Supplementary Appendix for "Multivariate Trend Filtering for Lattice Data"

Veeranjaneyulu Sadhanala[a] Yu-Xiang Wang[b] Addison J. Hu[c] Ryan J. Tibshirani[c]

[a]Google [b]University of California Santa Barbara [c]Carnegie Mellon University

A Proofs

A.1 Proof of Proposition 1

Abbreviating $D = D^{(k+1)}_{n,d}$, the null space of D is the number of its nonzero singular values or equivalently, the number of nonzero eigenvalues of $D^\mathsf{T} D$. Following from (8), and abbreviating $Q = D^{(k+1)}_N$ and $I = I_N$,

$$D^\mathsf{T} D = Q^\mathsf{T} Q \otimes I \otimes \cdots \otimes I \;+\; I \otimes Q^\mathsf{T} Q \otimes \cdots \otimes I \;+\; \ldots \;+\; I \otimes I \otimes \cdots \otimes Q^\mathsf{T} Q,$$

the Kronecker sum of $Q^\mathsf{T} Q$ with itself, a total of d times. Using a standard fact about Kronecker sums, if we denote by ρ_i, $i = 1, \ldots, N$ the eigenvalues of $Q^\mathsf{T} Q$ then

$$\rho_{i_1} + \rho_{i_2} + \cdots + \rho_{i_d}, \quad i_1, \ldots, i_d \in \{1, \ldots, N\}$$

are the eigenvalues of $D^\mathsf{T} D$. By counting the multiplicity of the zero eigenvalue, we arrive at a nullity for D of $(k+1)^d$. It is straightforward to check that the vectors specified in the proposition, given by evaluations of polynomials of max degree k, are in the null space, and that these are linearly independent, which completes the proof. □

A.2 Proof of Proposition 2

Let us define

$$B^{(k+1)}_N = \begin{bmatrix} C^{(k+1)}_N \\ D^{(k+1)}_N \end{bmatrix} \in \mathbb{R}^{N \times N},$$

where the first $k+1$ rows are given by a matrix $C^{(k+1)}_N \in \mathbb{R}^{(k+1) \times N}$ that completes the row space, as in Lemma 2 of Wang et al. (2014), or Section 6.2 of Tibshirani (2020). Now, again by Lemma 2 of Wang et al. (2014), or Section 6.3 of Tibshirani (2020),

$$\big(H^{(k+1)}_N\big)^{-1} = \frac{1}{k!} B^{(k+1)}_N \tag{S.1}$$

where $H^{(k+1)}_N \in \mathbb{R}^{N \times N}$ is the falling factorial basis matrix of order k, which has elements

$$\big[H^{(k+1)}_N\big]_{ij} = h^k_{N,j}(i/N), \quad i, j = 1, \ldots, N,$$

with $h^k_{N,i}$, $i = 1, \ldots, N$ denoting the falling factorial functions in (10) with respect to design points $1/N, 2/N, \ldots, 1$.

We now transform variables in (7) by defining

$$\theta = \Big(H^{(k+1)}_N \otimes \cdots \otimes H^{(k+1)}_N\Big)\alpha,$$

and using (S.1), this turns (7) into an equivalent basis form,

$$\underset{\alpha \in \mathbb{R}^n}{\text{minimize}} \;\; \frac{1}{2}\Big\|y - \Big(H^{(k+1)}_N \otimes \cdots \otimes H^{(k+1)}_N\Big)\alpha\Big\|_2^2 + \lambda k! \left\| \begin{bmatrix} I^0_N \otimes H^{(k+1)}_N \otimes \cdots \otimes H^{(k+1)}_N \\ H^{(k+1)}_N \otimes I^0_N \otimes \cdots \otimes H^{(k+1)}_N \\ \vdots \\ H^{(k+1)}_N \otimes H^{(k+1)}_N \otimes \cdots \otimes I^0_N \end{bmatrix} \alpha \right\|_1,$$

where $I_N^0 \in \mathbb{R}^{(N-k-1)\times N}$ denotes the last $N-k-1$ rows of the identity I_N. We can rewrite the problem once more by parametrizing the evaluations according to f in (11), which we claim yields (12). The equivalence between loss terms in the above problem and (12) is immediate (by definition of f); to see the equivalence between penalty terms, it can be directly checked that

$$k!\Big(I_N^0 \otimes H_N^{(k+1)} \otimes \cdots \otimes H_N^{(k+1)}\Big)\alpha$$

contains the differences of the function $\partial^k f/\partial x_1^k$ over all pairs of grid positions that are adjacent in the x_1 direction, where f is as in (11). This, combined with the fact that $\partial^k f/\partial x_1^k$ is constant in between lattice positions, means that

$$k!\left\|\Big(I_N^0 \otimes H_N^{(k+1)} \otimes \cdots \otimes H_N^{(k+1)}\Big)\alpha\right\|_1 = \sum_{x_{-1}} \mathrm{TV}\left(\frac{\partial^k f(\cdot, x_{-1})}{\partial x_1^k}\right),$$

the total variation of $\partial^k f/\partial x_1^k$ added up over all slices of the lattice $Z_{n,d}$ in the x_1 direction. Similar arguments apply to the penalty terms corresponding to dimensions $j = 2, \ldots, d$, and this completes the proof. □

A.3 Proof of Theorem 1

Denote by $f_\epsilon = \eta_\epsilon * f$ the mollified version of f, where $\eta_\epsilon(x) = \epsilon^{-d}\eta(x/\epsilon)$, and $\eta : \mathbb{R}^d \to \mathbb{R}$ is the standard mollifier, defined by

$$\eta(x) = \begin{cases} c \exp\left(\dfrac{1}{\|x\|_2^2 - 1}\right) & \text{if } \|x\|_2 \le 1, \\ 0 & \text{else,} \end{cases}$$

and $c > 0$ is a normalization constant so that η integrates to 1. By construction, for any $\epsilon > 0$, we have $f_\epsilon \in C^\infty(U)$. Now from (18), simply exchanging the sum over absolute partial derivatives and the integral, we have

$$\begin{aligned} \mathrm{TV}(f_\epsilon; U) &= \sum_{j=1}^d \int_U \left|\frac{\partial f_\epsilon(x)}{\partial x_j}\right| dx \\ &= \sum_{j=1}^d \int_{U_{-j}} \int_{I_{x_{-j}}} \left|\frac{\partial f_\epsilon(x_j, x_{-j})}{\partial x_j}\right| dx_j \, dx_{-j} \\ &= \sum_{j=1}^d \int_{U_{-j}} \mathrm{TV}\big(f_\epsilon(\cdot, x_{-j}); I_{x_{-j}}\big)\, dx_{-j}, \end{aligned}$$

where in the last line we applied the representation of TV for smooth functions in (18), but to the univariate function $x_j \mapsto f_\epsilon(x_j, x_{-j})$, for fixed x_{-j}. Recalling standard results on approximation of BV functions by smooth functions (see, for example, Theorem 5.22 in Evans and Gariepy (2015)), by sending $\epsilon \to 0$, we have that the left-most and right-most sides of the previous display approach those in (20), completing the proof.

A.4 Proof of Theorem 3

Abbreviate $N' = N - k - 1$. Let β_i, u_i, v_i be a triplet of nonzero singular value, left singular vector, and right singular vector of $D_{N,1}^{(k+1)}$, for $i \in [N']$ and let p_j, $j \in [k+1]$ form an orthogonal basis for the null space of $D_{N,1}^{(k+1)}$. From Lemma S.1 it suffices to show incoherence of u_i, v_i, $i \in [N']$, and p_i, $i \in [k+1]$. Incoherence of u_i, $i \in [N']$ and v_i, $i \in [N']$is established in Sadhanala et al. (2017) . Incoherence of p_i, $i \in [k+1]$ may be seen by choosing, e.g., these vectors to be the discrete Legendre orthogonal polynomials as in Neuman and Schonbach (1974). Applying Lemma S.1, we can see that $D_{n,d}^{(k+1)}$ satisfies the incoherence property, as defined in Theorem 2, say with a constant μ.

From the incoherence property and Theorem 2, the KTF estimator $\hat\theta$, satisfies

$$\frac{1}{n}\|\hat\theta - \theta_0\|_2^2 = O_{\mathbb{P}}\left(\frac{\kappa}{n} + \frac{|I|}{n} + \frac{\mu}{n}\sqrt{\frac{\log n}{n}\sum_{i \in [N]^d \setminus (I \cup [k+1]^d)} \frac{1}{\xi_i^2}} \cdot \|\Delta\theta_0\|_1\right), \tag{S.2}$$

where we abbreviate $\Delta = D_{n,d}^{(k+1)}$, $\xi_i, i \in [N]^d$ are eigenvalues of $\Delta^\mathsf{T}\Delta$ with $\xi_i = 0$ for $i \in [k+1]^d$. We reindexed the eigenvalues so that they correspond to grid positions.

Recall the shorthand $s = (k+1)/d$. For $s \le 1/2$, set $I = [k+2]^d \setminus [k+1]^d$. From Lemma S.6,

$$\sum_{i \in [N]^d \setminus (I \cup [k+1])^d} \frac{1}{\xi_i^2} \le c \begin{cases} n & s < 1/2 \\ n \log n & s = 1/2. \end{cases}$$

Plugging this into (S.2) gives the desired bounds when $s \le 1/2$:

$$\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2 = \begin{cases} O_{\mathbb{P}}\Big(\frac{(k+2)^d}{n} + \|\Delta\theta_0\|_1 \sqrt{\log n}\Big) & s < 1/2 \\ O_{\mathbb{P}}\Big(\frac{(k+2)^d}{n} + \|\Delta\theta_0\|_1 \log n\Big) & s = 1/2. \end{cases}$$

Now consider the case $s > 1/2$. Set $I = \{i \in [N]^d : \|(i-k-2)_+\|_2 < r\} \setminus [k+1]^d$ for an $r \in [1, \sqrt{d}N]$ to be chosen later. $|I| \le (r+k+1)^d$ because $I \subseteq \big[\lfloor r \rfloor + k + 1\big]^d$. Lemma S.6 shows that for a constant $c > 0$ depending only on k, d,

$$\sum_{i \in [N]^d \setminus (I \cup [k+1]^d)} \frac{1}{\xi_i^2} \le c n^{2s} / r^{d(2s-1)}.$$

Plug this bound in (S.2) and in order to minimize the resulting bound, choose r to balance

$$(r+k+1)^d \quad \text{with} \quad \frac{C_n}{\sqrt{n}} \sqrt{n^{2s}/r^{d(2s-1)} \log n}.$$

This leads us to take

$$(r+k+1)^d \asymp (C_n \sqrt{\log n})^{\frac{2}{2s+1}} n^{\frac{2s-1}{2s+1}}$$

when $C_n \sqrt{\log n}/n = O_{\mathbb{P}}(1)$ and $r = 1$ otherwise. With this choice (S.2) gives the desired bound for $s > 1/2$. This completes the proof. □

A.5 Incoherence of Kronecker product type operators

Let

$$\Delta = \begin{bmatrix} D \otimes I \otimes \cdots \otimes I \\ I \otimes D \otimes \cdots \otimes I \\ \vdots \\ I \otimes I \otimes \cdots \otimes D \end{bmatrix} \tag{S.3}$$

where each Kronecker product has d terms. With $D = D^{(k+1)}_{N,1}, I = I_N$, we get the KTF penalty operator $\Delta = D^{(k+1)}_{n,d}$.

Lemma S.1. *Let Δ be as defined in (S.3) for a matrix $D \in \mathbb{R}^{N' \times N}$ with $N' \le N$. Let $\gamma_i, u_i, v_i, i \in [N]$ denote the singular values of D, its left and right singular vectors. Note that $\gamma_i = 0, u_i = 0, v_i \in \mathrm{null}(D)$ for $i \in [p]$ where $p = \mathrm{nullity}(D)$. If these singular vectors are incoherent, that is $\|v_i\|_\infty \le \mu/\sqrt{N}, \|u_i\|_\infty \le \mu/\sqrt{N'}$ for a constant $\mu \ge 1$, then the left singular vectors ν of Δ are incoherent with a constant μ^d, that is, $\|\nu\|_\infty \le \mu^d/\sqrt{N^{d-1}N'}$.*

Note that $p = k+1$ when Δ is the KTF penalty operator with $D = D^{(k+1)}_{N,1}$.

Proof of Lemma S.1. Abbreviate $\rho_i = \gamma_i^2$ for $i \in [N]$. We are looking for a total of $N^d - p^d$ eigenvectors for $\Delta\Delta^\mathsf{T}$. Assume for exposition that $d = 3$. For any $(i,j,k) \in [N]^d \setminus [p]^d$ (where $\setminus$ is the set difference operator), the vectors

$$\nu_{i,j,k} := \frac{1}{\sqrt{\rho_i + \rho_j + \rho_k}} \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix} \tag{S.4}$$

are eigenvectors of $\Delta\Delta^\mathsf{T}$ as verified below.

$$\Delta\Delta^\mathsf{T} \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix} = \Delta \ \big(\gamma_i^2 + \gamma_j^2 + \gamma_k^2\big) \, v_i \otimes v_j \otimes v_k \tag{S.5}$$

$$= (\rho_i + \rho_j + \rho_k) \begin{bmatrix} \gamma_i \cdot u_i \otimes v_j \otimes v_k \\ \gamma_j \cdot v_i \otimes u_j \otimes v_k \\ \gamma_k \cdot v_i \otimes v_j \otimes u_k \end{bmatrix}$$

We see all $N^d - p^d$ eigenvectors of $\Delta\Delta^\mathsf{T}$ here. Notice that $\|z_{i,j,k}\|_2 = 1$ and the incoherence is readily available given that the left and right singular vectors of D are incoherent.

For general d, these $N^d - p^d$ eigenvectors are given by

$$\nu_{i_1,i_2,.,i_d} = \frac{1}{\sqrt{\sum_{j=1}^d \rho_{i_j}}} \begin{bmatrix} \gamma_{i_1} \cdot u_{i_1} \otimes v_{i_2} \otimes \dots v_{i_d} \\ \gamma_{i_2} \cdot v_{i_1} \otimes u_{i_2} \otimes \dots v_{i_d} \\ \vdots \\ \gamma_{i_d} \cdot v_{i_1} \otimes v_{i_2} \otimes \dots u_{i_d} \end{bmatrix} \tag{S.6}$$

with eigenvalues $\sum_{j=1}^d \rho_{i_j}$ and are easily seen to be incoherent. □

A.6 Upper bound for continuous KTV class

Recalling the continuous analog of KTF penalty from (12), define the class

$$\mathrm{KTV}^k_{n,d}(C) = \left\{ f : \sum_{j=1}^d \sum_{x_{-j}} \mathrm{TV}\left(\frac{\partial^k f(\cdot, x_{-j})}{\partial x_j^k} \right) \leq C \right\}$$

for $C > 0$. If the true signal θ_0 on the grid is an evaluation of a function $f \in \mathrm{KTV}^d_k(C)$, the rates in Theorem 3 hold with C_n replaced by C, due to the following result.

Lemma S.2. *Let $C > 0$ and let $d \geq 1, k \geq 0$ be integers. For all $f \in \mathrm{KTV}^k_{n,d}(C)$, if $\theta_f \in \mathbb{R}^n$ is the evaluation of f on the grid points $Z_{n,d}$, then*

$$\|D^{(k+1)}_{n,d} \theta_f\|_1 \leq c_1 C$$

for a constant c_1 that depends only on k and d.

Proof of Lemma S.2. Let f be an arbitrary function from $\mathrm{KTV}^k_{n,d}(C)$. Pick a $j \in [N]$ and an x_{-j} and consider the function $\phi(\cdot) = f(\cdot, x_{-j})$ (f with all but its j argument fixed to elements of x_{-j} appropriately in order). From Theorem 1 in Mammen (1991) and its proof, there exists a spline $\tilde{\phi}$ such that

$$\tilde{\phi}(i/N) = \phi(i/N), \quad i \in [N]$$
$$\mathrm{TV}(\tilde{\phi}^{(k)}) \leq \mathrm{TV}(\phi^{(k)})$$

Let $t_1, \dots, t_L$ be the knots of $\tilde{\phi}$, which are not necessarily in the set of input points. Because it is a spline, $\tilde{\phi}$ can be written as the sum of a polynomial and a linear combination of kth degree truncated power basis functions $g_t : x \mapsto (x-t)^k_+/k!$

$$\tilde{\phi}(u) = p(u) + \sum_{\ell=1}^L \beta_\ell g_{t_\ell}(u), \quad u \in [0,1]$$

where p is a polynomial of degree $\leq k$ and $\beta_\ell \in \mathbb{R}, \ell \in [L]$. Let $D^{(k+1)}_{1\mathrm{d}} = D^{(k+1)}_{N,1}$. Now

$$\left\| D^{(k+1)}_{1\mathrm{d}} \begin{bmatrix} \phi(1/N) \\ \vdots \\ \phi(N/N) \end{bmatrix} \right\|_1 = \left\| D^{(k+1)}_{1\mathrm{d}} \begin{bmatrix} \tilde{\phi}(1/N) \\ \vdots \\ \tilde{\phi}(N/N) \end{bmatrix} \right\|_1$$
$$= \left\| D^{(k+1)}_{1\mathrm{d}} \begin{bmatrix} p(1/N) \\ \vdots \\ p(N/N) \end{bmatrix} + \sum_{\ell=1}^L \beta_\ell \cdot D^{(k+1)}_{1\mathrm{d}} \begin{bmatrix} g_{t_\ell}(1/N) \\ \vdots \\ g_{t_\ell}(N/N) \end{bmatrix} \right\|_1$$

$$\leq \sum_{\ell=1}^{L} |\beta_\ell| \|D_{1\mathrm{d}}^{(k+1)} G_\ell^{(k)}\|_1 \tag{S.7}$$

where the vector $G_\ell^{(k)}$ is the evaluation of g_{t_ℓ} on $1/N, \ldots, N/N$, that is $(G_\ell^{(k)})_i = g_{t_\ell}(i/N), i \in [N]$. Here we used the fact that $D_{1\mathrm{d}}^{(k+1)}$ times the evaluations of a polynomial at the input points $1/N, \ldots, N/N$ is 0.

The terms in (S.7) can be bound as follows. For $\ell \in [L]$, let $i_\ell = \max_{i \in [N]}\{i/N \leq t_\ell\}$, that is, let i_ℓ/N be the largest input point that is not greater the knot t_ℓ. For any vector $v \in \mathbb{R}^N$, and $i \in [N-k-1]$, $(D_{1\mathrm{d}}^{(k+1)} v)_i = 0$ if $(v_i, \ldots, v_{i+k})$ is the evaluation of a polynomial at $i/N, \ldots, (i+k+1)/N$. g_{t_ℓ} is a polynomial on $[0, t_\ell]$ and on $[t_\ell, 1]$ for $\ell \in [L]$. Therefore, $D_{1\mathrm{d}}^{(k+1)} G_\ell^{(k)}$ is nonzero in at most $k+1$ elements. Letting A_i denote the ith row of matrix A, we can write

$$\begin{aligned}
\|D_{1\mathrm{d}}^{(k+1)} G_\ell^{(k)}\|_1 &= \sum_{i=1}^{N-k-1} \left| \left(D_{1\mathrm{d}}^{(k+1)}\right)_i G_\ell^{(k)} \right| \\
&= \sum_{i=(i_\ell-k)\vee 1}^{i_\ell} \left| \left(D_{1\mathrm{d}}^{(k+1)}\right)_i G_\ell^{(k)} \right| \\
&\leq \sum_{i=(i_\ell-k)\vee 1}^{i_\ell} \left\| \left(D_{1\mathrm{d}}^{(k+1)}\right)_i \right\|_1 g_{t_\ell}\left(\frac{i_\ell + k + 1}{N}\right) \\
&\leq \sum_{i=(i_\ell-k)\vee 1}^{i_\ell} \left\| \left(D_{1\mathrm{d}}^{(k+1)}\right)_i \right\|_1 \left(\frac{k+1}{N}\right)^k \frac{1}{k!} \\
&\leq (k+1) \cdot \left(\frac{k+1}{N}\right)^k \frac{1}{k!} \max_{i \in [N-k-1]} \left\| \left(D_{1\mathrm{d}}^{(k+1)}\right)_i \right\|_1 \\
&= (k+1) \cdot \left(\frac{k+1}{N}\right)^k \frac{1}{k!} \cdot 2^{k+1} N^k \\
&= b_k
\end{aligned}$$

where b_k is a constant depending only on k. Plugging this upper bound in (S.7),

$$\left\| D_{1\mathrm{d}}^{(k+1)} \begin{bmatrix} \phi(1/N) \\ \vdots \\ \phi(N/N) \end{bmatrix} \right\|_1 \leq b_k \sum_{\ell=1}^{L} |\beta_\ell| = b_k \mathrm{TV}(\tilde{\phi}^{(k)}) \leq b_k \mathrm{TV}(\phi^{(k)}).$$

This means,

$$\begin{aligned}
\|D_{n,d}^{(k+1)} \theta_f\|_1 &= \sum_{j=1}^{d} \sum_{x_{-j}} \left\| D_{1\mathrm{d}}^{(k+1)} \begin{bmatrix} f(1/N, x_{-j}) \\ \vdots \\ f(N/N, , x_{-j}) \end{bmatrix} \right\|_1 \\
&\leq \sum_{j=1}^{d} \sum_{x_{-j}} b_k \mathrm{TV}\left(\frac{\partial^k f(\cdot, x_{-j})}{\partial x_j^k}\right) \\
&\leq b_k \cdot C.
\end{aligned}$$

This completes the proof. □

A.7 Proof of Theorem 4

Here and henceforth, we use the notation $B_p(r) = \{x : \|x\|_p \leq r\}$ for the ℓ_p ball of radius r, where $p, r > 0$ (and the ambient dimension will be determined based on the context).

Lemma S.3 (Lemma 7 in Sadhanala et al. (2016)). *Let $\mathcal{T}(r) = \{\theta \in \mathbb{R}^n : \|D\theta\|_1 \leq r\}$ for a matrix D and $r > 0$. Recall that $\|D\|_{1,\infty} = \max_{i\in[n]} \|D_i\|_1$ where D_i is the ith column of D. Then for any $r > 0$, it holds that $B_1(r/\|D\|_{1,\infty}) \subseteq \mathcal{T}(r)$.*

From Lemma S.3 and the fact that $\|D^{(k+1)}_{n,d}\|_{1,\infty} = 2^{k+1}d$

$$B_1(r/(2^{k+1}d)) \subseteq \mathcal{T}^k_{n,d}(r). \tag{S.8}$$

for any $r > 0$, and integers $d \geq 1, k \geq 0$.

To prove Theorem 4 we will use the following result from Birge and Massart (2001), which gives a lower bound for the risk in a normal means problem, over ℓ_p balls. We state the result in our notation.

Lemma S.4 (Proposition 5 of Birge and Massart (2001)). *Assume i.i.d. observations $y_i \sim N(\theta_{0,i}, \sigma^2)$, $i = 1, \dots, n$, and $n \geq 2$. Then the minimax risk over the ℓ_p ball $B_p(r_n)$, where $0 < p < 2$, satisfies*

$$n \cdot R(B_p(r_n)) \geq c \cdot \begin{cases} \sigma^{2-p} r_n^p \left[1 + \log\left(\frac{\sigma^p n}{r_n^p}\right)\right]^{1-p/2} & \text{if } \sigma\sqrt{\log n} \leq r_n \leq \sigma n^{1/p}/\sqrt{\rho_p} \\ r_n^2 & \text{if } r_n < \sigma\sqrt{\log n} \\ \sigma^2 n/\rho_p & \text{if } r_n > \sigma n^{1/p}/\sqrt{\rho} \end{cases}.$$

Here $c > 0$ is a universal constant, and $\rho_p > 1.76$ is the unique solution of $\rho_p \log \rho_p = 2/p$.

Proof of Theorem 4. It suffices to show that the minimax optimal risk $R\big(\mathcal{T}^k_{n,d}(C_n)\big)$ is lower bounded by the three terms present in the statement's lower bound separately:

$$\begin{aligned} R\big(\mathcal{T}^k_{n,d}(C_n)\big) &= \Omega\bigg(\frac{\kappa\sigma^2}{n}\bigg), \\ R\big(\mathcal{T}^k_{n,d}(C_n)\big) &= \Omega\bigg(\frac{\sigma C_n}{n} \wedge \sigma^2\bigg), \\ R\big(\mathcal{T}^k_{n,d}(C_n)\big) &= \Omega\bigg(\bigg(\frac{C_n}{n}\bigg)^{\frac{2}{2s+1}} \sigma^{\frac{4s}{2s+1}} \wedge \sigma^2\bigg), \end{aligned} \tag{S.9}$$

where $\kappa = \mathrm{nullity}\big(D^{(k+1)}_{n,d}\big) = (k+1)^d$. First, as the null space of $D^{(k+1)}_{n,d}$ has dimension κ, we get the first lower bound:

$$\inf_{\hat\theta} \sup_{\theta_0 \in \mathcal{T}^k_{n,d}(C_n)} \frac{1}{n}\mathbb{E}\|\hat\theta - \theta_0\|_2^2 \;\geq\; \inf_{\hat\theta} \sup_{\theta_0 \in \mathrm{null}\left(D^{(k+1)}_{n,d}\right)} \frac{1}{n}\mathbb{E}\|\hat\theta - \theta_0\|_2^2 \;\geq\; \frac{\kappa\sigma^2}{n}.$$

We get the second lower bound in (S.9) by using the ℓ_1-ball embedding

$$B_1\left(C_n/d_{\max}\right) \subset \mathcal{T}^k_{n,d}(C_n)$$

from (S.8) and then using Lemma S.4. Finally, from Theorem 4 in Sadhanala et al. (2017), it follows that

$$R\big(\mathcal{C}^k_{n,d}(L_n)\big) = \Omega\bigg(\bigg(\frac{\sigma^2}{n}\bigg)^{\frac{2s}{2s+1}} L_n^{\frac{2}{2s+1}} \wedge \sigma^2\bigg) \tag{S.10}$$

with additional tracking for σ^2. Taking $L_n = C_n/n^{1-s}$ and applying the embedding in Proposition 3 would then give the third lower bound in (S.9). This completes the proof. □

A.8 Proof of Theorem 5

We use the following shorthand for the risk of an estimator $\hat\theta$ over a class $\mathcal{K}$:

$$\mathrm{Risk}(\hat\theta) = \sup_{\theta_0 \in \mathcal{K}} \frac{1}{n}\mathbb{E}\|\hat\theta - \theta_0\|_2^2.$$

For a matrix $S \in \mathbb{R}^{n\times n}$ let $\mathrm{Risk}(S)$ also denote the risk of the linear smoother $\hat\theta = Sy$.

Proof of Theorem 5. For brevity, denote $D = D_{n,d}^{(k+1)}$ and let S stand for a linear smoother in the context of this proof. The minimax linear risk for the class $\mathcal{T}_{n,d}^k(C_n)$ is

$$\begin{aligned} R_L(\mathcal{T}_{n,d}^k(C_n)) &= \inf_{S \in \mathbb{R}^{n\times n}} \sup_{\theta_0 \in \mathcal{T}_{n,d}^k(C_n)} \frac{1}{n}\mathbb{E}\|Sy - \theta_0\|_2^2 \\ &= \inf_{S} \sup_{\theta_0 \in \mathcal{T}_{n,d}^k(C_n)} \frac{1}{n}\mathbb{E}\|S(\theta_0 + \epsilon) - \theta_0\|_2^2 \\ &= \frac{1}{n}\inf_{S} \sup_{\theta_0 \in \mathcal{T}_{n,d}^k(C_n)} \sigma^2\|S\|_F^2 + \|(S-I)\theta_0\|_2^2 \end{aligned}$$

where in the last line we used the assumption that $\epsilon_i, i \in [n]$ are i.i.d. with mean zero and variance σ^2 and used the notation $\|A\|_F$ for the Frobenius norm of a matrix A. The infimum can be restricted to the set of linear smoothers

$$\mathbb{S} = \{S : \text{null}(S - I) \supseteq \text{null}(D)\}$$

because if for a linear smoother S, if there exists $\eta \in \text{null}(D)$ such that $(S-I)\eta \neq 0$, then the inner supremum above will be ∞, that is, its risk will be ∞. If the outer infimum is over $\mathbb{S}$, then the supremum can be restricted to $\{\theta_0 \in \text{row}(D) : \theta \in \mathcal{T}_{n,d}^k(C_n)\}$. We continue to lower bound minimax linear risk as follows:

$$\begin{aligned} R_L(\mathcal{T}_{n,d}^k(C_n)) &= \frac{1}{n}\inf_{S\in\mathbb{S}} \sigma^2\|S\|_F^2 + \sup_{\theta_0 \in \text{row}(D): \|D\theta_0\|_1 \le C_n} \|(S-I)\theta_0\|_2^2 \\ &= \frac{1}{n}\inf_{S\in\mathbb{S}} \sigma^2\|S\|_F^2 + \sup_{z:\|z\|_1 \le C_n} \|(S-I)D^+ z\|_2^2 \\ &= \frac{1}{n}\inf_{S\in\mathbb{S}} \sigma^2\|S\|_F^2 + C_n^2 \max_{i\in[m]} \left\|\left((S-I)D^+\right)_i\right\|_2^2 \qquad \text{(S.11)} \\ &\ge \frac{1}{n}\inf_{S\in\mathbb{S}} \sigma^2\|S\|_F^2 + \frac{C_n^2}{m}\sum_{i=1}^m \left\|\left((S-I)D^+\right)_i\right\|_2^2 \\ &= \inf_{S\in\mathbb{S}} \underbrace{\frac{\sigma^2}{n}\|S\|_F^2 + \frac{C_n^2}{mn}\left\|(S-I)D^+\right\|_F^2}_{=:r(S)} \qquad \text{(S.12)} \end{aligned}$$

In the third line, $(A)_i$ denotes the ith column of matrix A and m denotes the number of rows in D. In the fourth line, we used the fact that the maximum of a set is at least as much as their average. In the last line — within the context of this proof — we define the quantity $r(S)$ which is a lower bound on the risk of a linear smoother $S \in \mathbb{S}$.

Notice that $r(\cdot)$ is a quadratic in the entries of S and the constraint $S \in \mathbb{S}$ translates to linear constraints on the entries of S. Writing the KKT conditions, after some work, we see that $r(\cdot)$ is minimized at

$$S_0 = a_n\left(\sigma^2 L^{(k+1)} + a_n I\right)^{-1} \qquad \text{(S.13)}$$

where we denote $a_n = \frac{C_n^2}{m}$ and $L^{(k+1)} = D^\mathsf{T} D$ (the inverse is well defined because $a_n > 0$). Further, $S_0 \in \mathbb{S}$. Therefore,

$$R_L(\mathcal{T}_{n,d}^k(C_n)) \ge r(S_0). \qquad \text{(S.14)}$$

We simplify the expression for $r(S_0)$ now. Let $\lambda_i, i \in [n]$ be the eigenvalues of $L^{(k+1)}$. Then the eigenvalues of S_0 are

$$\frac{a_n}{\sigma^2\xi_i + a_n}, i \in [n]$$

and the non-zero squared singular values of $(S_0 - I)D^+$ are given by

$$\frac{\sigma^4\xi_i}{(\sigma^2\xi_i + a_n)^2}, \quad \kappa < i \le n.$$

Using the fact that the squared Frobenius norm of a matrix is the sum of squares of its singular values, substituting the above eigenvalues and singular values in (S.12), we have

$$\begin{aligned} r(S_0) &= \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{a_n}{\sigma^2\xi_i + a_n}\right)^2 + \frac{a_n}{n}\sum_{i=1}^{n}\frac{\sigma^4\xi_i}{(\sigma^2\xi_i + a_n)^2} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n}. \end{aligned} \tag{S.15}$$

Now we upper bound the risk $\mathrm{Risk}(S_0)$ of the linear smoother defined by S_0. From (S.11), we can write

$$\mathrm{Risk}(S_0) = \frac{\sigma^2}{n}\|S_0\|_F^2 + \frac{C_n^2}{n}\max_{i\in[m]}\left\|\left((S_0 - I)D^+\right)_i\right\|_2^2.$$

Let $D = U\Sigma V^\mathsf{T}$ be the singular value decomposition of D. Also let the eigen-decomposition of $S_0 - I = V\Lambda V^\mathsf{T}$. Then using incoherence of columns of U, that is, the fact that there exists a constant $c > 1$ that depends only on k, d such that $U_{ij}^2 \le \frac{c}{m}$ for all $i \in [m], j \in [n]$, we can write

$$\begin{aligned} \max_{i\in[m]}\left\|\left((S_0 - I)D^+\right)_i\right\|_2^2 &= \max_{i\in[m]}\left\|V\Lambda V^\mathsf{T} V\Sigma^+(U^\mathsf{T})_i\right\|_2^2 \\ &= \max_{i\in[m]}(U^\mathsf{T})_i^\mathsf{T}(\Lambda\Sigma^+)^2(U^\mathsf{T})_i \\ &\le \frac{c}{m}\mathrm{tr}\left((\Lambda\Sigma^+)^2\right) \\ &= \frac{c}{m}\sum_{i=1}^{n}\frac{\sigma^4\xi_i}{(\sigma^2\xi_i + a_n)^2}. \end{aligned}$$

Plugging this back in the previous display and also using the fact that the squared Frobenius norm of a matrix is equal to the sum of the squares of its eigenvalues,

$$\begin{aligned} \mathrm{Risk}(S_0) &= \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{a_n}{\sigma^2\xi_i + a_n}\right)^2 + \frac{c\cdot a_n}{n}\sum_{i=1}^{n}\frac{\sigma^4\xi_i}{(\sigma^2\xi_i + a_n)^2} \\ &\le c\cdot r(S_0) \end{aligned}$$

Combining this with the lower bound in (S.14), we have

$$r(S_0) \le R_L(\mathcal{T}_{n,d}^k(C_n)) \le \min\left\{\sigma^2, \mathrm{Risk}(S_0)\right\} \le \min\left\{\sigma^2,\ \ c\cdot r(S_0)\right\}. \tag{S.16}$$

In other words, the minimax linear rate is essentially $r(S_0)$ up to a constant factor. Further, one of the estimators $\hat{y} = S_0 y$, $\hat{y} = y$ achieves the minimax linear rate up to a constant factor.

Now we bound $r(S_0)$. Let $\kappa = (k+1)^d$ denote the nullity of D. Recall from (S.15)

$$r(S_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n} = \frac{\kappa\sigma^2}{n} + \frac{1}{n}\sum_{i=\kappa+1}^{n}\frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n}. \tag{S.17}$$

Lower bounding $r(S_0)$. We give three lower bounds on $r(S_0)$. By using the fact that arithmetic mean of positive numbers is at least as large as their harmonic mean, we have

$$\begin{aligned} r(S_0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n} \\ &\ge \frac{n\sigma^2 a_n}{\sum_{i=1}^{n}(\sigma^2\xi_i + a_n)} \\ &= \frac{n\sigma^2 a_n}{na_n + \sigma^2\|D\|_F^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{n\sigma^2 a_n}{na_n + \sigma^2 d n^{1-1/d}\|D_{\mathrm{1d}}^{(k+1)}\|_F^2} \\
&= \frac{\sigma^2 a_n}{a_n + \sigma^2 d n^{-1/d}(n^{1/d} - k - 1)\binom{2k+2}{k+1}} \\
&\geq \frac{\sigma^2 a_n}{a_n + \sigma^2 d 4^{k+1}} \qquad \text{(S.18)}
\end{aligned}
$$

Now we bound in $r(S_0)$ in a second way. Let n_1 be the cardinality of $\{i \in [n] : \sigma^2\xi_i \leq a_n\}$. Then

$$
r(S_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n} \geq \frac{1}{n}\sum_{i=1}^{n_1} \frac{\sigma^2 a_n}{a_n + a_n} = \frac{n_1\sigma^2}{2n}.
$$

Note that $n_1 = \lfloor nF(a_n/\sigma^2)\rfloor$ where F is the spectral distribution of $(D_{n,d}^{(k+1)})^{\mathsf{T}} D_{n,d}^{(k+1)}$ defined in Lemma S.10. Applying Lemma S.10, we get

$$
\begin{aligned}
r(S_0) &\geq \frac{\sigma^2}{2}\left(F\left(\frac{a_n}{\sigma^2}\right) - \frac{1}{n}\right) \\
&\geq c\sigma^2 \min\left\{1, (a_n/\sigma^2)^{\frac{1}{2s}}\right\} - \sigma^2/2n \\
&= \min\left\{c\sigma^2, c\sigma^{2-\frac{1}{s}} a_n^{\frac{1}{2s}} - \sigma^2/2n\right\} \qquad \text{(S.19)}
\end{aligned}
$$

In the special case $s = 1/2$, from Lemma S.9 we get a third bound:

$$
r(S_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n} \geq c_1 a_n \log\left(1 + c_2/a_n\right) \qquad \text{(S.20)}
$$

where c_1, c_2 constants that depend only on k, d.

From (S.17),(S.18), (S.19) and (S.20) we have the lower bound

$$
r(S_0) \geq \max\left\{\frac{\kappa\sigma^2}{n}, \frac{\sigma^2 a_n}{a_n + \sigma^2 d 2^{2k+2}}, \sigma^2 \wedge c\sigma^{2-\frac{1}{s}} a_n^{\frac{1}{2s}} - \frac{\sigma^2}{2n}\right\}. \qquad \text{(S.21)}
$$

and an additional lower bound of $c_1 a_n \log\left(1 + c_2/a_n\right)$ in the case $s = 1/2$. Substituting $a_n = C_n^2/m$, using the assumption that $C_n^2/n \leq 1$ and treating k, d, σ as constants , we get the stated lower bound.

Upper bounding $r(S_0)$. If $s < 1/2$, then

$$
\begin{aligned}
r(S_0) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n} \\
&\leq \frac{\kappa\sigma^2}{n} + \frac{1}{n}\sum_{i=\kappa+1}^{n} \frac{\sigma^2 a_n}{\sigma^2\xi_i} \\
&= \frac{\kappa\sigma^2}{n} + \frac{a_n}{n}\sum_{i=1}^{\kappa+1} \frac{1}{\xi_i} \\
&\leq \frac{\kappa\sigma^2}{n} + \frac{a_n}{n}(c_3 n) \\
&= \frac{\kappa\sigma^2}{n} + c_3 a_n \qquad \text{(S.22)}
\end{aligned}
$$

We used Lemma S.6 to control the second term in the third line. Similarly, if $s = 1/2$, $r(S_0) \leq \kappa\sigma^2/n + c_3 a_n \log n$. For the case $s > 1/2$, we can write

$$
r(S_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 a_n}{\sigma^2\xi_i + a_n}
$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n_1}\frac{\sigma^2 a_n}{a_n} + \frac{1}{n}\sum_{i=n_1+1}^{n}\frac{\sigma^2 a_n}{2\sigma^2\xi_i} \\
&= \frac{n_1\sigma^2}{n} + \frac{a_n}{2n}\sum_{i=n_1+1}^{n}\frac{1}{\xi_i} \\
&\leq c\frac{\sigma^2}{n} + c\sigma^2\left(\frac{a_n}{\sigma^2}\right)^{\frac{1}{2s}} + c\frac{a_n}{2n}n^{2s}\left(n(a_n/\sigma^2)^{\frac{1}{2s}}\right)^{1-2s} \\
&\leq c\frac{\sigma^2}{n} + c\sigma^{2-\frac{1}{s}}a_n^{\frac{1}{2s}}
\end{aligned} \tag{S.23}
$$

To get the fourth line, we used Lemma S.10 to bound n_1 and Lemma S.6 to bound the summation.

Upper bound with the polynomial projection estimator $\hat{\theta}^{\text{poly}}$. For brevity, let Π denote the matrix that projects on to the null space of D. Note that $(I-\Pi)D^+ = D^+$. From bias variance decomposition similar to that in (S.11),

$$
\begin{aligned}
\frac{1}{n}\sup_{\theta_0\in\mathcal{T}^k_{n,d}(C_n)}\mathbb{E}\left[\|\hat{\theta}^{\text{poly}}-\theta_0\|_2^2\right] &= \frac{\sigma^2}{n}\|\Pi\|_F^2 + \max_{i\in[m]}\|\left((\Pi-I)D^+\right)_i\|_2^2 \\
&= \frac{\kappa\sigma^2}{n} + \max_{i\in[m]}\|D_i^+\|_2^2
\end{aligned}
$$

Then using incoherence of columns of U, that is, the fact that there exists a constant $c>1$ that depends only on k,d such that $U_{ij}^2 \leq \frac{c}{m}$ for all $i\in[m], j\in[n]$, we can write

$$
\begin{aligned}
\max_{i\in[m]}\|D_i^+\|_2^2 &= \max_{i\in[m]}\left\|V\Sigma^+(U^\mathsf{T})_i\right\|_2^2 \\
&= \max_{i\in[m]}(U^\mathsf{T})_i^\mathsf{T}(\Sigma^+)^2(U^\mathsf{T})_i \\
&\leq \frac{c}{m}\text{tr}\left((\Sigma^+)^2\right) \\
&= \frac{c}{m}\sum_{i=\kappa+1}^{n}\frac{1}{\xi_i}
\end{aligned}
$$

Plugging this back in the above display and using the bound on $\sum_{i=\kappa+1}^n \frac{1}{\xi_i}$ from Lemma S.6, we get the desired result.

Upper bound with the projection estimator (34) when $s>1/2$. From (S.11), for the projection estimator $\hat{\theta}=S_Q y$ in (34),

$$
\text{Risk}(\hat{\theta}) = \frac{\sigma^2}{n}|Q| + \frac{C_n^2}{n}\max_{i\in[m]}\left\|\left((S_Q-I)D^+\right)_i\right\|_2^2. \tag{S.24}
$$

Set $Q=[\tau]^d$ for a τ to be chosen later from $(k+2,N]$. Also write $S_Q - I = V\Lambda_Q V^\mathsf{T}$. Again using incoherence of columns of U, we can write

$$
\begin{aligned}
\max_{i\in[m]}\left\|\left((S_Q-I)D^+\right)_i\right\|_2^2 &= \max_{i\in[m]}\left\|V\Lambda_Q V^\mathsf{T}V\Sigma^+(U^\mathsf{T})_i\right\|_2^2 \\
&= \max_{i\in[m]}(U^\mathsf{T})_i^\mathsf{T}(\Lambda_Q\Sigma^+)^2(U^\mathsf{T})_i \\
&\leq \frac{c}{m}\text{tr}\left((\Lambda_Q\Sigma^+)^2\right) \\
&= \frac{c}{m}\sum_{i\in[N]^d\setminus Q}\frac{1}{\xi_i}
\end{aligned}
$$

The summation in the last line can be bound using Lemma S.6 (recall $s>1/2$ here):

$$
\sum_{i\in[N]^d\setminus Q}\frac{1}{\xi_i} \leq \sum_{\|(i-k-2)_+\|_2\geq\tau-k-2}\frac{1}{\xi_i} \leq cn(n/(\tau-k-2)^d)^{2s-1}
$$

Tracing this back to (S.24),

$$\mathrm{Risk}(\hat{\theta}) \le \frac{\sigma^2}{n}\tau^d + \frac{cC_n^2}{m} \cdot (n/(\tau-k-2)^d)^{2s-1}$$

Minimize this bound by setting τ such that $\tau^d \asymp (C_n/\sigma)^{\frac{1}{s}} n^{1-\frac{1}{2s}}$ to get the desired bound. □

Remark 1. In Theorem 5, in the case $s \le 1/2$, the lower bound may also be obtained by embedding the ℓ_1-ball $B_1(C_n/(2^{k+1}d))$ into $\mathcal{T}^k_{n,d}(C_n)$.

A.9 Proof of Theorem 6

Proof of upper bound. Like in the proof of minimax linear rates for KTV class in Theorem 5, for the projection estimator $\hat{\theta} = S_Q y$ where $S_Q = V_Q V_Q^\mathsf{T}$, we can derive

$$\frac{1}{n}\sup_{\theta_0 \in \mathcal{W}^{k+1}_{n,d}(B_n)} \mathbb{E}\big[\|\hat{\theta}-\theta_0\|_2^2\big] = \frac{\sigma^2}{n}|Q| + \frac{1}{n}\sup_{\theta_0 \in \mathcal{W}^{k+1}_{n,d}(B_n)} \|(I-S_Q)\theta_0\|_2^2.$$

Denote $D = D^{(k+1)}_{n,d}$ for brevity. Set $Q = [\tau]^d$, where $\tau \in (k+2, N]$ is an integer (recall $N = n^{1/d}$) and analyze the maximum of the second term:

$$\begin{aligned}
\sup_{\theta_0:\|D\theta_0\|_2 \le B_n} \frac{1}{n}\|(I-S_q)\theta_0\|_2^2 &= \sup_{z:\|z\|_2 \le C_n} \frac{1}{n}\|(I-S_q)D^\dagger z\|_2^2 \\
&= \frac{B_n^2}{n}\sigma^2_{\max}\big((I-S_q)D^\dagger\big) \\
&\le \frac{B_n^2}{n}\frac{1}{4^{k+1}\sin^{2k+2}(\pi(\tau-k-2)/(2N))} \\
&\le \frac{B_n^2}{n}\frac{N^{2k+2}}{(\pi(\tau-k-2))^{2k+2}}.
\end{aligned}$$

Here we denote by $\sigma_{\max}(A)$ the maximum singular value of a matrix A. The last inequality above used the inequality $\sin(x) \ge x/2$ for $x \in [0, \pi/2]$. The earlier inequality used that $\sigma^2_{\max}((I-S_Q)D^\dagger)$ is the reciprocal of the smallest eigenvalue ρ_Q of $M = D^\mathsf{T}D$ with index in $[N]^d \setminus Q$. That is,

$$\rho_Q = \rho_{\tau+1,1,\ldots,1} \ge \big(4\sin^2(\pi(\tau-k-2)/(2N))\big)^{k+1},$$

where the last inequality is due to the relation in (S.37). Hence, we have established

$$\sup_{\theta_0:\|D\theta_0\|_2 \le B_n} \frac{1}{n}\mathbb{E}\big[\|\hat{\theta}-\theta_0\|_2^2\big] \le \frac{\sigma^2}{n}\tau^d + \frac{B_n^2}{n}\frac{N^{2k+2}}{(\pi(\tau-k-2))^{2k+2}}.$$

Choosing τ to balance the two terms on the right-hand side above results in $\tau^d \asymp (k+2)^d + \big(B_n^2 n^2 s/\sigma^2\big)^{\frac{1}{2s+1}}$. Also, in the edge case where $Q = [N]^d$, the risk is σ^2. Plugging this choice of τ gives the upper bound result.

Proof of lower bound. Similar to argument in the proof of Theorem 4, the nullity of $D^{(k+1)}_{n,d}$ implies the lower bound

$$R(\mathcal{W}^{k+1}_{n,d}(C_n)) = \Omega\big(\frac{\kappa\sigma^2}{n}\big). \tag{S.25}$$

The Holder ball embedding

$$\mathcal{W}^{k+1}_{n,d}(C_n) \supseteq \mathcal{C}^k_{n,d}(cC_n n^{s-\frac{1}{2}})$$

implies that

$$R(\mathcal{W}^{k+1}_{n,d}(C_n)) \ge R\big(\mathcal{C}^k_{n,d}(cC_n n^{s-\frac{1}{2}})\big) = \Omega\left(\frac{C_n^2}{n}\right)^{\frac{1}{2s+1}} \sigma^{\frac{4s}{2s+1}} \wedge \sigma^2,$$

where the second step follows from (S.10). Putting these two bounds together, we get the desired lower bound.

B Estimation theory for graph trend filtering on grids

We recall the GTF operator from Wang et al. (2016) for convenience. Let $G(V, E)$ be a graph with n vertices and m edges $(u_1, v_1), \ldots, (u_m, v_m) \in [n] \times [n]$. Assume that $u_i < v_i$ for $i \in [m]$ in the edges here for notational convenience. Let $D \in \mathbb{R}^{m \times n}$ be the incidence matrix of G satisfying

$$(Dx)_j = x_{u_j} - x_{v_j} \quad \text{for all } x \in \mathbb{R}^n$$

for all edges (u_j, v_j) for $j \in [m]$. The graph Laplacian is $L = D^\mathsf{T} D$ The GTF operators of all orders are defined by

$$\begin{aligned} S_{n,d}^{(1)} &= D, \quad S_{n,d}^{(2)} = L, \\ S_{n,d}^{(2k+1)} &= DL^k, \quad S_{n,d}^{(2k)} = L^k \text{ for } k \geq 0, k \in \mathbb{Z}. \end{aligned} \tag{S.26}$$

B.1 Upper bounds on estimation risk

Wang et al. (2016) used Theorem 2 (their Theorem 6) in order to derive error rates for GTF on 2d grids already; see their Corollary 8. Sadhanala et al. (2017) refine this result using a tighter upper bound for the partial sum of inverse eigenvalues. Here, we give a more general result that applies to not just 2d grids, but all $d \geq 2$ and $k \geq 0$. We further show that these rates are optimal by deriving a matching lower bound. Recall the abbreviation $s = (k+1)/d$.

Theorem S.1. *Assume that $d \geq 1$ and $k \geq 0$. Denote $C_n = \|S_{n,d}^{(k+1)} \theta_0\|_1$. Then GTF defined by the estimator in (30) with $D = S_{n,d}^{(k+1)}$ in (S.26) satisfies*

$$\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2 = O_\mathbb{P}\left(\frac{1}{n} + \frac{\lambda}{n} C_n\right)$$

with

$$\lambda \asymp \begin{cases} \sqrt{\log n} & s < 1/2 \\ \log n & s = 1/2 \\ (\log n)^{\frac{1}{2s+1}} \left(\frac{n}{C_n}\right)^{\frac{2s-1}{2s+1}} & s > 1/2. \end{cases}$$

With canonical scaling of C_n, we see the following error bound.

Corollary 1. *With canonical scaling $C_n = C_n^* = n^{1-s}$, the GTF estimator with λ scaling as in Theorem S.1 satisfies*

$$\sup_{\theta_0 \in \mathcal{S}_{n,d}^k(C_n)} \frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2 = \begin{cases} O_\mathbb{P}\left(n^{-s}\sqrt{\log n}\right) & s < 1/2 \\ O_\mathbb{P}\left(n^{-s} \log n\right) & s = 1/2 \\ O_\mathbb{P}\left(n^{-\frac{2s}{2s+1}} (\log n)^{\frac{1}{2s+1}}\right) & s > 1/2. \end{cases}$$

Remarks following Theorem 3 for KTF apply for GTF as well. The proof is in Appendix B.4.

B.2 Lower bounds on estimation risk

Similar to the lower bound in Theorem 4 for KTV class, we give a bound for the graph total variation (GTV) class

$$\mathcal{S}_{n,d}^k(C_n) = \{\theta \in \mathbb{R}^n : \|S_{n,d}^{(k+1)} \theta\|_1 \leq C_n\}. \tag{S.27}$$

Due to the lower order discrete derivatives on the boundary of the grid $Z_{n,d}$, the GTV class $\mathcal{S}_{n,d}^k(C_n)$ cannot contain the discrete Holder class with appropriate scaling $\mathcal{C}_{n,d}^k(C_n n^{s-1})$; see Lemma 4 in Sadhanala et al. (2017). However, by an alternative route Sadhanala et al. (2017) show a lower bound for $\mathcal{S}_{n,d}^k(C_n)$ that matches with the lower bound for the Holder class $\mathcal{C}_{n,d}^k(C_n n^{s-1})$. We further tighten their result by embedding an ℓ_1 ball of appropriate size.

Theorem S.2. *For any integers $k \geq 0$, $d \geq 1$, the minimax estimation error for the GTV class defined in (S.27) satisfies*

$$R\big(\mathcal{S}_{n,d}^k(C_n)\big) = \Omega\left(\frac{\sigma^2}{n} + \frac{\sigma C_n}{n} + \left(\frac{C_n}{n}\right)^{\frac{2}{2s+1}} \sigma^{\frac{4s}{2s+1}} \wedge \sigma^2\right).$$

Proof of Theorem S.2. Similar to the proof of Theorem 4, it is sufficient to show three lower bounds separately. We get the first two lower bounds just as in the proof of Theorem 4 by using the fact that $\text{nullity}\big(S^{(k+1)}_{n,d}\big) = 1$ and the ℓ_1-ball embedding

$$B_1(C_n/(2^{k+1}d)) \subseteq \mathcal{S}^k_{n,d}(C_n)$$

from Lemma S.3 and the fact that $\|S^{(k+1)}_{n,d}\|_{1,\infty} \leq 2^{k+1}d$. The third term is from Theorem 5 in Sadhanala et al. (2017). □

B.3 Minimax rates for linear smoothers

The minimax linear rate analysis for GTV class is very similar to that for KTV class. So we simply state the result and skip the proof.

Theorem S.3. *The minimax linear risk over the GTV class in (S.27) satisfies, for any sequence $C_n \leq \sqrt{n}$,*

$$R_L\big(\mathcal{S}^k_{n,d}(C_n)\big) = \begin{cases} \Omega(1/n + C_n^2/n) & \text{if } s < 1/2, \\ \Omega\big(1/n + C_n^2/n \log(1 + n/C_n^2)\big) & \text{if } s = 1/2, \\ \Omega\big(1/n + (C_n^2/n)^{\frac{1}{2s}}\big) & \text{if } s > 1/2. \end{cases} \tag{S.28}$$

This is achieved in rate by the projection estimator in (34), by setting $Q = [\tau]^d$ for $\tau^d \asymp (C_n n^{s-1/2})^{1/s}$, in the case $s > 1/2$. When $s < 1/2$, the simple mean estimator, $\hat{\theta}^{\text{mean}} = \bar{y}\mathbb{1}$, achieves the rate in (S.28). When $s = 1/2$, this estimator achieves the rate in (S.28) up to a log factor. Lastly, if $C_n^2 = O(n^\alpha)$ for $\alpha < 1$, and still $s = 1/2$, then the mean estimator achieves the rate in (S.28) without the additional log factor.

B.4 Proof of Theorem S.1

For $d = 2$, it is shown in the proof of Corollary 8 in Wang et al. (2016) that the GTF operator $S^{(k+1)}_{n,d}$ satisfies the incoherence property, as defined in Theorem 2, with a constant $\mu = 4$ when k is even and $\mu = 2$ when k is odd. Here we extend this incoherence property for $d > 2$ using Lemma S.1. We treat the cases where k is odd and even separately.

If k is odd we can extend the argument from Corollary 8 in Wang et al. (2016) in a straightforward manner. The GTF operator is $S^{(k+1)}_{n,d} = L^{(k+1)/2}$ where L is the Laplacian of the d-dimensional grid graph. Denoting the Laplacian of the chain graph of length N by $L_{1\text{d}}$, we note that L is given by

$$L = L_{1\text{d}} \otimes I \otimes I + I \otimes L_{1\text{d}} \otimes I + I \otimes I \otimes L_{1\text{d}}$$

for $d = 3$ and

$$L = L_{1\text{d}} \otimes I \cdots \otimes I + I \otimes L_{1\text{d}} \cdots \otimes I + \cdots + I \otimes \ldots I \otimes L_{1\text{d}}$$

for general d where each term in the summation is a Kronecker product of d matrices. Let $\alpha_i, u_i, i \in [N]$ be the eigenvalues and eigenvectors of $L_{1\text{d}}$. As shown in Wang et al. (2016), in 1d, we have the incoherence property $\|u_i\|_\infty \leq \sqrt{2/N}$ for all $i \in [N]$. The eigenvalues of L are $\sum_{j=1}^d \alpha_{i_j}$ and the corresponding eigenvectors are $u_{i_1} \otimes \cdots \otimes u_{i_d}$ for $i_1, \ldots, i_d \in [N]$. Clearly, incoherence holds for the eigenvectors of L with constant $\mu = 2^{d/2}$.

If k is even, then the left singular vectors of $S^{(k+1)}_{n,d}$ are the same as those of $S^{(1)}_{n,d}$. We know that both the left and right singular vectors of $D^{(1)}_{1\text{d}}$ satisfy the incoherence property with constant $\mu = \sqrt{2}$ (see the proof of Corollary 7 in Wang et al. (2016)). Setting $D = D^{(1)}_{1\text{d}}$ in Lemma S.1, we see that the left singular vectors of $S^{(1)}_{n,d}$ and hence those of $S^{(k+1)}_{n,d}$ satisfy incoherence property with constant $2^{d/2}$. Therefore, for all integers $k \geq 0$, the left singular vectors of $S^{(k+1)}_{n,d}$ are incoherent with constant $2^{d/2}$.

From the incoherence property and Theorem 2, the GTF estimator $\hat{\theta}$, satisfies

$$\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{|I|}{n} + \frac{\mu}{n}\sqrt{\frac{\log n}{n} \sum_{i \in [N]^d \setminus (I \cup \{1\}^d)} \frac{1}{\rho_i^2}} \cdot \|\Delta\theta_0\|_1\right), \tag{S.29}$$

where $\rho_i, i \in [N]^d$ are the eigenvalues $S^{(k+1)\mathsf{T}}_{n,d} S^{(k+1)}_{n,d}$ and $\mu = 2^{d/2}$.

Consider the set $I = \{i \in [N]^d : \|i-1\|_2 < r\} \setminus \{1\}^d$ for an $r \in [1, \sqrt{d}N]$ chosen later. Lemma S.5 gives the key calculation where it is shown that for large enough n,

$$\sum_{\|i-1\|\geq r} \frac{1}{\rho_i^2} = \sum_{\|i-1\|\geq r} \frac{1}{\lambda_i^{k+1}} \leq c \begin{cases} n & s < 1/2 \\ n\log(2\sqrt{d}N/r) & s = 1/2 \\ n(n/r^d)^{2s-1} & s > 1/2 \end{cases}$$

where $\lambda_i, i \in [N]^d$ are eigenvalues of the Laplacian L and $c > 0$ is a constant that depends only on k, d .

For $s \leq 1/2$, to minimize the upper bound in (S.29), set $r = 1$ so that I is empty and apply the above inequality. This gives the desired bound. Now consider $s > 1/2$. Note that $|I| \leq r^d$ because $I \subseteq [r]^d$. Therefore (S.29) reduces to

$$\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2 = O_{\mathbb{P}}\Big(\frac{r^d}{n} + \frac{\mu}{n}\sqrt{\log n (n/r^d)^{2s-1}}\|\Delta\theta_0\|_1\Big) \tag{S.30}$$

To minimize the upper bound in (S.30) balance

$$r^d \quad \text{with} \quad \frac{C_n}{\sqrt{n}}\sqrt{n(n/r^d)^{2s-1}\log n}.$$

This leads us to take

$$r^d \asymp (C_n\sqrt{\log n})^{\frac{2}{2s+1}} n^{\frac{2s-1}{2s+1}}$$

and plugging this in (S.30) gives the desired bound for $s > 1/2$. This completes the proof. □

C Technical lemmas

Lemma S.5. *Consider the eigenvalues $\{\lambda_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ of the d-dimensional grid graph Laplacian with $n = N^d$ nodes. Let k be a non-negative integer and $r_0 \in [1, \sqrt{d}N]$. Then,*

$$\sum_{i\in[N]^d : \|i-1\|_2^2 \geq r_0^2} \frac{1}{\lambda_i^k} \leq c \begin{cases} n & 2k < d \\ n\log(2\sqrt{d}N/r_0) & 2k = d \\ N^{2k} r_0^{d-2k} & 2k > d \end{cases}$$

for a constant $c > 0$ that depends on k, d but not on N, r_0.

Proof of Lemma S.5. Let I denote the summation on the left. Then

$$\begin{aligned} I = \sum_{i\in[N]^d : \|i-1\|_2 \geq r_0} \frac{1}{\lambda_i^k} &= \sum_{\|i-1\|_2 \geq r_0} \Big(\sum_{j=1}^d 4\sin^2\frac{\pi(i_j-1)}{2N}\Big)^{-k} \\ &\leq \sum_{\|i-1\|_2 \geq r_0} \Big(\sum_{j=1}^d \frac{\pi^2(i_j-1)^2}{4N^2}\Big)^{-k} \\ &= cN^{2k} \sum_{\|i-1\|_2 \geq r_0} \Big(\sum_{j=1}^d (i_j-1)^2\Big)^{-k} \\ &\leq cN^{2k} \sum_{i\in\{0,1,..,N-1\}^d : \|i\|_2 \geq r_0} \|i\|_2^{-2k} \end{aligned} \tag{S.31}$$

In the second line, we used the fact that $\sin x \geq x/2$ for $x \in [0, \pi/2]$.

Case $r_0 \geq 2\sqrt{d}$. In the last expression, upper bound $\|i\|_2^{-2k}$ with the integral of $f(x) = \|x\|_2^{-2k}$, $f : \mathbb{R}^d \to \mathbb{R}$ over the unit length cube whose top right corner is at i. Note that, the norm of any point in this cube is at least $\|i - \mathbb{1}\|_2 \geq \|i\|_2 - \|\mathbb{1}\|_2 = \|i\|_2 - \sqrt{d} \geq r_0 - \sqrt{d} \geq r_0/2$. Therefore, we can continue to bound

$$I \leq cN^{2k} \int_{r_0/2 \leq \|x\|_2 \leq \sqrt{d}N} \|x\|_2^{-2k}\, dx$$

$$\leq cN^{2k} \int_{r_0/2 \leq r \leq \sqrt{d}N} (r^2)^{-k} r^{d-1}\, dr$$

The last line is obtained by changing to polar coordinates and integrating out the angles. Recall that the constants c may change from line to line and they may depend on k, d, but not on N, r_0.
If $d = 2k$, then the integral

$$I \leq cN^{2k} \log(2\sqrt{d}N/r_0) = cn \log(2\sqrt{d}N/r_0).$$

If $2k < d$, then

$$I \leq cN^{2k}\big((N\sqrt{d})^{d-2k} - (r_0/2)^{d-2k}\big) \leq cN^d.$$

If $2k > d$, then

$$I \leq cN^{2k}\big((r_0/2)^{d-2k} - (N\sqrt{d})^{d-2k}\big).$$

Treating d, k as constants, we write

$$I \leq cN^{2k} r_0^{d-2k}.$$

Case $r_0 < 2\sqrt{d}$. Continuing from (S.31), write

$$I \leq cN^{2k} \sum_{i \in \{0,1,..,N-1\}^d : \|i\|_2 \in [r_0, 2\sqrt{d})} \|i\|_2^{-2k} + cN^{2k} \sum_{i \in \{0,1,..,N-1\}^d : \|i\|_2 \geq 2\sqrt{d}} \|i\|_2^{-2k} \tag{S.32}$$

From the previous case, the second summation can be upper bound with cn if $2k < d$, $cn \log n$ if $2k = d$ and cN^{2k} if $2k > d$. In the first summation (in the above display), the number of entries i is at most $(2\sqrt{d})^d$ and each entry is at most r_0^{-2k}. Therefore the first term is at most $cN^{2k}(2\sqrt{d})^d r_0^{-2k}$. Putting the two sums together, we can verify the stated bounds. □

Following lemma provides a result analogous to Lemma S.5 for KTF.

Lemma S.6. *Let $\{\xi_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ be the eigenvalues of $D_{n,d}^{(k+1)^\mathsf{T}} D_{n,d}^{(k+1)}$ and suppose $r_0 \in [1, \sqrt{d}N]$. Then,*

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

In the case $2k + 2 > d$,

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

Here $c > 0$ is a constant that depends on k, d but not on N, r_0.

Proof. Using Lemma S.7, we can write

$$\sum_{i \in [N]^d : \|(i-k-2)_+\|_2 \geq r_0} \frac{1}{\xi_i^2} \leq d^{2k} \sum_{i \in [N]^d : \|(i-k-2)_+\|_2 \geq r_0} \frac{1}{\lambda_{i-k-1}^{2k+2}} \leq d^{2k} \sum_{i \in [N]^d : \|i-1\|_2 \geq r_0} \frac{1}{\lambda_i^{2k+2}}. \tag{S.33}$$

Applying Lemma S.5 directly gives the desired result in the case $2k + 2 > d$. In the case $2k + 2 \leq d$, we get the bound by setting $r_0 = 1$ in (S.33) and then applying Lemma S.5. □

Lemma S.7. *Let $\{\xi_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ be the eigenvalues of $D_{n,d}^{(k+1)^\mathsf{T}} D_{n,d}^{(k+1)}$ for $k \geq 0, d \geq 1, N \geq 1, n = N^d$. Let $\alpha_i, i \in [N]$ be the eigenvalues of L, the Laplacian of chain graph of length N. Let $\lambda_{i_1,\ldots,i_d} = \sum_{j=1}^d \alpha_{i_j}, \quad i \leq N$ elementwise with the convention that $\alpha_\ell = 0$ for $\ell \leq 0$. Then*

$$\xi_i \geq d^{-k} \lambda_{i-k-1}^{k+1} \text{ for } i \in [N]^d.$$

Proof. Abbreviate $D = D_{N,1}^{(k+1)}$, and let G be the kth order GTF operator defined over a 1d chain of length N. Also let $N' = N - k - 1$, and $k' = \lfloor (k+1)/2 \rfloor$. Let

- $\beta_\ell, \ell \in [N']$ be the eigenvalues of DD^T
- $\gamma_\ell, \ell \in [N'']$ be the eigenvalues of GG^T where $N'' = N - 1\{k \text{ is even}\}$

GG^T and $G^\mathsf{T}G$ should have the same nonzero eigenvalues. From the definition of G, $G^\mathsf{T}G = L^{k+1}$. The first eigenvalue of L is 0 and the rest are nonzero. Putting these facts together, we see that

$$\gamma_\ell = \alpha_{\ell+N-N''}^{k+1} \text{ and } \alpha_\ell^{k+1} \le \gamma_\ell \quad \text{for } \ell \in [N''']. \tag{S.34}$$

Removing the top k' and bottom k' rows of G yields D, i.e.,

$$D = PG, \quad \text{where } P = \begin{bmatrix} 0_{N'\times k'} & I_{N'} & 0_{N'\times k'} \end{bmatrix}.$$

As $DD^\mathsf{T} = PGG^\mathsf{T}P^\mathsf{T}$ and $PP^\mathsf{T} = I_{N'}$, Cauchy interlacing theorem (Lemma S.8) tells us that

$$\gamma_i \le \beta_i \le \gamma_{i+N''-N'}, \quad \text{for } i \in [N']. \tag{S.35}$$

Thanks to the Kronecker sum structure, the eigenvalues of $(D_{n,d}^{(k+1)})^\mathsf{T} D_{n,d}^{(k+1)}$ are

$$\xi_{i_1,\dots,i_d} = \sum_{j=1}^d \rho_{i_j}, \quad i \in [N]^d,$$

where $\rho_1, \dots, \rho_N$ denote the eigenvalues of $D^\mathsf{T}D$, i.e., $\rho_1 = \cdots = \rho_{k+1} = 0$ and $\rho_{\ell+k+1} = \beta_\ell \ell \in [N']$. Similarly, we can write the eigenvalues of the Laplacian of the d-dimensional grid graph as

$$\lambda_{i_1,\dots,i_d} = \sum_{j=1}^d \alpha_{i_j}, \quad i \in [N]^d.$$

For arbitrary $i \in [N]^d$, we can write

$$\xi_{i_1,\dots,i_d} = \sum_{j=1}^d \beta_{i_j-k-1} \ge \sum_{j=1}^d \gamma_{i_j-k-1} \ge \sum_{j=1}^d \alpha_{i_j-k-1}^{k+1} \ge d^{-k} \lambda_{i_1-k-1,\dots,i_d-k-1}^{k+1},$$

with the convention $\alpha_\ell = \beta_\ell = \gamma_\ell = 0$ for $\ell \le 0$. The first inequality is due to (S.35), the second is due to (S.34), and the third is due to a simple application of Jensen's inequality: $(\frac{1}{d}\sum_{j=1}^d a_i)^k \le \frac{1}{d}\sum_{j=1}^d a_i^k$ if $k \ge 1$ and $a \ge 0$ elementwise. □

Lemma S.8 (Cauchy Interlacing theorem)**.** *Let A be an $n \times n$ symmetric matrix, $P \in \mathbb{R}^{m\times n}$ be an orthogonal projection matrix (satisfying $PP^\mathsf{T} = I_m$) with $m \le n$ and define $B = PAP^\mathsf{T}$. Let $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_n$ be the eigenvalues of A and $\beta_1 \le \beta_2 \le \cdots \le \beta_m$ be the eigenvalues of B. Then*

$$\alpha_i \le \beta_i \le \alpha_{i+n-m}, \quad \textit{for } i \in [m].$$

Lemma S.9. *Let $\{\xi_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ be the eigenvalues of $D_{n,d}^{(k+1)^\mathsf{T}} D_{n,d}^{(k+1)}$ for $k \ge 0, d \ge 1, N \ge 1, n = N^d$. Suppose $s = 1/2$ and $a > 0$. Then*

$$\sum_{i\in[N]^d} \frac{1}{\xi_i + a} \ge cn \log\left(1 + \pi^{2k+2} a^{-1}\right)$$

for a constant c that depends only on k, d.

Proof of Lemma S.9. From (S.37) and the inequality $\sin x \le x$ for $x \ge 0$, for any $i \in [N]^d$,

$$\xi_i = \sum_{j=1}^d \rho_{i_j} \le \sum_{j=1}^d 4^{k+1} \sin^{2k+2} \frac{\pi(i_j - 1)}{2N} \le \pi^{2k+2} n^{-2s} \|i-1\|_{2k+2}^{2k+2} \le \pi^{2k+2} n^{-2s} \|i-1\|_2^{2k+2}.$$

With this inequality,

$$\sum_{i\in[N]^d} \frac{1}{\xi_i + a} \ge \sum_{i\in[N]^d} \frac{1}{\pi^{2k+2} n^{-2s} \|i-1\|_2^{2k+2} + a}$$

$$\geq c \int_{r=0}^{N} \frac{1}{\pi^{2k+2} n^{-2s} r^{2k+2} + a} r^{d-1}\, dr. \tag{S.36}$$

In the second inequality is obtained as follows. Consider axis-parallel unit cubes with corners located at integer coordinates. Let $A_i \subset \mathbb{R}^d$ be the cube with its farthest corner from origin located at i, for $i \in [N]^d$. Clearly,

$$\frac{1}{\pi^{2k+2} n^{-2s} \|i-1\|_2^{2k+2} + a} \geq \int_{A_i} \frac{1}{\pi^{2k+2} n^{-2s} \|x\|_2^{2k+2} + a}\, dx.$$

Next observe that the set $\{x \in \mathbb{R}^d : \|x\|_2 \leq N, x \geq 0\}$ is contained in the cube $\{x \in \mathbb{R}^d : \|x\|_\infty \leq N, x \geq 0\}$ The former set is the non-negative orthant of the ℓ_2 ball of radius N in $\mathbb{R}^d$. So, for radially symmetric functions f, integral of f over this set is 2^{-d} times its integral over the ℓ_2 ball. This justifies (S.36) after a change to polar coordinates. In the integral (S.36), noting that $s = 1/2, 2k+2 = d$, put $u = r^d$ to get

$$\sum_{i \in [N]^d} \frac{1}{\xi_i + a} \geq c \int_0^n \frac{1}{\pi^{2k+2} u/n + a}\, du = cn \log\left(1 + \pi^{2k+2} a^{-1}\right). \qquad \square$$

Lemma S.10. *Let $\{\xi_i : i = (i_1, \cdots, i_d) \in [N]^d\}$ be the eigenvalues of $D_{n,d}^{(k+1)^\mathsf{T}} D_{n,d}^{(k+1)}$ for $k \geq 0, d \geq 1, N \geq 1, n = N^d$. Let $\alpha_i, i \in [N]$ be the eigenvalues of L, the Laplacian of chain graph of length N. Define*

$$F(t) = \frac{1}{n} \sum_{i \in [N]^d} 1\{\lambda_i \leq t\}, \quad \text{for } t \in [0, \lambda_n].$$

Then there exist constants $c_1, c_2, c_3 > 0$ independent of n, t such that

$$c_1 t^{\frac{d}{2k+2}} \leq F(t) \leq c_2 + c_3 t^{\frac{d}{2k+2}}$$

for all $t \in [0, \lambda_n]$.

Proof of Lemma S.10. Using the notation in the proof of Lemma S.6,

$$\lambda_{i_1,\ldots,i_d} = \rho_{i_1} + \cdots + \rho_{i_d}, \quad \text{for } (i_1, \ldots, i_d) \in [N]^d$$

where $\rho_\ell = \beta_{\ell-k-1}, \ell \in [N]$ with the convention that $\beta_\ell = 0$ for $\ell \leq 0$. From (S.35), (S.34) and the fact that the eigenvalues of chain Laplacian are given by $4 \sin^2 \frac{\pi(\ell-1)}{2N}$ for $\ell \in [N]$, we have

$$\left(4 \sin^2 \frac{\pi(\ell-k-2)_+}{2N}\right)^{k+1} \leq \rho_i \leq \left(4 \sin^2 \frac{\pi(\ell-1)}{2N}\right)^{k+1}, \quad \text{for } \ell \in [N] \tag{S.37}$$

where $(x)_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$. The upper bound can be argued as follows.

$$\begin{aligned}
nF(t) &= \sum_{i \in [N]^d} 1\Big\{\lambda_{i_1,\ldots,i_d} \leq t\Big\} \\
&= \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d \rho_{i_j} \leq t\Big\} \\
&\leq \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d 4^{k+1} \sin^{2k+2} \frac{\pi(i_j-k-2)_+}{2N} \leq t\Big\} \\
&\leq \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d \Big(\frac{\pi}{2}\Big)^{2k+2} (i_j-k-2)_+^{2k+2} \leq tN^{2k+2}\Big\}
\end{aligned}$$

In the third line, we use (S.37) and in the fourth line, we use the fact that $\sin x \geq x/2$ for $x \in [0, \pi/2]$. Observe that

$$\Big(\frac{\pi}{2}\Big)^{2k+2} \sum_{j=1}^d (i_j-k-2)_+^{2k+2} \leq tN^{2k+2} \Rightarrow \|i\|_\infty \leq k+2+\frac{2}{\pi} N t^{\frac{1}{2k+2}}.$$

Applying this fact to the previous bound on $nF(t)$,

$$nF(t) \le \sum_{i \in [N]^d} 1\Big\{\|i\|_\infty \le k+2+\frac{2}{\pi} N t^{\frac{1}{2k+2}}\Big\}$$
$$\le \Big(k+2+\frac{2}{\pi} N t^{\frac{1}{2k+2}}\Big)^d \le 2^{d-1}(k+2)^d + 2^{2d-1}\pi^{-d} n t^{\frac{d}{2k+2}}.$$

where in the last inequality we used the fact that $(a+b)^d \le 2^{d-1}(a^d+b^d)$ for $a, b \ge 0, d \ge 1$.

The lower bound can be derived as follows. Certainly, $F(t) \ge F(0) = \kappa/n$ for $t \ge 0$. We can write

$$\begin{aligned}
nF(t) &= \sum_{i \in [N]^d} 1\big\{\lambda_{i_1,\dots,i_d} \le t\big\} \\
&= \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d \rho_{i_j} \le t\Big\} \\
&= \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d 4^{k+1} \sin^{2k+2} \frac{\pi(i_j-1)}{2N} \le t\Big\} \\
&\ge \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d \pi^{2k+2}(i_j-1)^{2k+2} \le t N^{2k+2}\Big\} \\
&= \sum_{i \in [N]^d} 1\Big\{\sum_{j=1}^d \|i-1\|_{2k+2} \le r\Big\}
\end{aligned}$$

where $r = \frac{1}{\pi} N t^{\frac{1}{2k+2}}$. In the third line, we used (S.37) and in the fourth line, we used the fact that $\sin x \le x$ for $x \ge 0$. Note that, we can inscribe a cube $\{i : \|i-1\|_\infty \le r d^{\frac{-1}{2k+2}}\}$ in the ℓ_{2k+2} body $\{i : \|i-1\|_{2k+2} \le r\}$ and the cube contains $\big(1 + \lfloor r d^{\frac{-1}{2k+2}} \rfloor\big)^d$ lattice points in $[N]^d$. Therefore, continuing to bound from the previous display,

$$nF(t) \ge \big(1 + \lfloor \frac{1}{\pi} d^{\frac{-1}{2k+2}} N t^{\frac{1}{2k+2}} \rfloor\big)^d \ge \frac{1}{\pi^d} d^{\frac{-d}{2k+2}} n t^{\frac{d}{2k+2}}.$$

where in the last inequality we used the fact $(1 + \lfloor x \rfloor)^d \ge x^d$ for $x \ge 0$. □

D Fast algorithm for degrees of freedom

From (16), given a KTF estimate $\hat{\theta}$, $\mathrm{nullity}(D_{-A})$ is an unbiased estimate of degrees of freedom of KTF where A denotes the set of rows r in D for which $(D\hat{\theta})_r \neq 0$. We give an algorithm to compute $\mathrm{nullity}(D_{-A})$ in $O(ndk)$ time. This is described in Algorithm S.1, with Algorithm S.2 describing its core rubroutines. The notation we use is as follows: $N' = N - k - 1$, and $w \in \mathbb{R}^{k+2}$ is the order $(k+1)$ difference vector.

D.1 Time complexity and correctness of the algorithm

Let A denote the set of rows r in D for which $(D\hat{\theta})_r \neq 0$. Denote the null space of D_{-A} with $\mathcal{N}$.

In step 3 of DEGREES-OF-FREEDOM, we find line segments on the lattice where θ is a kth degree polynomial. In a bit more detail, for each straight line in the lattice between opposing faces, we find segments along the line where θ is a kth degree polynomial. We call these line segments (polynomial) *pieces* in the algorithm. In 2d, MAKE-POLYNOMIAL-PIECES is called on rows and columns separately, and a piece is a part of a row or a column. An important characterization of the null space $\mathcal{N}$ is the following:

A $\theta \in \mathcal{N}$ iff it is a kth degree polynomial on all the pieces found in step 3.

If a piece has fewer than $k+2$ elements, then any θ is trivially a kth degree polynomial on the piece. Otherwise, $k+1$ values on the piece determine a polynomial on the piece.

Algorithm S.1 DEGREES-OF-FREEDOM(θ, k)

Input: fitted values $\theta \in \mathbb{R}^n$, trend filtering order $k \geq 0$
Output: estimate of degrees of freedom

1. **struct** Piece: set = false, knowns = 0, start, end $\in [N]^d$
2. pieces : Piece[], pieces-containing[i]: Piece[], set[i] : bool for $i \in [N]^d$
3. for $d' \in [d]$ for $i' \in [N]^{d-1}$:
   (a) $i \leftarrow i'$ with an extra 1 inserted before d'th entry:
   $i_j = i'_j$ for $j < d'$, $i_j = 1$ for $j = d'$, $i_j = i'_{j-1}$ for $j > d'$
   (b) MAKE-POLYNOMIAL-PIECES(θ, i, d')
4. df = 0
5. for p in pieces:
   (a) if $p.set$: continue
   (b) df $+= \max\{0, \min\{p.length, k+1\} - p.\text{knowns}\}$
   (c) SPREAD(p)
6. return df

We pretend to build a vector $\eta \in \mathcal{N}$ by making sure that η is a kth degree polynomial on the pieces. In step 5, we pretend to set the values of η in a piece p. The number of new entries in η required to determine a polynomial on piece p is shown in 5(b). Once the new entries are picked arbitrarily, all the values on the piece are determined via the constraints in $D_{-A}\eta = 0$. Then we propagate the values from this piece to other adjoining pieces in a depth-first fashion. By the end of the procedure, `df` accumulates the total number of free parameters that we can use to build such a η. The dimension of $\mathcal{N}$ is equal to the number of free parameters in the algorithm.

Time complexity. It takes $O(nd(k+1))$ time to make the polynomial pieces in line 3 of DEGREES-OF-FREEDOM. The number of pieces is at most $nd(k+1)$. Therefore the for loop in line 5 is run at most as many times. SPREAD is called on a piece exactly once and SPREAD-VERTEX is called on a node exactly once. A node is contained in a maximum of $(k+2)d$ pieces. Therefore, the total time complexity is $O(nd(k+2))$.

Correctness. Suppose the number of free parameters returned by the algorithm is f. Given the values at the f free nodes $F \subset [n]$, the values are determined at all n nodes. Further this mapping from $\mathbb{R}^f \mapsto \mathbb{R}^n$ is linear. Therefore there exists a matrix C with size $n \times f$ such that $Cb \in \mathcal{N}$ for any $b \in \mathbb{R}^f$. Further, $(Cb)_F$ is a permutation of b, because the values at free nodes are not modified by C. Therefore, there are f rows in C corresponding to the free nodes F, which when vertically stacked together form a permutation of $f \times f$ identity matrix. Therefore, the column span of C has dimension f. Hence $f \leq \dim(\mathcal{N})$. Conversely, consider any $\eta \in \mathcal{N}$. Given the entries b of η at the locations of free parameters, then the rest of the entries of η are determined by $\eta = Cb$. Therefore η must lie in the column span of C. Therefore $f = \dim(\mathcal{N})$.

E More details on optimization

Generic quadratic programming on the dual. Recall that KTF solves the following convex optimization problem:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|D\theta\|_1. \tag{S.38}$$

with $D = D^{(k+1)}_{n,d}$. The corresponding Lagrange dual problem is

$$\begin{aligned} \max_u \quad & -\frac{1}{2} u D D^\mathsf{T} u + y^\mathsf{T} D^\mathsf{T} u \\ \text{subject to} \quad & -\lambda \leq u \leq \lambda. \end{aligned} \tag{S.39}$$

Note that the dual problem is a standard quadratic program (QP) and can be solved using the interior point method (IPM) to high precision. Then the primal solution can be constructed using $\hat{\theta} = y - D^\mathsf{T} u^*$ using the optimal solution

Algorithm S.2 Subroutines used in Algorithm S.1

SPREAD-VERTEX(i)
for piece q in pieces-containing[i] :
1. if $q.set$ continue
2. q.knowns++
3. if q.knowns$>k$: SPREAD(q)

SPREAD(p : Piece)
$p.set$ = true
for vertex i on the line $[p.start, p.end]$:
1. if set[i] : continue
2. set[i] $\leftarrow$ true
3. SPREAD-VERTEX(i)

MAKE-POLYNOMIAL-PIECES($\theta, i \in [N]^d, d' \in [d]$) makes polynomial pieces on the line containing i along axis d'

1. $a_j \leftarrow \theta[i$ with $i_{d'} = j]$ for $j \in [N]$
2. while $j \leq N$
   (a) start = j, end = j
   (b) while ($j \leq N'$ and $\langle w, a[j{:}j{+}k{+}1]\rangle{=}0$)
   j++
   (c) if $j \neq$ start: end $\leftarrow j{+}k$
   (d) Piece p(i with $i_{d'}$ = start, i with $i_{d'}$ = end)
   (e) for j' in [start, end]: pieces-containing[i with $i_{d'} = j'$].add(p)
   (f) pieces.add(p)

u^*. In practice, IPM takes only a few iterations to converge,[1] but each iteration involves solving a linear system. This linear system is sparse since D is sparse — it has only $O(dkn)$ non-zero elements. However, the condition number of the linear system grows as the weights on the barrier function increase, which makes it difficult to exploit the sparsity using methods such as preconditioned conjugates gradient method. On the other hand, direct solvers such as Gaussian elimination and Cholesky decomposition can take up to $O(n^3)$. Sometimes this can be improved by exploiting the banded structure of the linear system, we will describe a particular version of the interior point method using logarithmic-barrier function.

Primal-dual interior point method. The primal-dual version of the interior point solver proposed by (Kim et al., 2009) for ℓ_1 trend filtering can be straightforwardly applied to any generalized lasso problem, including KTF. The main idea is to trace a "central path" using Newton's method with an increasing weights t on the logarithmic barrier functions. The computation is dominated by computing the search direction of the Newton step, which boils down to solving the following system of linear equations

$$\begin{bmatrix} DD^\mathsf{T} & I & -I \\ I & J_1 & 0 \\ -I & 0 & J_2 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix} = - \begin{bmatrix} DD^T u - Dy + \mu_1 - \mu_2 \\ f_1 + (1/t)\mu_1^{-1} \\ f_2 + (1/t)\mu_2^{-1} \end{bmatrix} \tag{S.40}$$

where $\mu_1, \mu_2 \in \mathbb{R}^m$ are the dual variables of the dual problem (S.39), $f_1 = u - \lambda\mathbf{1}$, $f_2 = -u - \lambda\mathbf{1}$, $J_i = \mathrm{diag}{\mu_i}^{-1}\mathrm{diag}(f_i)$ are diagonal matrices and μ_i^{-1} denotes entrywise inversion. Following the derivation of (Kim et al., 2009), we can further eliminate $\Delta\mu_1$ and $\Delta\mu_2$ and solve a linear system of the form

$$(DD^\mathsf{T} - J_1^{-1}J_2^{-1})\Delta u = -(DD^\mathsf{T}u - Dy - (1/t)f_1^{-1} + (1/t)f_2^{-1}). \tag{S.41}$$

and then construct the remainder of the solutions using

$$\Delta\mu_1 = -(\mu_1 + (1/t)f_1^{-1} + J_1^{-1}\Delta u),$$

[1] In theory it could take up to $O(n^{1/2})$ iterations to converge.

$$\Delta\mu_2 = -(\mu_2 + (1/t)f_2^{-1} - J_2^{-1}\Delta u).$$

Unlike in the trend filtering problems in 1D where (S.41) is a banded linear system with a bandwidth at most $2k+3$, in a d-dimensional grid, the linear system is the following:

$$\begin{bmatrix} (D_{1\mathrm{d}}D_{1\mathrm{d}}^\mathsf{T}) \otimes I \otimes ... \otimes I, & D_{1\mathrm{d}} \otimes D_{1\mathrm{d}}^\mathsf{T} \otimes I \otimes ... \otimes I, & \cdots, & D_{1\mathrm{d}} \otimes I \otimes ... \otimes I \otimes D_{1\mathrm{d}}^\mathsf{T}, \\ D_{1\mathrm{d}}^\mathsf{T} \otimes D_{1\mathrm{d}} \otimes I \otimes ... \otimes I, & I \otimes (D_{1\mathrm{d}}D_{1\mathrm{d}}^\mathsf{T}) \otimes I \otimes ... \otimes I, & \cdots, & I \otimes D_{1\mathrm{d}} \otimes I \otimes ... \otimes I \otimes D_{1\mathrm{d}}^\mathsf{T} \\ \vdots & \vdots & \ddots & \vdots \\ D_{1\mathrm{d}}^\mathsf{T} \otimes I \otimes ... \otimes I \otimes D_{1\mathrm{d}} & I \otimes D_{1\mathrm{d}}^\mathsf{T} \otimes I \otimes ... \otimes I \otimes D_{1\mathrm{d}} & \cdots, & I \otimes ... \otimes I \otimes (D_{1\mathrm{d}}D_{1\mathrm{d}}^\mathsf{T}) \end{bmatrix} - J_1^{-1}J_2^{-1}$$

where $D_{1\mathrm{d}} = D_{N,1}^{(k+1)}$. This is still sparse, structured, but the bandedness is on the order of $O(n^{1-1/d} + k^2)$. Moreover, the above matrix is not full rank, and the condition number of the linear system blows up as the dual variables μ_1, μ_2 converge to 0 with $t \to \infty$.

Proximal Dykstra's algorithm. Proximal Dykstra's algorithm is an operator-splitting method for solving problems of the form

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|y - \theta\|_2^2 + r_1(\theta) + r_2(\theta) + ... + r_d(\theta) \tag{S.42}$$

where $r_1, ..., r_d$ are convex but possibly non-smooth functions. We can clearly see that the regularizer in KTF decomposes into this form with

$$r_i(\theta) = \sum_{x \in Z_{n,d}} |(\Delta_{x_j^{k+1}}\theta)(x)|$$

in the notation of Section 1.2. The proximal Dykstra algorithm (see, e.g., Tibshirani, 2017) initializes $\theta^{(0)} = y, z^{(-d+1)} = ... = z^{(0)} = 0$ and then iteratively applies the following update rule for $t = 1, 2, 3, ...$:

$$\begin{aligned} \theta^{(t)} &= \mathsf{prox}_{r_{t \bmod d+1}}(\theta^{(t-1)} + z^{(t-d)}) \\ z^{(t)} &= \theta^{(t-1)} + z^{(t-d)} - \theta^{(t)}. \end{aligned}$$

where $\cdot \bmod \cdot$ is the modulo operator, and the proximal operator

$$\mathsf{prox}_r(u) = \underset{\theta}{\text{argmin}} \ \frac{1}{2}\|u - \theta\|^2 + r(\theta).$$

Note that this is equivalent to a cyclic block coordinate descent in the dual.

For KTF, each proximal problem can be parallelized (Barbero and Sra, 2018). Specifically, on a d-dim regular grid, the proximal operator of r_i further splits into solving $O(n^{1-1/d})$ 1D-trend filters of size $n^{1/d}$ in parallel. Each subproblem can be solved efficiently in $O(n^{1.5/d})$ time with the primal-dual interior point method for $k \geq 1$ (Tibshirani, 2014) and in linear time when $k = 0$ using dynamic programming (Johnson, 2013).

Douglas-Rachford splitting. Another operator-splitting method for solving KTF is through the Douglas-Rachford (DR) algorithm (Eckstein and Bertsekas, 1992). For simplicity, we will focus our discussion on the case of 2D grids. The DR algorithm generically solves the following unconstrained problem:

$$\underset{\theta}{\text{minimize}} \ f(\theta) + g(\theta) \tag{S.43}$$

for convex functions f, g. The update rules include initializing an auxiliary variable $z^{(0)} = y$ and applying the following for $t = 0, 1, 2, ...$:

$$\begin{aligned} \theta^{(t+1)} &= \mathsf{prox}_f(z^{(t)}) \\ z^{(t+1)} &= z^{(t)} + \mathsf{prox}_g(2\theta^{(t+1)} - z^{(t)}) - \theta^{(t+1)}. \end{aligned}$$

There are multiple ways of applying this to our problem. We apply the DR algorithm to the dual of the following reformulation according to (Barbero and Sra, 2018, Algorithm 9):

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\theta\|_2^2 + \left(\lambda\|D_N^{(k+1)} \otimes I\theta\|_1 - \langle\theta, y\rangle\right) + \left(\lambda\|I \otimes D_N^{(k+1)}\theta\|_1\right). \tag{S.44}$$

We refer interested readers to (Barbero and Sra, 2018) for the derivation of the dual and the conversion of the problem into one that resembles (S.43). Ultimately, the proximal operator of the conjugate function (an indicator on a certain polytope) can be evaluated using the proximal operator of the r_1 and r_2 as in the proximal Dykstra updates via the Moreau decomposition:

$$\mathsf{prox}_{r_i}(u) + \mathsf{prox}_{r_i^*}(u) = u.$$

In other words, the Douglas-Rachford algorithm enjoys the same computational benefits of the proximal Dykstra's algorithm as each proximal operator evaluation involves only solving 1D trend filtering problems in parallel.

References

Alvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *Journal of Machine Learning Research*, 19(56):1–82, 2018.

Lucien Birge and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3): 203–268, 2001.

Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 2015. Revised edition.

Nicholas Johnson. A dynamic programming algorithm for the fused lasso and l_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2): 339–360, 2009.

Enno Mammen. Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, 19(2): 741–759, 1991.

Charles P. Neuman and Dave I. Schonbach. Discrete (Legendre) orthogonal polynomials—a survey. *International Journal for Numerical Methods in Engineering*, 8(4):743–770, 1974.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, 2016.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James Sharpnack, and Ryan J. Tibshirani. Higher-total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, 2017.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

Ryan J. Tibshirani. Dykstra's algorithm, ADMM, and coordinate descent: Connections, insights, and extensions. In *Advances in Neural Information Processing Systems*, 2017.

Ryan J. Tibshirani. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. arXiv: 2003.03886, 2020.

Yu-Xiang Wang, Alexander Smola, and Ryan J. Tibshirani. The falling factorial basis and its statistical applications. In *Proceedings of the International Conference on Machine Learning*, 2014.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.