The Voronoigram: Minimax Estimation of Bounded Variation Functions From Scattered Data

Addison J. Hu   Alden Green   Ryan J. Tibshirani

Abstract

We consider the problem of estimating a multivariate function f_0 of bounded variation (BV), from noisy observations $y_i = f_0(x_i) + z_i$ made at random design points $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. We study an estimator that forms the Voronoi diagram of the design points, and then solves an optimization problem that regularizes according to a certain discrete notion of total variation (TV): the sum of weighted absolute differences of parameters θ_i, θ_j (which estimate the function values $f_0(x_i), f_0(x_j)$) at all neighboring cells i, j in the Voronoi diagram. This is seen to be equivalent to a variational optimization problem that regularizes according to the usual continuum (measure-theoretic) notion of TV, once we restrict the domain to functions that are piecewise constant over the Voronoi diagram.

The regression estimator under consideration hence performs (shrunken) local averaging over adaptively formed unions of Voronoi cells, and we refer to it as the *Voronoigram*, following the ideas in Koenker (2005), and drawing inspiration from Tukey's *regressogram* (Tukey, 1961). Our contributions in this paper span both the conceptual and theoretical frontiers: we discuss some of the unique properties of the Voronoigram in comparison to TV-regularized estimators that use other graph-based discretizations; we derive the asymptotic limit of the Voronoi TV functional; and we prove that the Voronoigram is minimax rate optimal (up to log factors) for estimating BV functions that are essentially bounded.

1 Introduction

Consider a standard nonparametric regression setting, given observations $(x_i, y_i) \in \Omega \times \mathbb{R}$, $i = 1, \ldots, n$, for an open and connected subset Ω of $\mathbb{R}^d$, and with

$$y_i = f_0(x_i) + z_i, \quad i = 1, \ldots, n, \tag{1}$$

for i.i.d. mean zero stochastic errors z_i, $i = 1, \ldots, n$. We are interested in estimating the function f_0 under the working assumption that f_0 adheres to a certain notion of smoothness. A traditional smoothness assumption on f_0 involves its integrated squared derivatives, for example, the assumption that

$$\int_\Omega \sum_{\|\alpha\|_1 = 2} (D^\alpha f)^2(x) \, dx$$

is small, where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$ is a multi-index and $D^\alpha = (\frac{\partial}{\partial x_1})^{\alpha_1} \ldots (\frac{\partial}{\partial x_d})^{\alpha_d}$ denotes the associated mixed partial derivative operator. This is the notion of smoothness that underlies the celebrated *smoothing spline* estimator in the univariate case $d = 1$ (Schoenberg, 1964) and the *thin-plate spline* estimator when $d = 2$ or 3 (Duchon, 1977). We also note that assuming f_0 is smooth in the sense of the above display is known as second-order L^2 Sobolev smoothness (where we interpret $D^\alpha f$ as a weak derivative).

Smoothing splines and thin-plate splines are quite popular and come with a number of advantages. However, one shortcoming of using these methods, i.e., to using the working model of Sobolev smoothness, is that it does not permit f_0 to have discontinuities, which limits its applicability. More broadly, smoothing splines and thin-plate splines do not fare well when the estimand f_0 possesses heterogeneous smoothness, meaning that f_0 is more smooth at some parts of its domain Ω and more wiggly at others. This motivates us to consider regularity measured by the *total variation* (TV) seminorm

$$\text{TV}(f; \Omega) = \sup \left\{ \int_\Omega f(x) \operatorname{div} \phi(x) \, dx : \phi \in C_c^1(\Omega; \mathbb{R}^d), \, \|\phi(x)\|_2 \le 1 \text{ for all } x \in \Omega \right\}, \tag{2}$$

where $C_c^1(\Omega; \mathbb{R}^d)$ denotes the space of continuously differentiable compactly supported functions from Ω to $\mathbb{R}^d$, and we use $\operatorname{div} \phi = \sum_{i=1}^d \partial \phi_i / \partial x_i$ for the divergence of $\phi = (\phi_1, \ldots, \phi_d)$. Accordingly, we define the *bounded variation* (BV) class on Ω by

$$\mathrm{BV}(\Omega) = \{f \in L^1(\Omega) : \mathrm{TV}(f; \Omega) < \infty\},$$

to contain all $L^1(\Omega)$ functions with finite TV. The definition in (2) is often called the measure-theoretic definition of multivariate TV. For simplicity we will often drop the notational dependence on Ω and simply write this as $\mathrm{TV}(f)$. This definition may appear complicated at first, but it admits a few natural interpretations, which we present next to help build intuition.

1.1 Perspectives on total variation

Below are three perspectives on total variation. The first two reveal the way that TV acts on special types of functions; the third is a general equivalent form of TV.

Smooth functions. If f is (weakly) differentiable with (weak) gradient $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})$, then

$$\mathrm{TV}(f) = \int_\Omega \|\nabla f(x)\|_2 \, dx, \tag{3}$$

provided that the right-hand here is well-defined and finite. Consider the difference between this and the first-order L^2 Sobolev seminorm

$$\int_\Omega \sum_{\|\alpha\|_1 = 1} (D^\alpha f)^2(x) \, dx = \int_\Omega \|\nabla f(x)\|_2^2 \, dx.$$

The latter uses the squared ℓ_2 norm $\|\cdot\|_2^2$ in the integrand, whereas the former (3) uses the ℓ_2 norm $\|\cdot\|_2$. It turns out that this is a meaningful difference—one way to interpret this is as a difference between L^2 and L^1 regularity. Noting that $\|x\|_1 \leq \sqrt{d}\|x\|_2$ for all $x \in \mathbb{R}^d$, the space $\mathrm{BV}(\Omega)$ contains the first-order L^1 Sobolev space

$$W^{1,1}(\Omega) = \{f \in L^1(\Omega) : \int_\Omega \|\nabla f(x)\|_1 \, dx < \infty\},$$

which, loosely speaking, contains functions that can be more locally peaked and less evenly spread out (i.e., permits a greater degree of heterogeneity in smoothness) compared to the first-order L^2 Sobolev space

$$W^{1,2}(\Omega) = \{f \in L^2(\Omega) : \int_\Omega \|\nabla f(x)\|_2^2 \, dx < \infty\}.$$

It is important to emphasize, however, that $\mathrm{BV}(\Omega)$ is still much larger than $W^{1,1}(\Omega)$, because it permits functions to have sharp discontinuities. We discuss this next.

Indicator functions. If $S \subseteq \Omega$ is a set with locally finite perimeter, then the indicator function 1_S, which we define by $1_S(x) = 1$ for $x \in S$ and 0 otherwise, satisfies

$$\mathrm{TV}(1_S) = \mathrm{per}(S), \tag{4}$$

where $\mathrm{per}(S)$ is the perimeter of S (or equivalently, the codimension-1 Hausdorff measure of ∂S, the boundary of S). Thus, we see that that TV is tied to the geometry of the level sets of the function in question. Indeed, there is a precise sense in which this is true in full generality, as we discuss next.

Coarea formula. In general, for any $f \in \mathrm{BV}(\Omega)$, we have

$$\mathrm{TV}(f) = \int_{-\infty}^\infty \mathrm{per}(\{x \in \Omega : f(x) > t\}) \, dt. \tag{5}$$

This is known as the *coarea formula* for BV functions (see, e.g., Theorem 5.9 in Evans and Gariepy, 2015). It offers a highly intuitive picture of what total variation is measuring: we take a slice through the graph of a function f, calculate

the perimeter of the set of points (projected down to the Ω-axis) that lie above this slice, and add up these perimeters over all possible slices.

The coarea formula (5) also sheds light on why BV functions are able to portray such a great deal of heterogeneous smoothness: all that matters is the total integrated amount of function growth, according to the perimeter of the level sets, as we traverse the heights of level sets. For example, if the perimeter has a component ρ that persists for a range of level set heights $[t, t + h]$, then this contributes the same amount ρh to the TV as does a smaller perimeter component $\rho/100$ that persists for a larger range of level set heights $[t, t + 100h]$. To put it differently, the former might represent a local behavior that is more spread out, and the latter a local behavior that is more peaked, but these two behaviors can contribute the same amount to the TV in the end. Therefore, a ball in the BV space—all L^1 functions f such that $\mathrm{TV}(f) \leq r$—contains functions with a huge variety in local smoothness.

1.2 Why is estimating BV functions hard?

Now that we have motivated the study of BV functions, let us turn towards the problem of estimating a BV function from noisy samples. Given the centrality of penalized empirical risk minimization in nonparametric regression, one might be tempted to solve the TV-penalized variational problem

$$\underset{f \in \mathrm{BV}(\Omega)}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \, \mathrm{TV}(f), \tag{6}$$

given data (x_i, y_i), $i = 1, \ldots, n$ from the model (1), and under the working assumption that f has small TV. However, in short, solving (6) will "not work" in any dimension $d \geq 2$, in the sense that it does not yield a well-defined estimator regardless of the choice of tuning parameter $\lambda > 0$.

When $d = 1$, solving (6) produces a celebrated estimator known as the (univariate) *TV denoising* estimator (Rudin et al., 1992) or the *fused lasso* signal approximator (Tibshirani et al., 2005). (More will be said about this shortly, under the related work subsection.) But for any $d \geq 2$, problem (6) is ill-posed, as the criterion does not achieve its infimum. To see this, consider the function

$$f_\epsilon = \sum_{i=1}^{n} y_i \cdot 1_{B(x_i, \epsilon)},$$

where $B(x_i, \epsilon)$ denotes the closed ℓ_2 ball of radius ϵ centered at x_i, and $1_{B(x_i, \epsilon)}$ denotes its indicator function (which equals 1 on $B(x_i, \epsilon)$ and 0 outside of it). Now let us examine the criterion in problem (6): for small enough $\epsilon > 0$, the function f_ϵ has a squared loss equal to 0, and has TV penalty equal to $\lambda n c_d \epsilon^{d-1}$ (here $c_d > 0$ is a constant depending only on d). Hence, as $\epsilon \to 0$, the criterion value in (6) achieved by f_ϵ tends to 0. However, as $\epsilon \to 0$, the function f_ϵ itself trivially approaches the zero function, defined as $f(x) = 0$ for all x.[1] Note that this is true for any $\lambda > 0$, whereas the zero function certainly cannot minimize the objective in (6) for all $\lambda > 0$.

The problem here, informally speaking, is that the BV class is "too big" when $d \geq 2$; more formally, the evaluation operator is not continuous over the BV space—which means that convergence in BV norm[2] does not imply pointwise convergence—when $d \geq 2$. It is worth noting that this problem is not specific to BV spaces and it occurs also with the kth order L^p Sobolev space $W^{k,p}(\Omega) = \{f \in L^p(\Omega) : \int_\Omega \sum_{\|\alpha\|_1 = k} (D^\alpha f)^p(x) \, dx < \infty\}$ when $pk \leq d$, known as the the subcritical regime. In the supercritical regime, $pk > d$, convergence in norm does imply pointwise convergence,[3] but all bets are off when $pk \leq d$. Thus, just as the TV-penalized problem (6) is ill-posed for $d \geq 2$, the more familiar thin-plate spline problem

$$\underset{f \in W^{2,2}(\Omega)}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_\Omega \|\nabla^2 f(x)\|_F^2 \, dx$$

is itself ill-posed when $d \geq 4$. (Here we use $\nabla^2 f(x)$ for the weak Hessian of f, and $\| \cdot \|_F$ for the Frobenius norm, so that the second-order L^2 Sobolev seminorm can be written as $\int_\Omega \sum_{\|\alpha\|_1 = 2} (D^\alpha f)^2(x) \, dx = \int_\Omega \|\nabla^2 f(x)\|_F^2 \, dx$.)

---

[1] Just as with L^p classes, elements in $\mathrm{BV}(\Omega)$ are actually only defined up to equivalence classes of functions. Hence, to make point evaluation well-defined in the random design model (1), we must identify each equivalence class with a representative; we use the *precise representative*, which is defined at almost every point x by the limiting local average of a function around x; see Appendix A.1 for details. It is straightforward to see that the precise representative associated with f_ϵ converges to the zero function as $\epsilon \to 0$.

[2] Traditionally defined by equipping the TV seminorm with the L^1 norm, as in $\|f\|_{\mathrm{BV}} = \|f\|_{L^1} + \mathrm{TV}(f)$.

[3] This is effectively a statement about the everywhere continuity of the precise representative, which is a consequence of Morrey's inequality; see, e.g., Theorem 4.10 in Evans and Gariepy (2015).

What can we do to circumvent this issue? Broadly speaking, previous approaches from the literature can be stratified into two types. The first maintains the smoothness assumption on $\text{TV}(f_0)$ for the regression function f_0, but replaces the sampling model (1) by a white noise model of the form

$$dY(x) = f_0(x)dx + \frac{\sigma}{\sqrt{n}}dW(x), \quad x \in \Omega,$$

where dW is a Gaussian white noise process. Given this continuous-space observation model, we can then replace the empirical loss term $\sum_{i=1}^n (y_i - f(x_i))^2$ in (6) by the squared L^2 loss $\|Y - f\|_{L^2(\Omega)}^2 = \int_\Omega (Y(x) - f(x))^2\, dx$ (or some multiscale variant of this). The second type of approach keeps the sampling model (1), but replaces the assumption on $\text{TV}(f_0)$ by an assumption on discrete total variation (which is based on the evaluations of f_0 at the design points alone) of the form

$$\text{DTV}(f_0) = \sum_{\{i,j\} \in E} w_{ij}|f_0(x_i) - f_0(x_j)|,$$

for an edge set E and weights $w_{ij} \geq 0$. We then naturally replace the penalty $\text{TV}(f)$ in (6) by $\text{DTV}(f)$. More details on both types of approaches will be given in the related work subsection.

The approach we take in the current paper sits in the middle, between the two types. Like the first, we maintain a bona fide smoothness assumption on $\text{TV}(f_0)$, rather than a discrete version of TV. Like the second, we work in the sampling model (1), and define an estimator by solving a discrete version of (6) which is always well-posed, for any dimension $d \geq 2$. In fact, the connections run deeper: the discrete problem that we solve is not constructed arbitrarily, but comes from restricting the domain in (6) to a special finite-dimensional class of functions, over which the penalty $\text{TV}(f)$ in (6) takes on an equivalent discrete form.

1.3 The Voronoigram

This brings us to the central object in the current paper: an estimator defined by restricting the domain in the infinite-dimensional problem (6) to a finite-dimensional subspace, whose structure is governed by the Voronoi diagram of the design points $x_1, \ldots, x_n \in \Omega$. In detail, let $V_i = \{x \in \Omega : \|x_i - x\|_2 < \|x_j - x\|\}$ be the Voronoi cell[4] associated with x_i, for $i = 1, \ldots, n$, and define

$$\mathcal{F}_n^{\text{V}} = \text{span}\big\{1_{V_1}, \ldots, 1_{V_n}\big\},$$

where recall 1_{V_i} is the indicator function of V_i. In words, $\mathcal{F}_n^{\text{V}}$ is a space of functions from Ω to $\mathbb{R}$ that are piecewise constant on the Voronoi diagram of $x_1, \ldots, x_n$. (We remark that this is most certainly a subspace of $\text{BV}(\Omega)$, as each Voronoi cell has locally finite perimeter; in fact, as we will see soon, the TV of each $f \in \mathcal{F}_n^{\text{V}}$ takes a simple and explicit closed form.) Now consider the finite-dimensional problem

$$\underset{f \in \mathcal{F}_n^{\text{V}}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \, \text{TV}(f). \tag{7}$$

We call the solution to (7) the *Voronoigram* and denote it by $\hat{f}^{\text{V}}$. This idea—to fit a piecewise constant function to the Voronoi tessellation of the input domain Ω—dates back to (at least) Koenker (2005), where it was briefly proposed in Chapter 7 of this book (further discussion of related work will be given in Section 1.5). It does not appear to have been implemented or studied beyond this. Its name is inspired by Tukey's classic *regressogram* (Tukey, 1961).

Of course, there a many choices for a finite-dimensional subset of $\text{BV}(\Omega)$ that we could have used for a domain restriction in (7). Why $\mathcal{F}_n^{\text{V}}$, defined by piecewise constant functions on the Voronoi diagram, as in (7)? A remarkable feature of this choice is that it yields an equivalent optimization problem

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{\{i,j\} \in E^{\text{V}}} w_{ij}^{\text{V}} \cdot |\theta_i - \theta_j|, \tag{8}$$

for an edge set E^{V} defined by neighbors in the Voronoi graph, and weights w_{ij}^{V} that measure the "length" of the shared boundary between cells V_i and V_j, to be defined precisely later (in Section 2.1). The equivalence between problems (7)

---

[4]As we have defined it, each Voronoi cell is open, and thus a given function $f \in \mathcal{F}_n^{\text{V}}$ is not actually defined on the boundaries of the Voronoi cells. But this is a set of Lebesgue measure zero, and on this set it can be defined arbitrarily—any particular definition will not affect results henceforth.

and (8) sets $\theta_i = f(x_i)$, $i = 1, \ldots, n$, and is driven by the following special fact: for such a pairing, whenever $f \in \mathcal{F}_n^{\mathrm{V}}$, it holds (proved in Section 2.1) that

$$\mathrm{TV}(f) = \sum_{\{i,j\} \in E^{\mathrm{V}}} w_{ij}^{\mathrm{V}} \cdot |\theta_i - \theta_j|. \tag{9}$$

In this way we can view the Voronoigram as marriage between a purely variational approach, which maintains the use of a continuum TV penalty on a function f, and a purely discrete approach, which instead models smoothness using a discrete TV penalty on a vector θ defined over a graph. In short, the Voronoigram does both.

1.3.1 A first look at the Voronoigram

From its equivalent discrete problem form (8), we can see that the penalty term drives the Voronoigram to have equal (or "fused") evaluations at points x_i and x_j corresponding to neighboring cells in the Voronoi tessellation. Generally, the larger the value of $\lambda \geq 0$, the more neighboring evaluations will be fused together. Due to fact that each $f \in \mathcal{F}_n^{\mathrm{V}}$ is constant over an entire Voronoi cell, this means that the Voronoigram fitted function $\hat{f}^{\mathrm{V}}$ is constant over adaptively chosen unions of Voronoi cells. Furthermore, based on what is known about solutions of generalized lasso problems (details given in Section 2.2), we can express the fitted function here as

$$\hat{f}^{\mathrm{V}} = \sum_{k=1}^{\hat{K}} (\bar{y}_k - \hat{s}_k) \cdot 1_{\hat{R}_k}, \tag{10}$$

where $\hat{K}$ is the number of connected components that appear in the solution $\hat{\theta}^{\mathrm{V}}$ over the Voronoi graph, $\hat{R}_k$ denotes a union of Voronoi cells associated with the kth connected component, $\bar{y}_k$ denotes the average of response points y_i such that $x_i \in \hat{R}_k$; and $\hat{s}_k$ is a data-driven shrinkage factor. To be clear, each of $\hat{K}$, $\hat{R}_k$, $\bar{y}_k$, and $\hat{s}_k$ here are data-dependent quantities—they fall out of the structure of the solution in problem (8).

Thus, like the regressogram, the Voronoigram estimates the regression function by fitting (shrunken) averages over local regions; but unlike the regressogram, where the regions are fixed ahead of time, the Voronoigram is able to choose its regions *adaptively*, based on the geometry of the design points x_i (owing to the use of the Voronoi diagram) and on how much local variation is present in the response points y_i (a consideration inherent to the minimization in (8), which trades off between the loss and penalty summands).

Figure 1 gives a simple example of the Voronoigram and its adaptive structure in action.
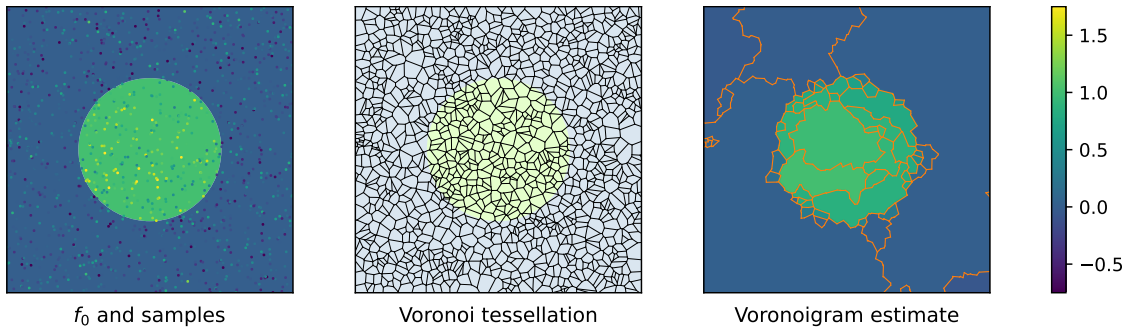

| f_0 and samples | Voronoi tessellation | Voronoigram estimate |

Figure 1: *A simple example of using the Voronoigram to estimate a function f_0, from noisy observations. Left: f_0 and noisy observations made at $n = 1274$ random points in $d = 2$ dimensions. Center: the Voronoi tessellation, whose cells constitute the piecewise constant basis for the Voronoigram. Right: the Voronoigram estimate (at a certain choice of λ), with the resulting adaptively chosen constant pieces—over which it performs averaging—outlined in orange.*

1.4 Summary of contributions

Our primary practical and methodological contribution is to motivate and study the Voronoigram as a nonparametric regression estimator for BV functions in a multivariate scattered data (random design) setting, including comparing and contrasting it to two related approaches: discrete TV regularization using ε-neighborhood or k-nearest neighbor graphs. A summary is as follows (a more detailed summary is given in Section 2.4).

- The graph used by Voronoigram—namely, the Voronoi adjacency graph—is *tuning-free*. This stands in contrast to ε-neighborhood or k-nearest neighbor graphs, which require a choice of a local radius ε or number of neighbors k, respectively.

- The Voronoigram penalty becomes *density-free* in large samples, which is term we use to describe the fact that it converges to "pure" total variation, independent of the density p of the (random) design points $x_1, \ldots, x_n$. This follows from one of our main theoretical results (reiterated below), and it stands in contrast to the TV penalties based on ε-neighborhood and k-nearest neighbor graphs, which are known to asymptotically approach particular p-weighted versions of total variation.

- The Voronoigram estimator yields a natural passage from a discrete set of fitted values $\hat{f}^{\mathrm{V}}(x_i)$, $i = 1, \ldots, n$ to a fitted function $\hat{f}^{\mathrm{V}}$ defined over the entire input domain Ω: this is simply given by local constant extrapolation of each fitted value $\hat{f}^{\mathrm{V}}(x_i)$ to its containing Voronoi cell V_i. (Equivalently, $\hat{f}^{\mathrm{V}}(x)$ is given by the 1-nearest neighbor prediction rule based on $(x_i, \hat{f}^{\mathrm{V}}(x_i))$, $i = 1, \ldots, n$.) Further, thanks to (9), we know that such an extrapolation method is *complexity-preserving*: the discrete TV of $\hat{\theta}_i^{\mathrm{V}}$, $i = 1, \ldots, n$ is precisely the same as the continuum TV of the extrapolant $\hat{f}^{\mathrm{V}}$. Other graph-based TV regularization methods do not come with this property.

On the theoretical side, our primary theoretical contributions are twofold, summarized below.

- We prove that the Voronoi penalty functional, applied to evaluations of f at i.i.d. design points $x_1, \ldots, x_n$ from a density p, converges to $\mathrm{TV}(f)$, as $n \to \infty$ (see Section 3 for details). The fact that its asymptotic limit here is independent of p is both important and somewhat remarkable.

- We carry out a comprehensive minimax analysis for L^2 estimation over $\mathrm{BV}(\Omega)$. The highlights (Section 5 gives details): for any fixed $d \geq 2$ and regression function f_0 with $\mathrm{TV}(f_0) \leq L$ and $\|f_0\|_{L^\infty} \leq M$ (where $L, M > 0$ are constants), a modification of the Voronoigram estimator $\hat{f}^{\mathrm{V}}$ in (7)—defined by simply clipping small weights w_{ij}^{V} in the penalty term—converges to f_0 at the squared L^2 rate $n^{-1/d}$ (ignoring log terms). We prove that this matches the minimax rate (up to log terms) for estimating a regression function f_0 that is bounded in TV and L^∞. Lastly, we prove that an even simpler *unweighted* Voronoigram estimator—defined by setting all edge weights in (7) to unity—also obtains the optimal rate (up to log terms), as do more standard estimators based on discrete TV regularization over ε-neighborhood and k-nearest neighbor graphs.

1.5   Related work

The work of Mammen and van de Geer (1997) marks an important early contribution promoting and studying the use of TV as a regularization functional, in univariate nonparametric regression. These authors considered a variational problem similar to (6) in dimension $d = 1$, with a generalized penalty $\mathrm{TV}(D^k f)$, the TV of the kth weak derivative $D^k f$ of f. They proved that the solution is always a spline of degree k (whose knots may lie outside the design points if $k \geq 2$) and named the solution the *locally adaptive regression spline* estimator. A related, more recent idea is *trend filtering*, proposed by Steidl et al. (2006); Kim et al. (2009) and extended by Tibshirani (2014) to the case of arbitrary design points. Trend filtering solves a discrete analog of the locally adaptive regression spline problem, in which the penalty $\mathrm{TV}(D^k f)$ is replaced by the discrete TV of the kth discrete derivative of f—based entirely on evaluations of f at the design points.

Tibshirani (2014) showed that trend filtering, like the Voronoigram, admits a special duality between discrete and continuum representations: the trend filtering optimization problem is in fact the restriction of the variational problem for locally adaptive regression splines to a particular finite-dimensional space of kth degree piecewise polynomials. The key fact underlying this equivalence is that for any function f in this special piecewise polynomial space, its continuum penalty $\mathrm{TV}(D^k f)$ equals its discrete penalty (discrete TV applied to its kth discrete derivatives), a result analogous to the property (9) of functions $f \in \mathcal{F}_n^{\mathrm{V}}$. Thus we can view the Voronoigram a generalization of this core idea, at the heart of trend filtering, to multiple dimensions—albeit restricted the case $k = 0$.

We note that similar ideas to locally adaptive regression splines and trend filtering were around much earlier; see, e.g., Schuette (1978); Koenker et al. (1994). Tibshirani (2022) provides an account of the history of these and related ideas in nonparametric smoothing, and also makes connections to numerical analysis—the study of discrete splines in particular. It is worth highlighting that when $k = 0$, the locally adaptive regression spline and trend filtering estimators coincide, and reduce to a method known as *TV denoising*, which has even earlier roots in applied mathematics (to be covered shortly).

Beyond the univariate setting, there is still a lot of related work to cover across different areas of the literature, and we break up our exposition into parts accordingly.

Continuous-space TV methods. The seminal work of Rudin et al. (1992) introduced TV regularization in the context of signal and image denoising, and has spawned to a large body of follow-up work, mostly in the applied mathematics community, where this is called the *Rudin-Osher-Fatemi* (ROF) model for TV denoising. See, e.g., Rudin and Osher (1994); Vogel and Oman (1996); Chambolle and Lions (1997); Chan et al. (2000), among many others. In this literature, the observation model is traditionally continuous-time (univariate), or continuous-space (multivariate)—this means that, rather than having observations at a finite set of design points, we have an entire observation process (deterministic or random), itself a function over a bounded and connected subset of $\mathbb{R}^d$. TV regularization is then used in a variational optimization context, and discretization usually occurs (if at all) as part of numerical optimization schemes for solving such variational problems.

Statistical analysis in continuous-space observation models traditionally assumes a white noise regression model, which has a history of study for adaptive kernel methods (via Lepski's method) or wavelet methods in particular, see, e.g., Lepski et al. (1997); Lepski and Spokoiny (1997); Neumann (2000); Kerkyacharian et al. (2001, 2008). In this general area of the literature, the recent paper of del Álamo et al. (2021) is most related to our paper: these authors consider a multiresolution TV-regularized estimator in a multivariate white noise model, and derive minimax rates for L^p estimation of TV and L^∞ bounded functions. When $p = 2$, they establish a minimax rate (ignoring log factors) of $n^{-1/d}$ on the squared L^2 error scale, for arbitrary dimension $d \geq 2$, which agrees with our results in Section 5.

Discrete, lattice-based TV methods. Next we discuss purely discrete TV-regularization approaches, in which both the observation model and the penalty are discrete, and are based on function values at a discrete sequence of points. Such approaches can be further delineated into two subsets: models and methods based on discrete TV over lattices (multi-dimensional grid graphs), and those based on discrete TV over geometric graphs (such as ε-neighborhood or k-nearest neighbor graphs constructed from the design points). We cover the former first, and the latter second.

For lattice-based TV approaches, Tibshirani et al. (2005) marks an early influential paper proposing discrete TV regularization over univariate and bivariate lattices, under the name *fused lasso*.[5] This generated much follow-up work in statistics, e.g., Friedman et al. (2007); Hoefling (2010); Tibshirani and Taylor (2011); Arnold and Tibshirani (2016), among many others. In terms of theory, we highlight Hutter and Rigollet (2016), who established sharp upper bounds for the estimation error of TV denoising over lattices, as well as Sadhanala et al. (2016), who certified optimality (up to log factors) by giving minimax lower bounds. The rate here (ignoring log factors) for estimating signals with bounded discrete TV, in mean squared error across the lattice points, is $n^{-1/d}$. This holds for an arbitrary dimension $d \geq 2$, and agrees with our results in Section 5. Interestingly, Sadhanala et al. (2016) also prove that the minimax linear rate over the discrete TV is class is constant—which means that the best estimator that is linear in the response vector $y \in \mathbb{R}^n$, of the form $\hat{f}(x) = w(x)^\mathsf{T} y$, is *inconsistent* in terms of its max risk (over signals with bounded discrete TV). We do not pursue minimax linear analysis in the current paper but expect a similar phenomenon to hold in our setting.

Lastly, we highlight Sadhanala et al. (2017, 2021), who proposed and studied an extension of trend filtering on lattices. Just like univariate trend filtering, the multivariate version allows for an arbitrary smoothness order $k \geq 0$, and reduces to TV denoising (or the fused lasso) on a lattice for $k = 0$. In the lattice setting, the theoretical picture is fairly complete: for general k, d, denoting by $s = (k + 1)/d$ the effective degree of smoothness, the minimax rate for estimating signals with bounded k^{th} order discrete TV is n^{-s} for $s \leq 1/2$, and $n^{-2s/(2s+1)}$ for $s > 1/2$. The minimax linear rates display a phase transition as well: constant for $s \leq 1/2$, and $n^{-(2s-1)/(2s)}$ for $s > 1/2$. In our setting, we do not currently have error analysis, let alone an estimator, for higher-order notions of TV smoothness (for general $k \geq 1$). With continuum TV and scattered data (random design), this is more challenging to formulate. However, the lattice-based world continues to provides goalposts for what we would hope to find in future work.

Discrete, graph-based TV methods. Turning to graph-based TV regularization methods, as explained above, much of the work in statistics stemmed from Tibshirani et al. (2005), and the algorithmic and methodological contributions cited above already considers general graph structures (beyond lattices). While one can view our proposal as a special instance of TV regularization over a graph—the Voronoi adjacency graph—it is important to recognize the independent

---

[5]The original work here proposed discrete TV regularization on the coefficients of regressor variables that obey an inherent lattice structure. If we denote the matrix of regressors by X, then a special case of this is simply $X = I$ (the identity matrix), which reduces to the TV denoising problem. In some papers, the resulting estimator is sometimes referred to as the fused lasso *signal approximator*.

and visionary work of Koenker and Mizera (2004), which served as motivation for us and intimately relates to our work. These authors began with a triangulation of scattered points in $d = 2$ dimensions (say, the Delaunay triangulation) and defined a nonparametric regression estimator called the *triogram* by minimizing, over functions f that are continuous and piecewise linear over the triangulation, the squared loss of f plus a penalty on the TV of the *gradient* of f. This is in some ways completely analogous to the problem we study, except with one higher degree of smoothness. As we mentioned in the introduction above, in Koenker (2005) the author actually proposes the Voronoigram as a lower-order analog of the triogram, but the method has not, to our knowledge, been studied beyond this brief proposal.

Outside of Koenker and Mizera (2004), existing work involving TV regularization on graphs relies on geometric graphs like ε-neighborhood or k-nearest neighbor graphs. In terms of theoretical analysis, the most relevant paper to discuss is the recent work of Padilla et al. (2020): they study TV denoising on precisely these two types of geometric graphs (ε-neighborhood and k-nearest neighbor graphs), and prove that it achieves an estimation rate in squared L^2 error of $n^{-1/d}$, but require that f_0 is more than TV bounded—they require it to satisfy a certain piecewise Lipschitz assumption. Although we primarily study TV regularization over the Voronoi adjacency graph, we build on some core analysis ideas in Padilla et al. (2020). In doing so, we are able to prove that the Voronoigram achieves the squared L^2 error rate $n^{-1/d}$, and we only require that $\mathrm{TV}(f_0)$ and $\|f_0\|_{L^\infty}$ are bounded (with the latter condition actually necessary for nontrivial estimation rates over BV spaces when $d \geq 2$, as we explain in Section 5.1). Furthermore, we are able to generalize the results of Padilla et al. (2020), and we prove that the TV-regularized estimator over ε-neighborhood and k-nearest neighbor graphs achieves the same rate under the same assumptions, removing the need for the piecewise Lipschitz condition. See Remark 9 for a more detailed discussion. We also mention that earlier ideas from Wang et al. (2016); Padilla et al. (2018) are critical analysis tools in Padilla et al. (2020) and critical for our analysis as well.

Finally, we would like to mention the recent and parallel work of Green et al. (2021a,b), which studies regularized estimators by discretizing Sobolev (rather than TV) functionals over neighborhood graphs, and establishes results on estimation error and minimaxity entirely complementary to ours, but with respect to Sobolev smoothness classes.

2 Methods and basic properties

In this section, we discuss some basic properties of our primary object of study, the Voronoigram, and compare these properties to those of related methods.

2.1 The Voronoigram and TV representation

We start with a discussion of the property behind (9)—we call this a *TV representation* property of functions in $\mathcal{F}_n^{\mathrm{V}}$, as their total variation over Ω can be represented exactly in terms of their evaluations over $x_1, \ldots, x_n$.

Proposition 1. *For any $x_1, \ldots, x_n$, with Voronoi tessellation $V_1, \ldots, V_n$, and any $f \in \mathcal{F}_n^{\mathrm{V}} = \mathrm{span}\{1_{V_1}, \ldots, 1_{V_n}\}$ of the form*

$$f = \sum_{i=1}^n \theta_i \cdot 1_{V_i},$$

it holds that

$$\mathrm{TV}(f) = \sum_{i,j=1}^n \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j) \cdot |\theta_i - \theta_j|, \tag{11}$$

where $\mathcal{H}^{d-1}$ denotes Hausdorff measure of dimension $d - 1$, and ∂V_i denotes the boundary of V_i.

The proof of this proposition follows from the measure-theoretic definition (2) of total variation, and we defer it to Appendix A.2. In a sense, the above result is a natural extension of the property that the TV of an indicator function is the perimeter of the underlying set, recall (4).

Note that (11) in Proposition 1 reduces to the property (9) claimed in the introduction, once we define weights

$$w_{ij}^{\mathrm{V}} = \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j), \quad i, j = 1, \ldots, n, \tag{12}$$

and define the edge set E^{V} to contain all $\{i, j\}$ such that $w_{ij}^{\mathrm{V}} \neq 0$. In words, each w_{ij}^{V} is the surface measure (length, in dimension $d = 2$) of the shared boundary between V_i and V_j. We say that i, j are adjacent with respect to the Voronoi diagram provided that $w_{ij}^{\mathrm{V}} \neq 0$. Using this nomenclature, we can think of E^{V} as the set of all adjacent pairs i, j. This

defines a weighted undirected graph on $\{1, \ldots, n\}$, which we call the *Voronoi adjacency graph* (the Voronoi graph for short). We denote this by $G^{\mathrm{V}} = ([n], E^{\mathrm{V}}, w^{\mathrm{V}})$, where here and throughout we write $[n] = \{1, \ldots, n\}$.

Backing up a little further, we remark that (12) also provides the remaining details needed to completely describe the Voronoigram estimator in (7). By the TV representation property (9), we see that we can equivalently express the penalty in (7) as that in (8), which certifies the equivalence between the two problems. Of course, since (9) is true of all functions in $\mathcal{F}_n^{\mathrm{V}}$, it is also true of the Voronoigram solution $\hat{f}^{\mathrm{V}}$. Hence, to summarize the relationship between the discrete (8) and continuum (7) problems, once we solve for the Voronoigram fitted values $\hat{\theta}_i^{\mathrm{V}} = \hat{f}^{\mathrm{V}}(x_i)$, $i = 1, \ldots, n$ at the design points, we extrapolate via

$$\hat{f}^{\mathrm{V}} = \sum_{i=1}^{n} \hat{\theta}_i^{\mathrm{V}} \cdot 1_{V_i}, \quad \text{which satisfies} \quad \mathrm{TV}(\hat{f}^{\mathrm{V}}) = \sum_{\{i,j\} \in E^{\mathrm{V}}} w_{ij}^{\mathrm{V}} \cdot |\hat{\theta}_i^{\mathrm{V}} - \hat{\theta}_j^{\mathrm{V}}|. \tag{13}$$

In other words, the continuum TV of the extrapolant $\hat{f}^{\mathrm{V}}$ is exactly the same as the discrete TV of the vector of fitted values $\hat{\theta}^{\mathrm{V}}$. This is perhaps best appreciated when discussed relative to alternative approaches based on discrete TV regularization on graphs, which do not generally share the same property. We revisit this in Section 2.4.

2.2 Insights from generalized lasso theory

Consider a generalized lasso problem of the form:

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \|D\theta\|_1, \tag{14}$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is a response vector and $D \in \mathbb{R}^{m \times n}$ is a penalty operator (as problem (14) has identity design matrix, it is hence sometimes also called a generalized lasso signal approximator problem). The Voronoigram is a special case of a generalized lasso problem: that is, problem (8) can be equivalently expressed in the more compact form (14), once we take $D = D^{\mathrm{V}}$, the edge incidence operator of the Voronoi adjacency graph. In general, given an weighted undirected graph $G = ([n], E, w)$, we denote its edge incidence operator $D(G) \in \mathbb{R}^{m \times n}$; recall that this is a matrix whose number of rows equals the number of edges, $m = |E|$, and if edge ℓ connects nodes i and j, then

$$\big[D(G)\big]_{\ell k} = \begin{cases} +w_{ij} & k = i \\ -w_{ij} & k = j \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Thus, to reiterate the equivalence using the notation just introduced, the penalty operator in the generalized lasso form (14) of the Voronoigram (8) is $D^{\mathrm{V}} = D(G^{\mathrm{V}})$, the edge incidence operator of the Voronoi graph G^{V}. And, as is clear from the discussion, the Voronoigram is not just an instance of an arbitrary generalized lasso problem, it is an instance of TV denoising on a graph. Alternative choices of graphs for TV denoising will be discussed in Section 2.3.

What does casting the Voronoigram in generalized lasso form do for us? It enables us to use existing theory on the generalized lasso to read off results about the structure and complexity of Voronoigram estimates. Tibshirani and Taylor (2011, 2012) show the following about the solution $\hat{\theta}$ in problem (14): if we denote by $A = \{i \in [n] : (D\hat{\theta})_i \neq 0\}$ the active set corresponding to $D\hat{\theta}$, and $s = \mathrm{sign}((D\hat{\theta})_A)$ the active signs, then we can write

$$\hat{\theta} = P_{\mathrm{null}(D_{-A})}(y - \lambda D_A^{\mathsf{T}} s). \tag{16}$$

where D_A is the submatrix of D with rows that correspond to A, D_{-A} is the submatrix with the complementary set of rows, and $P_{\mathrm{null}(D_{-A})}$ is the projection matrix onto $\mathrm{null}(D_{-A})$, the null space of D_{-A}. When we take $D = D(G)$, the edge incidence operator on a graph G, the null space D_{-A} has a simple analytic form that is spanned by indicator vectors on the connected components of the subgraph of G that is induced by removing the edges in A. This allows us to rewrite (16), for a generic TV denoising estimator $\hat{\theta} = \hat{\theta}(G)$, as

$$\big[\hat{\theta}(G)\big]_i = \sum_{k=1}^{\hat{K}} (\bar{y}_k - \hat{s}_k) \cdot 1\big\{i \in \hat{C}_k\big\}, \quad i = 1, \ldots, n \tag{17}$$

where $\hat{K}$ is the number of connected components of the subgraph of G induced by removing edges in A, $\hat{C}_k$ denotes the kth such connected component, $\bar{y}_k$ denotes the average of points y_i such that $i \in \hat{C}_k$, and $\hat{s}_k$ denotes the average of the values $\lambda(D_A^{\mathsf{T}} s)_i$ over $i \in \hat{C}_k$.

9

What is special about the Voronoigram is that (17), combined with the structure of $\mathcal{F}_n^{\mathrm{V}}$ (piecewise constant functions on the Voronoi diagram), leads to an analogous piecewise constant representation on the *original input domain* Ω, as written and discussed in (10) in the introduction. Here each $\hat{R}_k = \{V_i : i \in \hat{C}_k\}$, the union of Voronoi cells of points in connected component $\hat{C}_k$.

Beyond local structure, we can learn about the complexity of the Voronoigram estimator—vis-a-vis its *degrees of freedom*—from generalized lasso theory. In general, the (effective) degrees of freedom of an arbitrary estimator $\hat{\theta}$ is defined (Efron, 1986; Hastie and Tibshirani, 1990) as:

$$\mathrm{df}(\hat{\theta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{\theta}_i, y_i),$$

where $\sigma^2 = \mathrm{Var}(z_i)$ denotes the noise variance in the data model (1). Tibshirani and Taylor (2011, 2012) show using Stein's lemma (Stein, 1981) that when each $z_i \sim N(0, \sigma^2)$ (i.i.d. for $i = 1, \ldots, n$), it holds that

$$\mathrm{df}(\hat{\theta}) = \mathbb{E}[\mathrm{nullity}(D_{-A})], \tag{18}$$

where $\mathrm{nullity}(D_{-A})$ is the nullity (dimension of the null space) of D_{-A}, and recall A is the active set corresponding to $D\hat{\theta}$. For $D = D(G)$ and $\hat{\theta} = \hat{\theta}(G)$, the TV denoising estimator over a graph G, the result in (18) reduces to

$$\mathrm{df}(\hat{\theta}(G)) = \mathbb{E}[\text{\# of connected components in } \hat{\theta}(G)]. \tag{19}$$

As a short interlude, we note that this somewhat remarkable because the connected components are adaptively chosen in the graph TV denoising estimator, and yet it does not appear that we "pay extra" for this data-driven selection in (19). This is due to the ℓ_1 penalty that appears in the TV denoising criterion, which induces a "counterbalancing" shrinkage effect—recall we fit shrunkage averages, rather than averages, in (17). For more discussion, see Tibshirani (2015).

The result (19) is true of any TV denoising estimator, including the Voronoigram. However, what is special about the Voronoigram is that we are able to write this purely in terms of the fitted function $\hat{f}^{\mathrm{V}}$:

$$\mathrm{df}(\hat{\theta}^{\mathrm{V}}) = \mathbb{E}[\text{\# of locally constant regions in } \hat{f}^{\mathrm{V}}], \tag{20}$$

because by construction the number of locally constant regions in $\hat{f}^{\mathrm{V}}$ is equal to the number of connected components in $\hat{\theta}^{\mathrm{V}}$.[6]

2.3 Alternatives: ε-neighborhood and kNN graphs

We now review two more standard graph-based alternatives to the Voronoigram: TV denoising over ε-neighborhood and k-nearest neighbor (kNN) graphs. Discrete TV over such graphs has been studied by many, including Wang et al. (2016) (experimentally), and García Trillos and Slepčev (2016); García Trillos (2019); Padilla et al. (2020) (formally). The general recipe is to run TV denoising over a graph $G = ([n], E, w)$ formed using the design points $x_1, \ldots, x_n$. We note that it suffices to specify the weight function here, since the edge set is simply defined by all pairs of nodes that are assigned nonzero weights. For the ε-*neighborhood graph*, we take

$$w_{ij}^{\varepsilon} = \begin{cases} 1 & \|x_i - x_j\|_2 \leq \varepsilon \\ 0 & \text{otherwise,} \end{cases} \quad i, j = 1, \ldots, n, \tag{21}$$

where $\varepsilon > 0$ is a user-defined tuning parameter. For the (symmetrized) k-*nearest neighbor graph*, we take

$$w_{ij}^{k} = \begin{cases} 1 & \|x_i - x_j\|_2 \leq \max\{\|x_i - x_{(k)}(x_i)\|_2, \|x_j - x_{(k)}(x_j)\|_2\} \\ 0 & \text{otherwise,} \end{cases} \quad i, j = 1, \ldots, n, \tag{22}$$

where $x_{(k)}(x_i)$ denotes the element of $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ that is k^{th} closest in ℓ_2 distance to x_i (breaking ties arbitrarily, if needed), and $k \in [n]$ is a user-defined tuning parameter.

---

[6]For this to be true, strictly speaking, we require that for each i and j in different connected components with respect to the subgraph defined by the active set A of $\hat{\theta}^{\mathrm{V}}$, we have $\hat{\theta}_i \neq \hat{\theta}_j$. However, for any fixed λ, this occurs with probability one if the response vector y is drawn from a continuous probability distribution; see Tibshirani and Taylor (2012); Tibshirani (2013).

We denote the resulting graphs by G^ε and G^k, respectively, and the resulting graph-based TV denoising estimators by $\hat{\theta}^\varepsilon = \hat{\theta}(G^\varepsilon)$ and $\hat{\theta}^k = \hat{\theta}(G^k)$, respectively. To be explicit, these solve (14) when the penalty operators are taken to be the relevant edge incidence operators $D = D(G^\varepsilon)$ and $D = D(G^k)$, respectively.

It is perhaps worth noting that the ε-neighborhood graph is a special case of a *kernel graph* whose weight function is of the form $w_{ij} = K(\|x_i - x_j\|_2)$ for a kernel function K. Though we choose to analyze the ε-neighborhood graph for simplicity, much of our theoretical development for TV denoising on this graph carries over to more general kernel graphs, with suitable conditions on K. We remark that the kNN and Voronoi graphs do not fit neatly in kernel form, as the weight they assign to i, j depends not only x_i, x_j but also on $x_1, \ldots, x_n$. That said, in either case the graph weights are well-approximated by kernels asymptotically; see Appendix B for the effective kernel for the Voronoi graph.

2.4   Discussion and comparison of properties

We begin with some similarities, starting by recapitulating the properties discussed in the second-to-last subsection: all three of $\hat{\theta}^\varepsilon$, $\hat{\theta}^k$, and $\hat{\theta}^V$—the TV denoising estimators on the ε-neighborhood, kNN, and Voronoi graphs, respectively—have adaptively chosen piecewise constant structure, as per (17) (though to be clear, they will have generically different connected components for the same response vector y and tuning parameter λ). All three estimators also have a simple unbiased estimate for their degrees of freedom, as per (19). And lastly, all three are given by solving a highly structured convex optimization problems for which a number of efficient algorithms exist; see, e.g., Osher et al. (2005); Chambolle and Darbon (2009); Goldstein et al. (2010); Hoefling (2010); Chambolle and Pock (2011); Tibshirani and Taylor (2011); Landrieu and Obozinski (2015); Wang et al. (2016).

A further notable property that all three estimators share, which has not yet been discussed, is *rotational invariance*. This means that, for any orthogonal $U \in \mathbb{R}^{d \times d}$, if we were to replace each design point x_i by $\tilde{x}_i = U x_i$ and recompute the TV denoising estimate using the ε-neighborhood, kNN, or Voronoi graphs (and with the same response vector y and tuning parameter λ) then it will remain unchanged. This is true because the weights underlying these three graphs—as we can see from (12), (21), and (22)—depend on the design points only through the pairwise ℓ_2 distances $\|x_i - x_j\|_2$, which an orthogonal transformation preserves.

We now turn to a discussion of the differences between these graphs and their use in denoising.

Auxiliary tuning parameters.   TV denoising over the ε-neighborhood and k-nearest neighbor graphs each have an "extra" tuning parameter when compared the Voronoigram: a tuning parameter associated with learning the graph itself (ε and k, respectively). This auxiliary tuning parameter must be chosen carefully in order for the discrete TV penalty to be properly behaved; as usual, we can turn to theory (e.g., García Trillos and Slepčev, 2016; García Trillos, 2019) to prescribe the proper asymptotic scaling for such choices, but in practice these are really just guidelines. Indeed, as we vary ε and k, we can typically find an observable practical impact on the performance of TV denoising estimators using their corresponding graphs, especially for the ε-neighborhood graph (for which ε impacts connectedness). One may see this by comparing the results of Section 4 to those of Appendix C. All in all, the need to appropriately choose auxiliary tuning parameters when using these graphs for TV denoising is a complicating factor for the practitioner.

Connectedness.   A related practical consideration: only the Voronoi adjacency graph is guaranteed to be connected (cf. Lemma S.16 in the appendix), while the kNN and ε-neighborhood graphs have varying degrees of connectedness depending on their auxiliary parameter. In particular, the ε-neighborhood graph is susceptible to isolated points. This can be problematic in practice: having many connected components and in particular having isolated points prevents the estimator from properly denoising, leading to degraded performance. This phenomenon is studied in Section 4.3, where the ε-neighborhood graph, grown to have roughly the same average degree as the Voronoi adjacency and kNN graphs, sees worse performance when used in TV denoising. A workaround is to grow the ε-neighborhood graph to be denser; but of course this increases the computational burden in learning the estimator and storing the graph.

Computation.   On computation of the graphs themselves, the Voronoi diagram of n points in d dimensions has worst-case complexity of $O(n \log n + n^{\lceil d/2 \rceil})$ (Aurenhammer and Klein, 2000).[7] In applications, this worst-case complexity

---

[7]Note that the Voronoi adjacency graph as considered in Section 2.1 intersects the Voronoi diagram with the domain Ω on which the n points are sampled, which incurs the additional step of checking whether each vertex of the Voronoi diagram belongs in Ω. For simple domains (say, the unit cube), this can be done in constant time for each edge as they are enumerated during graph construction.

may be pessimistic; for example, Dwyer (1991) finds that the Voronoi diagram of n points sampled uniformly at random from the d-dimensional unit ball may be computed in linear expected time.

On the other hand, the $O(n \log n + n^{\lceil d/2 \rceil})$ runtime does not include calculation of the weights (12) on the edges of the Voronoi adjacency graph, which significantly increases the computational burden, especially in higher dimensions (it is essentially intractable for $d \geq 4$). One alternative is to simply use the unweighted Voronoi adjcacency graph for denoising—dropping the weights w_{ij}^{V} in the summands in (8) but keeping the same edge structure—which we will see, in what follows, has generally favorable practical and theoretical (minimax) performance.

Construction of the ε-neighborhood and kNN graphs, in a brute-force manner, has complexity $O(dn^2)$ in each case. The complexity of building the k-nearest neighbor graph can be improved to $O(dn \log n)$ by using k-d trees (Friedman et al., 1977). This is dominated by initial cost of building the k-d tree itself, so a practitioner seeking to tune over the number of nearest neighbors is able to build kNN graphs at different levels of density relatively efficiently. As far as we know, there is no analogous general-purpose algorithmic speedup for the ε-neighborhood graph, but practical speedups may be possible by resorting to approximation techniques (for example, using random projections or hashing).

Extrapolation. A central distinction between the Voronoigram and TV denoising methods based on ε-neighborhood and kNN graphs is that the latter methods are purely discrete, which means that—as defined—they really only produce fitted values (estimates of the underlying regression function values) at the design points, and not an entire fitted function (an estimate of the underlying function). Meanwhile, the Voronoigram produces a fitted function *via* the fitted values at the design points. Recall the equivalence between problems (7) and (8), and the central property between the discrete and continuum estimates highlighted in (13)—to rephrase once again, this says that $\hat{f}^{\mathrm{V}}$ is just as complex in continuous-space (as measured by continuum TV) as $\hat{\theta}^{\mathrm{V}}$ is in discrete-space (as measured by discrete TV).

We note that it would also be entirely natural to extend the fitted values $\hat{\theta}_i = \hat{f}(x_i)$, $i = 1, \ldots, n$ from TV denoising using the ε-neighborhood or kNN graph as a piecewise constant function over the Voronoi cells $V_1, \ldots, V_n$,

$$\hat{f} = \sum_{i=1}^{n} \hat{\theta}_i \cdot 1_{V_i}.$$

To see this, observe that this is nothing more than the ubiquitous 1-nearest neighbor (1NN) prediction rule performed on the fitted values,

$$\hat{f}(x) = \hat{f}(x_i), \quad \text{where } \|x - x_i\|_2 = \min_{j=1,\ldots,n} \|x - x_j\|_2.$$

However, this extension $\hat{f}$ does not generally satisfy the property that its continuum TV is equal to the graph-based TV of $\hat{\theta}$ (with respect to the original geometric graph, be it ε-neighborhood or kNN). The complexity-preserving property in (13) of the Voronoigram is truly special.[8]

We finish by summarizing two more points of comparison for discrete TV on the Voronoi graph versus ε-neighborhood and kNN graphs. These will come to light in the theory developed later, but are worth highlighting now. First, discrete TV on the Voronoi adjacency graph, the ε-neighborhood graph, and the kNN graph can be said to each track different population-level quantities—the most salient difference being that discrete TV on a Voronoi graph in the large-sample limit does not depend on the distribution of the design points, unlike the other two graphs (compare (24) to (25) and (26)). Second, while TV denoising on all three graphs obtains the minimax error rate for functions that are bounded in TV and L^∞, on the ε-neighborhood and the kNN graphs TV denoising is furthermore manifold adaptive, and it is not clear the same is true of the Voronoigram (see Remark 9 following Theorem 2).

3 Asymptotics for graph TV functionals

Having introduced, discussed, and compared graph-based formulations of total variation—with respect to the Voronoi, k-nearest neighbor, and ε-neighborhood graphs—a natural question remains: as we grow the number of design points n used to construct the graphs, do these discrete notions of TV approach particular continuum notions of TV? Answers

---

[8]In fact, this occurs for not one but two natural notions of complexity: TV, as in (13), and degrees of freedom, as in (20). The latter says that $\hat{f}^{\mathrm{V}}$ has just as many locally constant regions (connected subsets of Ω) as $\hat{\theta}^{\mathrm{V}}$ has connected components (with respect to the Voronoi adjacency graph). This is not true in general for the 1NN extensions fit to TV denoising estimates on ε-neighborhood or kNN graphs; see Section 4.4 and Figure 6 in particular.

to these questions, aside from being of fundamental interest, will help us better understand the effects of using these different graph-based TV regularizers in the context of nonparametric regression.

The asymptotic limits for the TV functional over the ε-neighborhood and k-nearest neighbor graphs have in fact already been derived by García Trillos and Slepčev (2016) and García Trillos (2019), respectively. These results are reviewed in Remark 2, following the presentation of our main result in this section, Theorem 1, on the asymptotic limit for the TV functional over the Voronoi graph. First, we introduce some helpful notation. Given $G = ([n], E, w)$, a weighted undirected graph, we denote its corresponding discrete TV functional by

$$\text{DTV}(\theta; w) = \sum_{\{i,j\} \in E} w_{ij} |\theta_i - \theta_j|. \tag{23}$$

Given $x_1, \ldots, x_n \in \Omega$, and $f : \Omega \to \mathbb{R}$, we also use the shorthand $f(x_{1:n}) = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$.

Next we introduce an assumption that we require on the sampling distribution of the random design points.

Assumption A1. The design distribution has density p (with respect to Lebesgue measure), which is bounded away from 0 and ∞ uniformly on $\Omega = (0, 1)^d$; that is, there exist constants $p_{\min}, p_{\max}$ such that

$$0 < p_{\min} \leq p(x) \leq p_{\max} < \infty, \quad \text{for all } x \in \Omega.$$

We are now ready to present our main result in this section.

Theorem 1. *Assume that $x_1, \ldots, x_n$ are i.i.d. from a distribution satisfying Assusmption A1, and additionally assume its density p is Lipschitz: $|p(y) - p(x)| \leq L\|y - x\|_2$ for all $x, y \in \Omega$ and some constant $L > 0$. Consider the Voronoi graph whose edge weights are defined in (12). For any fixed $d \geq 2$ and $f \in C^2(\Omega)$, as $n \to \infty$, it holds that*

$$\text{DTV}\left(f(x_{1:n}); w^{\text{V}}\right) \to c_d \int_\Omega \|\nabla f(x)\|_2 \, dx, \tag{24}$$

in probability, where c_d is the constant

$$c_d = \frac{\eta_{d-2}^2}{d-1} \int_0^\infty \int_0^\infty t^d s^{d-2} \exp\left(-\mu_d \left\{\frac{t^2}{4} + s^2\right\}^{d/2}\right) ds \, dt,$$

and η_{d-2} denotes the Hausdorff measure of the $(d-2)$-dimensional unit sphere, and μ_d the Lebesgue measure of the d-dimensional unit ball.

The proof of Theorem 1 is long and involved and deferred to Appendix B. A key idea in the proof is show that the weights (12) have an asymptotically equivalent kernel form, for a particular (closed-form) kernel that we refer to as the *Voronoi kernel*. We believe this result is itself significant and may be of independent interest.

We now make some remarks.

Remark 1. The assumption that f is twice continuously differentiable, $f \in C^2(\Omega)$, in Theorem 1 is used to simplify the proof; we believe this can be relaxed, but we do not attempt to do so. It is worth recalling that under this condition, the right-hand side in (24) is a scaled version of the TV of f, since in this case $\text{TV}(f) = \int_\Omega \|\nabla f(x)\|_2 \, dx$.

Remark 2. The fact that the asymptotic limit of the Voronoi TV functional is *density-free*, meaning the right-hand side in (24) is (a scaled version of) "pure" total variation and does not depend on p, is somewhat remarkable. This stands in contrast to the asymptotic limits of TV functionals defined over ε-neighborhood and kNN graphs, which turn out to be density-weighted versions of continuum total variation. We transcribe the results of García Trillos and Slepčev (2016) and García Trillos (2019) to our setting, to ease the comparison. From García Trillos and Slepčev (2016), for the ε-neighborhood weights (21) and any sequence $\varepsilon = \varepsilon_n$ satisfying certain scaling conditions, it holds as $n \to \infty$ that

$$\frac{1}{n^2 \varepsilon_n^{d+1}} \text{DTV}\left(f(x_{1:n}); w^\varepsilon\right) \to c_d' \int_\Omega \|\nabla f(x)\|_2 \, p^2(x) \, dx, \tag{25}$$

in a particular notion of convergence, for a constant $c_d' > 0$. From García Trillos (2019), for the kNN weights (22) and any sequence $k = k_n$ satisfying certain scaling conditions, defining $\bar{\varepsilon}_n = (k_n/n)^{1/d}$, it holds as $n \to \infty$ that

$$\frac{1}{n^2 \bar{\varepsilon}_n^{d+1}} \text{DTV}\left(f(x_{1:n}); w^k\right) \to c_d'' \int_\Omega \|\nabla f(x)\|_2 \, p^{1-1/d}(x) \, dx, \tag{26}$$

13

again in a particular notion of convergence, and for a constant $c_d'' > 0$.

These differences have interesting methodological interpretations. First, recall that traditional regularizers used in nonparametric regression—which includes those in smoothing splines, thin-plate splines, and locally adaptive regression splines, trend filtering, RKHS estimators, and so on—are not design-density dependent. In this way, the Voronoigram adheres closer to the statistical mainstream than TV denoising on ε-neighborhood or kNN graphs, since the regularizer in the Voronoigram tracks "pure" TV in large samples. Furthermore, by comparing (25) to (24) we see that, relative to the Voronoigram, TV denoising on the ε-neighborhood graph does not assign as strong a penalty to functions that are wiggly in low-density regions and smoother in high-density regions. TV denoising on the k-nearest neighbor graph lies in between the two: the density p appears in (26), but raised to a smaller power than in (25).

We may infer from this scenarios in which density-weighted TV denoising would be favorable to density-free TV denoising and vice versa. In a sampling model where the underlying regression function exhibits more irregularity in a low-density region of the input space, we would expect a density-weighted method to perform better since the density weighting provides a larger effective "budget" for the penalty, leading to greater regularization and variance reduction overall. Conversely, in a sampling model where the regression function exhibits greater irregularity in a high-density region, we would expect a density-free method to have a comparative advantage because the density weighting gives rise to a smaller "budget", hampering the ability to properly regularize. In Section 4, we consider sampling models that reflect these qualities and assess the performance of each method empirically.

Remark 3. It is worth noting that it should be possible to remove the density dependence in the asymptotic limits for the TV functionals over the ε-neighborhood and kNN graphs. Following seminal ideas in Coifman and Lafon (2006), we would first form an estimate $\hat{p}$ of the design density p, and then we would reweight the ε-neighborhood and kNN graphs to precisely cancel the dependence on p in their limiting expressions. Under some conditions (which includes consistency of $\hat{p}$) this should guarantee that the asymptotic limits are density-free, that is, in our case, the reweighted ε-neighborhood and kNN discrete TV functionals converge to "pure" TV.

4   Illustrative empirical examples

In this section, we empirically examine the properties elucidated in the last section. We first investigate whether the large sample behavior of the three graph-based TV functionals of interest matches the prediction from asymptotics. We then examine the use of each as a regularizer in nonparametric regression. Our experiments are not intended to be comprehensive, but are meant to tease out differences that arise from the interplay between the density of the design points and regions of wiggliness in the regression function.

4.1   Basic experimental setup

Throughout this section, our experiments center around a single function, in dimension $d = 2$: the indicator function of a ball of radius $r_0 = \frac{1}{4}$ centered at $x_0 = (\frac{1}{2}, \frac{1}{2}) \in \mathbb{R}^2$,

$$f_0 = 1\{x \in B(x_0, r_0)\}, \tag{27}$$

supported on $\Omega = (0, 1)^2$. This is depicted in the upper display of Figure 2 using a wireframe plot.

We also consider three choices for the distribution P of the design points $x_1, \ldots, x_n$, supported on Ω.

1. "Low inside tube": the sampling density p is 0.295 on an annulus A centered at x_0 that has inner radius $r_0 - 0.1$ and outer radius $r_0 + 0.1$. (The density on $\Omega \setminus A$ is set to a constant value such that p integrates to 1.)

2. "High inside tube": the sampling density p is 1.2 on A (with again a constant value chosen on $\Omega \setminus A$ such that p integrates to 1).

3. "Uniform": the sampling distribution is uniform on Ω.

We illustrate these sampling distributions empirically by drawing $n = 1274$ observations from each and plotting them on the lower set of plots in Figure 2. We note that the "high" density value of 1.2 for the "high inside tube" sampling distribution yields an empirical distribution that—by eye—is indistinguishable from the empirical distribution formed from uniformly drawn samples. However, as we will soon see, this departure from uniform is nonetheless large enough that the large sample behavior of the TV functionals on Voronoi adjacency, ε-neighborhood, and k-nearest neighbor graphs admit discernable differences.
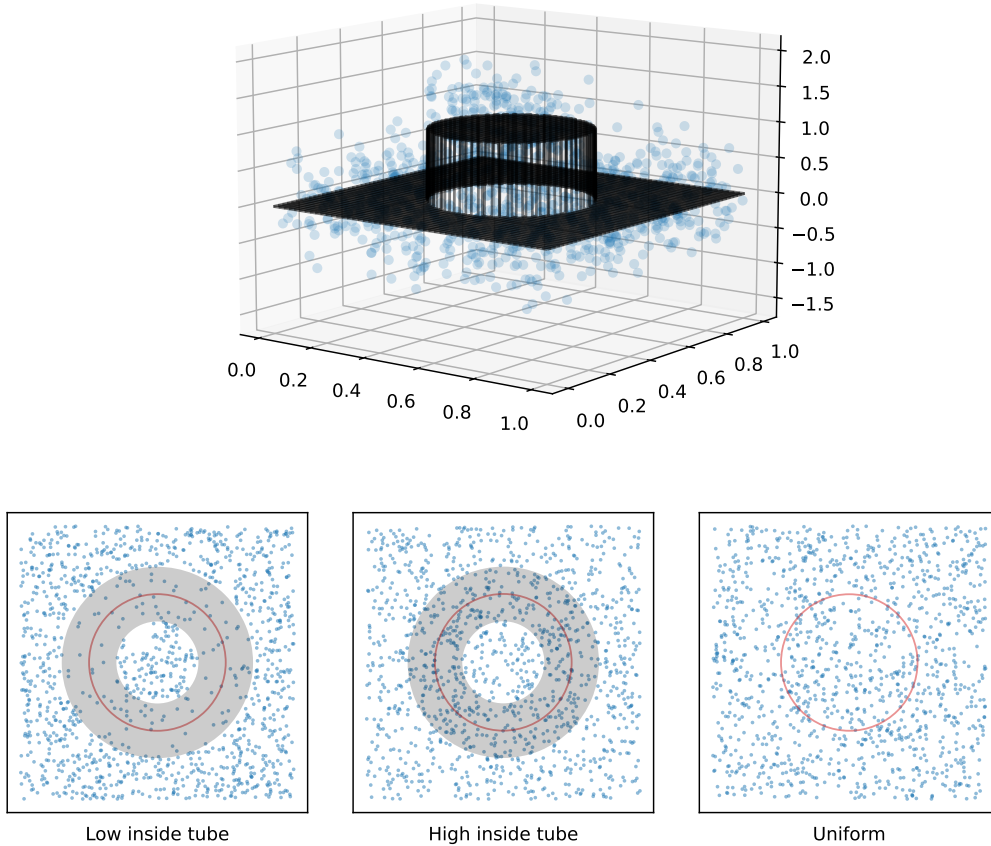
14

Figure 2: *Illustration of the basic experimental setup used in this section. Top: the function f_0 in (27) depicted using a wireframe plot, along with $n = 1274$ noisy evaluations of f_0 in blue (the noise level is set such that the signal-to-noise ratio is 1). Bottom row: $n = 1274$ samples from each of the three design distributions considered. The boundary of the set $B(x_0, r_0)$ is denoted in red, and the annulus A is shaded in translucent gray.*

4.2 Total variation estimation

We examine the of use of the Voronoi adjacency, k-nearest neighbor, and ε-neighborhood graphs, built from a random sample of design points, to estimate the total variation of the function f_0 in (27). To be clear, here we compute (using the notation (23) introduced in the asymptotic limits section):

$$\mathrm{DTV}\left(f_0(x_{1:n}); w\right) = \sum_{\{i,j\} \in E} w_{ij}|f_0(x_i) - f_0(x_j)|,$$

for three choices of edge weights w: Voronoi (12), ε-neighborhood (21), and kNN (22).

We let the number of design points n range from 10^2 to 10^5, logarithmically spaced, with 20 repetitions independently drawn from each design distribution for each n. The k-nearest neighbor graph is built using $k = \lfloor C_1 \log^{1.1} n \rfloor$, and the ε-neighborhood graph is built using $\varepsilon = C_2(\log^{1.1} n/n)^{1/2}$, where C_1, C_2 are constants chosen such that the average degree of these graphs is roughly comparable to the average degree of the Voronoi adjacency graph (which has no tuning parameter). We note that it is possible to obtain marginally more stable results for the k-nearest neighbor and ε-neighborhood graphs by taking C_1, C_2 to be larger, and thus making the graphs denser. These results are deferred to Appendix C, though we remark that the need to separately tune over such auxiliary parameters to obtain more stable results is a disadvantage of the kNN and ε-neighborhood methods (recall also the discussion in Section 2.4).

Figure 3 shows the results under the three design distributions outlined previously. For each sample size n and for each graph, we plot the average discrete TV, and its standard error, with respect to the 20 repetitions. We additionally plot the limiting asymptotic values predicted by the theory—recall (24), (25), (26)—as horizontal lines. Generally, we can see that the discrete TV, as measured by each of the three graphs, approaches its corresponding asymptotic limit. The standard error bars for the Voronoi graph tend to be the narrowest, whereas those for the kNN and ε-neighborhood graphs are generally wider. In the rightmost plot, showing the results under uniform sampling, the asymptotic limits of the discrete TV for the three methods match, since the density weighting is nullified by the uniform distribution.
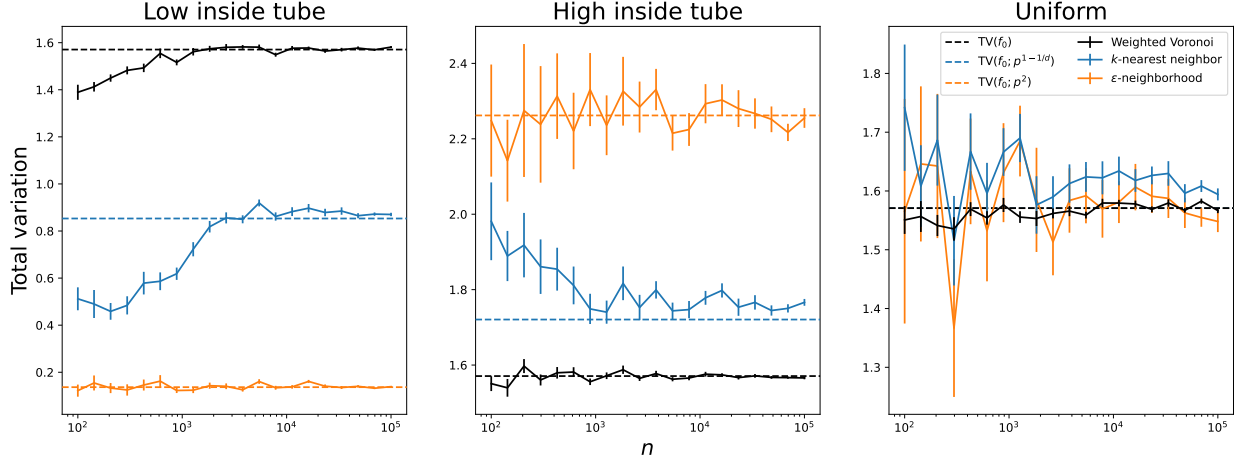

Figure 3: *Results from the TV estimation experiment ("weighted Voronoi" refers to the usual Voronoi adjacency graph, with weights in (12), and is used to distinguish it from the Voronoi adjacency graph with unit edge weights, which will appear in later experiments). We see that the discrete TV as measured by each graph converges to its asymptotic limit, drawn as a dashed horizontal line, as n grows (note that the x-axis is on a log scale).*

To give a qualitative sense of their differences, Figure 4 displays the graphs from each of the methods for a draw of $n = 1274$ samples under each sampling distribution. Note that the Voronoi adjacency and kNN graphs are connected (this is always the case for the former), whereas this is not true of the ε-neighborhood graph (recall Section 2.4), with the most noticable contrast being in the "low inside tube" sampling model. This relates to the notion that the Voronoi and kNN graphs effectively use an adaptive local bandwidth, versus the fixed bandwidth used by the ε-neighborhood graph. Comparing the former two (Voronoi and kNN graphs), we also see that there are fewer "holes" in the Voronoi graph as it has the quality that it seeks neighbors "in each direction" for each design point.

4.3 Regression function estimation

Next we study the use of discrete TV from the Voronoi, k-nearest neighbor, and ε-neighborhood graphs as a penalty in a nonparametric regression estimator. In other words, given noisy observations as in (1) of the function f_0 in (27), we solve the graph TV denoising problem (14) with penalty operator D equal to the edge incidence matrix corresponding to the Voronoi (12), ε-neighborhood (21), and kNN (22) graphs.

We fix $n = 1274$, and draw each $z_i \sim N(0, \sigma^2)$, where the noise level $\sigma^2 > 0$ is chosen so that the signal-to-noise ratio, defined as

$$\text{SNR} = \frac{\text{Var}(f_0(x_i))}{\sigma^2},$$

is equal to 1. (Here $\text{Var}(f_0(x_i))$ denotes the variance of $f_0(x_i)$ with respect to the randomness from drawing $x_i \sim P$.) Each graph TV denoising estimator is fit over a range of values for the tuning parameter λ, and at value of λ we record the $L^2(P_n)$ mean squared error

$$\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(x_i) - f_0(x_i)\right)^2.$$

Figure 5 shows the average of this $L^2(P_n)$ error, along with its standard error, across the 20 repetitions. The x-axis is parametrized by an estimated degrees of freedom for each λ value, to place the methods on common footing—that is,
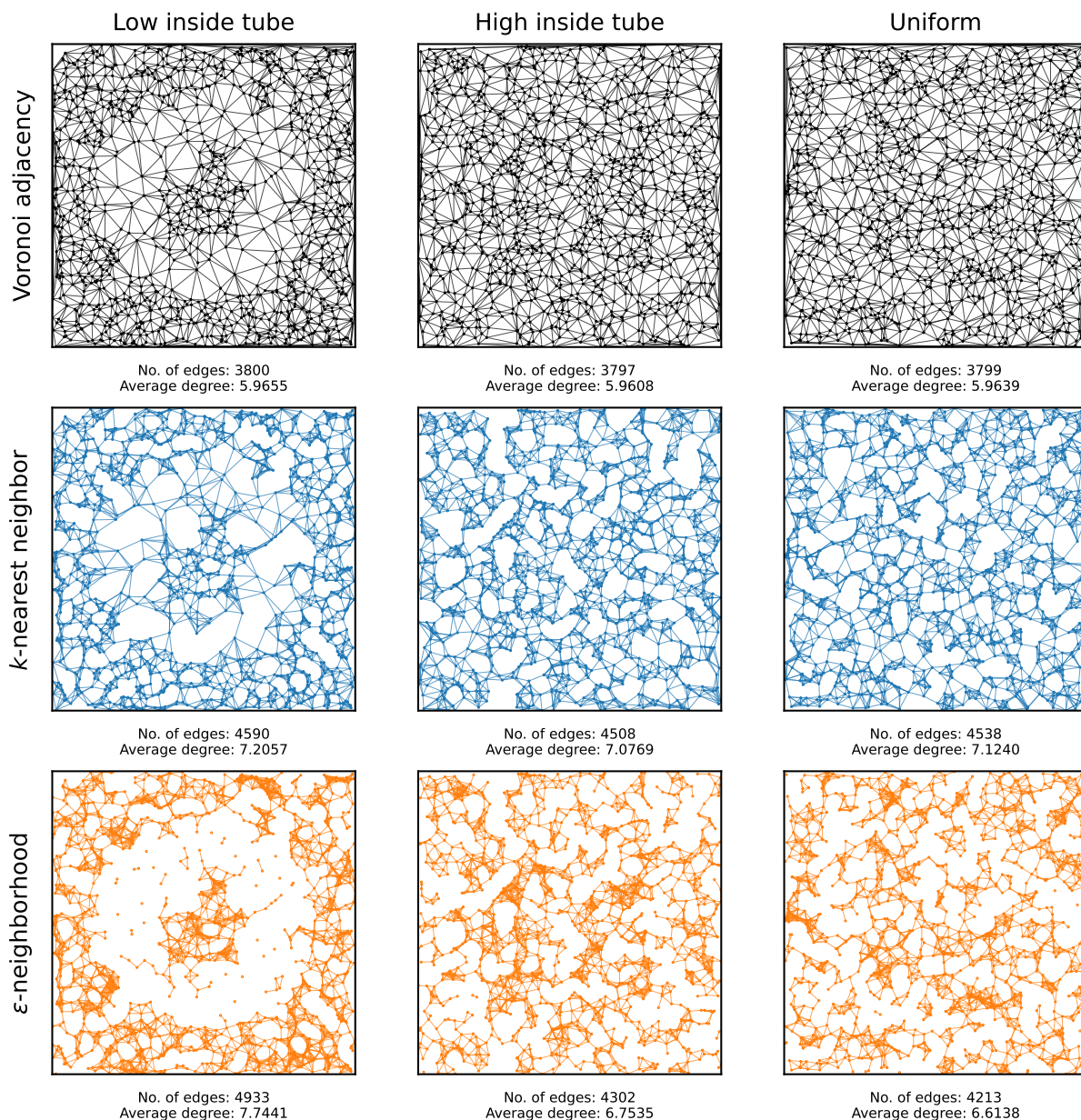
Figure 4: *Visualization of the Voronoi, kNN, and ε-neighborhood graphs for a sample of* $n = 1274$ *design points from each of the three sampling distributions considered. We see qualitatively very different behaviors in these three graph models, and we can also intuit the different asymptotic limits of their discrete TV functionals; for example, the strong dependence of the ε-neighborhood graph on the sampling density is quite noticeable in the "low inside tube" setting (bottom left plot).*

recalling the general formula in (19) for any TV denoising estimator, we convert each value of λ to the average number of resulting connected components over the 20 repetitions.

The results of Figure 5 broadly align with the expectations set forth at the end of Section 3: the density-weighted methods (using kNN and ε-neighborhood graphs) perform better when the irregularity is concentrated in a low density area ("low inside tube"), and the density-free method (the Voronoigram) does better when the irregularity is concentrated in a high density area ("high inside tube"). We also observe that across all settings, the best performing estimator tends to be the most parsimonious—the one that consumes the fewest degrees of freedom when optimally tuned.
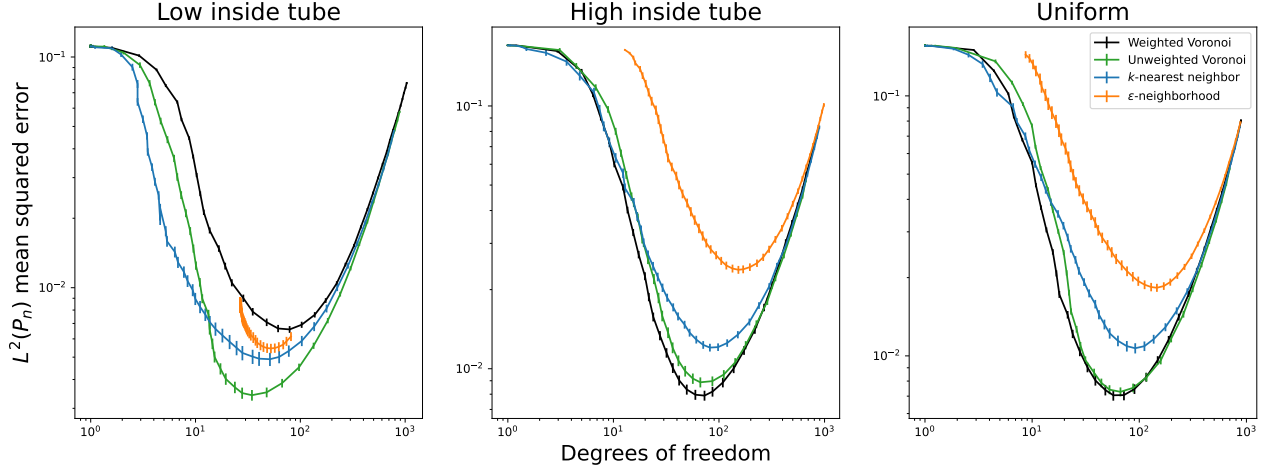

Figure 5: *Results from the function estimation experiment ("weighted Voronoi" refers to the usual Voronoi graph and "unweighted Voronoi" the graph with the same edge structure but unit edge weights). We see that the density-weighted methods—TV denoising over the kNN and ε-neighborhood graphs—generally do better in the "low inside tube" setting, where the irregularity in f_0 is concentrated in a low density region of the design distribution. Conversely, density-free method—TV denoising on the Voronoi graph, also known as the Voronoigram—does better in the "high inside tube" scenario, where irregularity is concentrated in a high density region. Lastly, TV denoising on the unweighted Voronoi graph does very well in each scenario.*

In the "low inside tube" setting (leftmost panel of Figure 5), we see that ε-neighborhood graph total variation does worse than its kNN counterpart, even though we would have expected the former to outperform the latter (because it weights the density more heavily; cf. (25) and (26)). The poor performance of TV denoising over the ε-neighborhood graph may be ascribed to the large number of disconnected points (see Figures 4 and 6), whose fitted values it cannot regularize. Such isolated points are also the reason why the minimal degrees of freedom obtained by this estimator (as $\lambda \to \infty$) is larger than that for TV denoising over the kNN and Voronoi graphs, across all settings. In Appendix C, we carry out a sensitivity analysis where we grow the ε-neighborhood and kNN graphs more densely, while retaining a comparable average degree (to each other). There we find that the performance of the estimators becomes comparable (the ε-neighborhood graph still has some disconnected points), which further emphasizes the peril of graph denoising methods that permit isolated points.

Interestingly, under the uniform sampling distribution (rightmost panel of Figure 5), where the asymptotic limits of the discrete TV functionals over the Voronoi, kNN, and ε-neighborhood graph are the same, we see that the Voronoigram performs best in mean squared error, which is encouraging empirical evidence in its favor.

Finally, Figure 5 also displays the error of the *unweighted Voronoigram*, which we use to refer to TV denoising on the unweighted Voronoi graph, obtained by setting each $w_{ij}^{\mathrm{V}} = 1$ in (8). This is somewhat of a "surprise winner"—it performs close to the best in each of the sampling scenarios, and is computationally cheaper than the Voronoigram (it avoids the expensive step of computing the Voronoi edge weights, which require surface area calculations). We lack an asymptotic characterization for discrete TV on the unweighted Voronoi graph, thus we cannot provide a strong a priori explanation for the favorable performance of the unweighted Voronoigram across our experimental suite. Nevertheless, in view of the example adjacency graphs in Figure 4, we hypothesize that its favorable performance is due in part to the adaptive local bandwidth inherent to the Voronoi graph, which seeks neighbors "in each direction" while avoiding edge crossings. Moreover, in Section 5 we show that the unweighted Voronoigram shares the property of minimax rate optimality (for estimating functions bounded in TV and L^∞), further strengthening its case.

4.4 Extrapolation: from fitted values to functions

As the last part of our experimental investigations, we consider extrapolating the graph TV denoising estimators, which represent a sequence of fitted values at the design points: $\hat{f}(x_i)$, $i = 1, \ldots, n$, to a entire fitted function: $\hat{f}(x)$, $x \in \Omega$. As discussed and motivated in Section 2.4, we use the 1NN extrapolation rule for each estimator. This is equivalently viewed as piecewise constant extrapolation over the Voronoi tessellation.
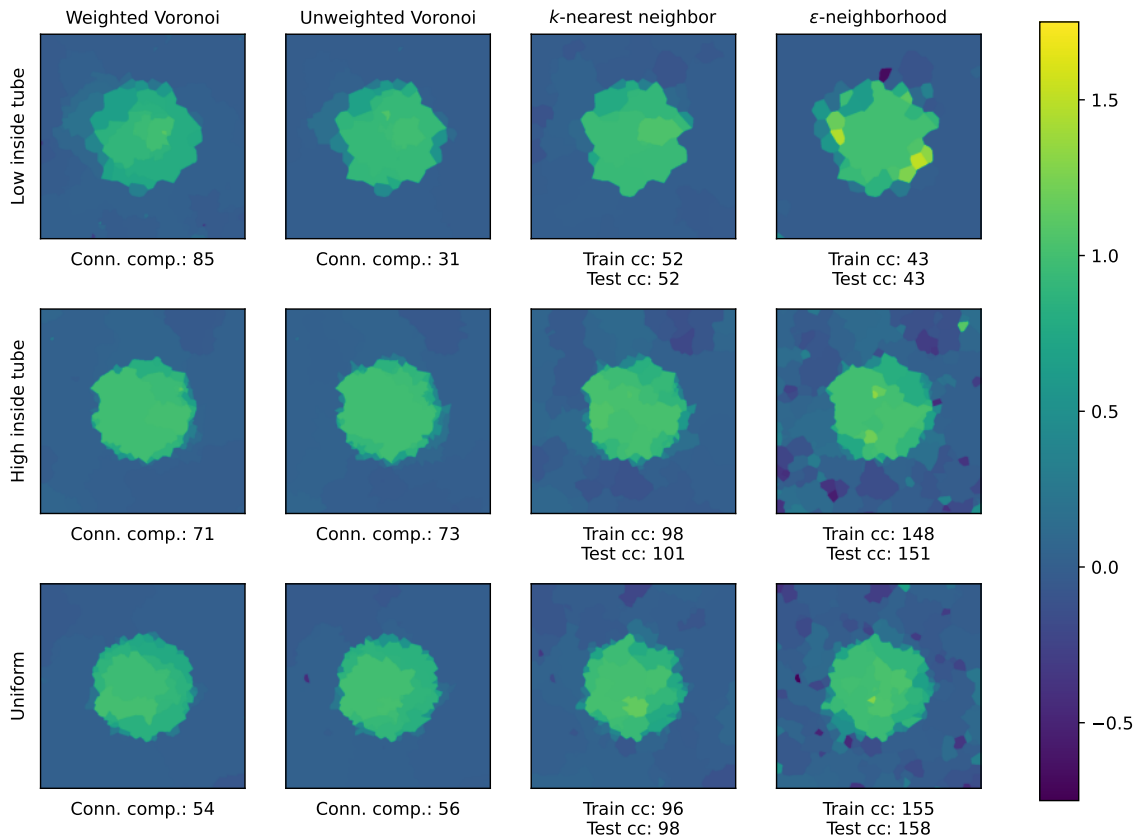

Figure 6: *Extrapolants from graph TV denoising estimates, using 1NN extrapolation. We can see several qualitative differences, for example, the issues posed by isolated points in the ε-neighborhood graph. We also note that the number of connected components in the graph used to learn the estimator (which gives an unbiased estimate of its degrees of freedom) is guaranteed to match the number of connected components in the extrapolant* only *for the Voronoi methods.*

Figure 6 plots the extrapolants for each TV denoising estimator, fitted over a particular sample of $n = 1274$ points from each design distribution. In each case, the estimator was tuned to have optimal mean squared error (cf. Figure 5). From these visualizations, we are able to clearly understand where certain estimators struggle; for example, we can see the effect of isolated components in the ε-neighborhood graph in the "low inside tube" setting, and to a lesser extent in the "high inside tube" and uniform sampling settings too. As for the Voronoigram, we previously observed (cf. Figure 5 again) that it struggles in the "low inside tube" setting due to the large weights placed on edges crossing the annulus, and in the upper left plot of Figure 6 we see "patchiness" around the annulus, where large jumps are heavily penalized, rather than sharper jumps made by other estimators (including its unweighted sibling). This is underscored by the large number of connected components in the Voronoigram versus others in the "low inside tube" setting.

Lastly, because the partition induced by the 1NN extrapolation rule is exactly the Voronoi diagram, we note that the number of connected components on the training set $\{x_1, \ldots, x_n\}$—as measured by connectedness of the fitted values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ over the Voronoi graph—always matches the number of connected components on the test set Ω—as measured by connectedness of the extrapolant $\hat{f}$ over the domain Ω. This is not true of TV denoising over the kNN and ε-neighborhood graphs, where we can see a mismatch between connectedness pre- and post-extrapolation.

5 Estimation theory for BV classes

In this section, we analyze error rates for estimating f_0 given data as in (1), under the assumption that f_0 has bounded total variation. Thus, of central interest will be a (seminorm) ball in the BV space, which we denote by

$$\mathrm{BV}(L) = \{f \in L^1(\Omega) : \mathrm{TV}(f) \le L\}.$$

For simplicity, here and often throughout this section, we suppress the dependence on the domain Ω when referring to various function classes of interest. We use P for the design distribution, and we will primarily be interested in error in the $L^2(P)$ norm, defined as

$$\|\hat{f} - f_0\|_{L^2(P)}^2 = \int \big(\hat{f}(x) - f_0(x)\big)^2 \, dP(x).$$

We also use P_n for the empirical distribution of sample $x_1, \dots, x_n$ of design points, and we will also be concerned with error in the $L^2(P_n)$ norm, defined as

$$\|\hat{f} - f_0\|_{L^2(P_n)}^2 = \frac{1}{n} \sum_{i=1}^n \big(\hat{f}(x_i) - f_0(x_i)\big)^2.$$

We will generally use the terms "error" and "risk" interchangeably. Finally, we will consider the following assumptions, which we refer to as *the standard assumptions*.

- The data (x_i, y_i), $i = 1, \dots, n$ are i.i.d. following (1), where each $z_i \sim N(0, \sigma^2)$.

- The design points are drawn from a distribution P that satisfies Assumption A1.

- The dimension satisfies $d \ge 2$ and remains fixed as $n \to \infty$.

Note that under Assumption A1, asymptotic statements about $L^2(P)$ and $L^2(\mu)$ errors are equivalent, with μ denoting Lebesgue measure (the uniform distribution) on Ω, since it holds that $p_{\min}\|g\|_{L^2(\mu)}^2 \le \|g\|_{L^2(P)}^2 \le p_{\max}\|g\|_{L^2(\mu)}^2$ for any function g.

5.1 Impossibility result without L^∞ boundedness

A basic issue to explain at the outset is that, when $d \ge 2$, consistent estimation over the BV class $\mathrm{BV}(L)$ is *impossible* in $L^2(P)$ risk. This is in stark contrast to the univariate setting, $d = 1$, in which TV-penalized least squares (Mammen and van de Geer, 1997; Sadhanala and Tibshirani, 2019), and various other estimators, offer consistency.

One way to see this is through the fact that $\mathrm{BV}(\Omega)$ does not compactly embed into $L^2(\Omega)$ for $d \ge 2$, which implies that L^2 estimation over $\mathrm{BV}(L)$ is impossible (see Section 5.5 of Johnstone, 2015 for a discussion of this phenomenon in the Gaussian sequence model). We now state this impossibility result and provide a more constructive proof, which sheds more light on the nature of the problem.

Proposition 2. *Under the standard assumptions, there exists a constant $c > 0$ (not depending on n) such that*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BV}(1) \cap L^2(\Omega)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \ge c > 0,$$

where the infimum is taken over all estimators $\hat{f}$ that are measurable functions of the data (x_i, y_i), $i = 1, \dots, n$.

Proof. As explained above, under Assumption A1 we may equivalently study $L^2(\mu)$ risk, which we do henceforth in this proof. We simply denote $\|\cdot\|_{L^2} = \|\cdot\|_{L^2(\mu)}$. Consider the two-point hypothesis testing problem of distinguishing

$$H_0 : f_0^\star = 0 \quad \text{versus} \quad H_1 : f_1^\star = \frac{\epsilon^{-d/2}}{2d} \cdot 1_{(0,\epsilon)^d},$$

where $0 < \epsilon < 1$. By construction, $f \in L^2(\Omega)$ and $\mathrm{TV}(f) \le 1$ for each of $f = f_0^\star$ and $f = f_1^\star$. Additionally, we have $\|f_0^\star - f_1^\star\|_{L^2} = \frac{1}{2d}$. It follows from a standard reduction that

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BV}(1) \cap L^2(\Omega)} \mathbb{E}\|\hat{f} - f_0\|_{L^2} \ge \inf_{\hat{f}} \sup_{f_0 \in \{f_0^\star, f_1^\star\}} \mathbb{E}\|\hat{f} - f_0\|_{L^2}$$

$$\ge \inf_{\psi} \Big(\mathbb{P}_{H_0}(\psi = 1) + \mathbb{P}_{H_1}(\psi = 0) \Big), \tag{28}$$

where the infimum in the rightmost expression is over all measurable tests ψ. Now, conditional on the event

$$\mathcal{E} = \{x_i \notin (0, \epsilon)^d, \, i = 1, \ldots, n\},$$

the distributions are the same under null and alternative hypotheses, $\mathbb{P}_{H_0}(\cdot|\mathcal{E}) = \mathbb{P}_{H_1}(\cdot|\mathcal{E})$. Additionally, note that we have $\mathbb{P}(\mathcal{E}) \geq (1 - p_{\max}\epsilon^d)^n$ under Assumption A1. Consequently, for any test ψ,

$$\begin{aligned} \mathbb{P}_{H_1}(\psi = 1) &= \mathbb{P}_{H_1}(\psi = 1|\mathcal{E})\mathbb{P}(\mathcal{E}) + \mathbb{P}_{H_1}(\psi = 1|\mathcal{E}^c)\mathbb{P}(\mathcal{E}^c) \\ &\leq \mathbb{P}_{H_1}(\psi = 1) + 1 - (1 - p_{\max}\epsilon^d)^n \\ &= \mathbb{P}_{H_0}(\psi = 1) + 1 - (1 - p_{\max}\epsilon^d). \end{aligned}$$

In other words, just rearranging the above, we have shown that

$$\mathbb{P}_{H_0}(\psi = 1) + \mathbb{P}_{H_1}(\psi = 0) \geq (1 - p_{\max}\epsilon^d).$$

Taking $\epsilon \to 0$, and plugging this back into (28), establishes the desired result. $\qquad\square$

The proof of Proposition 2 reveals one reason why consistent estimation over $\mathrm{BV}(L)$ is not possible: when $d \geq 2$, functions of bounded variation can have "spikes" of arbitrarily small width but large height, which cannot be witnessed by any finite number of samples. (We note that this has nothing to do with noise in the response, and the proposition still applies in the noiseless case with $\sigma = 0$.) This motivates a solution: in the remainder of this section, we will rule out such functions by additionally assuming that f_0 is bounded in L^∞.

5.2 Minimax error: upper and lower bounds

Henceforth we assume that f_0 has bounded TV *and* has bounded L^∞ norm, that is, we consider the class

$$\mathrm{BV}_\infty(L, M) = \{f \in L^1(\Omega) : \mathrm{TV}(f) \leq L, \, \|f\|_{L^\infty} \leq M\}.$$

Here $\|\cdot\|_{L^\infty} = \|\cdot\|_{L^\infty(\Omega)}$ is the essential supremum norm on Ω. Perhaps surprisingly, additionally assuming that f_0 is bounded in L^∞ dramatically improves prospects for estimation. The following theorem shows that two different and simple modifications of the Voronoigram, appropriately tuned, each achieve a $n^{-1/d}$ rate of convergence in its sup risk over $\mathrm{BV}_\infty(L, M)$, modulo log factors.

Theorem 2. *Under the standard assumptions, consider either of the following modified Voronoigram estimators $\hat{\theta}$:*

- *the minimizer in the Voronoigram problem (8), once we replace each weight w_{ij}^{V} by a clipped version defined as $\tilde{w}_{ij}^{\mathrm{V}} = \max\{c_0 n^{-(d-1)/d}, w_{ij}^{\mathrm{V}}\}$, for any constant $c_0 > 0$.*

- *the minimizer in the Voronoigram problem (8), once we replace each weight w_{ij}^{V} by 1.*

Let $\lambda = c\sigma\tau_n(\log n)^{1/2+\alpha}$ for any $\alpha > 1$ and a constant $c > 0$, where $\tau_n = n^{(d-1)/d}$ for the clipped weights estimator and $\tau_n = 1$ for the unit weights estimator. There exists another constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$, the estimated function $\hat{f} = \sum_{i=1}^n \hat{\theta}_i \cdot 1_{V_i}$ (which is piecewise constant over the Voronoi diagram) satisfies

$$\mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \leq C\left(\frac{\sigma L(\log n)^{5/2+\alpha+1/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}}\right). \tag{29}$$

We now certify that this upper bound is tight, up to log factors, by providing a complementary lower bound.

Theorem 3. *Under the standard assumptions, provided that n, L, M satisfy $c_0(M^2 n)^{-\frac{(d-1)}{d}} \leq L \leq C_0(M^2 n)^{1/d}$ for constants $C_0 > c_0 > 0$, the minimax risk satisfies*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BV}_\infty(L, M)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \geq CLM(M^2 n)^{-1/d}, \tag{30}$$

for another constant $C > 0$, where the infimum is taken over all estimators $\hat{f}$ that are measurable functions of the data (x_i, y_i), $i = 1, \ldots, n$.

Taken together, Theorems 2 and 3 establish that the minimax rate of convergence over $\mathrm{BV}_\infty(1,1)$ is $n^{-1/d}$, modulo log factors. Further, after subjecting it to minor modifications—either clipping small edge weights, or setting all edge weights to unity (the latter being particularly desirable from a computational point of view)—the Voronoigram is minimax rate optimal, again up to log factors.

The proof of the lower bound (30) is Theorem 3 is fairly standard and can be found in Appendix D. The proof of the upper bound (29) in Theorem 2 is much more involved, and the key steps are described over Sections 5.3 and 5.4 (with the details deferred to Appendix D). Before moving on to key parts of the analysis, we make several remarks.

Remark 4. It is not clear to us whether clipping small weights in the Voronoigram penalty as we do in Theorem 2 (via $\tilde{w}_{ij}^{\mathrm{V}} = \max\{c_0 n^{-(d-1)/d}, w_{ij}^{\mathrm{V}}\}$) is actually needed, or whether the unmodified estimator (8) itself attains the same or a similar upper bound, as in (29). In particular, it may be that under Assumption A1, the surface area of the boundaries of Voronoi cells (defining the weights) are already lower bounded in rate by $n^{-(d-1)/d}$, with high probability; however this is presently unclear to us.

Remark 5. The design points *must be random* in order to have nontrivial rates of convergence in our problem setting. If $x_1, \ldots, x_n$ were instead fixed, then for $d \geq 2$ and any n it is possible to construct $f_0 \in \mathrm{BV}_\infty(1,1)$ with $f_0(x_i) = 0$, $i = 1, \ldots, n$ and (say) $\|f\|_{L^2} = 1/2$. Standard arguments based on reducing to a two-point hypothesis testing problem (as in the proof of Proposition 2) reveal that the minimax rate in L^2 is trivially lower bounded by a constant, rendering consistent estimation impossible once again.

This is completely different from the situation for $d = 1$, where the minimax risks under fixed and random design models for TV bounded functions are basically equivalent. Fundamentally, this is because for $d \geq 2$ the space $\mathrm{BV}(\Omega)$ does not compactly embed into $C^0(\Omega)$, the space of continuous functions (whereas for $d = 1$, all functions in $\mathrm{BV}(\Omega)$ possess at least an approximate form of continuity). Note carefully that this is a different issue than the failure of $\mathrm{BV}(\Omega)$ to compactly embed into $L^2(\Omega)$, and that it is not fixed by intersecting a TV ball with an L^∞ ball.

Remark 6. We can generalize the definition of total variation in (2), by generalizing the norm we use to constrain the "test" function ϕ to an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. (See (S.1) in the appendix.) The original definition in (2) uses the ℓ_2 norm, $\|\cdot\| = \|\cdot\|_2$. What would minimax rates look if we used a different choice of norm to define TV? Suppose that we use an ℓ_p norm, for any $p \geq 1$; that is, suppose we take $\|\cdot\| = \|\cdot\|_p$ as the norm to constrain the "test" functions in the supremum. Then under this change, the minimax rate will still remain $n^{-1/d}$, just as in Theorems 2 and 3. This is simply due to the fact that ℓ_p norms are equivalent on $\mathbb{R}^d$ (thus a unit ball in the TV-ℓ_p seminorm will be sandwiched in between two balls in TV-ℓ_2 seminorm of constant radii).

Remark 7. The minimax rate for estimating a Lipschitz function, that is, the minimax rate over the class

$$\mathrm{Lip}(L) = \{f : \Omega \to \mathbb{R} \,:\, |f(x) - f(z)| \leq L\|x - z\|_2 \text{ for all } x, z \in \Omega\},$$

is $n^{-2/(2+d)}$ in squared L^2 risk, for constant $L > 0$ (not growing with n); see, e.g., Stone (1982). When $d = 2$, this is equal to $n^{-1/2}$, implying that the minimax rates for estimation over $\mathrm{Lip}(1)$ and $\mathrm{BV}_\infty(1,1)$ match (up to log factors). This is despite the fact that $\mathrm{Lip}(1)$ is a strict subset of $\mathrm{BV}_\infty(1,1)$, with the latter containing far more diverse functions, such as those with sharp discontinuities (indicator functions being a prime example). When $d \geq 3$, we can see that the minimax rates drift apart, with that for $\mathrm{BV}_\infty(1,1)$ being slower than $\mathrm{Lip}(1)$, increasingly so for larger d.

Remark 8. A related point worthy of discussion is about what types of estimators can attain optimal rates over $\mathrm{Lip}(1)$ and $\mathrm{BV}_\infty(1,1)$. For $\mathrm{Lip}(1)$, various *linear smoothers* are known to be optimal, which describes an estimator $\hat{f}$ of the form $\hat{f}(x) = w(x)^\mathsf{T} y$ for a weight function $w : \Omega \to \mathbb{R}^n$ (the weight function can depend on the design points but not on the response vector y). This includes kNN regression and kernel smoothing, among many other traditional methods. For $\mathrm{BV}_\infty(1,1)$, meanwhile, we have shown that the (modified) Voronoigram estimator is optimal (modulo log factors), which is highly *nonlinear* as a function of y. All other examples of minimax rate optimal estimators that we provide in Section 5.5 are nonlinear in y as well. In fact, we conjecture that no linear smoother can achieve the minimax rate over $\mathrm{BV}_\infty(1,1)$. There is very strong precedence for this, both from the univariate case (Donoho and Johnstone, 1998) and from the multivariate lattice case (Sadhanala et al., 2016). We leave a minimax linear analysis over $\mathrm{BV}_\infty(1,1)$ to future work.

Remark 9. Lastly, we comment on the relationship to the results obtained in Padilla et al. (2020). These authors study TV denoising over the ε-neighborhood and kNN graphs; our analysis also extends to cover these estimators, as shown in Section 5.5. They obtain a comparable squared L^2 error rate of $n^{-1/d}$, under a related but different set of assumptions.

In one way, their assumptions are more restrictive than ours, because they require conditions on f_0 that are stronger than TV and L^∞ boundedness: they require it to satisfy an additional assumption that generalizes piecewise Lipschitz continuity, but is difficult to assess, in terms of understanding precisely which functions have this property. (They also directly consider functions that are piecewise Lipschitz, but this assumption is so strong that they are able to remove the BV assumption entirely and attain the same error rates.)

In another way, the results in Padilla et al. (2020) go beyond ours, since they accomodate the case when the design points lie on a manifold, in which case their estimation rates are driven by the intrinsic (not ambient) dimension. Such manifold adaptivity is possible due to strong existing results on the properties of the ε-neighborhood and kNN graphs in the manifold setting. Is is unclear to us whether the Voronoi graph has similar properties. This would be an interesting topic for future work.

5.3 Analysis of the Voronoigram: $L^2(P_n)$ risk

We outline the analysis of the Voronoigram. The analysis proceeds in three parts. First, we bound the $L^2(P_n)$ risk of the Voronoigram in terms of the discrete TV of the underlying signal over the Voronoi graph. Second, we bound this discrete TV in terms of the continuum TV of the underlying function. This is presented in Lemmas 1 and 2, respectively. The third step is to bound the $L^2(P)$ risk after extrapolation (to a piecewise constant function on the Voronoi diagram), which is presented in Lemma 3 in the next subsection. All proofs are deferred until Appendix D.

For the first part, we effectively reduce the discrete analysis of the Voronoigram—in which we seek to upper bound its $L^2(P_n)$ risk in terms of its discrete TV—to the analysis of TV denoising on a grid. Analyzing this estimator over a grid is desirable because a grid graph has nice spectral properties (cf. the analyses in Wang et al. (2016); Hutter and Rigollet (2016); Sadhanala et al. (2016, 2017, 2021) which all leverage such properties). In the language of functional analysis, the core idea here is an *embedding* between the spaces defined by the discrete TV operators with respect to one graph G and another G', of the form

$$\|D(G')\,\theta\|_1 \leq C_n \|D(G)\,\theta\|_1, \quad \text{for all } \theta \in \mathbb{R}^n,$$

where $D(G), D(G')$ denote their respective edge incidence operators. This approach was pioneered in Padilla et al. (2018), who used it to study error rates for TV denoising in quite a general context. It is also the key behind the analysis of TV denoising on the ε-neighborhood and kNN graph in Padilla et al. (2020), who also perform a reduction to a grid graph. The next lemma, inspired by this work, shows that the analogous reduction is available for the Voronoi graph.

Lemma 1. *Under the standard assumptions, consider either of the two modified Voronoi weighting schemes defined in Theorem 2:*

- $\tilde{w}_{ij}^{\mathrm{V}} = \max\{c_0 n^{-(d-1)/d}, w_{i,j}^{\mathrm{V}}\}$ *for each i, j such that $w_{i,j}^{\mathrm{V}} > 0$;*

- $\check{w}_{ij}^{\mathrm{V}} = 1$ *for each i, j such that $w_{i,j}^{\mathrm{V}} > 0$.*

Let D denote the edge incidence operator corresponding to the modified graph, and $\hat{\theta}$ the solution in (14) (equivalently, it is the solution in (8) after substituting in the modified weights). Then there exists a matrix D', that can be viewed as a suitably modified edge incidence operator corresponding to a d-dimensional grid graph, such that

$$\|D'\theta\|_1 \leq C_n \tau_n \|D\theta\|_1, \quad \text{for all } \theta \in \mathbb{R}^n, \tag{31}$$

with probability at least $1 - 3/n^4$ (with respect to the distribution of design points), where $C_n > 0$ grows polylogarithmically in n and τ_n is the scaling factor defined in Theorem 2. Further, letting $\lambda = c\sigma\tau_n(\log n)^{1/2+\alpha}$ for any $\alpha > 1$ and a constant $c > 0$, there exists another constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}(\Omega)$,

$$\mathbb{E}\left[\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2\right] \leq C\left(\frac{\sigma\tau_n(\log n)^{1/2+\alpha}}{n}\mathbb{E}\|D\theta_0\|_1 + \frac{(\log n)^\alpha}{n}\right), \tag{32}$$

where we denote $\theta_0 = (f_0(x_1), \ldots, f_0(x_n)) \in \mathbb{R}^n$.

Notice that, in equivalent notation, we can write the left-hand side in (32) as $n^{-1}\|\hat{\theta} - \theta_0\|_2^2 = \|\hat{f} - f_0\|_{L^2(P_n)}^2$, for the estimated function satisfying $\hat{f}(x_i) = \hat{\theta}_i$, $i = 1, \ldots, n$; and for the ℓ_1 term on the right-hand side in (32) we can write $\|D\theta_0\|_1 = \mathrm{DTV}(f_0(x_{1:n}); w)$ for suitable edge weights w—either of the two choices defined in bullet points at the start of the theorem—over the Voronoi graph.

As we can see, the $L^2(P_n)$ risk of the Voronoigram depends on the discrete TV of the true signal over the Voronoi graph. A natural question to ask, then, is whether a function bounded in continuum TV is also bounded in discrete TV, when the latter is measured using the Voronoi graph. Our next result answers this in the affirmative. It is inspired by analogous results developed in Green et al. (2021a,b) for Sobolev functionals.

Lemma 2. *Under Assumption A1, there exists a constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}(\Omega)$, with w denoting either of the two choices of edge weights given at the start of Lemma 1,*

$$\mathbb{E}\Big[\,\mathrm{DTV}\left(f_0(x_{1:n});\, w\right)\Big] \leq C\bar{\tau}_n(\log n)^{1+1/d}\,\mathrm{TV}(f_0), \tag{33}$$

where $\bar{\tau}_n = n^{(d-1)/d}/\tau_n$ (which is 1 for the clipped weights estimator and $n^{(d-1)/d}$ for the unit weights estimator).

Lemmas 1 and 2 may be combined to yield the following result, which is the $L^2(P_n)$ analog of Theorem 2.

Corollary 1. *Under the standard assumptions, for either of the two modified Voronoigram estimators from Theorem 2, letting $\lambda = c\sigma\tau_n(\log n)^{1/2+\alpha}$ for any $\alpha > 1$ and a constant $c > 0$, there exists another constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}(L)$,*

$$\mathbb{E}\|\hat{f} - f_0\|_{L^2(P_n)}^2 \leq C\left(\frac{\sigma L(\log n)^{3/2+\alpha+1/d}}{n^{1/d}} + \frac{(\log n)^\alpha}{n}\right). \tag{34}$$

Note that for a constant L (not growing with n), the $L^2(P_n)$ bound in (34) converges at the rate $n^{-1/d}$, up to log factors. Interestingly, this $L^2(P_n)$ guarantee does *not* require f_0 to be bounded in L^∞, which we saw was required for consistent estimation in $L^2(P)$ error. Next, we will turn to an $L^2(P)$ upper bound, which does require L^∞ boundedness on f_0. That this is not needed for $L^2(P_n)$ consistency is intuitive (at least in hindsight): recall that we saw from the proof of Proposition 2 that inconsistency in $L^2(P)$ occurred due to tall spikes with vanishing width but non-vanishing L^2 norm, which could not be witnessed by a finite number of samples. To the $L^2(P_n)$ norm, which only measures error at locations witnessed by the sample points, these pathologies are irrelevant.

5.4   Analysis of the Voronoigram: $L^2(P)$ risk

To close the loop, we derive bounds on the $L^2(P)$ risk of the Voronoigram via the $L^2(P_n)$ bounds just established. For this, we need to consider the behavior of the Voronoigram estimator off of the design points. Recall that an equivalent interpretation of the Voronoigram fitted function, $\hat{f} = \sum_{i=1}^n \hat{f}(x_i) \cdot 1_{V_i}$, is that it is given by 1-nearest-neighbor (1NN) extrapolation, applied to $(x_i, \hat{f}(x_i))$, $i = 1, \ldots, n$. Our approach here is to define an analogous 1NN extrapolant $\bar{f}_0$ to $(x_i, f_0(x_i))$, $i = 1, \ldots, n$, and then use the triangle inequality, along with the fact that $\hat{f}, \bar{f}_0$ are piecewise constant on the Voronoi diagram, to argue that

$$\begin{aligned}
\|\hat{f} - f_0\|_{L^2(P)}^2 &\leq 2\|\hat{f} - \bar{f}_0\|_{L^2(P)}^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2 \\
&= 2\sum_{i=1}^n \left(\textstyle\int_{V_i} 1\,dP\right)\left(\hat{f}(x_i) - f_0(x_i)\right)^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2 \\
&\leq \underbrace{2p_{\max}n \cdot \left(\max_{i=1,\ldots,n} \mu(V_i)\right)}_{K_n} \|\hat{f} - f_0\|_{L^2(P_n)}^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2,
\end{aligned} \tag{35}$$

where $\mu(V_i)$ denotes the Lebesgue volume of V_i. The first term in (35) is the $L^2(P_n)$ error multiplied by a factor K_n that is driven by the maximum volume of a Voronoi cell, and we can show K_n is well-controlled (of order $\log n$) under Assumption A1. The second term is a kind of $L^2(P)$ approximation error from applying the 1NN extrapolation rule to evaluations of f_0 itself. When $f_0 \in \mathrm{BV}_\infty(L, M)$, this is also well-controlled, as we show next.

Lemma 3. *Assume that $x_1, \ldots, x_n$ are i.i.d. from a distribution satisfying Assumption A1. Then there is a constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$,*

$$\mathbb{E}\|\bar{f}_0 - f_0\|_{L^2(P)}^2 \leq C\left(\frac{LM(\log n)^{1+1/d}}{n^{1/d}}\right). \tag{36}$$

We make two remarks to conclude this subsection.

Remark 10. For nonparametric regression with random design, a standard approach is to use uniform concentration results that couple the $L^2(P)$ and $L^2(P_n)$ norms in order to obtain an error guarantee in one norm from a guarantee in the other; see, e.g., Chapter 14 of Wainwright (2019). In our setting, such an approach is not applicable—the simplest explanation being that for any $x_1, \ldots, x_n$, there will always exist a function $f \in \mathrm{BV}_\infty(1,1)$ for which $\|f\|_{L^2(P_n)} = 0$ but $\|f\|_{L^2(P)} = 1/2$. This is the same issue as that discussed in Remark 5.

Remark 11. The contribution of the extrapolation risk in (36) to the overall bound in (29) is not negligible. This raises the possibility that, for this problem, extrapolation from random design points with noiseless function values can be at least as hard as $L^2(P_n)$ estimation from noisy responses. This is in contrast with conventional wisdom which says that the noiseless problem is generally much easier. Of course, Lemma 3 only provides an upper bound on the extrapolation risk, without a matching lower bound. Resolving the minimax $L^2(P)$ error in the noiseless setting, and more broadly, studying its precise dependence on the noise level σ, is an interesting direction for future work.

5.5 Other minimax optimal estimators

Finally, we present $L^2(P)$ guarantees that show that other estimators can also obtain minimax optimal rates (up to log factors) for the class of functions bounded in TV and L^∞. First, we consider TV denoising on ε-neighborhood and kNN graphs, using 1NN extrapolation to turn them into functions on Ω. The analysis is altogether very similar to that for the Voronoigram outlined in the preceding subsections, and the details are deferred to Appendix D. A notable difference, from the perspective of methodology, is that these estimators require proper tuning in the graph construction itself.

Theorem 4. *Under the standard assumptions, consider the graph TV denoising estimator $\hat{\theta}^\varepsilon$ which solves problem (14) with $D = D(G^\varepsilon)$, the edge incidence operator of the ε-neighborhood graph G^ε, with edge weights as in (21). Letting $\varepsilon = c_1((\log n)^\alpha / n)^{1/d}$ and $\lambda = c_2 \sigma (\log n)^{1/2 - \alpha}$ for any $\alpha > 1$ and constants $c_1, c_2 > 0$, there is a constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$, the 1NN extrapolant $\hat{f}^\varepsilon = \sum_{i=1}^n \hat{\theta}_i^\varepsilon \cdot 1_{V_i}$ satisfies*

$$\mathbb{E}\|\hat{f}^\varepsilon - f_0\|_{L^2(P)}^2 \le C\left(\frac{\sigma L (\log n)^{3/2 + \alpha/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}} \right). \tag{37}$$

Consider instead the graph TV denoising estimator $\hat{\theta}^k$ which solves problem (14) with $D = D(G^k)$, the edge incidence operator of the kNN graph G^k, with edge weights as in (22). Letting $k = c_1'(\log n)^3$ and $\lambda = c_2' \sigma (\log n)^{1/2 - \alpha}$ for any $\alpha > 1$ and constants $c_1', c_2' > 0$, there is a constant $C' > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$, the 1NN extrapolant $\hat{f}^k = \sum_{i=1}^n \hat{\theta}_i^k \cdot 1_{V_i}$ satisfies

$$\mathbb{E}\|\hat{f}^k - f_0\|_{L^2(P)}^2 \le C'\left(\frac{\sigma L (\log n)^{9/2 - \alpha + 3/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}} \right). \tag{38}$$

Next, and last, we consider wavelet denoising. For this we assume that the design density is uniform on $\Omega = (0,1)^d$. The analysis is quite different from the preceding ones, but it relies on fairly standard techniques in wavelet theory, and we defer the details to Appendix D.

Theorem 5. *Under the standard conditions, further assume that $P = \mu$, the uniform measure on $\Omega = (0,1)^d$. For an estimator $\hat{f}^{\mathrm{wav}}$ based on hard-thresholding Haar wavelet coefficients, there exist constants $c, C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$, it holds that*

$$\mathbb{E}\|\hat{f}^{\mathrm{wav}} - f_0\|_{L^2}^2 \le \frac{CLM}{n^{1/d}} + C \cdot \begin{cases} L\delta_n^* \max\{1, 1/M, \log_2(M\sqrt{n})\} & d = 2 \\ L^{2/d}(\delta_n^*)^{4/(2+d)} + LM(\delta_n^*/M)^{2/d} & d \ge 3, \end{cases} \tag{39}$$

where $\delta_n^* = (c/\sqrt{n})((\log n)^{3/2} + M(\log n)^{1/2})$.

6 Discussion

In this paper, we studied total variation as it touches on various aspects of multivariate nonparametric regression, such as discrete notions of TV based on scattered data, the use of discrete TV as a regularizer in nonparametric estimators, and estimation theory over function classes where regularity is given by (continuum) TV.

We argued that a particular formulation of discrete TV, based on the graph formed by adjacencies with respect to the Voronoi diagram of the design points $x_1, \ldots, x_n$, has several desirable properties when used as the regularizer in a penalized least squares context—defining an estimator we call the Voronoigram. Among these properties:

- it is user-friendly (requiring no auxiliary tuning parameter unlike other geometric graphs, such as ε-neighborhood or k-nearest-neighbor graphs);

- it tracks "pure TV" in large samples, meaning that discrete TV on the Voronoi graph converges asymptotically to continuum TV, independent of the design density (as opposed to ε-neighborhood or kNN graphs, which give rise to certain types of density-weighted TV in the limit);

- it achieves the minimax optimal convergence rate in L^2 error over a class of functions bounded in TV and L^∞;

- it admits a natural duality between discrete and continuum formulations, so the fitted values $\hat{f}(x_i)$, $i = 1, \ldots, n$ have exactly the same variation (as measured by discrete TV) over the design points as the fitted function $\hat{f}$ (as measured by continuum TV) over the entire domain.

The last property here is completely analogous to the discrete-continuum duality inherent in trend filtering (Tibshirani, 2014, 2022), which makes the Voronoigram a worthy successor to trend filtering for multivariate scattered data, albeit restricted to the polynomial order $k = 0$ (piecewise constant estimation).

Several directions for future work have already been discussed throughout the paper. We conclude by mentioning one more: extension to the polynomial order $k = 1$, that is, adaptive piecewise linear estimation, in the multivariate scattered data setting. For this problem, we believe the estimator proposed by Koenker and Mizera (2004), defined in terms of the Delaunay triangulation (which is a dual graph to the Voronoi tessellation) of the design points, will enjoy many properties analogous to the Voronoigram, and is deserving of further study.

References

Taylor Arnold and Ryan J. Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

Franz Aurenhammer and Rolf Klein. Voronoi diagrams. *Handbook of Computational Geometry*, 5(10):201–290, 2000.

Antonin Chambolle and Jerome Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

Tony Chan, Antonio Marquina, and Pep Mulet. High-order total variation-based image restoration. *SIAM Journal on Scientific Computing*, 22(2):503–516, 2000.

Ronald R. Coifman and Stephane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Miguel del Álamo, Housen Li, and Axel Munk. Frame-constrained total variation regularization for white noise regression. *Annals of Statistics*, 49(3), 2021.

David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(8): 879–921, 1998.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, 1977.

Rex A. Dwyer. Higher-dimensional Voronoi diagrams in linear expected time. *Discrete & Computational Geometry*, 6 (3):343–367, 1991.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

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

Jerome Friedman, Trevor Hastie, Holger Hoefling, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226, 1977.

Nicolás García Trillos. Variational limits of k-NN graph-based functionals on data clouds. *SIAM Journal on Mathematics of Data Science*, 1(1):93–120, 2019.

Nicolás García Trillos and Dejan Slepčev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220(1):193–241, 2016.

Tom Goldstein, Xavier Bresson, and Stanley Osher. Geometric applications of the split Bregman method: Segmentation and surface reconstruction. *Journal of Scientific Computing*, 45(1–3):272–293, 2010.

Alden Green, Sivaraman Balakrishnan, and Ryan J. Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian regularization on neighborhood graphs. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021a.

Alden Green, Sivaraman Balakrishnan, and Ryan J. Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian eigenmaps on neighborhood graphs. arXiv:2111.07394, 2021b.

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. In *Proceedings of the Annual Conference on Learning Theory*, 2016.

Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Cambridge University Press, 2015. Draft version.

Gérard Kerkyacharian, Oleg V. Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170, 2001.

Gérard Kerkyacharian, Oleg V. Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. Sparse case. *Theory of Probability & Its Applications*, 52(1):58–77, 2008.

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

Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

Roger Koenker and Ivan Mizera. Penalized triograms: Total variation regularization for bivariate data. *Journal of the Royal Statistical Society: Series B*, 66(1):145–163, 2004.

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

Loic Landrieu and Guillaume Obozinski. Cut pursuit: fast algorithms to learn piecewise constant functions on general weighted graphs. HAL: 01306779, 2015.

Oleg V. Lepski and Vladimir G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, 25(6):2512–2546, 1997.

Oleg V. Lepski, Enno Mammen, and Vladimir G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics*, 25(3):929–947, 1997.

Enno Mammen and Sara van de Geer. Locally apadtive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

Michael H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statistica Sinica*, 10(2): 399–431, 2000.

Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

Oscar Hernan Madrid Padilla, James Sharpnack, James G. Scott, and Ryan J. Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:176–1, 2018.

Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela Witten. Adaptive non-parametric regression with the k-nn fused lasso. *Biometrika*, 107(2):293–310, 2020.

Leonid I. Rudin and Stanley Osher. Total variation based image restoration with free local constraints. In *Proceedings of the International Conference on Image Processing*, pages 31–35, 1994.

Leonid I. Rudin, Stanley Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models via trend filtering. *Annals of Statistics*, 47(6): 3032–3068, 2019.

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.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, Addison J. Hu, and Ryan J. Tibshirani. Multivariate trend filtering for lattice data. arXiv: 2112.14758, 2021.

Isaac J. Schoenberg. Spline functions and the problem of graduation. *Proceeding of the National Academy of Sciences*, 52(4):947–950, 1964.

Donald R. Schuette. A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30, 1978.

Gabriel Steidl, Stephan Didas, and Julia Neumann. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):214–255, 2006.

Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4): 1040–1053, 1982.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

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

Ryan J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, 25(3):1265–1296, 2015.

Ryan J. Tibshirani. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. *Foundations and Trends in Machine Learning*, 15(6):694–846, 2022.

Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3): 1335–1371, 2011.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232, 2012.

John W. Tukey. Curves as parameter, and touch estimation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 681–694, 1961.

Curtis R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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.