = \max_{\|z\|_1=1} z^\top g_j = \max_{\|z\|_1=1} z^\top V \dot{\Sigma}^\dagger U^\top e_j \leq \max_{\|z\|_1=1} \|V^\top z\|_\infty \|\dot{\Sigma}^\dagger U^\top e_j\|_1$$

using Hölder's inequality. Because every entry of V is at most μ/\sqrt{n} , we have

$$\max_{\|z\|_1=1} \|V^\top z\|_\infty \leq \frac{\mu}{\sqrt{n}}.$$

From the incoherence property of U ,

$$\|\dot{\Sigma}^\dagger U^\top e_j\|_1 \leq \frac{\mu}{\sqrt{n}} \sum_{i=\ell+1}^n \frac{1}{\xi_i}.$$

Therefore

$$\|g_j\|_\infty \leq \frac{\mu^2}{n} \sum_{i \in [N]^d \setminus J} \frac{1}{\xi_i} = L_{J,1}. \quad \square$$

B.9. Eigenvalue bounds

Lemma B.6. *Let $\{\xi_i^2 : i = (i_1, \dots, i_d) \in [N]^d\}$ be the eigenvalues of $D^\top D$ where $D = D_{n,d}^{(k+1)}$ and let $p \geq 1$, $\alpha = (k+1)/d$. Then*

$$\sum_{i \in [N]^d \setminus [k+1]^d} \frac{1}{\xi_i^p} \leq c \begin{cases} n & \text{if } p\alpha < 1 \\ n \log n & \text{if } p\alpha = 1 \end{cases}$$

for large enough n , where $c > 0$ is a constant depending only on k, d . In the case $p\alpha > 1$, for any $r_0 \in [1, \sqrt{d}N]$,

$$\sum_{i \in [N]^d : \|(i-k-2)_+\|_2 \geq r_0} \frac{1}{\xi_i^p} \leq cn(n/r_0^d)^{p\alpha-1}.$$

Proof of Theorem B.6. This is a generalization of Lemma 6 in [Sadhanala et al. \(2021\)](#), which states the bound for only $p = 2$. In their proof, if we change the power applied to the singular values in the summation to a general $p \geq 1$ we get (a) the bound in the second display and (b) a bound slightly weaker than the first display:

$$\sum_{i \in [N]^d \setminus [k+2]^d} \frac{1}{\xi_i^p} \leq c \begin{cases} n & p\alpha < 1 \\ n \log n & p\alpha = 1 \end{cases} \quad (29)$$

for large enough n , where $c > 0$ is a constant depending only on k, d . Notice that the summation excludes indices in $[k+2]^d$ whereas the statement in [Theorem B.6](#) requires only those in $[k+1]^d$ to be excluded. We claim that the additional terms from indices $[k+2]^d \setminus [k+1]^d$ do not change the rates in the bound. Thanks to the

Kronecker-sum structure of $D^\top D$, we can write $\xi_i^2 = \sum_{j=1}^d \rho_{i_j}$ where ρ_1, \dots, ρ_N are the eigenvalues of $(D_{N,1}^{(k+1)})^\top D_{N,1}^{(k+1)}$. Note that for $i \in [N]^d \setminus [k+1]^d$, we can write $\xi_i^2 \geq \rho_{k+2}$. Therefore,

$$\begin{aligned} \sum_{i \in [k+2]^d \setminus [k+1]^d} \frac{1}{\xi_i^p} &\leq \sum_{i \in [k+2]^d \setminus [k+1]^d} \frac{1}{\rho_{k+2}^{p/2}} \\ &\leq \sum_{i \in [k+2]^d \setminus [k+1]^d} N^{p(k+1)} \\ &\leq ((k+2)^d - (k+1)^d) c n^{p\alpha} \end{aligned}$$

where we used [Theorem B.7](#) for the second inequality. In the case $p\alpha \leq 1$, this and (29) are sufficient to prove the lemma. \square

Lemma B.7. *For $k \geq 1, N > 2k + 2$, the smallest eigenvalue of $D_{N,1}^{(k)} (D_{N,1}^{(k)})^\top$ is at least c/N^{2k} for some constant $c > 0$ depending only on k .*

Proof. For the purpose of this lemma, let $\lambda_i(A)$ denote the i th smallest eigenvalue of A .

Case: k is odd. By Cauchy interlacing argument in Lemma 7 of [Sadhanala et al. \(2021\)](#), we have $\lambda_1(D_{N,1}^{(k)} (D_{N,1}^{(k)})^\top) \geq \lambda_1(GG^\top)$ where G is the graph trend filtering operator of order k on a chain of length N . Recall that $G = D_{N,1}^{(1)} L^{(k-1)/2}$ where L is the graph Laplacian of a chain of length N . Note that, for odd k , $G^\top G = L^k$. The set of *nonzero* eigenvalues of GG^\top and $G^\top G$ should be the same. We know that $\lambda_1(L) = 0, \lambda_2(L) > 0$ and so $\lambda_1(G^\top G) = 0, \lambda_2(G^\top G) > 0$. GG^\top has full rank. Therefore,

$$\lambda_1(GG^\top) = \lambda_2(G^\top G) = \lambda_2(L^k) = (\lambda_2(L))^k.$$

Plugging in $\lambda_2(L) = 4 \sin^2 \pi/2N$ and using the inequality $\sin x \geq x/2$ for $x \in [0, \pi/2]$, we have $\lambda_1(GG^\top) \geq c/N^{2k}$. As $\lambda_1(D_{N,1}^{(k)} (D_{N,1}^{(k)})^\top) \geq \lambda_1(GG^\top)$, we get $\lambda_1(D_{N,1}^{(k)} (D_{N,1}^{(k)})^\top) \geq c/N^{2k}$.

Case: k is even. Apply [Theorem B.8](#) to get the bound in this case. \square

Lemma B.8. *let $\lambda_i(A)$ denote the i th smallest eigenvalue of A . For $k \geq 1$, and $N > 2k + 2$,*

$$\lambda_{2k+1}((D_{N,1}^{(2k)})^\top D_{N,1}^{(2k)}) \geq \left(4 \sin^2 \frac{\pi}{2N-2}\right)^{2k}.$$

Proof. Let L_m denote the Laplacian of cycle graph with m vertices. Its smallest nonzero eigenvalue is $4 \sin^2 \pi/m$. Its eigenvectors are given $(v_\ell)_j = e^{2\pi i \ell j/m}$.

Let $u \in \mathbb{R}^N$ be the eigenvector of $(D_{N,1}^{(2k)})^\top D_{N,1}^{(2k)}$ corresponding to its $(2k+1)$ th eigenvalue. By [Theorem B.9](#), there exists a $v \in \mathbb{R}^{2N-2}$ satisfying the following

properties:

$$\begin{aligned}\|L^k v\|_2^2 &\leq 2\|(D_{N,1}^{(2k)} u)\|_2^2, \\ \langle v, \mathbf{1} \rangle &= 0, \\ \|v\|_2^2 &\geq 2.\end{aligned}$$

With such a v ,

$$\lambda_{2k+1}((D_N^{(2k)})^\top D_N^{(2k)}) = \|D_N^{(2k)} u\|_2^2 \geq \frac{1}{2}\|L^k v\|_2^2 \geq \frac{1}{2}\lambda_2(L^{2k})\|v\|_2^2 \geq \lambda_2^{2k}(L).$$

The equality holds by definition of u . The three inequalities follow in order from the three properties satisfied by v above. This is sufficient to complete the proof because we know that $\lambda_2(L) = 4\sin^2 \frac{\pi}{2N-2}$. \square

Lemma B.9. *Let $u \in \mathbb{R}^N$ be the eigenvector of $(D_{N,1}^{(2k)})^\top D_{N,1}^{(2k)}$ corresponding to its $(2k+1)$ th eigenvalue. There exists a $v \in \mathbb{R}^{2N-2}$ satisfying the following properties:*

$$\begin{aligned}\|L^k v\|_2^2 &\leq 2\|(D_{N,1}^{(2k)} u)\|_2^2, \\ \langle v, \mathbf{1} \rangle &= 0, \\ \|v\|_2^2 &\geq 2.\end{aligned}$$

Proof. Define $\mathcal{U} = \{u \in \mathbb{R}^N : u_1 = u_N = 0\}$.

Δ and Δ^{-1} : Define the following truncated discrete difference operator,

$$\Delta u = (0, (D_{N,1}^{(2)} u)_1, D_{N,1}^{(2)} u)_2, \dots, D_{N,1}^{(2)} u)_{N-2}, 0)$$

for $u \in \mathcal{U}$ so that $\Delta : \mathcal{U} \rightarrow \mathcal{U}$. We can write

$$\Delta = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ & & \dots & & & \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (30)$$

Then we can construct the inverse as the following truncated discrete integral using the following: Let $u \in \mathcal{U}$, and define the cumulative sum operator,

$$(Iu)_i := \sum_{j=1}^{i-1} j u_{i-j}, \quad \text{and } a := \frac{1}{N-1}(Iu)_N.$$

Define

$$z_i := (i-1)a - (Iu)_i, \quad i = 1, \dots, N,$$

and note that $z_1 = z_N = 0$. Then we have that $\Delta z = u$ for $u \in \mathcal{U}$. To see this let $i = 2, \dots, N-1$,

$$\begin{aligned} -(\Delta z)_i &= -(2ia - (i-1)a - (i+1)a) + 2(Iu)_i - (Iu)_{i-1} - (Iu)_{i+1} \\ &= 2 \sum_{j=1}^{i-1} j u_{i-j} - \sum_{j=1}^{i-2} j u_{i-1-j} - \sum_{j=1}^i j u_{i-j+1} \\ &= 2 \sum_{j=1}^{i-1} j u_{i-j} - \sum_{j=2}^{i-1} (j-1) u_{i-j} - \sum_{j=0}^{i-1} (j+1) u_{i-j} \\ &= 2u_{i-1} - u_i - 2u_{i-1} = -u_i. \end{aligned}$$

Also, $(\Delta z)_1 = (\Delta z)_N = 0 = u_1 = u_N$.

Constructing v : Construct $\tilde{u} \in \mathbb{R}^N$ such that

$$\tilde{u}_i = u_i - u_1 - \frac{u_N - u_1}{N-1}(i-1), \quad i = 1, \dots, N$$

Define $w \in \mathbb{R}^N$ such that $w_i = (\Delta^k \tilde{u})_i$ for $i = 1, \dots, k$ and $i = N, N-1, N-k+1$; and $w_i = 0$ for other $i \in [N]$. Define $p = \Delta^{-k} w$ and note that $w, p \in \mathcal{U}$. Let $\text{ext}(x)$ denote the periodic extension of $x \in \mathbb{R}^N$, defined by $y \in \mathbb{R}^{2N-2}$ where $y_{1:N} = x, y_{N+i} = -x_{N-i}$ for $i = 1, \dots, N-2$. Set

$$v = \text{ext}(\tilde{u} - p).$$

Verifying the three properties: As $\tilde{u} - p \in \mathcal{U}$, by [Theorem B.10](#),

$$(L^k v)_{1:N} = \Delta^k(\tilde{u} - p) = [0_{k \times 1}; D_{N,1}^{(2k)} u; 0_{k \times 1}].$$

By construction of v via ext , $(L^k v)_{2:N} = -(L^k v)_{2N-2:N}$. So $\|(L^k v)_{N+1:2N-2}\|_2^2 = \|(L^k v)_{1:N}\|_2^2$. Therefore v satisfies the first desired property in the statement of the lemma:

$$\|L^k v\|_2^2 = 2\|D_{N,1}^{(2k)} u\|_2^2.$$

As $v = \text{ext}(\tilde{u} - p)$ and $\tilde{u} - p \in \mathcal{U}$, we get $\langle v, \mathbf{1} \rangle = 0$ from the definition of ext . Again due to the definition of ext , $\|v\|_2^2 = 2\|\tilde{u} - p\|_2^2$. Write $\tilde{u} - p = u + (\tilde{u} - u - p)$ and note that $u \perp \mathcal{N}(D_{N,1}^{(2k)})$, $\tilde{u} - u$ is linear and hence in $\mathcal{N}(D_{N,1}^{(2k)})$ and further $p \in \mathcal{N}(D_{N,1}^{(2k)})$ by construction. (Note that if strip out the top and bottom k rows from Δ^k , we get $D_{N,1}^{(2k)}$. So $D_{N,1}^{(2k)} p = (\Delta^k p)_{k+1:N-k} = w_{k+1:N-k} = 0$.) Therefore we get the third desired property for v :

$$\|v\|_2^2 \geq 2\|u\|_2^2 + 2\|\tilde{u} - u - p\|_2^2 \geq 2.$$

Therefore v satisfies all the three properties stated in the lemma. \square

Lemma B.10. Let $\mathcal{U} = \{u \in \mathbb{R}^N : u_1 = u_N = 0\}$. Let $\text{ext}(u)$ denote the periodic extension of $u \in \mathbb{R}^N$, defined by $v \in \mathbb{R}^{2N-2}$ where $v_{1:N} = u, v_{N+i} = -u_{N-i}$ for $i = 1, \dots, N-2$. Let L, Δ be as defined in [Theorem B.9](#) and (30) respectively. Then $(L^k \text{ext}(u))_{1:N} = \Delta^k u$ for $u \in \mathcal{U}$.

Proof. Let $v := \text{ext}(u)$ and let $\mathcal{S} = \{\text{ext}(u) : u \in \mathcal{U}\}$. We need to show that $(L^k v)_{1:N} = \Delta^k u$ for $k \geq 1$. First notice that $(\Delta u)_i = 2u_i - u_{i-1} - u_{i+1}$, $i = 2, \dots, N-1$. Furthermore, $(\Delta u)_1 = (\Delta u)_N = 0$ because the first and last rows of Δ are zeros and $(Lv)_1 = (Lv)_N = 0$ because $v \in \mathcal{S}$. (As $v \in \mathcal{S}$, v is anti-symmetric around index 1, that is: $v_1 = 0$, $v_i = -v_{2N-i}$ for $i = 2, 3, \dots, N$ and so $(Lv)_1 = 0$. Similarly $v_{N-i} = -v_{N+i}$ for $i = 0, 1, \dots, N-2$ and so $(Lv)_N = 0$.) So we have shown it for $k = 1$. Suppose the inductive hypothesis $\Delta^{k-1} u = (L^{k-1} v)_{1:N}$. We have for $i = 2, \dots, N-1$,

$$\begin{aligned} (\Delta^k u)_i &= 2(\Delta^{k-1} u)_i - (\Delta^{k-1} u)_{i-1} - (\Delta^{k-1} u)_{i+1} \\ &= 2(L^{k-1} v)_i - (L^{k-1} v)_{i-1} - (L^{k-1} v)_{i+1} = (L^k v)_i. \end{aligned}$$

Furthermore, $(\Delta^k u)_1 = (\Delta^k u)_N = 0$ by construction and $(L^k v)_1 = (L^k v)_N = 0$ because of anti-symmetry of v around indices 1 and N . Thus, $(L^k v)_{1:N} = \Delta^k u$. \square

Appendix C: Proofs for lower bounds

C.1. Proof of [Theorem 3.6](#)

Denote the ℓ_p balls

$$B_p(r; \mathbb{R}^n) = \{x \in \mathbb{R}^n : \|x\|_p \leq r\}$$

for $p \geq 1, r \geq 0, n \geq 1$. We simply refer to this $B_p(r)$ when the dimension n is clear from the context. Consider the set

$$B(r, m) = \{\beta \in \mathbb{R}^n : \|\beta\|_\infty \leq r, \|\beta\|_0 \leq m\} \quad (31)$$

which consists of signals with at most m non-zero components and with all entries at most r in magnitude.

For $\beta \in \mathbb{R}$ and $\sigma > 0$, let $\text{Lap}(\beta, \sigma)$ denote the Laplace distribution centered at β with scale σ . For $\beta \in \mathbb{R}^n$, let $\text{Lap}(\beta, \sigma)$ denote the product distribution of $\text{Lap}(\beta_1, \sigma), \dots, \text{Lap}(\beta_n, \sigma)$.

Proof of [Theorem 3.6](#). The null space of D has a dimension of κ . Using Fano's lemma, similar to the way it is applied in Example 15.8 in [Wainwright \(2019\)](#), we can show that

$$n \cdot R_M(T_{n,d}^k(C_n)) \geq \frac{\kappa \sigma^2}{128} \quad (32)$$

The main difference is in upper bounding for KL divergence, but from [Theorem C.2](#) we can show that

$$\text{KL}(\text{Lap}(a, \sigma), \text{Lap}(b, \sigma)) \leq \|a - b\|_2^2 / 2\sigma^2$$

for $a, b \in \mathbb{R}^n$. This is sufficient to apply the argument in Example 15.8 in [Wainwright \(2019\)](#).

Now we show the second lower bound. Note that

$$B_1(C_n/c_k) \subseteq T_{n,d}^k(C_n)$$

where c_k is the maximum ℓ_1 norm of columns of D . c_k depends only on k, d . Denote $r_1 = C_n/c_k$. For $q \in Q := \{1\} \cup \{2m : 2m \leq n/3\}$, set $r = C_n/(qc_k)$ so that $B(r, q)$ is contained in $B_1(C_n/c_k)$. From [Theorem C.3](#),

$$n \cdot R_M(B(r, q)) \geq \frac{1}{12} q a^2$$

where $a = r \wedge \sigma g^{-1}(\tau/6)$ where $\tau = \log(en/8q)$. Therefore, from the containment $B(r, q) \subset T_{n,d}^k(C_n)$,

$$\begin{aligned} n \cdot R_M(T_{n,d}^k(C_n)) &\geq \frac{1}{12} \sup_{q \in Q} q \min \left\{ r^2, \frac{\sigma^2}{3} \log \frac{en}{8q} \vee \frac{\sigma^2}{36} \log^2 \frac{en}{8q} \right\} \\ &= \frac{1}{12} \sup_{q \in Q} q \min \left\{ \frac{r_1^2}{q^2}, \frac{\sigma^2}{3} \log \frac{en}{8q} \vee \frac{\sigma^2}{36} \log^2 \frac{en}{8q} \right\} \end{aligned}$$

Choose $q \in Q$ that maximizes this bound. Set q to the closest number in Q to

$$q^* = \frac{r_1}{\sigma} \left(\sqrt{3} \log^{-1/2} \frac{\sigma n}{\sqrt{3} r_1} \vee 6 \log^{-1} \frac{\sigma n}{6 r_1} \right)$$

where $r_1 = C_n/c_k$. This gives a lower bound of

$$c_0 \sigma r_1 \left(\sqrt{\log \frac{c_1 \sigma n}{r_1}} \vee \log \frac{c_2 \sigma n}{r_1} \right) \quad (33)$$

provided q^* is within the range $[1, n/3]$. Two alternate bounds can be obtained by plugging in $q = 1$ and $q = 2\lfloor n/6 \rfloor$. With $q = 1$, the bound is $c \min \left\{ r_1^2, \sigma^2 \left(\log \frac{en}{8} \vee \log^2 \frac{en}{8} \right) \right\}$ and with $q = 2\lfloor n/6 \rfloor$, the bound is $c \min \left\{ \frac{r_1^2}{n}, \sigma^2 \right\}$.

Finally, we derive the third term in the lower bound by embedding a Hölder ball. We follow the proof of Theorem 2.5 in [Tsybakov \(2009\)](#). For $k \geq 0$ and $L > 0$, let $H(k+1, L; [0, 1]^d)$ denote the Hölder class of functions on $[0, 1]^d$ whose k th order partial derivatives $\partial^k f / \partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}$ with $\alpha_1 + \dots + \alpha_d = k$ are L -Lipschitz. Define the discrete Hölder set using evaluations of Hölder functions on the grid:

$$\mathcal{H}_{n,d}^k(L) = \{\theta \in \mathbb{R}^n : \theta_i = f(i_1/n, \dots, i_d/n), f \in H(k+1, L; [0, 1]^d)\}.$$

[Sadhanala et al. \(2017\)](#) shows that

$$\mathcal{H}_{n,d}^k(c C_n n^{\alpha-1}) \subset T_{n,d}^k(C_n)$$

for a constant c depending only k . Therefore, the minimax risk over $T_{n,d}^k(C_n)$ is at least the minimax risk over $\mathcal{H}_{n,d}^k(C_n)$. [Theorem C.1](#) gives a lower bound on this risk:

$$R_M(T_{n,d}^k(C_n)) = \Omega \left(\left(\frac{\sigma^2}{n} \right)^{\frac{2\alpha}{2\alpha+1}} (C_n n^{\alpha-1})^{\frac{2}{2\alpha+1}} \right).$$

This equation, together with [\(32\)](#), [\(33\)](#) gives the desired lower bound. \square

Lemma C.1. *On the d -dimensional grid, consider the observation model $y_i = f(x_i) + \epsilon_i$ for $i \in [N]^d$ where $f \in H(k+1, L; [0, 1]^d)$ and ϵ_i are i.i.d. $\text{Lap}(0, \sigma)$. Then*

$$\inf_{\hat{f}} \sup_{f_0 \in H(k+1, L; [0, 1]^d)} E \|\hat{f} - f_0\|_2^2 = \Omega \left(\left(\frac{\sigma^2}{n} \right)^{\frac{2\alpha}{2\alpha+1}} L^{\frac{2}{2\alpha+1}} \right). \quad (34)$$

Suppose there exists an $h_0 \geq 0$ such that, for any $h \geq h_0$, any ball of radius $ch/2$ in $[0, 1]^d$ contains at least $c_1 n (ch/2)^d$ grid points, where $c = \sqrt{\log_{2e} 2}$ and $c_1 > 0$ is a constant may depend on d . Then the following lower bound in terms of the empirical norm holds:

$$\inf_{\hat{f}} \sup_{f_0 \in H(k+1, L; [0, 1]^d)} E \|\hat{f} - f_0\|_n^2 = \Omega \left(\left(\frac{\sigma^2}{n} \right)^{\frac{2\alpha}{2\alpha+1}} L^{\frac{2}{2\alpha+1}} \right). \quad (35)$$

Proof of Theorem C.1. We adapt the proof of the univariate case in Section 2.6 of [Tsybakov \(2009\)](#). Partition $[0, 1]^d$ into $r = \lceil c_0 n^{1/(2\alpha+1)} \rceil$ hypercubes of equal size, where c_0 is to be determined later. The side length of each hypercube $h = (1/r)^{1/d}$. Let $z_i, i \in [r]$ be the centers of these hypercubes. Define the bump function

$$\varphi(x) = Lh^{k+1} K\left(\frac{\|x\|_2}{h}\right) \text{ for } x \in [0, 1]^d \quad \text{where } K(u) = ae^{\frac{-1}{1-4u^2}} 1\{|u| < \frac{1}{2}\}$$

for a constant a such that $\varphi \in H(k+1, 1)$. Note that $\varphi(x) = 0$ if $\|x\|_2 \geq h/2$. Define the bump functions $\varphi_i(x) = \varphi(x - z_i)$, centered around z_i for $i \in [r]$. These functions have disjoint support and so, they are orthogonal to each other with respect to the L_2 inner product and also the empirical inner product. Note that

$$\|\varphi\|_2^2 = L^2 h^{2k+2+d} \|K\|_2^2 \quad (36)$$

By Varshamov-Gilbert lemma (see Lemma 2.9 in [Tsybakov, 2009](#)), we can get $\omega^{(0)}, \dots, \omega^{(M)} \in \{0, 1\}^r$ such that $\omega^{(0)} = \mathbf{1}_r$, $M \geq 2^{r/8}$ and for $i \neq j \in \{0, \dots, M\}$, $d_H(\omega^{(i)}, \omega^{(j)}) \geq r/8$ where d_H calculates the Hamming distance between two binary vectors of same size. Let

$$f_i = \sum_{j=1}^r \omega_j^{(i)} \varphi_j$$

for $i = 0, \dots, M$. For $i \neq j$,

$$\begin{aligned} \|f_i - f_j\|_2^2 &= \sum_{\ell=1}^r 1\{\omega_\ell^{(i)} \neq \omega_\ell^{(j)}\} \|\varphi_\ell\|_2^2 \\ &= d_H(\omega^{(i)}, \omega^{(j)}) \|\varphi\|_2^2 \\ &\geq \frac{r}{8} \cdot L^2 h^{2k+2+d} \|K\|_2^2 \end{aligned} \quad (37)$$

The last line is true because (a) $d_H(\omega^{(i)}, \omega^{(j)}) \geq r/8$ by construction of the bump functions and (b) (36).

distribution $\Pi_{i=1}^n \text{Lap}(\mu_i, \sigma)$. Let $x_1, \dots, x_n \in [0, 1]^d$ denote the grid locations. For $j \in \{0, \dots, M\}$, let P_j denote the joint distribution of y_1, \dots, y_n given by $y_i = f_j(x_i) + \epsilon_i$ with ϵ_i i.i.d. $\text{Lap}(0, \sigma)$. Then

$$\begin{aligned}
\text{KL}(P_j, P_0) &= \sum_{i=1}^n \text{KL}(\text{Lap}(f_j(x_i), \sigma), \text{Lap}(0, \sigma)) \\
&\leq \sum_{i=1}^n \frac{1}{2\sigma^2} f_j^2(x_i) \\
&\leq \sum_{i=1}^n \frac{1}{2\sigma^2} L^2 a^2 h^{2k+2} \\
&= \frac{n}{2\sigma^2} L^2 a^2 h^{2k+2} \\
&= \frac{n}{2\sigma^2} L^2 a^2 r^{-2\alpha} \\
&= \frac{1}{2\sigma^2} L^2 a^2 r c_0^{-(2\alpha+1)} \tag{38}
\end{aligned}$$

The second line is from Lemma C.2 and the third line is from the fact that f_j is a summation of bump functions with (a) disjoint supports and (b) a maximum value of aLh^{k+1} . The last two lines follow from the relations $h = r^{-1/d}$, $r = \lceil c_0 n^{1/(2\alpha+1)} \rceil$.

Now we choose a c_0 (recall $r = \lceil c_0 n^{1/(2\alpha+1)} \rceil$) such that

$$\frac{1}{M} \sum_{j=1}^r \text{KL}(P_j, P_0) \leq \frac{1}{8 \log 4} \log M.$$

From (38) and the fact that $M \geq 2^{r/8}$, it is sufficient to choose c_0 such that $\frac{1}{2\sigma^2} L^2 a^2 r c_0^{-(2\alpha+1)} \leq \frac{r}{64}$. So we choose

$$c_0 = (32a^2 L^2 \sigma^{-2})^{1/(2\alpha+1)}.$$

With this choice of c_0 , and the lower bound in (37) we can apply Theorem 2.5 in Tsybakov (2009) to get the bound in (34).

Lower bound in empirical norm. We follow the same approach to show the lower bound in (35) in terms of the empirical norm. It is sufficient to show a bound analogous to (37) in terms of the empirical norm. Let $B(z, s)$ denote an ℓ_2 ball of radius s centered at z .

For any $\ell \in [r]$, by hypothesis, there are at least $c_1 n (ch/2)^d$ grid points in $B(z_\ell, ch/2)$. For $x \in B(z_\ell, ch/2)$, $\varphi(x) = Lh^{k+1} K(\|x - z_\ell\|_2/h) \geq Lh^{k+1} K(c/2)$. For our choice $c = \sqrt{\log_{2e} 2}$, $K(c/2) \geq K(0)/2e = a/2e$. Therefore, for all $x \in B(z_\ell, ch/2)$, $\varphi_\ell(x) \geq a/2e \cdot Lh^{k+1}$. Consequently,

$$\|\varphi_\ell\|_n^2 \geq \frac{1}{n} \cdot c_1 n (ch/2)^d \cdot (a/2e Lh^{k+1})^2 = c_2 L^2 h^{2k+2+d}.$$

Recall that

$$\|\varphi_\ell\|_2^2 = L^2 h^{2k+2+d} \|K\|_2^2$$

and therefore

$$\|\varphi_\ell\|_n^2 \geq c_3 \|\varphi_\ell\|_2^2 \quad (39)$$

for a constant c_3 that may depend on d .

$$\begin{aligned} \|f_i - f_j\|_n^2 &= \sum_{\ell=1}^r 1\{\omega_\ell^{(i)} \neq \omega_\ell^{(j)}\} \|\varphi_\ell\|_n^2 \\ &\geq \sum_{\ell=1}^r 1\{\omega_\ell^{(i)} \neq \omega_\ell^{(j)}\} c_3 \|\varphi_\ell\|_2^2 \\ &= \sum_{\ell=1}^r 1\{\omega_\ell^{(i)} \neq \omega_\ell^{(j)}\} c_3 \|\varphi\|_2^2 \\ &= d_H(\omega^{(i)}, \omega^{(j)}) c_3 \|\varphi\|_2^2 \\ &= c_3 \frac{r}{8} \cdot L^2 h^{2k+2+d} \|K\|_2^2 \end{aligned}$$

Second line follows from (39). Now (35) can be derived similar to (34), by applying Theorem 2.5 in [Tsybakov \(2009\)](#). \square

Lemma C.2. For $\mu_1, \mu_2 \in \mathbb{R}$, and $\sigma > 0$,

$$\text{KL}(\text{Lap}(\mu_1, \sigma), \text{Lap}(\mu_2, \sigma)) = e^{-\delta} + \delta - 1 \leq \frac{1}{2} \delta^2$$

where $\delta = |\mu_1 - \mu_2|/\sigma$. Let $g(x) = e^{-x} + x - 1$ for $x \geq 0$. Then for $y \geq 0$,

$$g^{-1}(y) \geq \max\{\sqrt{2y}, y\}.$$

Proof of Lemma C.2. From a direction integration, as shown in Appendix A in [Meyer \(2021\)](#),

$$\text{KL}(\text{Lap}(\mu_1, \sigma), \text{Lap}(\mu_2, \sigma)) = e^{-\delta} + \delta - 1 = g(\delta)$$

where $\delta = |\mu_1 - \mu_2|/\sigma$. We can verify with elementary calculus that, for all $y \geq 0$,

$$g(y) < y \text{ and } g(y) \leq \frac{y^2}{2}.$$

Therefore for all $y \geq 0$,

$$g(y) < y \text{ and } g(\sqrt{2y}) \leq y.$$

g is a strictly increasing function on $[0, \infty)$. Therefore,

$$y < g^{-1}(y), \sqrt{2y} \leq g^{-1}(y) \text{ for all } y \geq 0. \quad \square$$

Lemma C.3. Suppose $n \geq 6$. Suppose $q = 1$ or q is even with $q \leq n/3$. Then for $r > 0$, the minimax risk of $B(r, q)$ defined in (31) satisfies

$$n \cdot R_M(B(r, q)) \geq \frac{1}{12}q \min \left\{ r^2, \frac{\sigma^2}{3} \log \frac{en}{8q} \vee \frac{\sigma^2}{36} \log^2 \frac{en}{8q} \right\}$$

Proof of Theorem C.3. We will show a slightly stronger bound:

$$n \cdot R_M(B(r, q)) \geq \frac{1}{12}q \left(r \wedge \sigma g^{-1} \left(\frac{1}{6} \log \frac{en}{8q} \right) \right)^2$$

where $g(x) = e^{-x} + x - 1$ for $x \geq 0$. From this and Theorem C.2, we get the bound in Theorem C.3.

The proof is adapted from that of Theorem 5 in Birge and Massart (2001) for Gaussian error model. We use Fano's lemma from information theory.

Abbreviate $\tau = \log \frac{en}{8q}$.

- Let

$$\mathcal{M}_q = \{S \subseteq [n] : |S| = q\}$$

Here $|S|$ denotes the cardinality of a set S . Consider signals $\beta_S \in \mathbb{R}^n$

$$(\beta_S)_i = \mathbf{1}\{i \in S\}a$$

where $a = r \wedge \sigma g^{-1}(\tau/6)$. As $q \leq n/3$, $\tau = \log \frac{en}{8q}$ should be positive. g is strictly increasing over $x \geq 0$, $\lim_{x \rightarrow \infty} g(x) = \infty$ and so $g^{-1}(\tau/6)$ is well-defined.

We will pick sufficiently separated elements from \mathcal{M}_q to construct signals for Fano's lemma.

- Suppose q is even with $q \leq n/3$. From Lemma 4 Birge and Massart (2001) we can find a subset \mathcal{S} of \mathcal{M}_q such that
 - for any distinct $S, S' \in \mathcal{S}$, $|S \cap S'| < q/2$
 -

$$\log |\mathcal{S}| > \frac{q\tau}{2} \tag{40}$$

Note that when $q = 1$, $\mathcal{S} = \mathcal{M}_q$ satisfies these two requirements.

Denote $\delta(S, S') = |S \cup S'| - |S \cap S'| = |S| + |S'| - 2|S \cap S'|$. For $S, S' \in \mathcal{S}$ we have $\delta(S, S') = 2q - 2|S \cap S'|$. Therefore for distinct $S, S' \in \mathcal{S}$, as $|S \cap S'| < q/2$,

$$q < \delta(S, S') \leq 2q.$$

- Consider the signals $\{\beta_S : S \in \mathcal{S}\}$. For any distinct $S, S' \in \mathcal{S}$
 - From Theorem C.2,

$$\begin{aligned} \text{KL}(\text{Lap}(\beta_S, \sigma), \text{Lap}(\beta_{S'}, \sigma)) &= \delta(S, S') \text{KL}(\text{Lap}(0, \sigma), \text{Lap}(a, \sigma)) \\ &\leq 2q \cdot g(a/\sigma) \end{aligned} \tag{41}$$

where $g(x) = e^{-x} + x - 1$ for $x \geq 0$.

- From Proposition 9 of [Birge and Massart \(2001\)](#) and the KL divergence bound in (41),

$$n \cdot R_M(B(r, q)) \geq \frac{1}{4}qa^2 \left[1 - \left(\frac{2}{3} \vee \frac{2qg(a/\sigma)}{\log |\mathcal{S}|} \right) \right].$$

Applying the bound on $\log |\mathcal{S}|$ from (40),

$$n \cdot R_M(B(r, q)) \geq \frac{1}{4}qa^2 \left[1 - \left(\frac{2}{3} \vee \frac{4g(a/\sigma)}{\tau} \right) \right]$$

By definition of a , $\frac{4g(a/\sigma)}{\tau} \leq \frac{2}{3}$. Therefore

$$n \cdot R_M(B(r, q)) \geq \frac{1}{12}qa^2$$

Plug in the expression for a and then for τ to arrive at the desired bound.

□

C.2. Proof of [Theorem 3.7](#)

Proof of [Theorem 3.7](#). We apply Le Cam's method to derive the lower bound. Define $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^n$ as follows. $\beta_i^{(1)} = \beta_i^{(2)} = 1$ for all $i \in [n-1]$ and $\beta_n^{(1)} = 1 + C_n/4, \beta_n^{(2)} = 1 + C_n/2$. Observe that

$$\frac{1}{n} \|\beta^{(1)} - \beta^{(2)}\|_2^2 = \frac{C_n^2}{16n}.$$

Verify that $\beta^{(1)}, \beta^{(2)} \in \Theta(C_n)$. From equation (15.14) in [Wainwright \(2019\)](#), we can write

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta(C_n)} E \|\hat{\beta} - \beta\|_n^2 \geq \frac{C_n^2}{64n} (1 - \|\mathbb{P}_1 - \mathbb{P}_2\|_{\text{TV}}) \quad (42)$$

where \mathbb{P}_j is the product distribution of y_1, \dots, y_n with $y_i \sim \text{Exp}(\text{mean} = \beta_i^{(j)})$ for $i \in [n]$. We can calculate $\|\mathbb{P}_1 - \mathbb{P}_2\|_{\text{TV}}$ as follows.

$$\begin{aligned} & \|\mathbb{P}_1 - \mathbb{P}_2\|_{\text{TV}} \\ &= \frac{1}{2} \int \left| p_1^{(1)}(x_1) p_2^{(1)}(x_2) \dots p_n^{(1)}(x_n) - p_1^{(2)}(x_1) p_2^{(2)}(x_2) \dots p_n^{(2)}(x_n) \right| dx \\ &= \frac{1}{2} \int p_1^{(1)}(x_1) p_2^{(1)}(x_2) \dots p_{n-1}^{(1)}(x_{n-1}) |p_n^{(1)}(x_n) - p_n^{(2)}(x_n)| dx_1 \dots dx_n \\ &= \frac{1}{2} \int |p_n^{(1)}(x_n) - p_n^{(2)}(x_n)| dx_n \\ &= \frac{1}{4} \end{aligned}$$

Here $p_i^{(j)}$ is the density of the exponential distribution with mean $\beta_i^{(j)}$ for $i \in [n]$. The second line above is true because $p_i^{(1)} = p_i^{(2)}$ for $i \in [n-1]$. The calculation for the last line is given in [Theorem C.4](#). Plugging this back into (42), we get the lower bound

$$\inf_{\hat{\beta}} \sup_{\beta \in \Theta(C_n)} E \|\hat{\beta} - \beta\|_n^2 \geq \frac{3C_n^2}{256n}. \quad \square$$

Lemma C.4. *The total variation distance between two exponential distributions with means β and 2β is $\frac{1}{4}$, for any $\beta > 0$.*

Proof of Theorem C.4. The stated total variation distance is

$$\begin{aligned} & \frac{1}{2} \int_0^\infty \left| \frac{1}{\beta} e^{-x/\beta} - \frac{1}{2\beta} e^{-x/2\beta} \right| dx \\ &= \frac{1}{2} \int_0^\infty |2e^{-2y} - e^{-y}| dy \\ &= \frac{1}{2} \int_0^{\log 2} (2e^{-2y} - e^{-y}) dy + \frac{1}{2} \int_{\log 2}^\infty (e^{-y} - 2e^{-2y}) dy \\ &= \frac{1}{4}. \end{aligned}$$

In the first line, the variable is changed ($x \rightarrow 2\beta y$). \square

Appendix D: Algorithmic details

This section expands on the algorithmic implementation for the MLE trend filter described in [Section 4](#). First, rewrite Equation (2) (substituting x for θ) as

$$\min_{Dx=z} \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1.$$

This is equivalent to (2) but with additional variables. The Lagrangian for this constrained minimization is given by

$$L(x, z, w) = \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1 + w^\top (Dx - z), \quad (43)$$

and the augmented Lagrangian is

$$L_\rho(x, z, w) = \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1 + w^\top (Dx - z) + \frac{\rho}{2} \|Dx - z\|_2^2.$$

The augmented Lagrangian effectively adds a quadratic term that penalizes infeasibility. So for any feasible solution with $Dx = z$, the augmented Lagrangian will be equal to (43). Rather than this form, we instead use the “scaled” form for the augmented Lagrangian, as it makes the update steps a little simpler. Defining $u = w/\rho$, then the augmented Lagrangian becomes

$$L_\rho(x, z, u) = \frac{1}{n} \sum \psi(x_i) - y_i x_i + \lambda \|z\|_1 + \frac{\rho}{2} \|Dx - z + u\|_2^2 - \frac{\rho}{2} \|u\|_2^2.$$

The scaled ADMM algorithm iteratively solves this problem by minimizing over x then z then a dual ascent update on u :

$$\begin{aligned} x &\leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \frac{\rho}{2} \|Dx - z + u\|_2^2, \\ z &\leftarrow \operatorname{argmin}_z \lambda \|z\|_1 + \frac{\rho}{2} \|Dx - z + u\|_2^2, \\ u &\leftarrow u + Dx - z. \end{aligned} \quad (44)$$

The x update involves a matrix inversion which is best avoided when n is large. So we linearize that problem (the x update only) around the current value x^o

$$x \leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \rho (D^\top Dx^o - D^\top z + D^\top u)^\top x + \frac{\mu}{2} \|x - x^o\|_2^2. \quad (45)$$

To include the null space penalty, the changes only impact the x update. Therefore, (44) becomes

$$x \leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \frac{\rho}{2} \|Dx - z + u\|_2^2 + \lambda_2 \|P_{\mathcal{N}}x\|_2,$$

and (45) becomes

$$\begin{aligned} x &\leftarrow \operatorname{argmin}_x \frac{1}{n} \sum \psi(x_i) - y_i x_i + \rho (D^\top Dx^o - D^\top z + D^\top u)^\top x \\ &\quad + \lambda_2 (g(x^o))^\top x + \frac{\mu}{2} \|x - x^o\|_2^2. \end{aligned}$$

where $g(v)$ is a subgradient of the function $v \mapsto \|P_{\mathcal{N}}v\|_2$ given by $g(v) = \frac{P_{\mathcal{N}}v}{\|P_{\mathcal{N}}v\|_2}$ when $P_{\mathcal{N}}v \neq 0$ and $g(v) = 0$ when $P_{\mathcal{N}}v = 0$.

The z -update is easily shown to be given by elementwise soft-thresholding,

$$z_i \leftarrow \operatorname{sign}(z_i) (|z_i| - (Dx - u)_i)_+;$$

and the u -update is simply vector addition. The x -update is potentially more challenging. Note first that the x -update is the same for each i , so we can solve n 1-dimensional problems. The KKT stationarity condition requires

$$\begin{aligned} 0 &= (\psi'(x_i) - y_i) + \rho (D^\top (Dx^o - z + u))_i + \mu(x_i - x_i^o). \\ \implies \psi'(x_i) + \mu x_i &= y_i - \rho (D^\top Dx^o - D^\top z + u)_i + \mu x_i^o. \end{aligned}$$

Therefore, for any loss function as given by ψ , we want to solve $\psi'(x_i) + \mu x_i = b_i$, for each $i \in [n]$. For many functions ψ , the solution has a closed form. The Binomial distribution with $\psi(x) = \log(1 + e^x)$ is a family without a simple solution, though standard root finding methods implemented in low-level languages have no difficulties. To include the nullspace penalty, the x update changes slightly, but the logic is the same.

Appendix E: Degrees of freedom and tuning parameter selection

Here, we provide further details of the tuning parameter selection procedure described in [Section 5](#). If $Y \sim N(\theta^*, \sigma^2)$, a now common method of risk estimation makes use of Stein's Lemma.

Lemma E.1 (Stein's Lemma). *Assume $f(Y)$ is weakly differentiable with essentially bounded weak partial derivatives on \mathbb{R}^n , then*

$$\text{tr Cov}(Y, f(Y)) = E[\langle Y, f(Y) \rangle] = \sigma^2 E \left[\text{tr } Df(Y) \Big|_y \right].$$

The utility of this result comes from examining the decomposition of the mean squared error of $\hat{\theta}(Y)$ as an estimator of θ^* .

$$\begin{aligned} E \left[\|\theta^* - \hat{\theta}(Y)\|_2^2 \right] &= E \left[\|Y - \hat{\theta}(Y)\|_2^2 \right] - n\sigma^2 + 2 \text{tr Cov}(Y, \hat{\theta}(Y)) \\ &= E \left[\|Y - \hat{\theta}(Y)\|_2^2 \right] - n\sigma^2 + 2\sigma^2 E \left[\text{tr } J\hat{\theta}(z) \Big|_Y \right]. \end{aligned}$$

This characterization motivates the definition of degrees-of-freedom for linear predictors ($\text{df} := \frac{1}{\sigma^2} \text{tr } J\hat{\theta}(z) \Big|_y$) ([Efron, 1986](#)), where $\hat{\theta}(y) = Hy$. Using Stein's Lemma, assuming σ^2 is known, we have Stein's Unbiased Risk Estimator

$$\text{SURE}(\hat{\theta}) = \|y - \hat{\theta}\|_2^2 - n\sigma^2 + 2\sigma^2 \text{tr} \left(J\hat{\theta}(z) \Big|_y \right),$$

which satisfies $E[\text{SURE}(\hat{\theta})] = E\|\theta^* - \hat{\theta}(Y)\|_2^2$. Note that this is the risk for estimating the n -dimensional parameter θ^* . The following result generalizes this idea to certain continuous exponential families.

Lemma E.2 (Generalized Stein Lemma; [Eldar, 2009](#)). *Assume $\hat{\theta}(y)$ is weakly differentiable in y with essentially bounded weak partial derivatives on \mathbb{R}^n . Let Y be distributed according to a natural exponential family and assume that the base measure h is weakly differentiable. Then,*

$$E \left[\theta^{*\top} \hat{\theta}(Y) \right] = -E \left[\left\langle \frac{\nabla h(Y)}{h(Y)}, \hat{\theta}(Y) \right\rangle + \text{tr } J\hat{\theta}(y) \Big|_Y \right].$$

Note that $\nabla h(Y)$ here means the vector $[d/dy \ h(y)|_{y_i}]$ and $h(Y)$ means the vector $[h(y_i)]$.

Therefore we define the Generalized SURE ([Eldar, 2009](#)) along the lines of the multivariate Gaussian case.

Lemma E.3. *Assume h is weakly differentiable, $\hat{\theta}(y)$ is weakly differentiable with essentially bounded partial derivatives. Then*

$$\text{SURE}(\hat{\theta}) = \|\hat{\theta}(y)\|_2^2 + 2 \left\langle \frac{\nabla h(y)}{h(y)}, \hat{\theta}(y) \right\rangle + 2 \text{tr} \left(J\hat{\theta}(z) \Big|_y \right) + \frac{1}{h(y)} \text{tr} \frac{\partial^2 h(z)}{\partial z^2} \Big|_y$$

is an unbiased estimator for the MSE of an estimator $\hat{\theta}(Y)$ of θ : $E\|\hat{\theta}(Y) - \theta\|_2^2$.

Proof. We have

$$E \left[\|f(Y) - \theta(\beta)\|_2^2 \right] = E \left[\|f(Y)\|_2^2 \right] + E \left[\|\theta\|_2^2 \right] - 2E \left[\langle \theta(\beta), f(Y) \rangle \right].$$

Now, the first term is a function of the data only, and to the last term, we simply apply [Theorem E.2](#). For the second term,

$$\begin{aligned} E \left[\|\theta\|_2^2 \right] &= E \left[\langle \theta, \theta \rangle \right] = -E \left[\left\langle \frac{\nabla h(Y)}{h(Y)}, \theta \right\rangle \right] \\ &= E \left[\left\langle \frac{\nabla h(Y)}{h(Y)}, \frac{\nabla h(Y)}{h(Y)} \right\rangle \right] + E \left[\text{tr} \frac{\partial}{\partial y} \frac{\nabla h(y)}{h(y)} \Big|_Y \right] \\ &= E \left[\frac{\|\nabla h(Y)\|_2^2}{h(Y)^2} \right] + E \left[\text{tr} \frac{\|\nabla h(Y)\|_2^2 + h(Y) \partial^2 / \partial y^2 h(y)}{h(Y)^2} \Big|_Y \right] \\ &= E \left[\frac{1}{h(Y)} \text{tr} \frac{\partial^2 h(y)}{\partial y^2} \Big|_Y \right], \end{aligned}$$

by applying [Theorem E.2](#) twice along with the quotient rule. \square

For the MLE trend filter with a continuous exponential family, we have that h is weakly differentiable. Furthermore, Theorem 1 in [Vaiteer et al. \(2017\)](#) applies. Thus the special case in [Theorem 5.1](#) holds generally for this estimator.

However, we would prefer to estimate the Kullback–Leibler Divergence between the density under $\theta = \hat{\theta}(y)$ and that under $\theta = \theta^*$. For exponential families,

$$E \left[\text{KL}(\hat{\theta}(Y) \parallel \theta^*) \right] = E \left[\left\langle \hat{\theta}(Y) - \theta^*, \hat{\beta}(Y) \right\rangle + \psi(\theta^*) - \psi(\hat{\theta}(Y)) \right],$$

and, an application of [Theorem E.2](#) provides an unbiased estimator of this quantity. The result is given in [Theorem 5.1](#) in the main body.

Finally, we conclude this section with the proof of [Theorem 5.2](#).

Proof of Theorem 5.2. The proof follows from [Vaiteer et al. \(2017, Theorem 2\)](#). We have

$$\begin{aligned} X_T &= P_{\mathcal{N}(\check{D})} \\ \nabla^2 F_0(\hat{\mu}(y), y) &= \text{diag} \left(\psi''(\hat{\theta}) \right) \\ \mathfrak{A}_\beta &= 0 \\ \nabla_{\mathcal{M}}^2 J \left(\hat{\beta}(y) \right) &= \lambda_2 P_{\mathcal{N}} \\ D(\nabla F_0)(\hat{\mu}(y), y) &= \text{diag} \left(\psi''(\hat{\theta}) \right). \square \end{aligned}$$

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

DJM was partially supported by the National Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2021-02618.

References

- BABY, D. and WANG, Y.-X. (2021). Optimal Dynamic Regret in Exp-Concave Online Learning. In *Proceedings of Thirty Fourth Conference on Learning Theory* (M. BELKIN and S. KPOTFE, eds.). *Proceedings of Machine Learning Research* **134** 359–409.
- BARBERO, A. and SRA, S. (2018). Modular Proximal Optimization for Multidimensional Total-variation Regularization. *Journal of Machine Learning Research* **19** 2232–2313.
- BASSETT, R. and SHARPNACK, J. (2019). Fused Density Estimation: Theory and Methods. *Journal of Royal Statistical Society, Series B* **81** 839–860.
- BAUSCHKE, H. H. and BORWEIN, J. M. (1997). Legendre Functions and the Method of Random Bregman Projections. *Journal of Convex Analysis* **4** 27–67.
- BIRGE, L. and MASSART, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3** 203–268.
- BROWN, L. D. (1986). *Fundamentals of statistical exponential families with applications in statistical decision theory. Lecture Notes-Monograph Series* **9**. Institute of Mathematical Statistics.
- BROWN, L. D., CAI, T. T. and ZHOU, H. H. (2010). Nonparametric regression in exponential families. *The Annals of Statistics* **38** 2005–2046.
- CHATTERJEE, S. and GOSWAMI, S. (2021). New Risk Bounds for 2D Total Variation Denoising. *IEEE Transactions on Information Theory* **67** 4060–4091.
- CONDAT, L. (2013). A Direct Algorithm for 1-D Total Variation Denoising. *IEEE Signal Processing Letters* **20** 1054–1057.
- DELEDALLE, C.-A. (2017). Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Electronic Journal of Statistics* **11** 3141–3164.
- DÜMBGEN, L. and KOVAC, A. (2009). Extensions of Smoothing via Taut Strings. *Electronic Journal of Statistics* **3** 41–75.
- EFRON, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association* **81** 461–470.
- ELDAR, Y. C. (2009). Generalized SURE for Exponential Families: Applications to Regularization. *IEEE Transactions on Signal Processing* **57** 471–481.
- GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2020). Adaptive Risk Bounds in Univariate Total Variation Denoising and Trend Filtering. *Annals of Statistics* **48** 205–229.
- HANSEN, J., SATO, M. and RUEDY, R. (2012). Perception of climate change. *Proceedings of the National Academy of Sciences* **109** E2415–E2423.

- HARCHAOUI, Z. and LEVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* **105** 1480–1493.
- HUNTINGFORD, C., JONES, P. D., LIVINA, V. N., LENTON, T. M. and COX, P. M. (2013). No increase in global temperature variability despite changing regional patterns. *Nature* **500** 327–330.
- HÜTTER, J.-C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In *29th Annual Conference on Learning Theory* (V. FELDMAN, A. RAKHLIN and O. SHAMIR, eds.). *Proceedings of Machine Learning Research* **49** 1115–1146.
- JOHNSON, N. (2013). A dynamic programming algorithm for the fused lasso and L_0 -segmentation. *Journal of Computational and Graphical Statistics* **22** 246–260.
- KAKADE, S., SHAMIR, O., SRIDHARAN, K. and TEWARI, A. (2010). Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. TEH and M. TITTERINGTON, eds.). *Proceedings of Machine Learning Research* **9** 381–388.
- KHODADADI, A. and McDONALD, D. J. (2019). Algorithms for Estimating Trends in Global Temperature Volatility. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence* (P. V. HENTENRYCK and Z.-H. ZHOU, eds.). *Association for the Advancement of Artificial Intelligence* **33** 614–621.
- KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). ℓ_1 Trend Filtering. *SIAM Review* **51** 339–360.
- LIN, K., SHARPNACK, J. L., RINALDO, A. and TIBSHIRANI, R. J. (2017). A Sharp Error Analysis for the Fused Lasso, with Application to Approximate Changepoint Screening. In *Advances in Neural Information Processing Systems* (I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN and R. GARNETT, eds.) **30**. Curran Associates, Inc.
- MADRID PADILLA, O. H. and CHATTERJEE, S. (2021). Risk Bounds for Quantile Trend Filtering. *Biometrika* forthcoming.
- MADRID PADILLA, O. H., SHARPNACK, J., SCOTT, J. G. and TIBSHIRANI, R. J. (2018). The DFS Fused Lasso: Linear-Time Denoising over General Graphs. *Journal of Machine Learning Research* **18** 1–36.
- MADRID PADILLA, O. H., SHARPNACK, J., CHEN, Y. and WITTEN, D. M. (2020). Adaptive nonparametric regression with the K-nearest neighbour fused lasso. *Biometrika* **107** 293–310.
- MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Annals of Statistics* **25** 387–413.
- MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, Boca Raton, FL.
- MEYER, G. P. (2021). An Alternative Probabilistic Interpretation of the Huber Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5261–5269.
- NUÑO, M., GARCÍA, Y., RAJASEKAR, G., PINHEIRO, D. and SCHMIDT, A. J. (2021). COVID-19 hospitalizations in five California hospitals: A retrospective

- cohort study. *BMC Infectious Diseases* **21** 938. <https://doi.org/10.1186/s12879-021-06640-4>
- ORTELLI, F. and VAN DE GEER, S. (2020). Adaptive Rates for Total Variation Image Denoising. *Journal of Machine Learning Research* **247** 1–38.
- ORTELLI, F. and VAN DE GEER, S. (2021). Prediction bounds for higher order total variation regularized least squares. *The Annals of Statistics* **49** 2755–2773. <https://doi.org/10.1214/21-AOS2054>
- POLI, P., HERBACH, H., DEE, D. P., BERRISFORD, P., SIMMONS, A. J., VITART, F., LALOYAU, P., TAN, D. G. H., PEUBEY, C., THÉPAUT, J.-N., TRÉMOLET, Y., HÓLM, E. V., BONAVITA, M., ISAKSEN, L. and FISHER, M. (2016). ERA-20C: An Atmospheric Reanalysis of the Twentieth Century. *Journal of Climate* **29** 4083–4097.
- PRASAD, A., SUGGALA, A. S., BALAKRISHNAN, S. and RAVIKUMAR, P. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B* **82** 601–627.
- RAMDAS, A. and TIBSHIRANI, R. J. (2016). Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics* **25** 839–858.
- RINALDO, A. (2009). Properties and refinements of the fused lasso. *Annals of Statistics* **37** 2922–2952.
- RUDIN, L. I., OSHER, S. and FATERNI, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* **60** 259–268.
- SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers. In *Advances in Neural Information Processing Systems* (D. LEE, M. SUGIYAMA, U. LUXBURG, I. GUYON and R. GARNETT, eds.) **29**. Curran Associates, Inc.
- SADHANALA, V., WANG, Y.-X., SHARPNACK, J. L. and TIBSHIRANI, R. J. (2017). Higher-Order Total Variation Classes on Grids: Minimax Theory and Trend Filtering Methods. In *Advances in Neural Information Processing Systems*, **30** 5800–5810.
- SADHANALA, V., WANG, Y.-X., HU, A. and TIBSHIRANI, R. (2021). Multivariate Trend Filtering on Lattice Data.
- STEIDL, G., DIDAS, S. and NEUMANN, J. (2006). Splines in higher order TV regularization. *International Journal of Computer Vision* **70** 214–255.
- TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics* **42** 285–323.
- TIBSHIRANI, R. J. (2022). Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems. *Foundations and Trends in Machine Learning* **15** 694–846.
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67** 91–108.
- TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.
- VAITER, S., DELEDALLE, C., FADILI, J., PEYRÉ, G. and DOSSAL, C. (2017).

- The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics* **69** 791–832.
- VAN DE GEER, S. (2020). Logistic regression with total variation regularization. *Transactions of A. Razmadze Mathematical Institute* **174** 217 - 233.
- VASSEUR, D. A., DELONG, J. P., GILBERT, B., GREIG, H. S., HARLEY, C. D. G., MCCANN, K. S., SAVAGE, V., TUNNEY, T. D. and O’CONNOR, M. I. (2014). Increased temperature variation poses a greater risk to species than climate warming. *Proceedings of the Royal Society of London B: Biological Sciences* **281**. <https://doi.org/10.1098/rspb.2013.2612>
- VERSHYNIN, R. (2018). *High-Dimensional Probability*. Cambridge University Press, Cambridge, UK.
- VU, V. Q. (2018). Computational Sufficiency, Reflection Groups, and Generalized Lasso Penalties. <https://doi.org/10.48550/ARXIV.1809.02857>
- WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* **1** 1-305.
- WANG, Y.-X., SHARPNAK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend Filtering on Graphs. *Journal of Machine Learning Research* **17** 1-41.
- ZHANG, T. and CHATTERJEE, S. (2023). Element-wise estimation error of generalized Fused Lasso. *Bernoulli* **29** 2691–2718.