Estimation of BV^k functions from scattered data

Addison J. Hu

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Ryan Tibshirani, Chair
Aarti Singh
Sivaraman Balakrishnan
Dejan Slepčev
Robert Nowak (UW-Madison)
Adityanand Guntuboyina (UC Berkeley)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Support: This research was supported by the National Science Foundation Graduate Research Fellowship (Award DGE175016).

Disclaimer: Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or any sponsoring institution.

Keywords: nonparametric regression, bounded variation, gradient variation, minimax analysis, generalized lasso, total variation, local adaptivity, trend filtering, Voronoi, Delaunay, penalized empirical risk minimization

Abstract

The study of bounded variation (BV) functions, and its higher-order generalizations (BV^k functions), is rooted in many fields: statistics, signal processing, functional analysis, approximation theory, among others. From these diverse perspectives has emerged a comprehensive theory of BV^k functions and their estimation from noisy data in dimension $d = 1$. In dimension $d > 1$, the statistical picture is much less clear. Existing statistical theory for BV^k functions can broadly broken down into two categories: replacing the continuous-time BV^k function class with a discrete analogue; or retaining the continuous-time function class and using continuous-time loss (e.g., the white noise model). A gap in the literature lies in the estimation of functions from the continuous-time function class under a sampling model. This thesis is motivated by that gap.

The second and third chapters address the problem of estimating BV^k functions, for index values $k = 0$ and $k = 1$, using scattered data. These cases correspond to functions of bounded variation and functions whose gradient are bounded variation, respectively. For the $k = 0$ case, we study an estimator, the Voronoigram, which fits piecewise constant functions using the Voronoi tessellation of the sample locations. Using the Voronoigram, we establish that the minimax rate (up to log terms) over bounded variation classes is $n^{-1/d}$. For the $k = 1$ case, we study an estimator, the Delaunaygram, which fits continuous piecewise linear functions using the Delaunay tessellation of the sample locations. We find that the Delaunaygram has a $n^{-4/(4+d)}$ rate of convergence when $d < 4$ and $n^{-2/d}$ rate when $d \geq 4$ over discrete gradient variation classes, and obtain matching minimax lower bounds over continuous gradient variation classes. We address the discrete-to-continuous gap, which we expect to be resolved in following work. Along the way, we explore methodological, computational, and practical properties of the two estimators.

The final chapter addresses the broader goal of estimating of BV^k functions, $k \geq 2$. Special attention is called to the following topics: bounded variation classes for $k \geq 2$; anticipated rates of convergence for BV^k classes; challenges specific to dimension $d > 1$; and the desired properties of higher-order generalizations of the estimators studied in this thesis.

Contents

List of Figures

Acknowledgments

The path to a PhD has sometimes felt like a series of lucky breaks, and maybe my luckiest break of all is to have Ryan Tibshirani as my advisor. Thanks for always being generous with your time and for having enough faith in a bright-eyed, bushy-tailed first-year to suggest this as a research topic. You have taught me plenty of lessons inside and outside the classroom, and the greatest privilege of my time as a graduate student is to call you not just my advisor but also my friend.

During the past five years I have been lucky to collaborate with many brilliant people. Of those, I would especially like to thank Alden Green. I can always count on you to provide a healthy dose of skepticism and to show me different ways of thinking. Thank you for helping me navigate the thicket of research. I have also had the privilege of working closely on several applied problems in epidemiological forecasting with Maria Jahja, Jacob Bien, Daniel McDonald, Valérie Ventura, Larry Wasserman, and Rob Tibshirani. Thanks for making the early days of Covid-19 work manageable and for upholding high standards for scientific research even during that frenzied time. Finally, I would like to thank the members of my thesis committee for their guidance and insight over the past two years.

Completing a PhD does take a lot of work, but I've been lucky to have a bit of fun during that time too. Thank you, Tudor, Vinni, Maya, Matteo, Siddhaarth, Kayla, Anni, Konrad, Catherine, Beomjo, and Galen, for tennis, bouldering, running, and countless hours spent just hanging out. You have been the best friends and classmates one could ask for. Pittsburgh will always be a special place to me because of all the wonderful, small moments with you all that have made graduate school such a happy time.

Graduate school was full of surprises, including the opportunity to spend the last year and a half in Berkeley. Even more surprising was to have made such good friends in so short a time. Thanks to Seunghoon, Tiffany, Arisa, Melody, Sizhu, Mogeng, Yaxuan, Karissa, and Michael for indulging me in my hobbies and introducing me to new ones: swimming, biking, foosball, pickleball.

From the very beginning, my parents, Norman and Tina, and my sister, Emily, have been there, providing me with love, support, and encouragement. Thank you, Mom and Dad, for the daily lessons on how to be patient, curious, and kind. Thank you, Emily, for showing me how to take risks. It is only through your work and sacrifice that I have had the opportunity to pursue a PhD.

Chapter 1

Introduction

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

$$y_i = f_0(x_i) + z_i, \quad i = 1, \dots, n, \tag{1.1}$$

for i.i.d. mean zero stochastic errors z_i, $i = 1, \dots, n$. 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 $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}_+^d$ is a multi-index and $D^\alpha = (\frac{\partial}{\partial x_1})^{\alpha_1} \dots (\frac{\partial}{\partial x_d})^{\alpha_d}$ 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.

1.1 Bounded variation functions

This motivates us to consider regularity measured by the *total variation* (TV) seminorm

$$\mathrm{TV}(f;\Omega) = \sup\left\{\int_\Omega f(x)\,\mathrm{div}\,\phi(x)\,dx : \phi \in C_c^1(\Omega;\mathbb{R}^d),\ \|\phi(x)\|_2 \le 1 \text{ for all } x \in \Omega\right\}, \tag{1.2}$$

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 $\mathrm{div}\,\phi = \sum_{i=1}^d \partial\phi_i/\partial_{x_i}$ for the divergence of $\phi = (\phi_1,\dots,\phi_d)$. 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 (1.2) 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.

1.2 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 $\nabla f = (\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_d})$, then

$$\mathrm{TV}(f) = \int_\Omega \|\nabla f(x)\|_2\,dx, \tag{1.3}$$

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 (1.3) 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 $\|x\|_1 \le \sqrt{d}\|x\|_2$ 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

$$\mathrm{TV}(1_S) = \mathrm{per}(S), \tag{1.4}$$

where $\mathrm{per}(S)$ is the perimeter 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

$$\mathrm{TV}(f) = \int_{-\infty}^{\infty} \mathrm{per}\big(\{x \in \Omega : f(x) > t\}\big) \, dt. \tag{1.5}$$

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 (1.5) 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.

1.3 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

$$\underset{f \in \mathrm{BV}(\Omega)}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \, \mathrm{TV}(f), \tag{1.6}$$

given data (x_i, y_i), $i = 1, \ldots, n$ from the model (1.1), under the working assumption that f has small TV. However, in short, solving (1.6) 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 (1.6) 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 (1.6) 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 (1.6): 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 (1.6) 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 (1.6) 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 $d \geq$

[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 (1.1), 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.1.1 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 $\|f\|_{\mathrm{BV}} = \|f\|_{L^1} + \mathrm{TV}(f)$.

2. It is worth noting that this problem is not specific to BV spaces and it occurs also with the k^{th} 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 $pk < d$, 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 $pk < d$. Thus, just as the TV-penalized problem (1.6) 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 $\int_\Omega \sum_{\|\alpha\|_1=2} (D^\alpha f)^2(x)\, dx = \int_\Omega \|\nabla^2 f(x)\|_F^2 \, dx$.) An analogous phenomenon occurs when $d \geq 3$ for bounded gradient variation functions, to be described next.

1.4 Bounded gradient variation functions

A function $f : \Omega \to \mathbb{R}$ is *bounded gradient variation* (BGV) if it satisfies

$$f_0 \in \mathrm{BGV}(\Omega) = \{f : \nabla f \in L^1(\Omega); \mathrm{TV}(\nabla f; \Omega) < \infty\},$$

where the total variation (TV) seminorm for a vector-valued function $g : \mathbb{R}^d \to \mathbb{R}^m$ is defined

$$\mathrm{TV}(g; \Omega) := \sup \left\{ \sum_{k=1}^{m} \int_\Omega g_k(x) \operatorname{div} \phi_k(x) : \phi \in C_c^1(\Omega; \mathbb{R}^{d \times m}), \|\phi(x)\|_F \leq 1 \forall x \in \Omega \right\}. \tag{1.7}$$

The use of the Frobenius norm in (1.7) can be substituted for any other matrix norm to yield the same function space, due to the equivalence of norms on finite-dimensional spaces (i.e., equivalence of matrix norms). Our choice of the Frobenius norm yields an equivalent expression of (1.7) as a weighted sum of ℓ_2 norms of differences in the case where g is piecewise constant.

For a weakly differentiable function f of bounded gradient variation, we refer to the quantity $\mathrm{TV}(\nabla f)$ as the *gradient variation*. This space coincides with the space of bounded Hessian functions first discussed by Demengel [1984] and whose functional analytic properties have been more recently explored by Ambrosio et al. [2023]. We acknowledge that our terminology diverges from previous usage, and we have conscientiously chosen our terminology for consistency with the forthcoming notion of *discrete gradient variation*, whose construction does not necessitate the use of a Hessian (which anyways would not exist except in a weak, measure-theoretic sense).

[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].

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 (1.6) in dimension $d = 1$, with a generalized penalty $\mathrm{TV}(D^k f)$, the TV of the k^{th} 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 Kim et al. [2009], Steidl et al. [2006] 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 k^{th} 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 k^{th} 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 k^{th} discrete derivatives), a result analogous to the property (2.3) 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., Koenker et al. [1994], Schuette [1978]. 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.,

Chambolle and Lions [1997], Chan et al. [2000], Rudin and Osher [1994], Vogel and Oman [1996], 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., Kerkyacharian et al. [2001, 2008], Lepski and Spokoiny [1997], Lepski et al. [1997], Neumann [2000]. 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 2.5.

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*.[4] This generated much follow-up work in statistics, e.g., Arnold and Tibshirani [2016], Friedman et al. [2007], Hoefling [2010], Tibshirani and Taylor [2011], 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

[4]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*.

$n^{-1/d}$. This holds for an arbitrary dimension $d \geq 2$, and agrees with our results in Section 2.5. 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 this thesis 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 an estimator, let alone error analysis, for higher-order notions of TV smoothness (for general $k \geq 2$). 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, and we discuss this problem further in the final chapter of the thesis.

Graph- and discretization-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). Both estimators considered in this thesis were first proposed by the visionary work of Koenker [2005], Koenker and Mizera [2004]. The first-order, continuous piecewise linear estimator came first, when Koenker and Mizera [2004] began with a triangulation of scattered points in $d = 2$ dimensions (say, the Delaunay triangulation) and defined a nonparametric regression estimator called the *penalized 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. Some basic properties for penalized triograms is provided in their work, which we expand upon in addition to providing estimation theory. In the subsequent work of Koenker [2005], the Voronoigram is proposed as a lower-order analog of the peanlized triogram, but to our knowledge this method has not been studied beyond this brief proposal.

Outside of this work, 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 $d \geq 2$, and BGV spaces when $d \geq 4$, as we explain in Sections 2.5.1 and 3.3.3, respectively). 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 Padilla et al. [2018], Wang et al. [2016] are critical analysis tools in Padilla et al. [2020] and critical for our analysis as well.

The 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.

BGV functions and CPWL estimation. Beyond the penalized triogram proposal of Koenker and Mizera [2004], the fitting of continuous piecewise linear functions has received both historical and contemporary attention. The spiritual forerunner to triogram models is the proposal of the tent basis by Courant [1943] in the context of the finite element method. The tent basis was introduced to the statistical literature by Hansen et al. [1998], who also introduced the term "triogram model". Their approach fits a CPWL function by triangulating the entire domain $\Omega \subset \mathbb{R}^2$, adding and deleting knots (vertices in the triangulation) in a stepwise fashion. Therefore the fitted function has all of Ω as its domain, but the vertices of the triangulation over which the function is CPWL are generally not located at the design points $x_{1:n}$. More recently, Pourya et al. [2023] proposes using the penalized triogram with the Delaunay triangulation of some fixed vertex set in $d \geq 2$. Their proposal, which appears to have been made independently of Koenker and Mizera [2004], is motivated by recent study of bounded gradient variation functions and its functional analytic and approximation theoretic properties, due to its representation properties vís-a-vís CPWL functions; see, e.g., [Ambrosio et al., 2022, Aziznejad et al., 2023], who refer to this space as functions with bounded "Hessian total variation." Preceding all of this contemporary interest, the study of these functions was initiated by Demengel [1984], who uses the term "bounded Hessian functions." We acknowledge that our chosen terminology, "bounded

gradient variation functions," diverges from previous usage, but we believe this is a reasonable choice in order to draw analogy to bounded variation functions and to a discrete notion of gradient variation (in which the discussion of a Hessian is not necessary or perhaps even sensible).

Chapter 2

$k = 0$: estimation of bounded variation functions

2.1 Introduction

This chapter is dedicated to providing estimation theory for bounded variation functions. The principal estimator analyzed in this chapter is the Voronoigram, which fits a piecewise constant function on an adaptively chosen partition of the input space. The results of this chapter are joint work with Alden Green and Ryan J. Tibshirani and appear in Hu et al. [2022].

Recall from Chapter 1 the TV-penalized variational problem

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

given data (x_i, y_i), $i = 1, \ldots, n$ from the model (1.1), under the working assumption that f has small TV. In Section 1.3, we discussed the functional analytic reason why estimating BV functions from scattered noisy data is difficult. 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 $\mathrm{TV}(f_0)$ for the regression function f_0, but replaces the sampling model (1.1) 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 (1.6) by the

squared L^2 loss $\|Y - f\|^2_{L^2(\Omega)} = \int_\Omega (Y(x) - f(x))^2 \, dx$ (or some multiscale variant of this). The second type of approach keeps the sampling model (1.1), but replaces the assumption on $\mathrm{TV}(f_0)$ by an assumption on discrete total variation (which is based on the evaluations of f_0 at the design points alone) of the form

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

for an edge set E and weights $w_{ij} \geq 0$. We then naturally replace the penalty $\mathrm{TV}(f)$ in (1.6) by $\mathrm{DTV}(f)$. More details on both types of approaches is given in the related work subsection of the introductory chapter.

The approach we take in this thesis sits in the middle, between the two types. Like the first, we maintain a bona fide smoothness assumption on $\mathrm{TV}(f_0)$, rather than a discrete version of TV. Like the second, we work in the sampling model (1.1), and define an estimator by solving a discrete version of (1.6) 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 (1.6) to a special finite-dimensional class of functions, over which the penalty $\mathrm{TV}(f)$ in (1.6) takes on an equivalent discrete form.

2.1.1 The Voronoigram

This brings us to the central object of this chapter: an estimator defined by restricting the domain in the infinite-dimensional problem (1.6) 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[1] associated with x_i, for $i = 1, \ldots, n$, and define

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

where recall 1_{V_i} is the indicator function of V_i. In words, $\mathcal{F}^{\mathrm{V}}_n$ 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 $\mathrm{BV}(\Omega)$, as each Voronoi cell has locally finite perimeter; in fact, as we will see soon, the TV of each $f \in \mathcal{F}^{\mathrm{V}}_n$ takes a simple and explicit closed form.) Now consider the finite-dimensional problem

$$\underset{f \in \mathcal{F}^{\mathrm{V}}_n}{\text{minimize}} \;\; \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \, \mathrm{TV}(f). \tag{2.1}$$

[1]As we have defined it, each Voronoi cell is open, and thus a given function $f \in \mathcal{F}^{\mathrm{V}}_n$ 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.

We call the solution to (2.1) the *Voronoigram* and denote it by $\hat{f}^{\mathrm{V}}$. 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 $\mathrm{BV}(\Omega)$ that we could have used for a domain restriction in (2.1). Why $\mathcal{F}_n^{\mathrm{V}}$, defined by piecewise constant functions on the Voronoi diagram, as in (2.1)? A remarkable feature of this choice is that it yields an equivalent optimization problem

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{\{i,j\} \in E^{\mathrm{V}}} w_{ij}^{\mathrm{V}} \cdot |\theta_i - \theta_j|, \tag{2.2}$$

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 2.2.1). The equivalence between problems (2.1) and (2.2) sets $\theta_i = f(x_i)$, $i = 1, \dots, n$, 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 2.2.1) that

$$\mathrm{TV}(f) = \sum_{\{i,j\} \in E^{\mathrm{V}}} w_{ij}^{\mathrm{V}} \cdot |\theta_i - \theta_j|. \tag{2.3}$$

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.

A first look at the Voronoigram

From its equivalent discrete problem form (2.2), 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 2.2.2), we can express the fitted function here as

$$\hat{f}^{\mathrm{V}} = \sum_{k=1}^{\hat{K}} (\bar{y}_k - \hat{s}_k) \cdot 1_{\hat{R}_k}, \tag{2.4}$$

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 k^{th} 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 (2.2).

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 (2.2), which trades off between the loss and penalty summands).

Figure 2.1 gives a simple example of the Voronoigram and its adaptive structure in action.

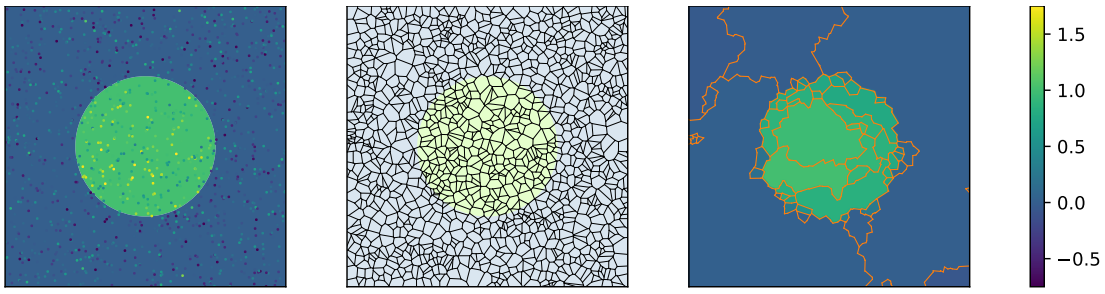

Figure 2.1: *A simple example using the Voronoigram to estimate a function f_0, from noisy observations.* Left: *f_0 and noisy observations made at $n = 1274$ random points in $d = 2$ dimensions.* Center: *the Voronoi tessellation, whose cells constitute the piecewise constant basis for the Voronoigram.* 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.*

2.1.2 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 2.2.4).

- 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 $x_1, \dots, x_n$. 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 $\hat{f}^{\mathrm{V}}(x_i)$, $i = 1, \dots, n$ 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 $(x_i, \hat{f}^{\mathrm{V}}(x_i))$, $i = 1, \dots, n$.) Further, thanks to (2.3), we know that such an extrapolation method is *complexity-preserving*: the discrete TV of $\hat{\theta}^{\mathrm{V}}_i$, $i = 1, \dots, n$ 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 $x_1, \dots, x_n$ from a density p, converges to $\mathrm{TV}(f)$, as $n \to \infty$ (see Section 2.3 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 2.5 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 (2.1)—defined by simply clipping small weights w^{V}_{ij} 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 (2.1) 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.

2.2 The Voronoigram: 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.

2.2.1 The Voronoigram and TV representation

We start with a discussion of the property behind (2.3)—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 $x_1, \dots, x_n$.

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

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

it holds that

$$\mathrm{TV}(f) = \sum_{i,j=1}^n \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j) \cdot |\theta_i - \theta_j|, \tag{2.5}$$

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 (1.2) of total variation, and we defer it to Appendix A.1.2. 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 (1.4).

Note that (2.5) in Proposition 1 reduces to the property (2.3) claimed in the introduction, once we define weights

$$w_{ij}^{\mathrm{V}} = \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j), \quad i, j = 1, \dots, n, \tag{2.6}$$

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 $\{1, \dots, n\}$, 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 $[n] = \{1, \dots, n\}$.

Backing up a little further, we remark that (2.6) also provides the remaining details needed to completely describe the Voronoigram estimator in (2.1). By the TV representation property (2.3), we see that we can equivalently express the penalty in (2.1)

as that in (2.2), which certifies the equivalence between the two problems. Of course, since (2.3) 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 (2.2) and continuum (2.1) problems, once we solve for the Voronoigram fitted values $\hat{\theta}_i^{\mathrm{V}} = \hat{f}^{\mathrm{V}}(x_i)$, $i = 1, \dots, n$ at the design points, we extrapolate via

$$\hat{f}^{\mathrm{V}} = \sum_{i=1}^{n} \hat{\theta}_i^{\mathrm{V}} \cdot 1_{V_i}, \quad \text{which satisfies} \quad \mathrm{TV}(\hat{f}^{\mathrm{V}}) = \sum_{\{i,j\} \in E^{\mathrm{V}}} w_{ij}^{\mathrm{V}} \cdot |\hat{\theta}_i^{\mathrm{V}} - \hat{\theta}_j^{\mathrm{V}}|. \tag{2.7}$$

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 2.2.4.

2.2.2 Insights from generalized lasso theory

Consider a generalized lasso problem of the form:

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|y - \theta\|_2^2 + \lambda \|D\theta\|_1, \tag{2.8}$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ is a response vector and $D \in \mathbb{R}^{m \times n}$ is a penalty operator (as problem (2.8) 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 (2.2) can be equivalently expressed in the more compact form (2.8), 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

$$\big[D(G)\big]_{\ell k} = \begin{cases} +w_{ij} & k = i \\ -w_{ij} & k = j \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Thus, to reiterate the equivalence using the notation just introduced, the penalty operator in the generalized lasso form (2.8) of the Voronoigram (2.2) 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 2.2.3.

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 (2.8): 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

$$\hat{\theta} = P_{\mathrm{null}(D_{-A})}(y - \lambda D_A^{\mathsf{T}} s), \tag{2.10}$$

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 (2.10), for a generic TV denoising estimator $\hat{\theta} = \hat{\theta}(G)$, as

$$\big[\hat{\theta}(G)\big]_i = \sum_{k=1}^{\hat{K}} (\bar{y}_k - \hat{s}_k) \cdot 1\big\{i \in \hat{C}_k\big\}, \quad i = 1, \ldots, n \tag{2.11}$$

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 k^{th} 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 $i \in \hat{C}_k$.

What is special about the Voronoigram is that (2.11), 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 (2.4) 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 (1.1). 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

$$\mathrm{df}(\hat{\theta}) = \mathbb{E}[\mathrm{nullity}(D_{-A})], \tag{2.12}$$

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 (2.12) reduces to

$$\mathrm{df}\big(\hat{\theta}(G)\big) = \mathbb{E}\big[\#\text{ of connected components in } \hat{\theta}(G)\big]. \tag{2.13}$$

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 (2.13). 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 (2.11). For more discussion, see Tibshirani [2015].

The result (2.13) 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 $\hat{f}^{\mathrm{V}}$:

$$\mathrm{df}(\hat{\theta}^{\mathrm{V}}) = \mathbb{E}\big[\#\text{ of locally constant regions in } \hat{f}^{\mathrm{V}}\big], \tag{2.14}$$

because by construction the number of locally constant regions in $\hat{f}^{\mathrm{V}}$ is equal to the number of connected components in $\hat{\theta}^{\mathrm{V}}$.[2]

2.2.3 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 [2019], García Trillos and Slepčev [2016], 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

$$w_{ij}^{\varepsilon} = \begin{cases} 1 & \|x_i - x_j\|_2 \leq \varepsilon \\ 0 & \text{otherwise,} \end{cases} \quad i, j = 1, \ldots, n, \tag{2.15}$$

[2]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 [2013], Tibshirani and Taylor [2012].

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

$$w_{ij}^k = \begin{cases} 1 & \|x_i - x_j\|_2 \leq \max\big\{\|x_i - x_{(k)}(x_i)\|_2,\ \|x_j - x_{(k)}(x_j)\|_2\big\} \\ 0 & \text{otherwise,} \end{cases} \quad i, j = 1, \ldots, n, \tag{2.16}$$

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.

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 (2.8) 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 A.2 for the effective kernel for the Voronoi graph.

2.2.4 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 $\hat{\theta}^{\mathrm{V}}$—the TV denoising estimators on the ε-neighborhood, kNN, and Voronoi graphs, respectively—have adaptively chosen piecewise constant structure, as per (2.11) (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 (2.13). 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., Chambolle and Darbon [2009], Chambolle and Pock [2011], Goldstein et al. [2010], Hoefling [2010], Landrieu and Obozinski [2015], Osher et al. [2005], Tibshirani and Taylor [2011], 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 $\tilde{x}_i = Ux_i$ 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 (2.6), (2.15), and (2.16)—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, 2019, García Trillos and Slepčev, 2016) 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 2.4 to those of Appendix A.3. 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 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 2.4.3, 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].[3] In applications, this worst-case complexity may be pessimistic;

[3]Note that the Voronoi adjacency graph as considered in Section 2.2.1 intersects the Voronoi diagram with the domain Ω on which the n points are sampled, which incurs the additional step of checking

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 (2.6) 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 (2.2) 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 (2.1) and (2.2), and the central property between the discrete and continuum estimates highlighted in (2.7)—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 $V_1, \ldots, V_n$,

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

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.

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

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

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 (2.7) of the Voronoigram is truly special.[4]

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 (2.18) to (2.19) and (2.20)). 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).

2.3 Asymptotics for graph TV functionals

> *The material in this Section regarding the asymptotic limit of the Voronoi TV functional was derived by my collaborator, Alden Green, and is included in this thesis for completeness and due to its relationship to accompanying results.*

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 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.

[4]In fact, this occurs for not one but two natural notions of complexity: TV, as in (2.7), and degrees of freedom, as in (2.14). 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 2.4.4 and Figure 2.6 in particular.

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

$$\mathrm{DTV}(\theta; w) = \sum_{\{i,j\} \in E} w_{ij} |\theta_i - \theta_j|. \tag{2.17}$$

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

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 $x_1, \dots, x_n$ 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* (2.6)*. For any fixed $d \geq 2$ and $f \in C^2(\Omega)$, as $n \to \infty$, it holds that*

$$\mathrm{DTV}\left(f(x_{1:n});\, w^{\mathrm{V}}\right) \to c_d \int_\Omega \|\nabla f(x)\|_2 \, dx, \tag{2.18}$$

in probability, where c_d is the constant

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

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 A.2. A key idea in the proof is show that the weights (2.6) 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 (2.18) is a scaled version of the TV of f, since in this case $\mathrm{TV}(f) = \int_\Omega \|\nabla f(x)\|_2 \, dx$.

Remark 2. The fact that the asymptotic limit of the Voronoi TV functional is *density-free*, meaning the right-hand side in (2.18) 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 (2.15) and any sequence $\varepsilon = \varepsilon_n$ satisfying certain scaling conditions, it holds as $n \to \infty$ that

$$\frac{1}{n^2 \varepsilon_n^{d+1}} \mathrm{DTV}\left(f(x_{1:n});\, w^\varepsilon\right) \to c'_d \int_\Omega \|\nabla f(x)\|_2 \, p^2(x) \, dx, \tag{2.19}$$

in a particular notion of convergence, for a constant $c'_d > 0$. From García Trillos [2019], for the kNN weights (2.16) 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

$$\frac{1}{n^2 \bar{\varepsilon}_n^{d+1}} \mathrm{DTV}\left(f(x_{1:n});\, w^k\right) \to c''_d \int_\Omega \|\nabla f(x)\|_2 \, p^{1-1/d}(x) \, dx, \tag{2.20}$$

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 (2.19) to (2.18) 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 (2.20), but raised to a smaller power than in (2.19).

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 2.4, 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.

2.4 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.

2.4.1 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 $x_0 = (\frac{1}{2}, \frac{1}{2}) \in \mathbb{R}^2$,

$$f_0 = 1\{x \in B(x_0, r_0)\}, \tag{2.21}$$

supported on $\Omega = (0, 1)^2$. This is depicted in the upper display of Figure 2.2 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.)

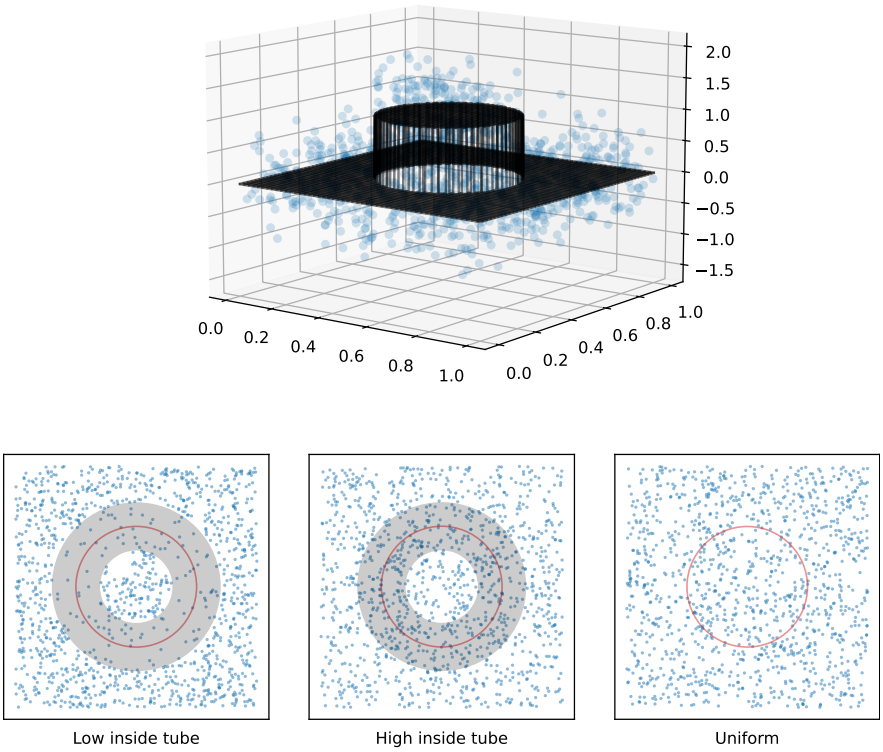

Figure 2.2: *Illustration of the basic experimental setup used in this section.* Top row: *the function f_0 in* (2.21) *depicted using a wireframe plot, along with $n = 1274$ noisy evaluations of f_0 in blue (the noise level is set such that the signal-to-noise ratio is 1).* Bottom row: $n = 1274$ *samples 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.*

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 2.2. 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.

2.4.2 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 (2.21). To be clear, here we compute (using the notation (2.17) 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 (2.6), ε-neighborhood (2.15), and kNN (2.16).

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 A.3, 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 2.2.4).

Figure 2.3 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 (2.18), (2.19), (2.20)—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.

To give a qualitative sense of their differences, Figure 2.4 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

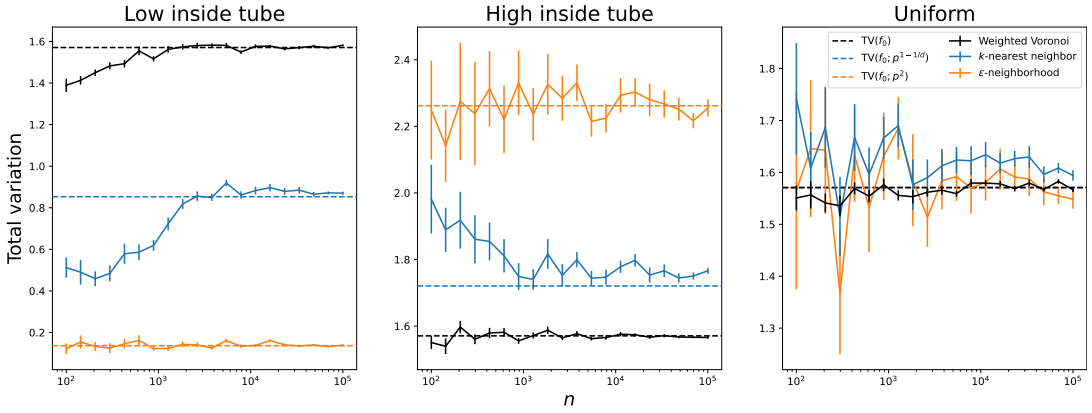

Figure 2.3: *Results from the TV estimation experiment ("weighted Voronoi" refers to the usual Voronoi adjacency graph, with weights in* (2.6)*, and is used to distinguish it from the Voronoi adjacency graph with unit edge weights, which will appear in later experiments). We see that the discrete TV 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).*

Section 2.2.4), 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.

2.4.3 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 (1.1) of the function f_0 in (2.21), we solve the graph TV denoising problem (2.8) with penalty operator D equal to the edge incidence matrix corresponding to the Voronoi (2.6), ε-neighborhood (2.15), and kNN (2.16) 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

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

is equal to 1. (Here $\mathrm{Var}(f_0(x_i))$ 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

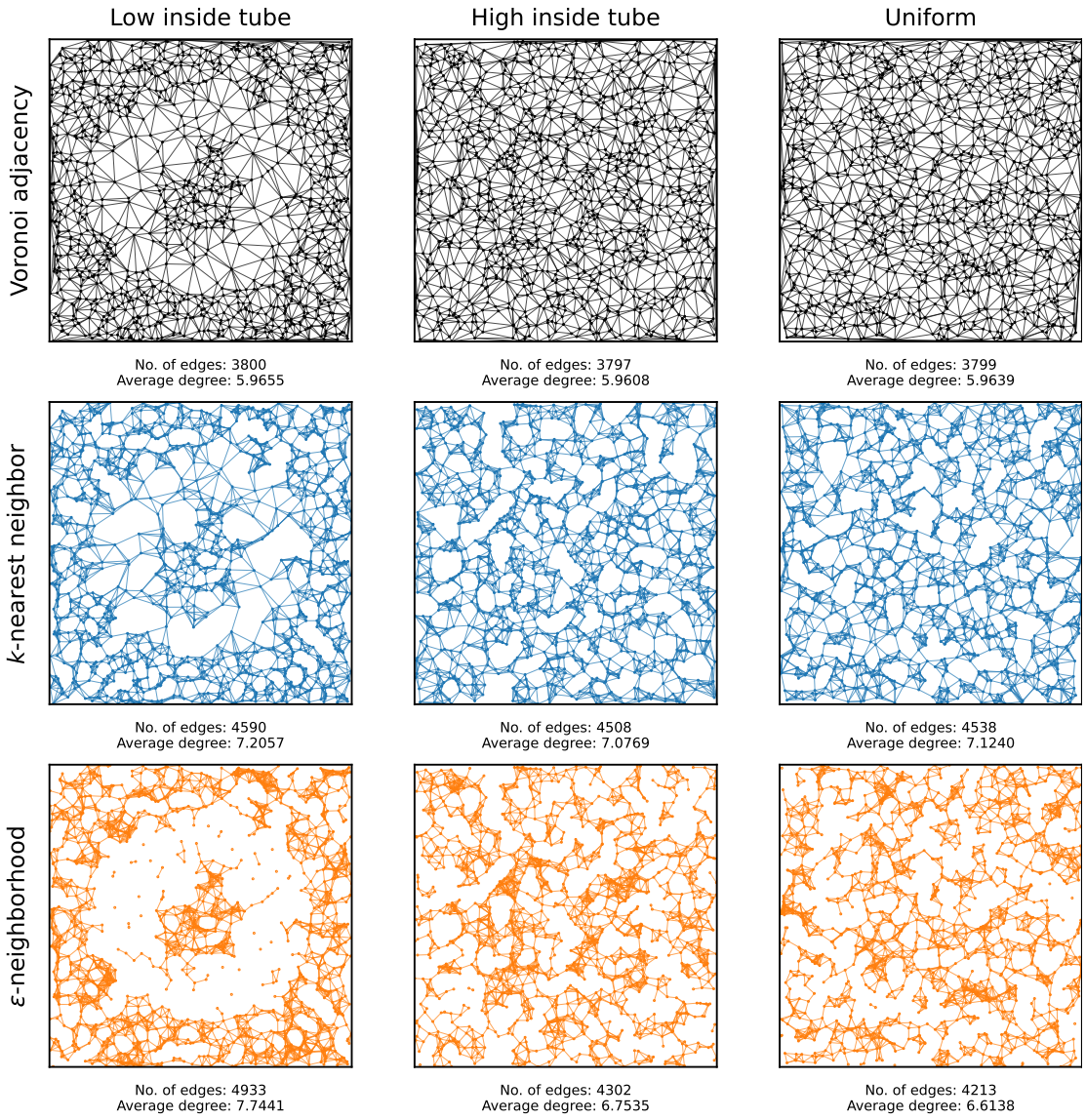

Figure 2.4: *Visualization of the Voronoi, kNN, and ε-neighborhood graphs for a sample of $n = 1274$ design points from each of the three sampling distributions considered. We see qualitatively very different behaviors in these three graph models, and we can also intuit the different asymptotic limits 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).*

range of values for the tuning parameter λ, and at value of λ we record the $L^2(P_n)$ mean squared error

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

Figure 2.5 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, recalling the general formula in (2.13) 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 2.5 broadly align with the expectations set forth at the end of Section 2.3: 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.

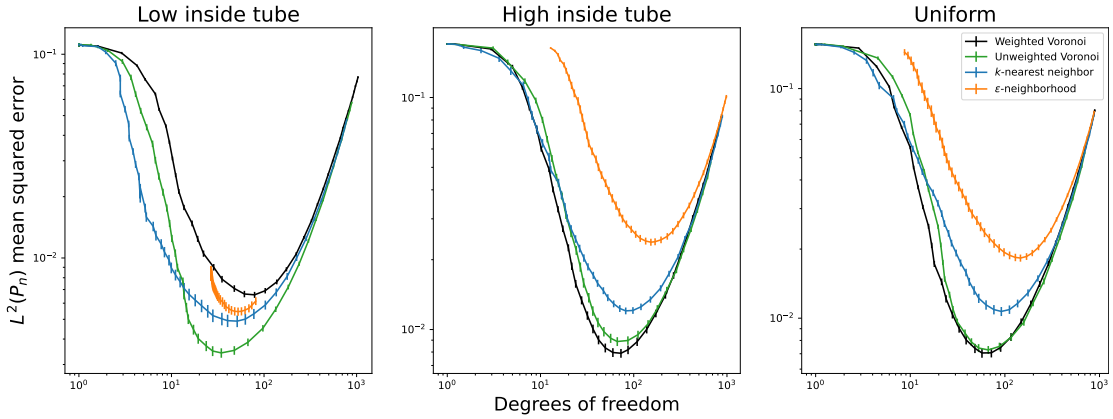

Figure 2.5: *Results from the function estimation experiment ("weighted Voronoi" refers to the usual Voronoi graph and "unweighted Voronoi" the graph with the same edge structure but unit edge weights). We see that the density-weighted methods—TV denoising over the kNN and ε-neighborhood graphs—generally 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 2.5), 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. (2.19) and (2.20)). The poor performance of TV denoising over the ε-neighborhood graph may be ascribed to the large number of disconnected points (see Figures 2.4 and 2.6), 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 A.3, 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 2.5), 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 2.5 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 (2.2). 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 2.4, 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 2.5 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.

2.4.4 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: $\hat{f}(x_i)$, $i = 1, \dots, n$, to a entire fitted function: $\hat{f}(x)$, $x \in \Omega$. As discussed and motivated in Section 2.2.4, we use the 1NN extrapolation rule for each estimator. This is equivalently viewed as piecewise constant extrapolation over the Voronoi tessellation.

Figure 2.6 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

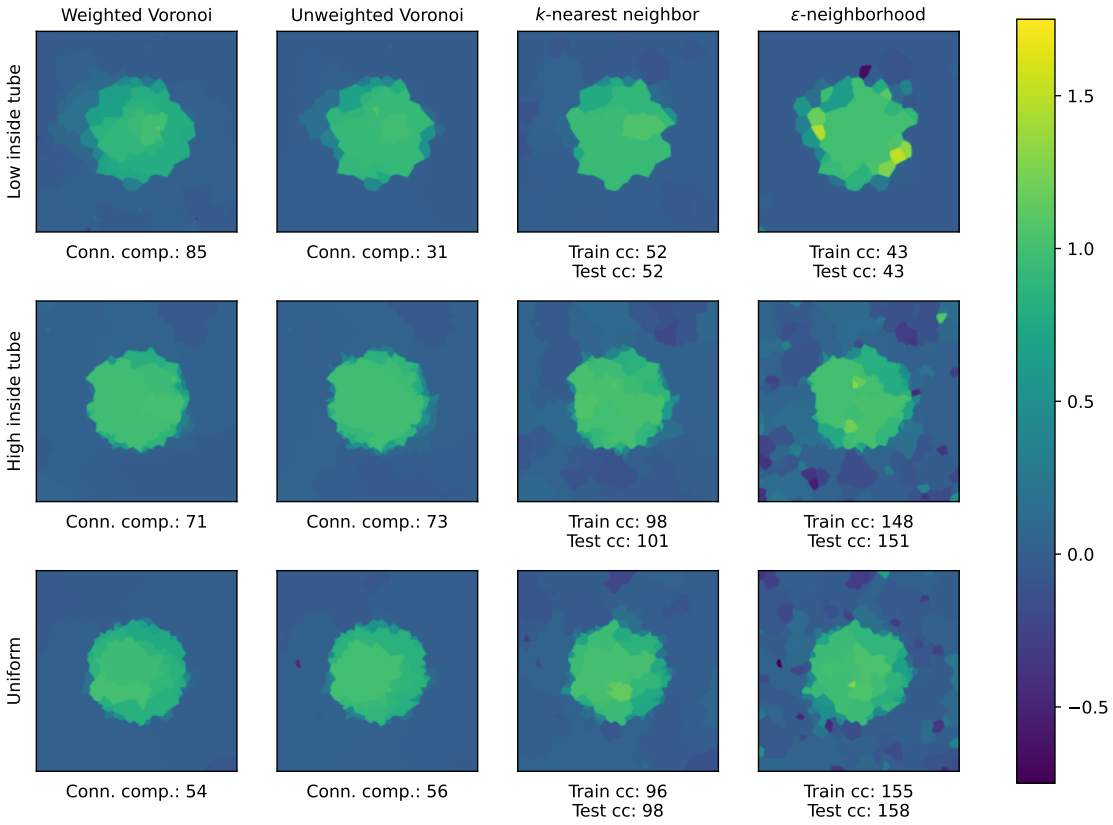

Figure 2.6: *Extrapolants from graph TV denoising estimates, using 1NN extrapolation. We can see several qualitative differences, for example, the issues posed by isolated points in the ε-neighborhood graph. We also note that the number of connected components in the graph used to learn the estimator (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.*

estimator was tuned to have optimal mean squared error (cf. Figure 2.5). 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 2.5 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 2.6 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 $\{x_1, \dots, x_n\}$—as measured by connectedness of the fitted values $\hat{f}(x_1), \dots, \hat{f}(x_n)$ 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.

2.5 Estimation theory for BV classes

In this section, we analyze error rates for estimating f_0 given data as in (1.1), 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

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

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 (1.1), 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 $d \geq 2$ 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 $p_{\min}\|g\|_{L^2(\mu)}^2 \leq \|g\|_{L^2(P)}^2 \leq p_{\max}\|g\|_{L^2(\mu)}^2$ for any function g.

2.5.1 Impossibility result without L^∞ boundedness

A basic issue to explain at the outset is that, when $d \geq 2$, 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 $d \geq 2$, 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

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

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

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 $\mathrm{TV}(f) \leq 1$ 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

$$\begin{aligned}
\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BV}(1) \cap L^2(\Omega)} \mathbb{E}\|\hat{f} - f_0\|_{L^2} &\geq \inf_{\hat{f}} \sup_{f_0 \in \{f_0^\star, f_1^\star\}} \mathbb{E}\|\hat{f} - f_0\|_{L^2} \\
&\geq \inf_{\psi} \Big(\mathbb{P}_{H_0}(\psi = 1) + \mathbb{P}_{H_1}(\psi = 0) \Big), \qquad (2.22)
\end{aligned}$$

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,

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

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

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

Taking $\epsilon \to 0$, and plugging this back into (2.22), establishes the desired result. □

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^∞.

2.5.2 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* (2.2)*, 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* (2.2)*, 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

$$\mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \leq C\left(\frac{\sigma L(\log n)^{5/2+\alpha+1/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}}\right). \quad (2.23)$$

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*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BV}_\infty(L,M)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \geq CLM(M^2 n)^{-1/d}, \tag{2.24}$$

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 (2.24) is Theorem 3 is fairly standard and can be found in Appendix A.4. The proof of the upper bound (2.23) in Theorem 2 is much more involved, and the key steps are described over Sections 2.5.3 and 2.5.4 (with the details deferred to Appendix A.4). 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 (2.2) itself attains the same or a similar upper bound, as in (2.23). 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 (1.2), by generalizing the norm we use to constrain the "test" function ϕ to an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$. (See (A.1) in the appendix.) The original definition in (1.2) 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 2.5.5 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 2.5.5. 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.

2.5.3 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 A.4.

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 Hutter and Rigollet [2016], Sadhanala et al. [2016, 2017, 2021], Wang et al. [2016] 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:*

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

Let D *denote the edge incidence operator corresponding to the modified graph, and* $\hat{\theta}$ *the solution in* (2.8) *(equivalently, it is the solution in* (2.2) *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*

$$\|D'\theta\|_1 \leq C_n \tau_n \|D\theta\|_1, \quad \text{for all } \theta \in \mathbb{R}^n, \tag{2.25}$$

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* $f_0 \in \mathrm{BV}(\Omega)$*,*

$$\mathbb{E}\left[\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2\right] \leq C\left(\frac{\sigma\tau_n(\log n)^{1/2+\alpha}}{n}\mathbb{E}\|D\theta_0\|_1 + \frac{(\log n)^{\alpha}}{n}\right), \tag{2.26}$$

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 (2.26) 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 (2.26) 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,*

$$\mathbb{E}\Big[\,\mathrm{DTV}\left(f_0(x_{1:n});\, w\right)\Big] \leq C\bar{\tau}_n(\log n)^{1+1/d}\,\mathrm{TV}(f_0), \tag{2.27}$$

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* $f_0 \in \mathrm{BV}(L)$*,*

$$\mathbb{E}\|\hat{f} - f_0\|_{L^2(P_n)}^2 \leq C\bigg(\frac{\sigma L(\log n)^{3/2+\alpha+1/d}}{n^{1/d}} + \frac{(\log n)^\alpha}{n}\bigg). \tag{2.28}$$

Note that for a constant L (not growing with n), the $L^2(P_n)$ bound in (2.28) 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.

2.5.4 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

$$\begin{aligned}
\|\hat{f} - f_0\|_{L^2(P)}^2 &\leq 2\|\hat{f} - \bar{f}_0\|_{L^2(P)}^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2 \\
&= 2\sum_{i=1}^n \Big(\textstyle\int_{V_i} 1 dP\Big)\big(\hat{f}(x_i) - f_0(x_i)\big)^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2 \\
&\leq \underbrace{2p_{\max} n \cdot \Big(\max_{i=1,\ldots,n} \mu(V_i)\Big)}_{K_n} \|\hat{f} - f_0\|_{L^2(P_n)}^2 + 2\|\bar{f}_0 - f_0\|_{L^2(P)}^2, \quad (2.29)
\end{aligned}$$

where $\mu(V_i)$ denotes the Lebesgue volume of V_i. The first term in (2.29) 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 $x_1, \dots, x_n$ are i.i.d. from a distribution satisfying Assusmption A1. Then there is a constant $C > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}_\infty(L, M)$,*

$$\mathbb{E}\|\bar{f}_0 - f_0\|^2_{L^2(P)} \leq C\bigg(\frac{LM(\log n)^{1+1/d}}{n^{1/d}}\bigg). \tag{2.30}$$

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 $x_1, \dots, x_n$, 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 (2.30) to the overall bound in (2.23) 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.

2.5.5 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 A.4. 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 (2.8) *with $D = D(G^\varepsilon)$, the edge incidence operator of the ε-neighborhood graph G^ε, with edge weights as in* (2.15)*. 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*

$$\mathbb{E}\|\hat{f}^\varepsilon - f_0\|_{L^2(P)}^2 \leq C\bigg(\frac{\sigma L(\log n)^{3/2+\alpha/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}}\bigg). \quad (2.31)$$

Consider instead the graph TV denoising estimator $\hat{\theta}^k$ which solves problem (2.8) *with $D = D(G^k)$, the edge incidence operator of the kNN graph G^k, with edge weights as in* (2.16)*. 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*

$$\mathbb{E}\|\hat{f}^k - f_0\|_{L^2(P)}^2 \leq C'\bigg(\frac{\sigma L(\log n)^{9/2-\alpha+3/d}}{n^{1/d}} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}}\bigg). \quad (2.32)$$

2.6 Discussion

In this chapter, 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). Having thoroughly consider the $k = 0$ case, the subsequent chapter will discuss extension to polynomial order $k = 1$, that is, adaptive piecewise linear estimation in the multivariate scattered data setting.

Chapter 3

$k = 1$: estimation of bounded gradient variation functions

3.1 Introduction

This chapter works towards the goal of providing estimation theory for bounded gradient variation (BGV) functions. The principal estimator analyzed in this chapter is the Delaunaygram, which fits a continuous piecewise linear (CPWL) function on an adaptively chosen partition of the input space. The results of this chapter are joint work with Alden Green and Ryan J. Tibshirani and have not yet appeared separately in preprint or publication.

Let $\mathrm{GV} := \mathrm{TV}(\nabla f)$ be the gradient variation (GV) of a function that is weakly differentiable with a measure-theoretic second derivative. We consider the GV-penalized variational problem

$$\underset{f \in \mathrm{BGV}(\Omega)}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \, \mathrm{GV}(f), \tag{3.1}$$

given data (x_i, y_i), $i = 1, \ldots, n$, from the model (1.1), under the working assumption that f has small total gradient variation (TGV). For reasons previously discussed in the introductory chapter, the variational problem (3.1) lacks a solution when $d \geq 3$; and while it is possible that the problem has a solution when $d = 2$, this solution (if it exists) is difficult to characterize.

Chapter 2, which dealt with the estimation of bounded variation (BV) functions, outlined the typical avenues for estimation when the variational problem lacks a solution in this way. The first, "continuous-time" approach, is to retain the continuous-time smoothness assumption on f_0 and the continuous-time penalty, but replace the

sampling model by a white noise observational model. The second, "discrete-time" approach, replaces the smoothness assumption on f_0 and the penalty in the variational problem with a discrete-time analogue. Ultimately, Chapter 2 proposed an estimator which operates under the sampling model (1.1) and uses a discrete-time smoothness penalty; however, the continuous-time smoothness assumption on f_0 was retained and the discrete-time penalty was obtained by applying the continuous-time penalty to a special finite-dimensional class of functions.

We follow this blueprint to develop methodology and theory for estimating BGV functions. We study an estimator, the Delaunaygram, which performs empirical risk minimization while penalizing a discrete notion of gradient variation. Like the Voronoigram, this discrete problem can be obtained by restricting the domain of (3.1) to a finite-dimensional class of functions, for which $\mathrm{GV}(f)$ reduces to an equivalent discrete form. And as in Chapter 2, we develop theory with the goal of consistently estimating functions that are BGV in a continuous-time sense. The results of this chapter provide several important results en route to a rate of estimation for BGV functions, including a rate of estimation in terms of the discrete notion of gradient variation and minimax lower bounds.

3.1.1 The Delaunaygram

We now introduce the principal object of this Chapter: an estimator obtained by restricting the domain of the infinite-dimensional problem (3.1) to a finite-dimensional subspace. As in Chapter 2, the finite-dimensional subspace is defined in reference to data-dependent geometric object, but unlike Chapter 2 the subspace of functions is (continuous) piecewise linear (CPWL), rather than piecewise constant. In particular, let $\mathcal{DT} = \{s_1, \ldots, s_m\}$ be the Delaunay triangulation[1] of the input points x_i, $i = 1, \ldots, n$, and define

$$\mathcal{F}_n^{\mathcal{DT}} = \big\{ f : \mathrm{conv}(x_{1:n}) \to \mathbb{R} \ : \ f \in C(\Omega), f|_{s_i} \text{ is linear for } s_i \in \mathcal{DT} \big\},$$

the set of continuous piecewise linear functions on the Delaunay triangulation of the inputs. We then restrict the variational problem to this finite dimensional subspace,

$$\underset{f \in \mathcal{F}_n^{\mathcal{DT}}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \, \mathrm{GV}(f). \tag{3.2}$$

[1]For points $x_1, \ldots, x_n \in \mathbb{R}^d$, the Delaunay triangulation $\mathcal{DT}$ is an open partition of $\mathrm{conv}(x_{1:n})$ whose elements consist of d-simplices s and which satisfies the property that, for each simplex s, (a) the vertices of s are contained in $x_{1:n}$, and (b) the ball whose boundary circumscribes s does not contain any point x_i in its interior. When the points $x_1, \ldots, x_n$ are drawn from a continuous distribution, the Delaunay triangulation is unique almost surely.

We call the solution to (3.2) the *Delaunaygram* and denote it by $\hat{f}^{\mathcal{DT}}$. Koenker and Mizera [2004] first proposed the Delaunaygram as a special instance of the ℓ_1-penalized triogram in two dimensions; subsequently Pourya et al. [2023] proposed this estimator in multiple dimensions.

A special property of the Delaunaygram (and of triograms generally), is that the finite-dimensional variational problem reduces to an equivalent discrete problem

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \, \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{\{i,j\} \in E^{\mathcal{DT}}} w_{ij}^{\mathcal{DT}} \cdot \|G_{s_i}\theta - G_{s_j}\theta\|_2, \tag{3.3}$$

for an edge set $E^{\mathcal{DT}}$ corresponding to neighboring simplices in the Delaunay triangulation, weights $w_{ij}^{\mathcal{DT}}$ measuring the "length" of the shared boundary between neighboring simplices, and linear operators G_{s_i} depending only on $\mathcal{DT}$ such that

$$\nabla(f|_{s_i}) = G_{s_i}\theta,$$

for $i = 1, \ldots, m$. We denote the solution to this discrete problem by $\hat{\theta}^{\mathcal{DT}}$. Importantly, the equivalent formulation (3.3) implies that the gradient evaluations in the penalty may be obtained as a linear transformation of the parameter values θ. The equivalence between the variational problem (3.2) and (3.3) is due to the fact that for any $f \in \mathcal{F}_n^{\mathcal{DT}}$,

$$\mathrm{GV}(f) = \sum_{\{i,j\} \in E^{\mathcal{DT}}} w_{ij}^{\mathcal{DT}} \cdot \left\| \nabla(f|_{s_i}) - \nabla(f|_{s_j}) \right\|_2 = \sum_{\{i,j\} \in E^{\mathcal{DT}}} w_{ij}^{\mathcal{DT}} \cdot \|G_{s_i}\theta - G_{s_j}\theta\|_2. \tag{3.4}$$

The statements in (3.4) say, first, that the gradient variation of any $f \in \mathcal{F}_n^{\mathcal{DT}}$ can be represented by a finite number of gradient evaluations; and second, that those sufficient gradient evaluations may be obtained merely by evaluations at the vertices $x_1, \ldots, x_n$ of $\mathcal{DT}$. These properties will be made rigorous in Section 3.2. As with the Voronoigram of Chapter 2, this ability of the Delaunaygram to represent a continuous-time penalty in equivalent discrete form allows it to unite the variational and discrete worlds, this time for functions of one higher polynomial degree.

A first look at the Delaunaygram

From its equivalent discrete form (3.3) and the gradient representation property of $f \in \mathcal{F}_n^{\mathcal{DT}}$ (3.4), we see that the penalty term allows the Delaunaygram to set the gradient on neighboring simplices of the Delaunay triangulation. In general, larger values of the penalty parameter $\lambda \geq 0$ will result in more simplices being "fused together," that is, to have matching gradients. Because the Delaunaygram uses this penalty while searching over the set CPWL functions $\mathcal{DT}$, the function with "fused simplices" is still

CPWL, but over a coarsened partition of $\mathrm{conv}(x_{1:n})$. And importantly, this coarsened partition is data-adaptive and chosen implicitly simply by fitting the Delaunaygram estimator (3.2). The exact coarsening of $\mathcal{DT}$ chosen by the Delaunaygram depends on how much local variation is present in the response points y_i, by trading off between the loss and penalty summands.

Figure 3.1 gives a simple example of the Delaunaygram and its adaptive structure in action.

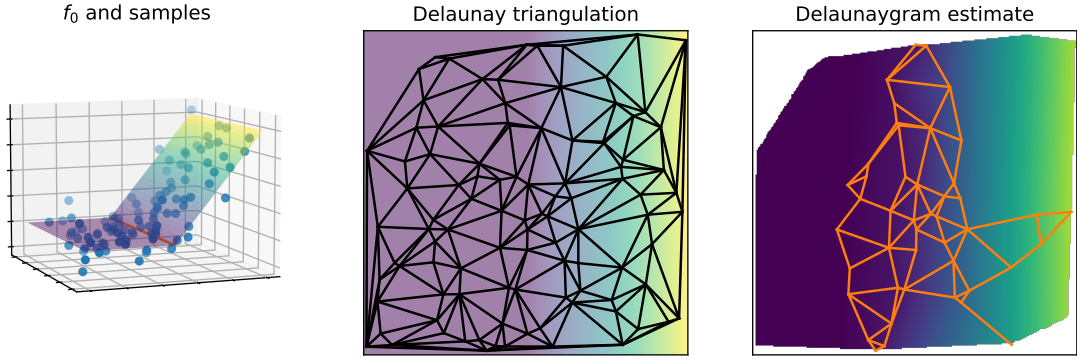

Figure 3.1: *A simple example using the Delaunaygram to estimate a function f_0, from noisy observations.* Left: *f_0 and noisy observations made at $n = 100$ random points in $d = 2$ dimensions.* Center: *the Delaunay triangulation, over which CPWL functions are considered to yield the penalized empirical risk minimization defining the Delaunaygram.* Right: *the Delaunaygram estimate (at a certain choice of λ), with the resulting adaptively-chosen gradient discontinuities outlined in orange (recall that the fitted function satisfies continuity).*

3.1.2 Summary of contributions

The results in this chapter cover methodological properties of the Delaunaygram and theoretical properties of the Delaunaygram and the problem of estimating BGV functions. Although the Delaunaygram estimator analyzed in this chapter was first proposed nearly two decades ago by Koenker and Mizera [2004], its basic properties (beyond its ability to fuse neighboring simplices into affine pieces) are not well understood. Our first set of results illuminate methodological properties of the Delaunaygram when the design points are drawn from a continuous distribution. In the following, will use the terminology "simplex" for the elements of the Delaunay triangulation, and "cell" for elements of a coarsened partition.

- We find that although the Delaunaygram it indeed is able to merge simplices in the Delaunay triangulation together to fit a CPWL function on a coarsening

of the original partition, when the input points are in general position (as they are almost surely whenever they are drawn from a continuous distribution), the structure of the boundary between any two cells is greatly restricted. Specifically, the boundary between any two cells (connected components) contains at most d vertices.

- We derive a characterization of the degrees of freedom of the Delaunaygram estimator with tuning parameter λ, which can be interpreted as the number of free parameters in the fitted function after accounting for linearity constraints on each cell and continuity constraints in between neighboring cells.
- Viewing the Delaunaygram through the lens of generalized lasso theory, we propose variations of the Delaunaygram which allow it to enforce sparsity in individual directional derivative differences (as opposed to sparsity in the differences of entire gradients); or which allow it to fit a piecewise linear function that is not necessarily continuous.

We also provide a set of theoretical results that serve as stepping stones towards establishing a rate of convergence for the Delaunaygram in estimating BGV functions and certification of optimality of this rate of convergence.

- We provide an error analysis for penalized triogram estimators, obtaining a rate of convergence for estimating signals whose discrete gradient variation (as measured by the triangulation defining the triogram) obeys a certain scaling in n. The rate of convergence is $n^{-\frac{4}{4+d}}$ when $d < 4$ and $n^{-2/d}$ when $d \geq 4$. This rate of convergence (and the phase transition when $d = 4$) was previously observed in the lattice-based variation class setting by Sadhanala et al. [2021].
- We show that the Delaunaygram, which is a penalized triogram estimator which uses the Delaunay triangulation as its partition of the input space, satisfies the aforementioned rate of convergence over bounded discrete gradient variation (DGV) signals when the input points come from a density that is bounded above and below.
- We provide minimax lower bounds for the estimation of functions from BGV function classes, which match in rate (up to log terms) with the estimation rate over bounded DGV signal classes. There is a "gap" between these results, in that one needs to reason about the relationship between bounded DGV signal classes and BGV function classes; we provide a partial result which suggests that function evaluations from a BGV function should give rise to a signal that is bounded in DGV, and indicate current difficulties in completing this result.

3.2 The Delaunaygram: methods and basic properties

In this section, we develop basic properties of our estimator, the Delaunaygram. The results of this section are novel except where otherwise noted. Although the subsequent results are phrased in terms of the Delaunaygram, they are applicable to any ℓ_1-penalized triogram estimator [Koenker and Mizera, 2004].

3.2.1 The Delaunaygram and GV representation

The equivalence between the continuous and discrete forms of the gradient variation penalty in (3.4) is a special case of the following result, applied to the weak gradient of any function in $\mathcal{F}_n^{\mathcal{DT}}$; it is the vector-valued extension of Proposition 8. A narrower version of this result, corresponding to (3.6) with $p = 2$, was given in $d = 2$ by Koenker and Mizera [2004] and subsequently in $d \geq 2$ by Pourya et al. [2023].

Proposition 3. Let $V_1, \ldots, V_m$ be an open partition of $\Omega \subseteq \mathbb{R}^{d_1}$ such that each V_i is semialgebraic. Let g be of the form

$$f = \sum_{i=1}^{m} \gamma_i \cdot 1_{V_i},$$

for arbitrary $\gamma_1, \ldots, \gamma_m \in \mathbb{R}^{d_2}$. Then, for any matrix norm $\|\cdot\|$, we have

$$\mathrm{TV}(g; \Omega, \|\cdot\|) = \sum_{i,j=1}^{m} \left(\int_{\partial V_i \cap \partial V_j} \|(\gamma_i - \gamma_j)^\top \otimes n_i(t)\| d\mathcal{H}^{d-1}(t) \right), \tag{3.5}$$

where $n_i(t)$ is the measure-theoretic unit outer normal for V_i. In particular, when $\|\cdot\|$ is the $\ell_{2,p}$ norm,

$$\mathrm{TV}(g; \Omega, \|\cdot\|) = \sum_{i,j=1}^{m} \|\gamma_i - \gamma_j\|_p \cdot \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j). \tag{3.6}$$

Because any function $f \in \mathcal{F}_n^{\mathcal{DT}}$ is linear on each simplex $s \in \mathcal{DT}$ and the design points $x_1, \ldots, x_n$ are the vertices of the simplices in $\mathcal{DT}$, the gradient $\nabla(f|_s)$ on any simplex s is fully determined by evaluations $f(x_i)$, $x_i \in v(s)$, where $v(s)$ is the vertex set of simplex s. This allows one to define a linear transformation G_s for each s such that

$$G_s \theta = G_s f(x_{1:n}) = \nabla(f|_s),$$

where we have identified the coefficient vector $\theta = f(x_{1:n})$. This allows the gradient variation of a function $f \in \mathcal{F}_n^{\mathcal{DT}}$ to be obtained using only its evaluations at $x_1, \ldots, x_n$,

and we call this the *GV representation* property of functions in $\mathcal{F}_n^{\mathcal{DT}}$, analogous to the TV representation property of functions in $\mathcal{F}_n^{\mathrm{V}}$.

In order to complete the equivalence between the continuous- and discrete-time penalties in (3.4), define the weights

$$w_{ij}^{\mathcal{DT}} := \mathcal{H}^{d-1}(\partial s_i \cap \partial s_j), \quad i, j = 1, \dots, m,$$

where $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure, and define the edge set $E^{\mathcal{DT}}$ as the all $\{i, j\}$ such that $w_{ij}^{\mathcal{DT}} > 0$. We say that i, j are adjacent with respect to the Delaunay triangulation when $w_{ij}^{\mathcal{DT}} > 0$, and we can think of $E^{\mathcal{DT}}$ as the set of all adjacent $\{i, j\}$. This defines a weighted, undirected graph on $\mathcal{DT}$, which we call the *Delaunay adjacency graph*, and we denote this graph by $G^{\mathcal{DT}} = (\mathcal{DT}, E^{\mathcal{DT}}, w^{\mathcal{DT}})$. To complete the equivalence between the continuous (3.2) and discrete (3.3) problems, we note that there is a basis $\{h_i : i = 1, \dots, n\}$ for $\mathcal{F}_n^{\mathcal{DT}}$ such that

$$\hat{f}^{\mathcal{DT}} = \sum_{i=1}^{n} \hat{\theta}_i^{\mathcal{DT}} \cdot h_i;$$

moreover, the solutions to the continuous and discrete problems satisfy

$$\mathrm{GV}(\hat{f}^{\mathcal{DT}}) = \sum_{\{i,j\} \in E^{\mathcal{DT}}} w_{ij}^{\mathcal{DT}} \cdot \|G_{s_i}\theta - G_{s_j}\theta\|,$$

certifying the equivalence between problems. This special basis consisting of functions h_i is called the tent basis, and we explore its properties next.

3.2.2 Tent basis

A function f in $\mathcal{F}_n^{\mathcal{DT}}$ is parameterized by its values at the vertices $x_{1:n}$; these values $f(x_{1:n})$ are sufficient to define f off of $x_{1:n}$ via linear interpolation between the vertices of each simplex. Alternatively, the values of f off of the design points may be recovered via expansion in the *tent basis* of $\mathcal{F}_n^{\mathcal{DT}}$. The tent basis, which was proposed as early as Courant [1943] for the purposes of the finite element method, consists of functions $h_i, i = 1, \dots, n$, which are continuous piecewise linear on $\mathcal{DT}$ and satisfy the property that

$$h_i(x) = \begin{cases} 1 & x = x_i, \\ 0 & x = x_j, j \neq i. \end{cases} \tag{3.7}$$

Dual basis and interpolation. It is immediately apparent from its definition that the tent basis has as its dual basis the point evaluation functionals at the vertices,

$$L_i f := f(x_i), \qquad i = 1, \dots, n.$$

Any function $f \in \mathcal{F}_n^{\mathcal{DT}}$ may then be expressed in terms of its evaluations at the vertices,

$$f = \sum_{i=1}^{n} \big(L_i f\big) \cdot h_i = \sum_{i=1}^{n} f(x_i) \cdot h_i.$$

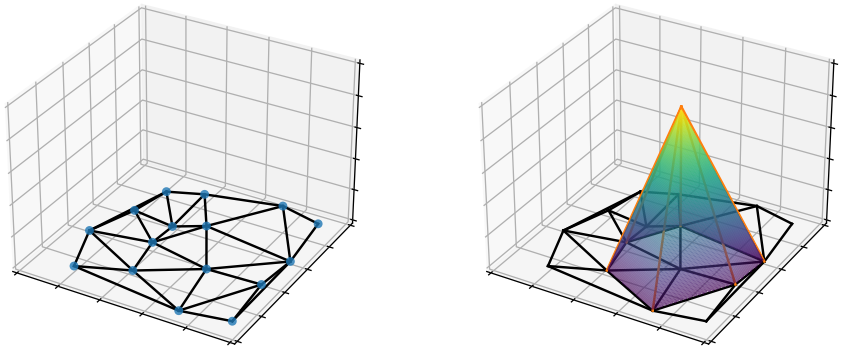

Figure 3.2: *A member of the tent basis induced by the Delaunay triangulation of input points. The tent function h_i is uniquely specified as the function, continuous piecewise linear on $\mathcal{DT}$, which takes on value 1 at x_i and 0 and x_j, $j \neq i$.*

Connection to barycentric coordinates. It turns out that the coefficients defined by the tent basis have a special structure intrinsic to the fact that it is defined over a triangulation $\mathcal{DT}$. For a point $x \in \mathrm{conv}(x_{1:n})$ and allow s such that $x \in \bar{s}$. Relabeling $x_1, \ldots, x_{d+1}$ to denote the vertices of s,

$$(h_1(x), h_2(x), \ldots, h_{d+1}(x))$$

are the barycentric coordinates of x relative to $x_1, \ldots, x_{d+1}$, and $h_j(x) = 0$ for $j = d+2, \ldots, n$. An example of a tent basis element is illustrated in Figure 3.2, which also provides visual confirmation that the tent functions are supported only on $\mathrm{conv}(x_{1:n})$. Approaches for extending a function in the tent basis onto all of Ω are discussed in Section 3.2.7.

3.2.3 Structure of Delaunaygram estimates

The discrete form (3.3) of the Delaunaygram reveals that it performs *group lasso* penalization on the gradient differences across simplices in $\mathcal{DT}$ that share a facet. It follows that for sufficiently large values of the regularization parameter λ, the gradients on certain neighboring simplices will match, yielding a linear function across the two

simplices. Past a certain maximum value of λ, the Delaunaygram will simply perform ordinary least squares (OLS), fitting a single linear function over the entire domain. On the other end of the spectrum, at $\lambda = 0$, the Delaunaygram will interpolate the values y_i, producing $\hat{f} = \sum_{i=1}^{n} y_i \cdot h_i$.

Remark 12. That the Delaunaygram performs OLS past a certain maximum value of λ is a direct consequence of (1) the null space of the penalty functional, which is the linear functions on $\mathbb{R}^d$, and (2) the existence of a maximum "hitting time" in the generalized lasso path (Tibshirani and Taylor, 2011; see also Section 3.2.4). The thin-plate spline penalty functional shares a null space with the Delaunaygram penalty functional. However, due to its "generalized ridge" structure, the thin-plate spline only attains the OLS estimate in the limiting case $\lambda \to \infty$.

Consider the Delaunaygram $\hat{f}$ at a fixed regularization parameter λ, and consider the connected components $\mathcal{C}_j$ of a graph $G = (\mathcal{DT}, E)$, where

$$E = \big\{(s_i, s_j) : \nabla \hat{f}|_{s_i} = \nabla \hat{f}|_{s_j}, \mathcal{H}^{d-1}(\partial s_i \cap \partial s_j) > 0\big\}.$$

We now list some properties of these connected components:

- The Delaunaygram fits a single linear function on each of the $\mathcal{C}_i$.
- Each simplex s_i belongs to a single connected component $\mathcal{C}_j$. As a result, the collection of connected components forms a coarsening of the original partition $\mathcal{DT}(x_{1:n})$ (abbreviated $\mathcal{DT}$). Use $\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda) = \{\mathcal{C}_1, \dots, \mathcal{C}_{\tilde{m}}\}$ to denote the coarsened partition, with the abbreviation $\mathcal{DT}(\lambda)$.
- Each connected component $\mathcal{C}_j$ is a polyhedron whose facets are a subset of facets of $s_i \subset \mathcal{C}_j$[2].

We now give a result on the structure of the connected components.

Proposition 4. Suppose $x_1, \dots, x_n \in \Omega$ are in general position, and fit the Delaunaygram $\hat{f}$ using values $y_1, \dots, y_n \in \mathbb{R}$ and some regularization parameter $\lambda > 0$. The connected components $\mathcal{C}_1, \dots, \mathcal{C}_{\tilde{m}} \in \mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)$ satisfy the property that

$$|\partial \mathcal{C}_i \cap \partial \mathcal{C}_j \cap x_{1:n}| \leq d$$

for any two $\mathcal{C}_i \neq \mathcal{C}_j$, and when $\mathcal{H}^{d-1}(\partial \mathcal{C}_i \cap \partial \mathcal{C}_j) > 0$, the connected components satisfy the stronger property that

$$|\partial \mathcal{C}_i \cap \partial \mathcal{C}_j \cap x_{1:n}| = d.$$

[2]One may ask whether a facet of $\mathcal{C}_i$ may be the union of facets of its simplices. This can only happen if the vertices of the simplices are not in general position, which occurs with probability zero when the design points are drawn from a continuous distribution.

In other words, any two connected components sharing a boundary of dimension $d-1$ overlap on exactly d design points.

Proof. First, we show that any two distinct connected components cannot share more than d design points. We proceed by contradiction; suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ share $d+1$ vertices in common. Because they are distinct connected components, $\nabla \hat{f}|_{\mathcal{C}_1} \neq \nabla \hat{f}|_{\mathcal{C}_2}$. On the other hand, the values at the $d+1$ shared vertices in general position prescribe the same linear function on both connected components.

To show the latter, stronger, statement, simply observe that a shared boundary of Hausdorff dimension $d-1$ must contain at least d vertices. □

Remark 13. The linchpin to the above proposition is continuity, which allows the function $\hat{f}$ to be properly defined at the boundary between connected components. If $\hat{f}$ only required a piecewise affine (without continuity), then one could fashion connected components with more complex boundaries.

Proposition 4 states that (as long as the design points are in general position), the shared boundary between any two connected components must be a single facet of a simplex in $\mathcal{DT}$; complex boundaries between connected components composed of several facets are not possible. In a sense, this is a negative result: connected components cannot be assigned to simplices arbitrarily while also satisfying the requirements that (1) the function takes on distinct linear structure on neighboring connected components and (2) the function is continuous over Ω. Satisfying both aforementioned requirements dictates that only certain coarsenings of $\mathcal{DT}$ are possible.

In particular, given a function with a "simple" gradient discontinuity set, like a ramp function, we would hope that an estimator like the Delaunaygram would produce two connected components, with a single boundary between the two connected components at which the gradient changes. Proposition 4 dictates that this is not possible (with probability zero, if the design points x are drawn from a continuous distribution).

Instead, Figure 3.3 shows that while the Delaunaygram is able to accommodate the ramp structure in producing fitted values, it must do so by introducing many discontinuities in its gradient; it is unable to fit the ramp function using a single "clean ridge." It may seem unfortunate that in order to estimate a function f_0 that has relatively simple structure, the Delaunaygram uses a function $\hat{f}$ that is—by eye—comparatively complex, with its many gradient discontinuities. Luckily, we will see in the sequel that the very condition that precludes a simpler discontinuity set—continuity—restricts the degrees of freedom of the fitted $\hat{f}$ to a relatively small quantity, providing reassurance that the complexity of the fitted function $\hat{f}$ is indeed on a similar order to the underlying function f_0.

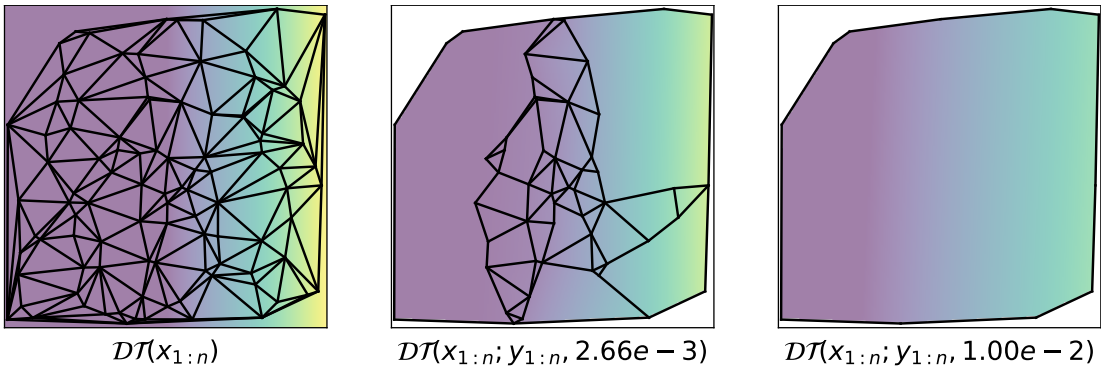

Figure 3.3: *A closer look at the use of the Delaunaygram in estimating the ramp function $f_0 = 2(x_1 - 0.5)_+$ introduced in Figure 3.1.* Left: *the Delaunay triangulation built from points at $n = 100$ random locations on the domain, with the f_0 in background.* Center: *the Delaunaygram estimate using noisy observations, at a certain penalty parameter λ. Observe that the Delaunaygram is able to fit the ramp structure, but is not able to do so using a "clean ridge".* Right: *the Delaunaygram estimate using the same noisy observations, using a sufficiently large λ. Note that there are no discontinuities in the gradient and therefore the Delaunaygram matches the OLS estimate.*

Remark 14. This "defect" of the Delaunaygram is reminiscent of, but distinct from, a limitation we observed with the Voronoigram. With the Voronoigram, the partition is fixed to be the Voronoi diagram of the design points x, and so it would not, for example, fit an "exact" axis-aligned boundary between constant pieces. Here, the issue is more subtle—not only is the partition fixed here, too, but also the requirement of continuity on the fitted function restricts the possible coarsenings of the partition, whereas for the Voronoigram any coarsening of the partitioning was possible.

3.2.4 Degrees of freedom and generalized lasso theory

In the preceding Section 3.2.3, we referred to the estimated *degrees of freedom* of the Delaunaygram estimator. We now formalize this notion through generalized lasso theory. A penalized estimator is an instance of the *generalized lasso* if it may be written in the form

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - B\theta\|_2^2 + \lambda \|D\theta\|_1 \tag{3.8}$$

for some basis matrix $B \in \mathbb{R}^{n \times p}$ and penalty matrix $D \in \mathbb{R}^{r \times p}$. At first blush, the Delaunaygram does not fall into this categorization, since (3.3) demonstrates that it performs group lasso penalization on gradient differences, and generally a group lasso estimator is not an instance of the generalized lasso. That is, an ℓ_2 penalty cannot generically be rewritten as an ℓ_1 penalty to satisfy the form (3.8). It is remarkable, then,

to find that for the Delaunaygram can in fact be cast in generalized lasso form.

Proposition 5. The Delaunaygram estimator (3.3) has an equivalent generalized lasso form,

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \|y - \theta\|_2^2 + \lambda \|D\theta\|_1, \tag{3.9}$$

for a penalty matrix $D \in \mathbb{R}^{r \times n}$ which depends only on the design points $x_1, \ldots, x_n$.

Results of this form have previously been presented by Koenker and Mizera [2004], Pourya et al. [2023][3]; we provide its proof in the Appendix for completeness.

Remark 15. The linchpin to Proposition 5 is again continuity. While the equivalent penalty in (3.3) may be viewed as a generalized group lasso penalty, because it sets groups of $d + 1$ coefficients to zero (corresponding to differences in gradient), the continuity condition—which is enforced at d distinct points between each pair of simplices sharing a facet—reduces the number of effective parameters in each gradient difference to only one. The fact that each gradient difference in the penalty lies in a subspace of dimension 1 allows the reduction to the generalized (non-group) lasso form in (3.9), which sets individual (transformed) coefficients to zero.

An immediate consequence of Proposition 5 is improved computational tools for obtaining the Delaunaygram. A group lasso estimator is generically a quadratic program with quadratic constraints (QCQP), whereas a generalized lasso estimator is a quadratic program (QP) with linear constraints. This opens the doors to a variety of efficient methods for solving the Delaunaygram.

Another consequence of Proposition 5 is an interpretable estimate of the degrees of freedom for the Delaunaygram. Recall that given data $y \sim N(\mu, \sigma^2 I_n)$, the degrees of freedom of a smoother $g : \mathbb{R}^n \to \mathbb{R}^n$ is given [Efron, 1986, Hastie and Tibshirani, 1990] by

$$\operatorname{df}(g) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \operatorname{Cov}\big(g_i(y), y_i\big).$$

For a signal-approximator generalized lasso problem like (3.9), we obtain the simplified form [Tibshirani and Taylor, 2012, Theorem 3]

$$\operatorname{df}(\hat{\theta}) = \mathbb{E}\big[\operatorname{nullity}(D_{\mathcal{A}^c})\big], \tag{3.10}$$

where $\mathcal{A}$ is the active set corresponding to the solution $\hat{\theta}$, i.e., the indices i such that $(D\hat{\theta})_i \neq 0$.

[3]Koenker and Mizera [2004] exclusively consider the $d = 2$ case and state this result without proof. Pourya et al. [2023] consider the $d \geq 2$ case, as we do, and use a Schatten 1-norm $\|\cdot\|_{\mathcal{S}_1}$ in lieu of the Frobenius norm in $\operatorname{TV}(\nabla f; \Omega, \|\cdot\|)$. However, for continuous piecewise linear functions, $\operatorname{TV}(\nabla f; \Omega; \|\cdot\|_{\mathcal{S}_p})$ coincide for all $p \in [1, +\infty)$ (this is apparent from the fact that the matrix in the integrand of Proposition 3 has rank 1), and $\|\cdot\|_F = \|\cdot\|_{\mathcal{S}_2}$.

Figure 3.4 illustrates the use of degrees of freedom to quantify the complexity of the Delaunaygram estimate at different penalty parameter values λ.

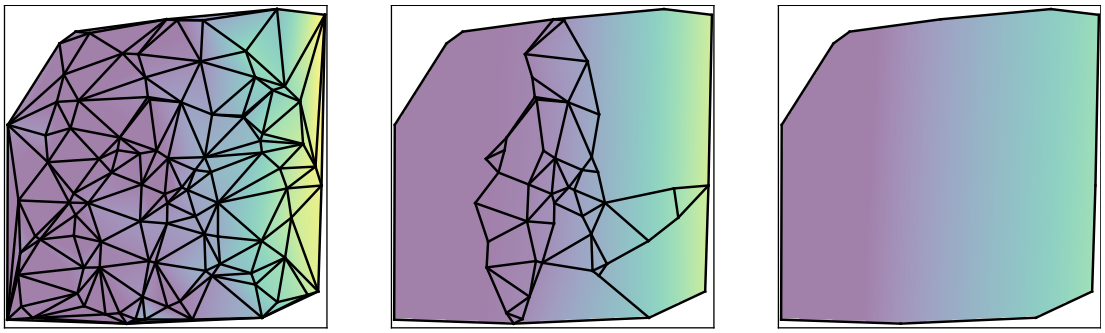

Figure 3.4: *An illustration of the estimated degrees of freedom using the ramp function and noisy observations introduced in Figure 3.1.* Left: *using penalty parameter $\lambda = 0$, an interpolator is fit using the tent basis, and the resulting function has n degrees of freedom (equal to the dimension of the tent basis).* Center: *at a certain value for λ, the Delanuaygram adequately estimates the ramp structure, and although there are many gradient discontinuities, the (estimated) degrees of fredom is still quite small.* Right: *for a sufficiently large penalty parameter λ, all gradients are set to match and the Delaunaygram produces an OLS estimate, which in $d = 2$ has three degrees of freedom.*

Interpreting the degrees of freedom. Casting the Delaunaygram in generalized lasso form allows us to obtain the following characterization of its degrees of freedom. The proof of the result is deferred to the Appendix.

Proposition 6. Suppose $x_1, \dots, x_n \in \Omega$ are in general position, and fit the Delaunaygram $\hat{f}$ using values $y_1, \dots, y_n \in \mathbb{R}$ and regularization parameter $\lambda > 0$. The degrees of freedom for the Delaunaygram is given by

$$\mathrm{df}\big(\hat{\theta}(x_{1:n}, y_{1:n}, \lambda)\big) = \mathbb{E}\bigg[\tilde{m}(d+1) - \sum_{i=1}^{n}\Big(\sum_{j=1}^{\tilde{m}} 1\{x_i \in \bar{\mathcal{C}}_j\} - 1\Big)\bigg], \tag{3.11}$$

where recall $\tilde{m} := |\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)|$.

The degrees of freedom has a simple interpretation: it counts the number of parameters necessary to describe a piecewise linear function on $\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)$. This consists of the $d + 1$ parameters on each of the $\tilde{m}$ linear pieces, minus the number of parameters made redundant by enforcing continuity across pieces—this is the number of times a vertex x_i appears in more than one $\mathcal{C}_i$, since the redundancy only occurs if x_i lies on the shared boundary between $\mathcal{C}_i \neq \mathcal{C}_j$.

At this point we make a few remarks:

- *Null active set.* When the active set $\mathcal{A}$ is empty, there is only one connected component corresponding to the $\mathrm{conv}(x_{1:n})$. This results in $d + 1$ degrees of freedom, the same number of parameters necessary to describe a function that is linear on the entire domain.
- *Saturated active set.* When the active set $\mathcal{A}$ is saturated, $\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda) = \mathcal{DT}(x_{1:n})$, and each simplex s_i is its own connected component. A simple calculation verifies that

$$\begin{aligned}\hat{\mathrm{df}}(\hat{\theta}) &= m(d+1) - \sum_{i=1}^{n} \Big(\sum_{j=1}^{m} 1\{x_i \in \bar{s}_j\} - 1 \Big) \\ &= m(d+1) - \sum_{j=1}^{m} \sum_{i=1}^{n} 1\{x_i \in \bar{s}_j\} + n \\ &= m(d+1) - m(d+1) + n \\ &= n,\end{aligned}$$

  which is the number of parameters necessary to describe the interpolator in the tent basis.
- *Comparison to the Voronoigram.* The estimated degrees of freedom in the current continuous piecewise linear case is exactly analogous to the previously considered piecewise constant case (Voronoigram). For the Voronoigram, the estimated degrees of freedom is the number of constant pieces in the fitted function; each constant piece requires only one parameter. For the Delaunaygram, the CPWL structure of the fitted function yields degrees of freedom that is more intricate but which can still be interpreted as the number of parameters required to describe the fitted function.
- *Degrees of freedom without general position.* If $x_1, \ldots, x_n$ are not in general position, (3.11) may underestimate the degrees of freedom by introducing too large of a continuity correction. This is because without general position, Proposition 4 no longer holds, and if two connected components share more than d points in common, some of the continuity correction in (3.11) will in fact redundant. More careful handling of condition (b ii) in the proof of Proposition 6 would give a correct estimate of the degrees of freedom even without the general position assumption.
- *Relaxing the continuity constraint.* The block variable form (B.6) suggests a version of the Delaunaygram which allows for an adaptively chosen discontinuity set. A sketch of the idea is to rather than enforce the "hard constraints"

$$\beta_{ij_i} - \beta_{i\ell} = 0 \qquad x_i \in \partial \mathcal{C}_\ell, \ell > j_i,$$

introduce "soft constraints"

$$\beta_{ij_i} - \beta_{i\ell} \leq \epsilon \qquad x_i \in \partial\mathcal{C}_\ell, \ell > j_i.$$

We return to this idea and develop it more thoroughly in Section 3.2.6.

Structure of Delaunaygram estimates, revisited. In Section 3.2.3, we used the group lasso formulation of the Delaunaygram to conclude that the Delaunaygram produces continuous piecewise linear estimates, where (given a sufficiently large regularization parameter λ) the estimated function is piecewise linear over a coarsened partition $\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)$ of the original Delaunay partition $\mathcal{DT}(x_{1:n})$. Generalized lasso theory allows us another viewpoint from which to understand the structure of the Delaunaygram estimator. Tibshirani and Taylor [2011, 2012] show that

$$\hat{\theta} = P_{(D_{\mathcal{A}^c})}(y - \lambda D_{\mathcal{A}}^\top s), \tag{3.12}$$

where $s = \mathrm{sign}((D\hat{\theta})_{\mathcal{A}})$ are the active signs and $P_{(D_{\mathcal{A}^c})}$ is the projection matrix onto the null space of $D_{\mathcal{A}^c}$. Recalling that the null space of $D_{\mathcal{A}^c}$ consists of the continuous piecewise linear functions on $\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)$, we may rewrite (3.12), and hence reinterpret the original Delaunaygram problem (3.2), as a shape-constrained least squares problem

$$\hat{f} = \operatorname*{argmin}_f \sum_{i=1}^n \big((y_i - \hat{s}_i) - f(x_i)\big)^2 \quad \text{subject to} \quad f \in \mathrm{CPWL}(\mathcal{DT}(x_{1:n}; y_{1:n}, \lambda)), \tag{3.13}$$

where $\hat{s}_i$, $i = 1, \ldots, n$, is a data-determined shrinkage factor.

3.2.5 ℓ_2 versus ℓ_1 penalty on gradient differences

Recall from the discrete form of the Delaunaygram (3.3) that this estimator may be viewed as penalizing the ℓ_2 norm of the gradient differences across neighboring simplices in the Delaunay triangulation. This is a consequence of Proposition 3 and the default choice of $\mathrm{TV}(\nabla; \Omega) = \mathrm{TV}(\nabla; \Omega, \|\cdot\|_F)$ for the total variation penalty in the original estimator definition (3.2).

However, $\mathrm{TV}(\nabla f; \Omega, \|\cdot\|_F)$ is not the only "flavor" of total variation that gives rise to a sensible estimator. Substituting $\mathrm{TV}(\nabla f; \Omega, \|\cdot\|_{2,1})$ for the penalty in (3.2), we obtain the estimator

$$\tilde{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{i,j=1}^m \|G_{s_i}\theta - G_{s_j}\theta\|_1 \cdot \mathcal{H}^{d-1}(\partial s_i \cap \partial s_j). \tag{3.14}$$

Note carefully that whereas $\hat{\theta}$ penalizes the ℓ_2 norm of gradient differences, $\tilde{\theta}$ penalizes the ℓ_1 norm of gradient differences. This gives rise to important structural differences between these two estimators. The group lasso structure of $\hat{\theta}$ dictates that all partial derivative differences between two neighboring simplices s_i, s_j must be set to zero simultaneously, and when this happens the affine functions on s_i, s_j "snap together" and form a single affine function across both simplices. On the other hand, $\tilde{\theta}$ has the additional liberty to set individual partial derivatives between simplices to be equal, without setting the entire gradients to be equal. This can be advantageous in a setting where the coordinate system has special meaning and the target function is expected to obey an additive piecewise linear structure. On the other hand, the expressivity of the Delaunaygram is limited by the Delaunay triangulation of the design points, and the boundaries between simplices will generically not be axis aligned, limiting somewhat the efficacy of the Delaunaygram (with either ℓ_2 or ℓ_1 penalty on the gradient differences) in estimating functions with axis-aligned structure.

3.2.6 Delaunaygram without continuity

In Section 3.2.3 and Proposition 4, we observed that the continuity constraint imposed by the Delaunaygram places severe restrictions on the on the shapes of the boundaries of the locally linear pieces in the estimate. We now introduce a variant of the Delaunaygram which relaxes the requirement of continuity, which allows for more flexible behavior at the boundaries between the locally linear pieces.

The core of this idea is to decouple the value assigned at each vertex x_i across the simplices, penalizing differences in value at the vertex rather than constraining them to be equal. More formally, introduce a parameter vector $\beta \in \mathbb{R}^{m(d+1)}$ indexed β_{ij}, $j : x_i \in \bar{s_j}$ and let $\tilde{G}_{s_j} : \mathbb{R}^{md} \to \mathbb{R}^d$ be a linear operator which calculates the gradient on the simplex s_j using values $\beta_{\cdot j}$. β and $\tilde{G}_{s_j}$ differ from θ and G_{s_j} in Section 3.1.1 in that the vertex values are no longer shared across simplices.

Further introduce an averaging operator $A \in \mathbb{R}^{n \times md}$ which is a block diagonal matrix,

$$A = \begin{bmatrix} m_1^{-1} 1_{m_1} & 0 & \cdots & 0 \\ 0 & m_2^{-1} 1_{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n^{-1} 1_{m_n} \end{bmatrix},$$

where $m_i := |\{s_j : x_i \in \bar{s}_j\}|$. The non-continuous Delaunaygram is obtained by

solving

$$\tilde{\theta} = \underset{\beta \in \mathbb{R}^{md}}{\operatorname{argmin}} \frac{1}{2}\|y - A\beta\|_2^2 + \lambda_1 \Big(\sum_{i,j=1}^{m} \|\tilde{G}_{s_i}\beta - \tilde{G}_{s_j}\beta\|_2 \cdot \mathcal{H}^{d-1}(\partial s_i \cap \partial s_j) \Big) + \lambda_2 \Big(\sum_{i=1}^{n} \sum_{\ell : x_i \in \bar{s}_\ell} |\beta_{i_j} - \beta_{i\ell}| \Big), \tag{3.15}$$

where $j_i := \min\{j : x_i \in \bar{s}_j\}$. Observe that the usual Delaunaygram (3.2) is a special case of the non-continuous Delaunaygram, obtained by taking λ_2 sufficiently large. In view of this etimator, we make some remarks:

- *Mixed trend filtering.* The mixed penalty structure in (3.15), which includes both first-order and second-order differences, recalls univariate mixed trend filtering [Tibshirani, 2014, Section 8.2].
- *Generalized lasso form.* The non-continuous Delaunaygram (3.15) lacks an equivalent generalized lasso form, since the decoupling of vertex values across simplices also breaks the special structure of gradient differences across neighboring simplices at the heart of Proposition 5. This points to a dual blessing and curse of continuity: it gives provides us with a reduction of a difficult computational problem (QCQP) to one that is easily solved (QP), at the cost of restricting the expressivity of the fitted estimates.
- *ℓ_1-penalized gradient differences.* Section 3.2.5 introduced a version of the Delaunaygram which penalizes the ℓ_1 difference in gradients rather than the ℓ_2 difference in gradients. This estimator admits a generalized lasso form *without* requiring continuity across simplices (i.e., without Proposition 5) and therefore a non-continuous Delaunaygram with an ℓ_1 penalty on gradient differences is still of generalized lasso form. This is worth investigating as an alternative to (3.15) in settings where non-continuity or complex component boundary structures are desired. Unfortunately, in this case the connected components will also be much more complex: an ℓ_1 penalty on gradient differences can set individual directional derivatives to be equal without setting the entire gradient to be equal, and it is not clear how this sparsity structure interacts with the continuity sparsity structure.

3.2.7 Extension beyond design points

As discussed previously, the Delaunaygram estimates a function $f \in \mathcal{F}_n^{\mathcal{DT}}$, which is supported only on $\operatorname{conv}(x_{1:n})$. We now discuss schemes for extending a function $f : \operatorname{conv}(x_{1:n}) \to \mathbb{R}$ to $\tilde{f} : \Omega \to \mathbb{R}$.

Remark 16 (Extension versus extrapolation)**.** Here, we purposefully use the term *extension* to refer to the definition of a function $\tilde{f}$ matching f on $\mathrm{conv}(x_{1:n})$ which is defined on all of Ω. We reserve the term *extrapolation* to refer to the more general act of assigning values to f off of the design points $x_{1:n}$. Over $\mathrm{conv}(x_{1:n})$, extrapolation is accomplished by use of $\mathcal{F}_n^{\mathcal{DT}}$; extension is concerned with continuing the extrapolation process onto all of Ω.

In the following, we discuss a "zeroth-order extension" and a "first-order extension." The former is well-defined in all dimensions and readily implementable. We define and implement the latter in $d = 1, 2$, and discuss how it can be extended to higher dimensions. We then compare properties of the two extension schemes.

Zeroth-order extension. A zeroth-order extension of the function f defines $\tilde{f}$ as

$$\tilde{f}(x) = \begin{cases} f(x) & x \in \mathrm{conv}(x_{1:n}), \\ f(\Pi_{\mathrm{conv}(x_{1:n})} x) & x \in \Omega \setminus \mathrm{conv}(x_{1:n}), \end{cases} \tag{3.16}$$

where for a set A, $\Pi_A(x) := \mathrm{argmin}_{x' \in A} \|x' - x\|_2$. It is easy to check that $\tilde{f}$ is continuous piecewise linear and that it is "piecewise constant" where the pieces are the normal cones to each point on $\partial\mathrm{conv}(x_{1:n})$. Coincidentally, this is the extension scheme proposed by Pourya et al. [2023].

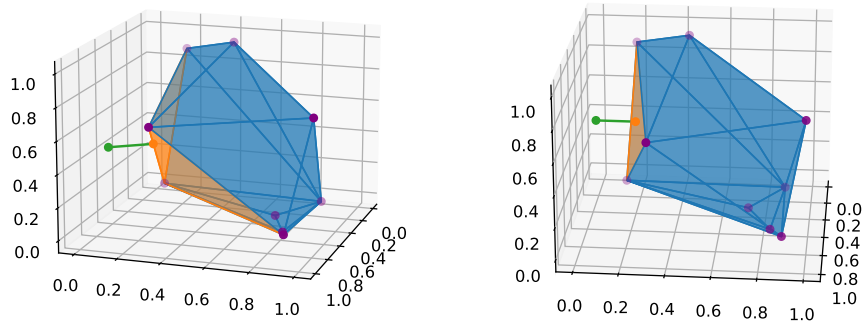

Figure 3.5: *In zeroth-order extension, $\tilde{f}$ is defined at a new point (green) by obtaining its projection (orange point) onto the convex hull of the design points. We shade the facets of the convex hull to which the projection belongs in orange. In the LHS plot, the green point is projected onto a 1-face of the convex hull, so it belongs to two facets.*

First-order extension. A first-order extension of the function f uses both the "value and gradient" of the function f at $\partial\mathrm{conv}(x_{1:n})$, rather than merely the value (as in a zeroth-order extension).

- In $d = 1$, the extension is simple: $\partial\mathrm{conv}(x_{1:n})$ consists of two points x_1 and x_n, and the extended function is simply

$$\tilde{f}(x) = \begin{cases} f(x_1) + f'(x_1) \cdot (x - x_1) & x < x_1, \\ f(x_n) + f'(x_n) \cdot (x - x_n) & x > x_n. \end{cases} \tag{3.17}$$

- In $d = 2$, we may proceed by analogy. Partition $\Omega \setminus \mathrm{conv}(x_{1:n})$ into

$$\Omega \setminus \mathrm{conv}(x_{1:n}) = \cup\{V_{s_i^j} : i = 1, \ldots, m_j, j = 1, \ldots, d-1\},$$

  where $\{s_i^j\}_i$ enumerates the j-dimensional facets of the simplices in the Delaunay triangulation which appear on $\partial\mathrm{conv}(x_{1:n})$, and

$$V_i^j = \{x + tw : x \in s_i^j, w \in \mathcal{C}(s_i^j), t > 0\},$$

  where $\mathcal{C}(s_i^j)$ is the normal cone to s_i^j. We will define an extension of $\tilde{f}$ such that it is continuous piecewise linear, where V_i^j constitute the linear pieces.

  For extension into V_i^1, we can extend the linear function on the boundary as usual,

$$\tilde{f}(x) = f(x_i^1) + \nabla f(x_i^1)^\top (x - x_i^1), \qquad x \in V_i^1, x_1 \in s_i^1.$$

  It remains to define $\tilde{f}$ on V_i^0, the areas of $\Omega \setminus \mathrm{conv}(x_{1:n})$ that are projected onto vertices of the convex hull. Each of these regions borders two regions $V_{i_1}^1, V_{i_2}^1$ associated with facets, on which we have already defined the extension $\tilde{f}$. Evaluations of $\tilde{f}$ at $x_i^0 \in s_i^0$, $x_{i_1}^1 \in \partial V_{i_1}^1 \cap \partial V_i^0$, $x_{i_2}^1 \in \partial V_{i_2}^1 \cap \partial V_i^0$, uniquely specify a linear function with gradient g_i^0. The extension $\tilde{f}$ on V_i^0 may then be expressed,

$$\tilde{f}(x) = f(x_i^0) + {g_i^0}^\top (x - x_i^0), \qquad x \in V_i^0.$$

- In $d \geq 3$, the details of a first-order extension have not been worked out yet. Extension onto V_i^{d-1} is straightforward, but extension onto the regions of $\Omega \setminus \mathrm{conv}(x_{1:n})$ corresponding to the lower-dimensional faces require more thought. Do continuity and linearity provide enough constraints to uniquely specify a function?

A devil's advocate view: is $\{V_i^j\}$ the "appropriate" partition over which to define linear pieces? For example, in the $d = 2$ case, should we eliminate the V_i^0 pieces and instead divide them in half and assign them to V_i^1? This would result in roughly half as many linear pieces. Would continuity still be guaranteed if we did this? See Figure 3.7 for an example of the partition of $\Omega \setminus \mathrm{conv}(x_{1:n})$ using the normal fan of the convex hull.

Comparison of properties. We now make some remarks on the methodological and computational properties of the two extension approaches outlined above.

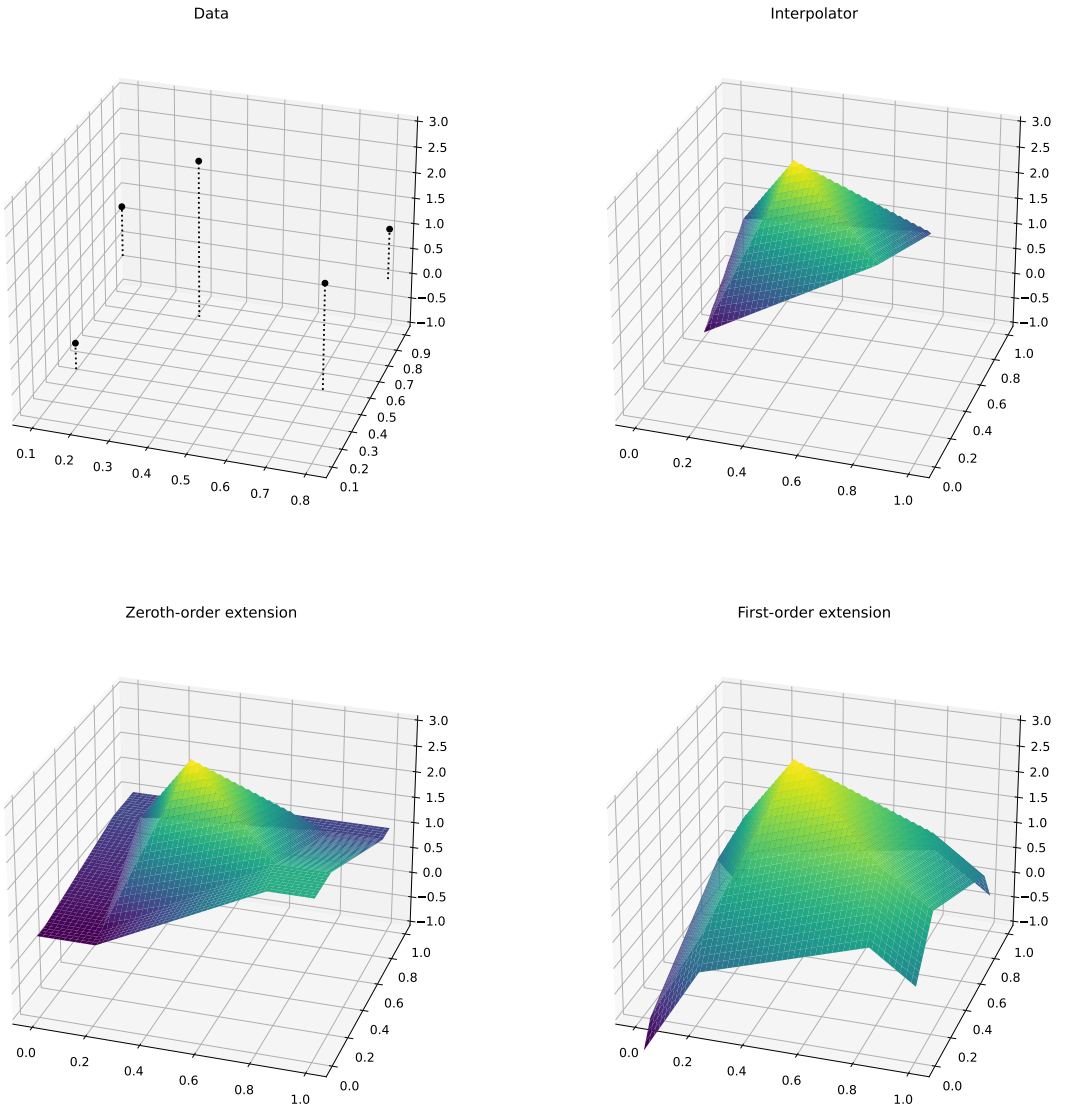

Figure 3.6: *Extension schemes in $d = 2$ applied to a toy dataset with five samples. The zeroth-order scheme simply propogates the value at each point on the boundary along the normal cone anchored at that point. The extended function satisfies continuity and piecewise linear structure; however, there is a sharp and noticable change in gradient across $\partial \mathrm{conv}(x_{1:n})$. In contrast, the first-order extension more gracefully extends the linear structure of the fitted function beyond $\mathrm{conv}(x_{1:n})$, yielding a function with lower gradient variation (in this example) than the zeroth-order extension.*

Practical aspects. To demonstrate the structure of the extension schemes, the zeroth- and first-order approaches are applied to a toy dataset with five samples in Figures

3.6 and 3.7. In Figure 3.6, we see that the first-order approach gracefully extends the fitted function f by continuing the linear structure of the boundary simplices beyond $\mathrm{conv}(x_{1:n})$, whereas the zeroth-order approach introduces abrupt changes in the gradient across $\partial\mathrm{conv}(x_{1:n})$. This point is underscored by Figure 3.7, which outlines the locations across which the gradient of the extended function changes.

The difference in behavior between the zeroth- and first-order extensions can be summarized by observing that the former only uses the value of f along $\partial\mathrm{conv}(x_{1:n})$, whereas the latter uses both the value and the gradient[4]. Another way to view the lower complexity allowed by the zeroth-order extension is to note that the directional derivative along the outer normal of $\mathrm{conv}(x_{1:n})$ is constrained to be zero.

Based on these qualities, one would expect the zeroth-order approach to fare worse than the first-order method in general, except in cases where the function to be estimated to expected to have a smaller gradient near the boundary of the domain. In practice, however, we anticipate that as long as n is sufficiently large, any difference in the overall performance of the estimators (i.e., in terms of integrated error over all of Ω) will be negligible, as $\Omega \setminus \mathrm{conv}(x_{1:n})$ constitutes a vanishing fraction of the domain).

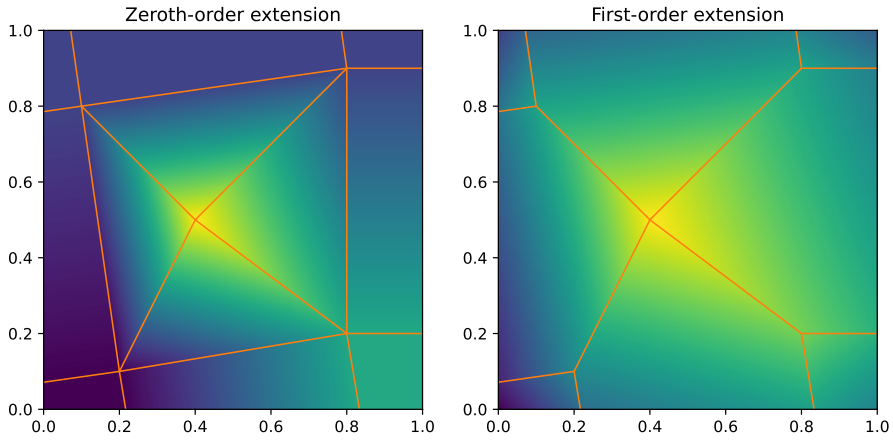

Figure 3.7: *Extension schemes in $d = 2$ applied to a toy dataset with five samples. Locations across which the fitted function experiences a change in gradient are marked in orange. Whereas zeroth-order extension requires a change in gradient across $\partial\mathrm{conv}(x_{1:n})$, first-order extension continues the linear function of the simplices on the boundary into $\Omega \setminus \mathrm{conv}(x_{1:n})$. Both schemes partition $\Omega \setminus \mathrm{conv}(x_{1:n})$ using the normal fan of the convex hull.*

[4]At the vertices, the gradient is not well-defined, but we work around this by first extending from the facets and then from the vertices.

Computation. Both zeroth- and first-order extension require a Euclidean projection onto $\mathrm{conv}(x_{1:n})$, which may be posed as a quadratic program. Write the convex hull as an intersection of halfspaces, i.e., $\mathrm{conv}(x_{1:n}) = \{x : Ax \leq b\}$ for some A, b. The projection of a point $x_0 \in \mathbb{R}^d$ onto $\mathrm{conv}(x_{1:n})$ is then obtained by solving

$$\begin{aligned} &\underset{x}{\text{minimize}} \;\; \|x - x_0\|_2^2 \\ &\text{subject to} \;\; Ax \leq b. \end{aligned}$$

In two dimensions, the projection may be solved directly by calculating the distance to each facet (line segment).

Projection onto $\mathrm{conv}(x_{1:n})$ is sufficient to perform zeroth-order extension (we simply evaluate f at the projection). For first-order extension, the projection is used to determine onto which face (including all lower-dimensional faces) x_0 is projected, which in turn affects which linear function to evaluate for $\tilde{f}$. In two dimensions, points projected onto a facet of the convex hull simply requires evaluation of the linear function from the simplex containing that facet. Points projected onto a vertex of the convex hull requires evaluation of the linear functions of the two simplices to which the vertex belongs.

Minimal energy extension? The zeroth- and first-order extension schemes described above both define the extended function in an ad hoc fashion. In contrast, a more principled approach (and a natural candidate) for the extension is to solve the problem

$$\begin{aligned} &\underset{\tilde{f} \in \mathrm{CPWL}(\Omega)}{\text{minimize}} \;\; \mathrm{TV}(\nabla \tilde{f}) \\ &\text{subject to} \;\; \big(\tilde{f} - f\big)|_{\mathrm{conv}(x_{1:n})} = 0. \end{aligned} \tag{3.18}$$

In the one-dimensional case, this coincides with the customary first-order extension of f onto Ω. When $d > 1$, it is unknown what a solution to (3.18) might be, or if one even exists.

3.3 Estimation theory for BGV classes

This section is dedicated to estimation theory for bounded gradient variation (BGV) functions, of both the discrete- and continous-time flavors. First, we introduce a statistical analysis of the penalized triogram, of which the Delaunaygram is a special case. A worst-case risk upper bound for the penalized triogram is obtained for estimating functions with bounded discrete-time gradient variation. We then show that under Assumption A1, the Delaunaygram obtains this rate as well. A phase transition in the rate of estimation at $d = 4$ is observed, mirroring the findings of Sadhanala et al. [2021]

in the lattice-based gradient variation setting. Minimax lower bounds for estimation over continuous-time gradient variation classes are obtained, with rates that match the risk upper bounds for the Delaunaygram. We conclude with a discussion of how to relate the upper bounds, which work over the discrete-time gradient variation class, and the lower bounds, which work over the continuous-time gradient variation class.

3.3.1 Discrete analysis of penalized triogram estimators

This section provides a statistical analysis of the penalized triogram in recovering a function f_0 of bounded discrete gradient variation, given noisy values at fixed design points $x_1, \dots, x_n \in \Omega = (0,1)^d$ and a triangulation of these points. The analysis in this section assumes the triangulation is pre-specified and is agnostic to how the triangulation is formed. Rates of estimation are given under assumptions on the geometry of the triangulation. In a subsequent section, random design on $x_{1:n}$ is considered, and the geometry of the triangulation is controlled probabilistically.

The penalized triogram. We first consider the ℓ_1*-penalized triogram* of Koenker and Mizera [2004]. Given points $x_1, \dots, x_n \in \Omega$, a triogram model fits a CPWL function on a conforming triangulation $\mathcal{T}$ (to be formally defined shortly) of $x_{1:n}$. In other words, it is just like the Delaunaygram, except that the triangulation of $x_{1:n}$ need not be the Delaunay triangulation. Formally, a penalized triogram fits the estimator

$$\hat{f}^{\mathcal{T}} = \operatorname*{argmin}_{f \in \mathcal{F}_n^{\mathcal{T}}} \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{i,j=1}^{m} w_{ij} \cdot \left\| \nabla f|_{s_i} - \nabla f|_{s_j} \right\|_2, \tag{3.19}$$

for weights $w_{ij} \geq 0$, where $\mathcal{F}_n^{\mathcal{T}}$ is the space of functions on $\mathrm{conv}(x_{1:n})$ which are CPWL on $\mathcal{T}$.

Assumptions on the triangulation. We analyze the penalized triogram using a general triangulation $\mathcal{T}$ under the following assumptions. A triangulation $\mathcal{T}$ of points $x_1, \dots, x_n \in \Omega$ is an open partition $\{s_1, \dots, s_m\}$ of $\mathrm{conv}(x_{1:n})$ whose simplices s satisfy the property that $v(s) \subset x_{1:n}$.

(T1) The triangulation $\mathcal{T} = \{s_1, \dots, s_m\}$ is conforming. A triangulation is *conforming* of the nonempty intersection of the closures of any two simplices in $\mathcal{T}$ is an entire common face (of dimension $0, 1, \dots, d = 1$).

(T2) The graph $(\mathcal{T}, E_{\mathcal{T}})$,

$$E_{\mathcal{T}} = \{\{s_i, s_j\} : \mathcal{H}^{d-1}(\partial s_i \cap \partial s_j) > 0\},$$

is connected.

(T3) $x_{1:n}$ form the vertices, also known as the 0-dimensional faces, of the triangulation $\mathcal{T}$.

We also define certain functionals of the design points $x_{1:n}$ and the triangulation $\mathcal{T}$ relative to a resolution-N lattice partition

$$\Gamma = \{\gamma_i : i \in [N]^d\}, \tag{3.20}$$

i.e., Γ forms an open partition using hypercubes of sidelength $1/N$. In particular, we will take

$$N \asymp \left(\frac{p_{\min} n}{\log^{\alpha_\Gamma} n}\right)^{1/d} \tag{3.21}$$

for a user-chosen parameter α_Γ. Consider now the following functionals of $x_{1:n}$, $\mathcal{T}$.

- $n_\Gamma(s) := |\{\gamma \in \Gamma : \gamma \cap s \neq \emptyset\}|$ is the number of grid cells γ which overlap a simplex s.
- $n_\Gamma(\mathcal{T}) := \max_{s \in \mathcal{T}} n_\Gamma(s)$ is the maximum overlap of grid cells with any one simplex.
- $n_\mathcal{T}(\gamma) := |\{s \in \mathcal{T} : s \cap \gamma \neq \emptyset\}|$ is the number of simplices s which overlap a grid cell γ.
- $n_\mathcal{T}(\Gamma) := \max_{\gamma \in \Gamma} n_\mathcal{T}(\gamma)$ is the maximum overlap of simplices with any one grid cell.
- $n_x(\gamma) := |\{x_{1:n} \cap \gamma\}|$ is the empirical content of a grid cell γ.
- $n_x(\Gamma) := \max_{\gamma \in \Gamma} n_x(\gamma)$ is the maximum empirical content of any one grid cell.

Discrete gradient variation. We now additionally define a discrete notion of gradient variation for a function f based on evaluations of f at locations $x_1, \ldots, x_n \in \Omega$, and a triangulation $\mathcal{T}(x_{1:n}) = s_1, \ldots, s_m$ of $x_{1:n}$. On each simplex s_i, with vertices $x_{i_1}, \ldots, x_{i_{d+1}}$, take $\hat{g}(s_i)$ to be the gradient of the unique linear function passing through the points $(x_{i_1}, f(x_{i_1})), \ldots, (x_{i_{d+1}}, f(x_{i_{d+1}}))$[5]. The discrete gradient variation is defined

$$\mathrm{DGV}(f; \mathcal{T}(x_{1:n}), w) = \sum_{i,j=1}^{m} w_{ij} \cdot \|\hat{g}(s_i) - \hat{g}(s_j)\|_2 \tag{3.22}$$

for a set of weights $w_{ij} \geq 0, i, j = 1, \ldots, m$. Note that the discrete gradient variation is defined for *any* function f such that $x_{1:n} \subset \mathrm{supp}(f)$ without requiring any additional structure.

Theorem 5. *Consider points $x_1, \ldots, x_n \in \Omega$, a triangulation $\mathcal{T}$ of these points satisfying Assumptions (T1)–(T3) and weights $w_{ij} \asymp n^{1/d-1}$, $\{s_i, s_j\} \in E_\mathcal{T}$. Suppose a function*

[5]Equivalently, the expansion of x_i in the tent basis.

$f_0 : \Omega \to \mathbb{R}$ *satisfying*

$$\mathrm{DGV}(f_0; \mathcal{T}, w) = \tilde{O}(L), \tag{3.23}$$

observe noisy responses $y_i = f_0(x_i) + z_i$, $z_i \sim N(0, \sigma^2)$ *independently. If the design points and triangulation satisfy*

$$n_\Gamma(\mathcal{T}), n_{\mathcal{T}}(\Gamma), n_x(\Gamma) = \tilde{O}(1), \tag{3.24}$$

the penalized triogram (3.19) *using triangulation* $\mathcal{T}$*, weights* w*, and properly chosen* λ *satisfies average squared error*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}^{\mathcal{T}}(x_i) - f_0(x_i)\right)^2 = \begin{cases} \tilde{O}_{\mathbb{P}}\left(Ln^{-\frac{4}{4+d}}\right) & d < 4, \\ \tilde{O}_{\mathbb{P}}\left(Ln^{-\frac{2}{d}}\right) & d \geq 4, \end{cases} \tag{3.25}$$

for $d \in \{1, 2, 3, \dots\}$.

Remark 17. The scaling condition on the weights $w \asymp n^{1/d-1}$ corresponds to the shared surface area between neighboring grid cells in the lattice discretization, or the expected shared surface area between neighboring simplices of the Delaunay triangulation of uniformly sampled points. This scaling condition was also found in the $k = 0$ setting of Chapter 2.

Remark 18. The scaling condition on the discrete gradient variation (3.23) can be rewritten as

$$\frac{\mathrm{DGV}(f; \mathcal{T}, w)}{n^{1/d} w_{\min}} = \tilde{O}(Ln^{1-2/d})$$

to correspond to the canonical scaling $C_n \asymp n^{1-s}$ introduced by Sadhanala et al. [2021] for lattice-based discrete variation classes. Their work also observed the phase transition in the rates of estimation (3.25) in the lattice setting.

3.3.2 In-sample rate for the Delaunaygram

In this and subsequent subsections, we use the Assumption A1 and the *standard assumptions* from Chapter 2.

Background & setup. The analysis of Section 3.3.1 considered estimation of bounded discrete gradient variation functions under fixed design, using a penalized triogram estimator with a fixed triangulation assumed to be given. Rates of estimation were derived assuming the design and triangulation satisfied certain conditions (3.24) formalizing the notion that the input points are evenly spread about the domain and the simplices in the triangulation were not too large.

We now consider the random design setting, where the function f_0 is now observed at locations $x_1, \dots, x_n \sim P$, where P follows Assumption A1. We consider the

penalized triogram with the Delaunay triangulation $\mathcal{DT}$ of the input points $x_{1:n}$, which we previously introduced as the Delaunaygram in Section 3.1,

$$\hat{f}^{\mathcal{DT}} = \underset{f \in \mathcal{F}_n^{\mathcal{DT}}}{\operatorname{argmin}} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \sum_{i,j=1}^{m} w_{ij} \|\nabla f|_{s_i} - \nabla f|_{s_j}\|_2,$$

for $w_{ij} \geq 0$.

Challenges of the Delaunay triangulation with random inputs. In order to obtain rates of estimation for the Delaunaygram via Theorem 5, it is necessary to control certain functionals of the triangulation (3.24). Unfortunately, it turns out these functionals cannot be controlled at the proper order for the Delaunay triangulation using points randomly sampled from a uniform-like distribution on Ω. This is due to well-known boundary effects suffered by forming the Delaunay triangulation from points on a bounded set; see, e.g., Bern et al. [1991], which rather than triangulating randomly sampled points on a bounded set, triangulates all of $\mathbb{R}^d$ using points from a Poisson process, and then analyzes the subtriangulation on a bounded subset of $\mathbb{R}^d$. This approach for the elimination of boundary effects is known as minus-sampling [Miles, 1974].

Inspired by minus-sampling, we analyze an modified version of the Delaunaygram which restricts estimation to a subset $\tilde{\Omega} \subset \Omega$. The subset $\tilde{\Omega}$ is formed by excluding a tube around the boundary of the domain where boundary effects are observed

$$\tilde{\Omega} := \Omega \setminus B(\partial\Omega, r), \tag{3.26}$$

where

$$r \asymp \Big(\frac{p_{\min}^{-1}(\alpha + d + 1)\log n}{n}\Big)^{1/d} \tag{3.27}$$

for a user-chosen parameter $\alpha > 0$. Consider the subtriangulation

$$\widetilde{\mathcal{DT}} = \{s \in \mathcal{DT} : s \in \tilde{\Omega}\}, \tag{3.28}$$

and denote the subset of design points

$$\{v(s) : s \in \widetilde{\mathcal{DT}}\} =: \tilde{x}_{1:n} \subset x_{1:n}. \tag{3.29}$$

Finally, let $\tilde{n} := |\tilde{x}_{1:n}|$. Although boundary effects prohibit the high-probability geometric control of Delaunay simplices in the full Delaunay triangulation $\mathcal{DT}$, the results of Appendix B.2.3 reveal that probabilistic control is possible over $\widetilde{\mathcal{DT}}$.

Main result. The preceding discussion discussed the idea of learning a penalized triogram using the Delaunay triangulation of randomly sampled points, as well as issues such an estimator may encounter on the boundary of the sampling domain. We propose an estimator which uses a subtriangulation $\widetilde{\mathcal{DT}}$ of the Delaunay triangulation and analyze the in-sample error properties of such an estimator. Since we are only learning estimates on a subset $\tilde{x}_{1:n}$ of the original data, we measure error only on this subset of data. As n grows with dimension fixed, the share of the data used to learn the estimator approaches the full dataset.

Theorem 6. *Consider points $x_1, \dots, x_n \sim P$ following Assumption A1 and noisy responses $y_i = f_0(x_i) + z_i$, where $f_0 : \Omega \to \mathbb{R}$ and $z_i \sim N(0, \sigma^2)$ independently. Form the Delaunay triangulation $\mathcal{DT}$ from $x_{1:n}$ and consider the estimator $\hat{f}^{\widetilde{\mathcal{DT}}}$ obtained by forming the penalized triogram* (3.19) *with the Delaunay subtriangulation* (3.28)*. If the estimand f_0 satisfies*

$$\mathrm{DGV}(f_0; \widetilde{\mathcal{DT}}, w) = \tilde{O}_{\mathbb{P}}(L)$$

with weights $w_{ij} \asymp \tilde{n}^{1/d-1}$, $\{s_i, s_j\} \in E_{\widetilde{\mathcal{DT}}}$, then the estimator $\hat{f}^{\widetilde{\mathcal{DT}}}$ with weights w and properly chosen λ satisfies average squared error

$$\frac{1}{\tilde{n}} \sum_{x_i \in \tilde{x}_{1:n}} \big(\hat{f}^{\widetilde{\mathcal{DT}}}(x_i) - f_0(x_i)\big)^2 = \begin{cases} \tilde{O}_{\mathbb{P}}\big(Ln^{-\frac{4}{4+d}}\big) & d < 4, \\ \tilde{O}_{\mathbb{P}}\big(Ln^{-\frac{2}{d}}\big) & d \geq 4, \end{cases} \tag{3.30}$$

for $d \in \{1, 2, 3, \dots\}$.

Remark 19. Pruning the Delaunay triangulation $\mathcal{DT}$ by removing simplices outside $\tilde{\Omega}$ does remove a nontrivial fraction of the data, and so we are exploring methodological adjustments to use more of the data, including approaches based on the Voronoi neighbor relationship and minus-sampling.

- The Voronoi neighbor relationship, derived from the Voronoi tessellation of Ω (rather than all of $\mathbb{R}^d$, in which case it coincides with the Delaunay triangulation), provides good geometric properties with high probability in a way that (we suspect) removes fewer simplices.
- A minus-sampling-based approach would introduce artifically sampled points on $\mathbb{R}^d \setminus \Omega$, triangulate all of $\mathbb{R}^d$, and then use the subtriangulation Ω to form the estimator. Both these approaches are promising but introduce subtleties that complicate the analysis, and so they are outside the scope of this document.

Remark 20. The extension approaches of Section 3.2.7 assume a function that is fit on a convex polyhedron. The pruned Delaunay triangulation $\widetilde{\mathcal{DT}}$ is still a polyhedron, but it is not longer necessarily convex, unlike the original Delaunay triangulation $\mathcal{DT}$. To extend the CPWL fit on $\widetilde{\mathcal{DT}}$ onto the rest of Ω, one may first extend the function

from $\widetilde{\mathcal{DT}}$ onto $\mathcal{DT}$ using the tent basis on $\mathcal{DT}$ (which is a superset of the tent basis for $\widetilde{\mathcal{DT}}$), and then applying any of the extension approaches from Section 3.2.7.

3.3.3 Minimax lower bounds

In the preceding subsection, we derived an in-sample rate of convergence for the Delaunaygram in terms of the discrete-time gradient variation of the underlying regression function. For an appropriately scaled bounded discrete gradient variation function class, we obtained a rate of estimation with a phase transition: in the supercritical regime $d < 4$, the rate (up to log factors) of $n^{-\frac{4}{4+d}}$, and in the subcritical regime $d \geq 4$, the rate (again up to log factors) of $n^{-2/d}$. We now present minimax lower bounds for the rate of estimation for the closely related continuous-time bounded gradient variation function classes. These rates match the worst-case rates for the Delaunaygram, up to the phase transition between the supercritical and subcritical regimes. Note carefully that there is a mismatch between the function classes — for the worst-case risk bounds for the Delaunaygram, we consider a discrete-time notion of complexity, whereas for the minimax lower bounds we consider a continuous-time notion of complexity. These two function classes are intimately related, and their exact relationship is further discussed in the subseqeuent subsection.

Supercritical and subcritical lower bounds

For the following result, we introduce a class of bounded gradient variation functions which are also essentially bounded, i.e., the class

$$\mathrm{BGV}_\infty(L, M) := \mathrm{BGV}(L) \cap L^\infty(M).$$

The following theorem lower bounds the rate of estimation over $\mathrm{BGV}(L)$ in the supercritical regime of $d = 1, 2, 3$ and over $\mathrm{BGV}_\infty(L, M)$ is the subcritical regime of $d \geq 4$.

Theorem 7. *Under the standard assumptions,*

- *when $d = 1, 2, 3$, the minimax risk satisfies*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BGV}(L)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \geq C_1 L^{\frac{2d}{4+d}} n^{-\frac{4}{4+d}}, \tag{3.31}$$

  for a constant $C_1 > 0$; and

- *when $d \geq 4$, provided that n, L, M satisfy $c_0(M^2 n)^{-\frac{(d-2)}{d}} \leq L \leq C_0(M^2 n)^{2/d}$ for constants $C_0 > c_0 > 0$, the minimax risk satisfies*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BGV}_\infty(L,M)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \geq C_2 L M^{\frac{d-4}{d}} n^{-2/d}, \tag{3.32}$$

for another constant $C_2 > 0$,

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

The proof of Theorem 7 follows a classical approach also used to obtain a $n^{-1/d}$ lower bound over the class $\mathrm{BV}_\infty(L, M)$ (cf. Theorem 3), but with a different construction of test functions. Slight differences in the analysis account for the fact that a tighter lower bound of $n^{-4/(4+d)}$ is obtained in the supercritical regime $d = 1, 2, 3$.

Impossibility result when $d \geq 4$

We now discuss why, for Theorem 7, we considered the essentially bounded BGV function class rather than the usual BGV function class for the subcritical regime. It turns out that (analogously to the $k = 0$ case), in the subcritical regime estimation under the sampling model is impossible for functions in the BGV class.

Proposition 7. Under the standard assumptions, exists a constant $c > 0$ not depending n such that

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

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

The proof of this result follows the argument of Proposition 2, with an adjustment to the construction of the test functions.

Proof. Recalling Assumption A1, we will equivalently study $L^2(\mu)$ risk. Consider the two-point hypothesis testing problem of distinguishing

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

where $0 < \epsilon < 1$ and C_d is a constant that depends only on dimension. Note that f_1 is a continuous and piecewise linear function that takes the value 0 on $\mathbb{R}^d \setminus (0, \epsilon)^d$ and the value $C_d^{-1}\epsilon^{-d/2}$ at $\epsilon/2 \cdot 1$, where 1 is the length-d all-ones vector. Direct calculations reveal that

$$\|f_1\|^2_{L^2(\Omega)} = \frac{2}{(d+1)(d+2)C_d} \qquad \text{and} \qquad \mathrm{TV}(\nabla f_1) = \epsilon^{d/2-2},$$

implying that $f_1 \in L^2(\Omega)$, and $f_1 \in \mathrm{BGV}(1)$ for $d \geq 4$. A standard reduction provides

that

$$\inf_{\hat{f}} \sup_{f_0 \, \mathrm{BGV}(1) \cap L^2(\Omega)} \mathbb{E}\|\hat{f} - f_0\|_{L^2(\Omega)} \geq \inf_{\hat{f}} \sup_{f_0 \in \{f_0, f_1\}} \mathbb{E}\|\hat{f} - f_0\|_{L^2(\Omega)}$$
$$\geq \inf_{\psi} \Big(\mathbb{P}_{H_0}\{\psi = 1\} + \mathbb{P}_{H_1}\{\psi = 1\} \Big), \qquad (3.33)$$

where the infimum in the final line is over all measurable tests ψ. Conditional on the event

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

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

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

Rearranging the above, we find that

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

Take $\epsilon \to 0$ and substitute into (3.33) to obtain the result. □

3.3.4 Discussion: discrete- versus continuous-time gradient variation

In the preceding development, we have provided a framework for discrete analysis of triogram estimators; an in-sample rate of convergence for the Delaunaygram in terms of discrete-time gradient variation of the underlying function; and minimax lower bounds for estimation of functions in terms of their continuous-time gradient variation. This playbook follows the same brushstrokes of the analysis of the Voronoigram in Chapter 2, in which the discrete-time analysis of the Voronoigram is connected to rates of estimation for continous-time function classes by Lemma 2, which relates the discrete-time and continuous-time variation measures of complexity for a function. In this final subsection, we discuss a partial result of similar flavor for the gradient variation and the current shortcomings of our analysis. Obtaining this result would yield an $L^2(P_n)$ rate of convergence for the Delaunaygram over (continuous-time) bounded gradient variation function classes, and a $L^2(P)$ minimax estimation lower bound with the same rate up to log factors.

For a triangulation $\mathcal{T}$, define the quantity $\lambda_{\min}(\mathcal{T})$ to be

$$\min_{s \in \mathcal{T}} \ \lambda_{\min}(M(s)^\top M(s)),$$

where

$$M(s) = \begin{bmatrix} 1 & 0_{1\times d} \\ 1 & x_2 - x_1 \\ \vdots & \vdots \\ 1 & x_{d+1} - x_1 \end{bmatrix}$$

for $s = (x_1, \ldots, x_{d+1})$.

Lemma 4. *Sample $x_1, \ldots, x_n$ i.i.d. from P following Assumption A1. Form the Delaunay triangulation $\mathcal{DT}$ and consider the subtriangulation $\widetilde{\mathcal{DT}}$ formed from $\tilde{n}$ points $\tilde{x}_{1:n}$ as constructed in Section 3.3.2. There exists a constant $C > 0$ depending only on d and a user-chosen parameter $\alpha > 1$ such that for all sufficiently large n and $f \in \mathrm{BGV}(\Omega)$, the discrete gradient variation measured by $\widetilde{\mathcal{DT}}$ using weights $w_{ij} \asymp \tilde{n}^{1/d-1}$ is bound above by*

$$\begin{aligned} \mathrm{DGV}(f; \widetilde{\mathcal{DT}}, w) &:= \sum_{\{s_i, s_j\} \in E_{\widetilde{\mathcal{DT}}}} w_{ij} \left\| \hat{g}(s_i) - \hat{g}(s_j) \right\|_2 \\ &\leq \frac{C}{\delta} (\log \tilde{n})^{d+1+1/d} \cdot \frac{\lambda_{\min}^{-1/2}(\widetilde{\mathcal{DT}})}{(\log \tilde{n}/\tilde{n})^{1/d}} \cdot \mathrm{TV}\left(\nabla f|_{\tilde{\Omega}}\right) \end{aligned} \tag{3.34}$$

with probability at least $1 - \delta - \tilde{n}^{-\alpha}$.

The proof of this result in deferred to Appendix B.2.6. We now make some remarks:

- We conjecture that
$$\frac{\lambda_{\min}^{-1/2}(\widetilde{\mathcal{DT}})}{(\log \tilde{n}/\tilde{n})^{1/d}} = \tilde{O}_{\mathbb{P}}(1), \tag{3.35}$$
in which case Lemma 4, along with Theorem 6, provides an in-sample rate of estimation for the Delaunaygram in terms of the continuous-time gradient variation of the estimand f_0.
- The scaling (3.35) calls for uniform control of a functional of the Delaunay simplices. We are able to establish non-uniform control, i.e.,
$$\lambda_{\min}^{-1/2}(s) \lesssim n^{1/d}$$
up to log terms, for each $s \in \widetilde{\mathcal{DT}}$, as a sanity check. The proof this result is given in Appendix B.2.6.

3.4 Illustrative empirical examples

In this section we illustrate the practical usage of the Delaunaygram through numerical experiments. The first experiment performs extensive comparisons of the Delaunaygram to the thin-plate spline, a linear smoother, in a structured function recovery

setup against a set of canonical signal functions meant to represent heterogeneous and homogeneous notions of smoothness. The second experiment considers real data from a meteorological setting, where the goal is to estimate the ocean thermal response to the passage of a tropical cyclone.

3.4.1 Comparisons on synthetic data

The first set of experiments contrasts the performance of the Delaunaygram, a nonlinear estimator, to the thin-plate spline, a linear estimator, in estimating functions of heterogeneous and homogeneous smoothness. Functions of heterogeneous smoothness constitute a larger function class than functions of homogeneous smoothness, and it is generally hypothesized that while linear estimators are able to recover functions of homogeneous smoothness well, they suffer in recovering functions of heterogeneous smoothness as compared to nonlinear estimators. This notion is formalized in certain settings, such as the univariate problem and the multivariate problem on grids, by deriving a minimax rate of convergence over the heterogeneous function class which is achieved by a nonlinear estimator, and then deriving a minimax linear rate of estimation which is observed to be slower. While we do not have minimax linear theory for the multivariate scattered data case, this set of experiments illustrates some of the deficiencies of linear estimators in recovering functions of heterogeneous smoothness.

Experimental setup. We consider three signal functions supported on $(0, 1)^2$, which are depicted in Figure 3.8. They consist of:

- the *Bumps* function, which is the addition of two spherical normal probability density functions centered at $(1/4, 3/4)$ and $(3/4, 1/4)$, with standard deviation $7/40$;
- the *Pyramids* function, which follows a similar structure to Bumps, except that it is piecewise linear;
- the *Sine* function, which is the outer product of a sine wave with period 1.

The Bumps and Pyramids functions exhibit heterogeneous smoothness: the roughness for those functions is entirely localized to the upper-left and lower-right quadrants. The Sine function, on the other hand, is a classic model of homogeneous smoothness.

All three signal functions f_0 are normalized to have signal strength one, i.e., $\mathrm{Var}(f_0) = 1$, with respect to the Lebesgue measure. For each f_0, we observe samples

$$y_i = f_0(x_i) + z_i, \qquad i = 1, \ldots, 2000, \tag{3.36}$$

where $x_i \sim \mathrm{Unif}((0, 1)^2)$ and $z_i = N(0, 1)$, giving a signal-to-noise ratio of 1.

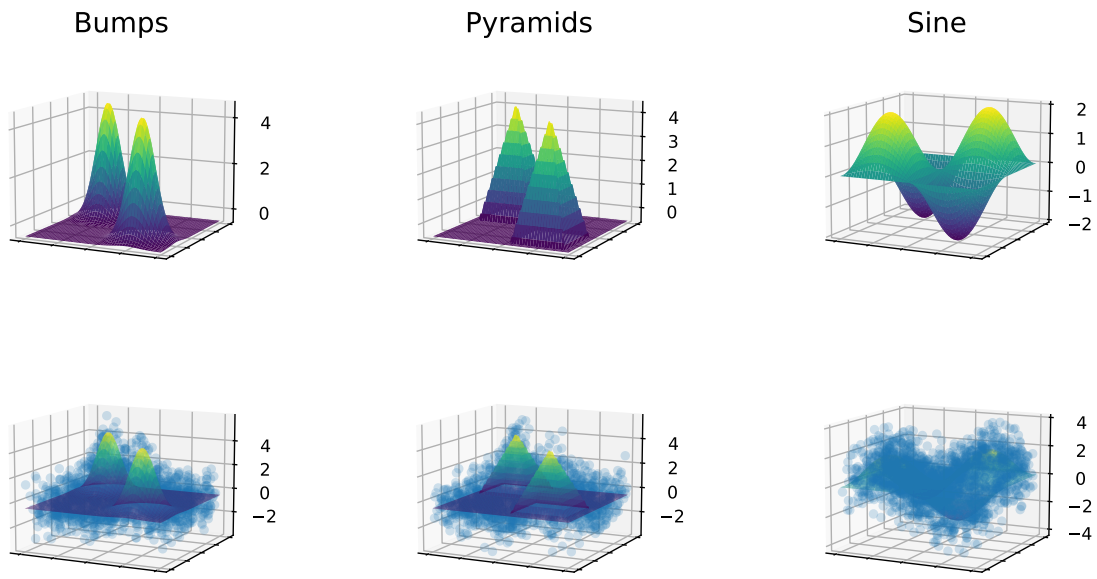

Figure 3.8: *The Bumps, Pyramids, and Sine functions are depicted in the first row, and in the second row we add* $n = 2000$ *evaluations* $f_0(x_i) + z_i$*, where* $z_i \sim N(0, 1)$*. The heterogeneous smoothness of the Bumps and Pyramids functions are apparent, with high signal in upper-left and lower-right quadrants, and low signal in the lower-left and upper-right quadrants.*

Estimators. Using data drawn following (3.36) from each f_0, we fit the Delaunaygram and a thin-plate spline, over a wide range of complexity for both estimators. This process is repeated using ten independent sets of $n = 2000$ observations for each signal function.

Results. In Figure 3.9, we report the mean squared error (MSE) performance of the Delaunaygram and the thin-plate spline in recovering each of the signal functions. At each tuning parameter value, the average MSE over the ten repetitions is reported, with error bars given by the standard error of the average MSE. The values of the tuning parameters themselves have been re-parameterized as the estimated degrees of freedom for the estimators (see Section 3.2, as well as Tibshirani and Taylor, 2012). For the Delaunaygram, an unbiased estimate of the degrees of freedom is the nullity of the generalized lasso penalty operator after removing the rows corresponding to the active set[6], whereas the degrees of freedom for the thin-plate spline, being a linear smoother,

[6]While mathematically the definition of expected degrees of freedom for the Delaunaygram at a fixed λ is clear, the determination of the degrees of freedom via the nullity of a subsetted penalty operator in the presence of numerical error requires two thresholding values which affect the reported degrees of freedom: the first is used to determine the active set (i.e., i such that $(D\theta)_i \neq 0$), and the second is

is the trace of the smoother matrix which takes the observed y to the smoothed $\hat{y}$ at the data points (e.g., Hastie and Tibshirani, 1990).

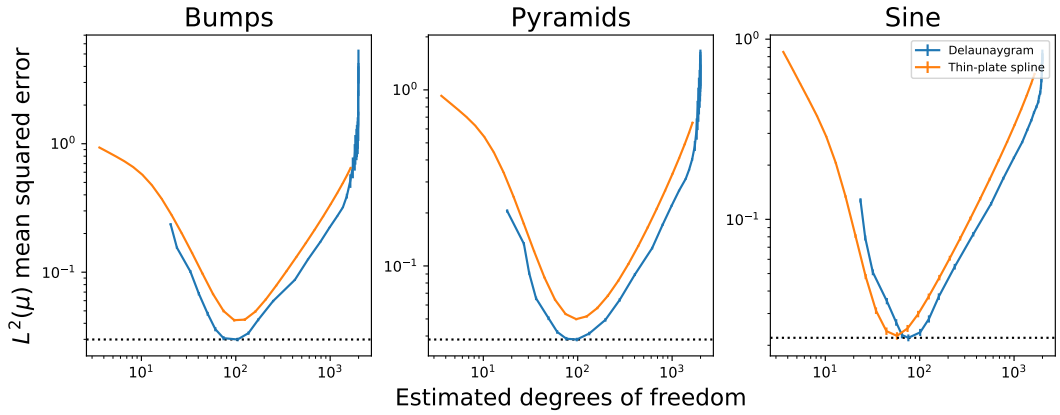

Figure 3.9: *The average MSE (evaluated against the Lebesgue measure) over the ten experimental repetitions is reported over a range of complexity levels for each estimator, with the standard error of the average MSE indicated using vertical bars. While the Delaunaygram and the thin-plate spline perform comparably in recovering the homogeneously smooth Sine function, the Delaunaygram outperforms the thin-plate spline in recovering the heterogeneously smooth Bumps and Pyramids functions in by a noticeable margin.*

We see that for the Bumps and Pyramids functions, which exhibit heterogeneous smoothness, the Delaunaygram outperforms the thin-plate spline by a considerable margin. For the Sine function, which exhibits homogeneous smoothness, the Delaunaygram and the thin-plate spline perform comparably. The gap in performance between the Delaunaygram and the thin-plate spline in the heterogeneous smoothness regime is in keeping with the hypothesized behavior of linear versus nonlinear smoothers under heterogeneous and homogeneous smoothness. We also observe that in the heterogeneously smooth settings, the Delaunaygram is a more efficient estimator than the thin-plate spline in another sense: the best performing Delaunaygram estimator consumes fewer degrees of freedom than the best performing thin-plate spline. In fact, at every level of complexity, holding estimated degrees of freedom fixed, the Delaunaygram is able to use the data more efficiently when estimating the heterogeneously smooth functions.

In Figure 3.10, we depict the Delaunaygram and thin-plate spline estimates from one

used to determine the numerical matrix rank from the calculated singular values. The former source of numerical error arises in the optimization step of determining $\hat{\theta}$ for a fixed λ, and the latter source of numerical error is inherent in the process of diagonalizing a (large, sparse) subsetted penalty matrix. The second source of numerical error may be eliminated by using an equivalent characterization of the degrees of freedom which counts the number of piecewise affine constraints implied by $D_{\mathcal{A}^c}\theta = 0$ to determine $\mathrm{nullity}(D_{\mathcal{A}^c})$. For more details, see Section 3.2 (methods and basic properties).

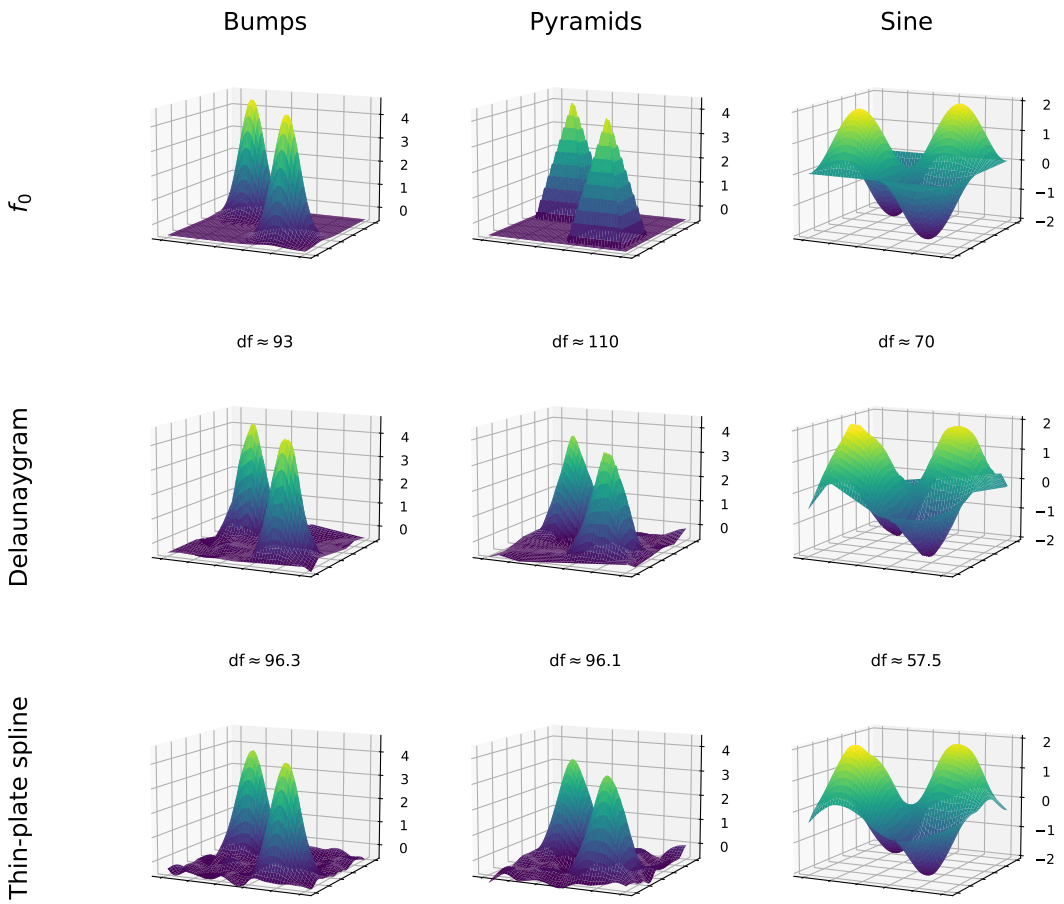

Figure 3.10: *The estimates produced by the Delaunaygram (second row) and the thin-plate spline (third row), each using $n = 2000$ noisy evaluations of the signal function (first row). For the Bumps and Pyramids functions, the thin-plate spline undersmooths in the low-signal regions (lower-left and upper-right quadrants), whereas the Delaunaygram adaptively enforces greater regularization over those regions.*

repetition of the experimental setup, at the minimizing tuning parameter values from Figure 3.9. In assessing the thin-plate spline estimates for the Bumps and Pyramids function, we can clearly see the pitfalls of using a linear estimator (which is not expected to achieve spatial adaptivity) in estimating functions exhibiting heterogeneous smoothness. For both of these functions, the thin-plate spline captures the high-peaked regions well, but at the price of noticeably undersmoothing in the flat regions. The Delaunaygram, in comparison, automatically enforces greater regularization in the low-signal regions and less regularization in the high-signal regions. For the Sine function, the estimates produced by the Delaunaygram and the thin-plate spline are

similar, as anticipated by their comparable performance in average MSE. The estimated degrees of freedom for each of these estimators is also reported; in keeping with Figure 3.9, we see that the oracle-tuned Delaunaygram and thin-plate spline estimators consume roughly the same degrees of freedom. (The apparent discrepancy between reported degrees of freedom in Figures 3.9 and 3.10 is due to Figure 3.9 reporting the degrees of freedom *averaged* over the ten experimental repetitions, whereas Figure 3.10 reports the degrees of freedom as calculated for the estimators for the repetition that is plotted.)

3.4.2 Comparisons on real data

In this second experiment, we consider the estimation of the ocean thermal response to the passage of a tropical cyclone (TC). The data consist of measured temperature differences at a depth of ten meters belong the ocean surface, parameterized in terms of the time since TC passage τ and cross-track angle (distance) d from the TC, referred to as the "TC-centric coordinate system." Prevailing scientific theory suggests that energy flux from the ocean is a driver of TCs, through a process called wind-induced surface heat exchange (see, e.g., Emanuel 1986, 1999). This energy flux can be observed through negative differences in the subsurface ocean temperature, measured before and after TC passage.

Statistical model. The statistical problem under consideration is to estimate the change in temperature at each location relative to the TC, using scattered temperature difference observations. The statistical model for the data is

$$y_i = f(d_i, \tau_i) + z_i, \tag{3.37}$$

where y_i, $i = 1, \ldots, n$, is the observed temperature difference at location (d_i, τ_i) in the TC-centric coordinate system, and $z_i \in N(0, \sigma_i^2)$ is an observation-specific noise term which depends on external factors including the location of the observation in Earth (latitude and longitude) coordinates, time of year, etc. In the scientific setting, σ_i must be estimated using other data; for the purposes of our illustrative example, these variances have already been estimated through a separate procedure. Therefore we pose our estimators as penalized weighted least squares estimators, and use the estimated variances $\hat{\sigma}_i$ as plug-in weights.

Experimental setup. Formally, we have $n = 4202$ observations, which we use to set up a prediction problem comparing the Delaunaygram, the thin-plate spline, and multivariate adaptive regression splines (also known as Mars; see Friedman, 1991). The experimental setup is as follows:

1. Of the 4202 observations, 80% are placed into a training set and 20% are placed into a held-out test set.
2. The training set is used to fit a Delaunaygram model using five-fold cross-validation, a thin-plate spline again using five-fold cross-validation, and a multivariate adaptive regression spline using the default automatic model selection settings in the `earth` package (citation here).
3. Quantitive performance is assessed by model performance on the held-out test set. Qualitative performance is assessed through visual examination of the estimated surface $f(d, \tau)$ over the its domain $(-8, 8) \times (-2, 20)$.

Results. Figure 3.11 provides a comparison of the three estimators in terms of predictive performance. We observe that the Delaunaygram and thin-plate spline achieve comparable predictive performance, as assessed by cross-validation error on the training set and by held-out test set error. The Mars estimator (which performs its own internal model selection process on the training set) performs markedly worse in terms of test set error.

	Test Error
Delaunaygram	0.577
Thin-plate spline	0.580
Mars	0.746

Figure 3.11: Left panel: *Five-fold cross-validation errors for the Delaunaygram and thin-plate spline on the train dataset. The predictive performance of the two estimators largely match, although at each level of model complexity (degrees of freedom), the Delaunaygram does no worse than the thin-plate spline, suggesting a more efficient representation at the same number effective parameters.* Right panel: *The three estimators, learned using the train dataset, are evaluated on the held-out test set. The Delaunaygram and thin-plate spline perform comparably, the error of Mars is markedly worse.*

Figure 3.12 illustrates the ocean thermal response to TC passage, as estimated by the Delaunaygram, thin-plate spline, and Mars. The local adaptivity property of the

Delaunaygram allows it to capture a large decrease in subsurface ocean temperature in the immediate wake of the TC, while smoothing away variation in the rest of the domain. On the other hand, the thin-plate spline, which as a linear smoother enforces the same amount of smoothing on the entire domain, estimates a smaller decrease in ocean temperature in the wake of the TC, while yielding "lumpier" estimates on the rest of the domain. Examination of the Mars estimator suggests that it is not well-suited to a problem of this structure. Mars is able to capture the general notion that there is a decrease in temperature in the wake of the TC and correctly localizes it to a few degrees of zero in the cross-track angle coordinate, but is unable to localize the largest magnitude of effect to about ten days of TC passage, as the Delaunaygram and thin-plate spline do. On the other hand, it is able to enforce greater regularization of the estimates away from the zero cross-track angle axis, unlike the thin-plate spline.

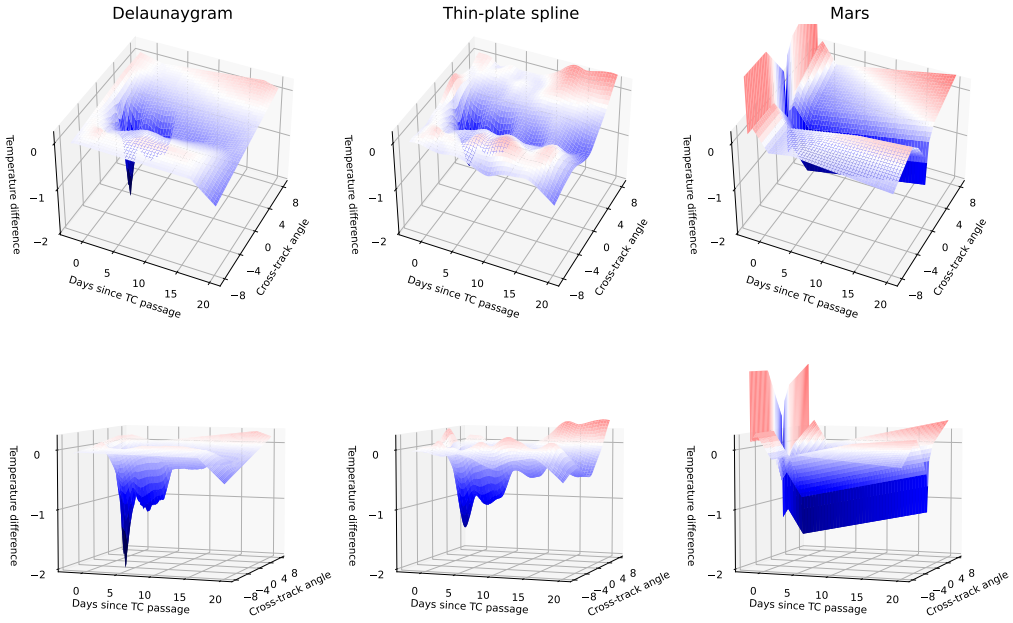

Figure 3.12: *The mean ocean thermal response to the TC passage, as learned using the Delaunaygram, thin-plate spline, and Mars estimators. The Delaunaygram estimator is able to capture a large decrease in temperature in the wake of the TC, while smoothing away variability in the region away from the TC path. The thin-plate spline captures a smaller mean effect, while undersmoothing away from the TC path. Mars generally is unable to accurately capture the ocean thermal response in the wake of the TC.*

In summary, the thin-plate spline is not able to localize the signal, produces a

low-variance estimate that allow it to do well enough in predictive performance. Mars performs some localization of the signal, but yields a high-variance estimate that performs poorly in prediction. The Delaunaygram finds a sweet spot, localizing the signal properly while doing as well as the thin-plate spline in prediction.

3.5 Discussion

This chapter has provided a first-order extension to the results of Chapter 2. We have studied an estimator, the Delaunaygram, which fits a CPWL function on a data-dependent, adaptively chosen partition of the input space.

Although the core idea of the Delaunaygram existed previously in the literature [Koenker and Mizera, 2004, Pourya et al., 2023], we have provided several novel results that elucidate basic properties of the Delaunaygram, including:

- structure of estimates, in particular the types of boundary between fused cells of the Delaunay partition which are possible when the design points lay in general position;
- an estimate of the degrees of freedom, via generalized lasso theory, which has a natural interpretation in terms of the number of free parameters;
- a view of the Delaunaygram as a generalized group lasso estimator, which opens the door to a number of methodological extensions based on penalizing ℓ_p differences of gradients or relaxing the requirement that the fit be continuous.

We have also provided theoretical results which build towards a rate of convergence for estimating BGV functions, including:

- a framework for analyzing penalized triogram estimators, which give a rate of convergence in terms of the discrete gradient variation of the underlying signal when the triangulation upon which the triogram estimator is posed is sufficiently well-behaved;
- probabilistic results ensuring that the Delaunay triangulation using points sampled randomly satisfies the well-behavedness conditions sufficient to provide a rate of convergence in terms of the discrete gradient variation of the underlying signal;
- analysis that provides a path towards connecting the discrete gradient variation of a function to the continuous-time gradient variation of that same function.

In the final chapter to come, we will discuss the general problem of adaptive piecewise polynomial estimation of order $k \geq 2$ in the multivariate scattered data setting.

Chapter 4

Discussion: $k = 2$ and beyond

In this final chapter, we briefly outline the research landscape that remains in the estimation of k^{th}-order bounded variation functions using scattered data. As mentioned in the introductory chapter, when $d = 1$ we have a complete understanding of this problem through the locally adaptive regression spline [Mammen and van de Geer, 1997] and trend filtering [Tibshirani, 2014, 2022] estimators, which penalize using k^{th}-order total variation and achieve optimal rate over the corresponding bounded variation classes. The trend filtering estimator also exhibits several properties that connect discrete- and continuous-time notions of complexity, which we will return to at the end of this chapter.

4.1 Bounded variation classes for $k \geq 2$

In one dimension, the k^{th}-order bounded variation class over an interval $I = (a, b)$ is

$$\mathrm{BV}^k(I) = \Big\{ f : I \to \mathbb{R} \ : \ f \text{ is } k\text{-times weakly differentiable, and } \mathrm{TV}(f^{(k)}) < \infty \Big\}. \tag{4.1}$$

In this thesis, we studied estimation of functions in the bounded variation and bounded gradient variation spaces, corresponding to order indices $k = 0$ and 1, posed over $(0, 1)^d$. These function spaces can be generalized to the k^{th}-order case by defining a k^{th}-order notion of total variation[1] for a function $f : \Omega \to \mathbb{R}$,

$$\mathrm{TV}(D^k f) := \sum_{|\alpha|=k} \mathrm{TV}(D^\alpha f), \tag{4.2}$$

[1]Strictly speaking, this does exactly match our notion of gradient variation as given in Section 1.4. For these notions first-order total variation to match exactly, one would need to replace the Frobenius norm constraint on the test functions ϕ in (1.7) with an $\ell_{2,\infty}$-norm constraint. However, these spaces are equivalent in terms of rates.

where D^α is the partial derivative operator corresponding to the multi-index α. The k^{th}-order bounded variation class is then defined,

$$\mathrm{BV}^k(\Omega) = \Big\{ f : (0,1)^d \to \mathbb{R} \ : f \text{ is } k\text{-times weakly differentiable, and } \mathrm{TV}(D^k f) < \infty \Big\}. \tag{4.3}$$

Note that the BV^k space matches the $W^{k+1,1}$ space while admitting functions whose $(k+1)^{\text{st}}$ derivatives are Radon measures.

4.2 Anticipated rates of convergence

With the k^{th}-order bounded variation classes in hand, we now outline the anticipated rates of convergence for estimating functions from these classes in mean squared error. Over the class BV^k, we anticipate the $L^2(P_n)$ rate of estimation to be

$$\big\|\hat{f} - f_0\big\|_{L^2(P_n)}^2 = \begin{cases} \tilde{O}_{\mathbb{P}}\big(n^{-\frac{2(k+1)}{2(k+1)+d}}\big) & d < 2(k+1), \\ \tilde{O}_{\mathbb{P}}\big(n^{-\frac{k+1}{d}}\big) & d \geq 2(k+1). \end{cases} \tag{4.4}$$

These anticipated rates are based on the discrete-time, lattice-based case, which was comprehensively analyzed by Sadhanala et al. [2017, 2021]. This thesis obtained the result (4.4) for $k = 0$ and (pending resolution of the uniform control issue outlined in Section 3.3.4) for $k = 1$.

We further anticipate that the optimal rate of convergence for estimation in $L^2(P)$ to be

$$\big\|\hat{f} - f_0\big\|_{L^2(P)}^2 = \begin{cases} \tilde{O}_{\mathbb{P}}\big(n^{-\frac{2(k+1)}{2(k+1)+d}}\big) & d < 2(k+1),\ f_0 \in \mathrm{BV}^k(L) \\ \tilde{O}_{\mathbb{P}}\big(n^{-\frac{k+1}{d}}\big) & d \geq 2(k+1),\ f_0 \in \mathrm{BV}^k(L) \cap L^\infty(M), \\ C & d \geq 2(k+1),\ f_0 \in \mathrm{BV}^k(L). \end{cases} \tag{4.5}$$

The additional casing on whether f_0 belongs to the full BV^k space or if it adheres to an additional essential boundedness assumption reflects the impossibility of estimation in the subcritical regime, which we show in this thesis for $k = 0, 1$, and which we anticipate to persist for higher orders[2]. This thesis obtained the result (4.5) for $k = 0$.

4.3 Closely related function spaces and methods

The multivariate k^{th}-order notion of total variation provided in (4.2) is by no means the only way to generalize univariate k^{th}-order total variation. A stream of contempo-

[2]We expect that an appropriate choice of test functions to obtain minimax lower bounds for all of these orders is the family of box splines [de Boor et al., 1993].

raneous work considers various multivariate notions of smoothness which collapse into k^{th}-order total variation in one dimension (under suitable regularity), including:

- *Hardy-Krause, Sectional, and Mars variation*, as studied by Bibaut and van der Laan [2019], Fang et al. [2021], Ki et al. [2021]. These notions of smoothness involve integrating the L^1 norm of mixed partial derivatives of higher order than coordinate-aligned derivatives. Optimal rates of estimation over these classes are derived using estimators which perform constrained least squares, where the constraint is in terms of the Hardy-Krause or Mars complexity measure.
- *Radon variation*, as studied by Parhi and Nowak [2021, 2023]. This notion of smoothness applies the total variation norm (in the sense of measures) to a transformation of the function f into the Radon domain. It has a special "sparsifying" properties for ridge splines, which in fact are the solution to the Radon total variation-penalized empirical risk minimization problem. Optimal rates of estimation over the second-order Radon bounded variation class (which corresponds to $k = 1$ in the indexing of this thesis) are derived using this estimator.
- *Vitali variation*, as studied by Ortelli and van de Geer [2021]. This notion of smoothness applies an L^1 penalty to the mixed partial derivative of order k (along every coordinate index) of the function f. This stands in contrast to Kronecker variation [Sadhanala et al., 2021], which penalizes the sum of variations of the k^{th} (non-mixed) partial derivative along *each* coordinate index. Like Kronecker variation, this notion of smoothness is defined on tensors, and the analysis is entirely in discrete time. A trend filtering-type estimator is defined by applying empirical risk minimization penalizing this notion of smoothness, and rates of estimation over the Vitali variation class are given.

Each of these notions of smoothness yields a function class over which the rate of estimation is *dimension-independent*. These results seem remarkable, in the sense that they apparently defy the "curse of dimensionality," especially when compared to the anticipated rates of convergence over the k^{th}-order bounded variation classes outlined in Section 4.2. It turns out that the differing dependence on dimension in rates between these function spaces is due to the relative sizes of the function classes and complexity of the functions admitted by each of them, and we believe that BV^k classes, which admit a broader set of functions, are interesting to study in their own right.

4.4 What is a multivariate trend filter?

Up until now, we have focused on the estimation setting for the $k \geq 2$ case—bounded variation function spaces and rates of estimation—without explicit reference to an estimator which we might hope attains these rates. We now describe the desired

properties of a hypothetical estimator, a k^{th}-order *multivariate trend filter*, which would generalize the result of this thesis:

1. The estimator takes the form,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \, \mathrm{TV}(D^k f), \tag{4.6}$$

   where $\mathcal{F}_n$ is a finite-dimensional subspace of BV^k.
2. The estimator $\hat{f}$ is a piecewise polynomial of order k.
3. The estimator $\hat{f}$ has an equivalent discrete form,

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \theta_i\big)^2 + \lambda \|D\theta\|_1, \tag{4.7}$$

   where $\|D\theta\|_1$ measures the k^{th}-order discrete total variation of θ.
4. The immediately preceding property implies the existence of an extrapolator taking the fitted parameter vector $\hat{\theta}$ from the discrete problem to a fitted function $\hat{f}$ (coinciding with the solution of the variational problem) with the property that

$$\|D\hat{\theta}\|_1 = \mathrm{TV}(D^k \hat{f});$$

   that is, the k^{th}-order discrete TV of $\hat{\theta}$ matches the k^{th}-order TV of $\hat{f}$.
5. The estimator achieves the $L^2(P_n)$ rate for estimating $f_0 \in \mathrm{BV}^k$,

$$\mathbb{E}\big\|\hat{f} - f_0\big\|_{L^2(P_n)}^2 \lesssim \begin{cases} n^{-\frac{2(k+1)}{2(k+1)+d}} & d < 2(k+1), \\ n^{-\frac{k+1}{d}} & d \geq 2(k+1), \end{cases}$$

   and the minimax rate in $L^2(P)$.
6. The estimator is adaptive to the local degree of smoothness in practice.

The estimators analyzed in Chapters 2 and 3 of this thesis satisfy all of these properties (with the anticipated rates of convergence for the Delaunaygram). We look forward to future work with the hope that all of these properties may be satisfied for orders $k \geq 2$.

Bibliography

Luigi Ambrosio, Shayan Aziznejad, Camillo Brena, and Michael Unser. Linear inverse problems with Hessian-Schatten total variation. *arXiv preprint arXiv:2210.04077*, 2022. 1.5

Luigi Ambrosio, Camillo Brena, and Sergio Conti. Functions with bounded Hessian-Schatten variation: density, variational and extremality properties. *arXiv preprint arXiv:2302.12554*, 2023. 1.4

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. 1.5

Franz Aurenhammer and Rolf Klein. Voronoi diagrams. *Handbook of Computational Geometry*, 5(10):201–290, 2000. 2.2.4

Shayan Aziznejad, Joaquim Campos, and Michael Unser. Measuring complexity of learning schemes using hessian-schatten total variation. *SIAM Journal on Mathematics of Data Science*, 5(2):422–445, 2023. 1.5

Marshall Bern, David Eppstein, and Frances Yao. The expected extremes in a Delaunay triangulation. *International Journal of Computational Geometry & Applications*, page 13, 1991. 3.3.2, B.2.3

Aurélien F. Bibaut and Mark J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. arXiv: 1907.09244, 2019. 4.3

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011. B.1.1

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. 2.2.4

Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997. 1.5

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. 2.2.4

Tony Chan, Antonio Marquina, and Pep Mulet. High-order total variation-based image restoration. *SIAM Journal on Scientific Computing*, 22(2):503–516, 2000. 1.5

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, 2010. A.7.1

Ronald R. Coifman and Stephane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. 3

Richard Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49(12):1–23, 1943. 1.5, 3.2.2

Carl de Boor, Klaus Höllig, and Sherman Riemenschneider. *Box Splines*. Springer, 1993. 2

Miguel del Álamo, Housen Li, and Axel Munk. Frame-constrained total variation regularization for white noise regression. *Annals of Statistics*, 49(3), 2021. 1.5, A.4.2

Françoise Demengel. Fonctions à Hessian borné. *Annales de l'Institut Fourier*, 34(2): 155–190, 1984. 1.4, 1.5

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

Jean Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, 1977. 1

Rex A. Dwyer. Higher-dimensional Voronoi diagrams in linear expected time. *Discrete & Computational Geometry*, 6(3):343–367, 1991. 2.2.4

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988. 9

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986. 2.2.2, 3.2.4

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society: Graduate Studies in Mathematics, 2010. Second edition. A.4.4

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 2015. Revised edition. 1.2, 3, A.1.1, A.4.4, B.1.1

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *Annals of Statistics*, 49(2):769–792, 2021. 4.3

Jerome Friedman, Trevor Hastie, Holger Hoefling, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007. 1.5

Jerome H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, pages 1–67, 1991. 3.4.2

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. 2.2.4

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. 2.2.3, 2.2.4, 2.3, 2, 2

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. 2.2.3, 2.2.4, 2.3, 2, A.2.4

Nicolás García Trillos, Dejan Slepčev, and James Von Brecht. Estimating perimeter using graph cuts. *Advances in Applied Probability*, 49(4):1067–1090, 2017. A.2.3

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. 2.2.4

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. 1.5, 2.5.3, A.1.1, A.4.4

Alden Green, Sivaraman Balakrishnan, and Ryan J. Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian eigenmaps on neighborhood graphs. arXiv:2111.07394, 2021b. 1.5, 2.5.3, A.1.1

Mark Hansen, Charles Kooperberg, and Sylvain Sardy. Triogram models. *Journal of the American Statistical Association*, 93(441):101–119, 1998. 1.5

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990. 2.2.2, 3.2.4, 3.4.1

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. A.2.3

Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010. 1.5, 2.2.4

Addison J. Hu, Alden Green, and Ryan J. Tibshirani. The Voronoigram: Minimax estimation of bounded variation functions from scattered data. *arXiv preprint arXiv:2212.14514*, 2022. 2.1

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. In *Proceedings of the Annual Conference on Learning Theory*, 2016. 1.5, 2.5.3, A.5, A.5

Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Cambridge University Press, 2015. Draft version. 2.5.1

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. 1.5

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. 1.5

Dohyeong Ki, Billy Fang, and Adityanand Guntuboyina. MARS via LASSO. arXiv: 2111.11694, 2021. 4.3

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. 1.5

Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005. 1.5, 2.1.1

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. 1.5, 1.5, 3.1.1, 3.1.2, 3.2, 3.2.1, 3.2.4, 3, 3.3.1, 3.5

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994. 1.5

Loic Landrieu and Guillaume Obozinski. Cut pursuit: fast algorithms to learn piecewise constant functions on general weighted graphs. HAL: 01306779, 2015. 2.2.4

Oleg V. Lepski and Vladimir G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, 25(6):2512–2546, 1997. 1.5

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. 1.5

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1):387–413, 1997. 1.5, 2.5.1, 4

Frederik Riis Mikkelsen and Niels Richard Hansen. Degrees of freedom for piecewise Lipschitz estimators. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 54(2):819–841, 2018. 21, B.1.1

Roger E. Miles. On the elimination of edge effects in planar sampling. *Stochastic geometry*, 10:228–247, 1974. 3.3.2

Michael H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statistica Sinica*, 10(2):399–431, 2000. 1.5

Francesco Ortelli and Sara van de Geer. Tensor denoising with trend filtering. arXiv: 2101.10692, 2021. 4.3

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. 2.2.4

Alexander M. Ostrowski. A quantitative formulation of Sylvester's law of inertia. *Proceedings of the National Academy of Sciences*, 45(5):740–744, 1959. 10

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. 1.5, 2.5.3, A.5, B.2.1

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. 1.5, 2.2.3, 9, 2.5.3, A.5, A.5, B.2.1

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021. 4.3

Rahul Parhi and Robert D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *Transactions on Information Theory*, 69(2):1125–1140, 2023. 4.3

Mehrsa Pourya, Alexis Goujon, and Michael Unser. Delaunay-triangulation-based learning with Hessian total-variation regularization. *IEEE Open Journal of Signal Processing*, 4:167–178, 2023. 1.5, 3.1.1, 3.2.1, 3.2.4, 3, 3.2.7, 3.5

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. 1.5

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 1.3, 1.5

Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models via trend filtering. *Annals of Statistics*, 47(6):3032–3068, 2019. 2.5.1

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. 1.5, 8, 2.5.3, A.5

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. 1.5, 2.5.3, 4.2, A.5

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, Addison J. Hu, and Ryan J. Tibshirani. Multivariate trend filtering for lattice data. arXiv: 2112.14758, 2021. 1.5, 2.5.3, 3.1.2, 3.3, 18, 4.2, 4.3, B.2.1, B.2.1, B.2.1, B.2.2, 23

Isaac J. Schoenberg. Spline functions and the problem of graduation. *Proceeding of the National Academy of Sciences*, 52(4):947–950, 1964. 1

Donald R. Schuette. A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30, 1978. 1.5

Gabriel Steidl, Stephan Didas, and Julia Neumann. Splines in higher order TV regularization. *International Journal of Computer Vision*, 70(3):214–255, 2006. 1.5

Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981. 2.2.2

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982. 7

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. 1.3, 1.5, 1.5

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013. 2

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014. 1.5, 2.6, 3.2.6, 4

Ryan J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, 25(3): 1265–1296, 2015. 2.2.2

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. 1.5, 2.6, 4

Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011. 1.5, 2.2.2, 2.2.2, 2.2.4, 12, 3.2.4

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232, 2012. 2.2.2, 2.2.2, 2, 3.2.4, 3.2.4, 3.4.1

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. A.4.2, 9, 29

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. 2.1.1

Curtis R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996. 1.5

Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(1): 1751–1798, 2014. A.6

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. 10

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. 1.5, 2.2.3, 2.2.4, 2.5.3, A.5, B.2.1, B.2.1, B.2.2

Appendix A

Supplement to Chapter 2

A.1 Added details and proofs for Sections 2.1 and 2.2

A.1.1 Discussion of sampling model for BV functions

We clarify what is meant by the sampling model in (1.1), since, strictly speaking, each element $f \in \mathrm{BV}(\Omega)$ is really an equivalence class of functions, defined only up to sets of Lebesgue measure zero. This issue is not simply a formality, and becomes a genuine problem for $d \geq 2$, as in this case the space $\mathrm{BV}(\Omega)$ does not compactly embed into $C^0(\Omega)$, the space of continuous functions on Ω (equipped with the L^∞ norm). A key implication of this is that the point evaluation operator is not continuous over $\mathrm{BV}(\Omega)$.

In order to make sense of the evaluation map, $x \mapsto f(x)$, we will pick a representative, denoted $f^\star \in f$, and speak of evaluations of this representative. Our approach here is the same as that taken in Green et al. [2021a,b], who study minimax estimation of Sobolev functions in the subcritical regime (and use an analogous random design model). We let $f^\star$ be the *precise representative*, defined [Evans and Gariepy, 2015] as:

$$f^\star(x) = \begin{cases} \lim\limits_{\epsilon \to 0} \dfrac{1}{\mu(B(x,\epsilon))} \displaystyle\int_{B(x,\epsilon)} f(z)\, dz & \text{if the limit exists} \\ 0 & \text{otherwise.} \end{cases}$$

Here μ denotes Lebesgue measure and $B(x, \epsilon)$ is the ball of radius ϵ centered at x.

Now we explain why the particular choice of representative is not crucial, and any choice of representative would have resulted in the same interpretation of function evaluations in (1.1), *almost surely, assuming that each x_i is drawn from a continuous*

distribution on Ω. Recall that for a locally integrable function f on Ω, we say that a given point $x \in \Omega$ is a *Lebesgue point* of f provided that $\lim_{\epsilon \to 0}(\int_{B(x,\epsilon)} f(z)\,dz)/\mu(B(x,\epsilon))$ exists and equals $f(x)$. By the Lebesgue differentiation theorem (e.g., Theorem 1.32 of Evans and Gariepy, 2015), for any $f \in L^1(\Omega)$, almost every $x \in \Omega$ is a Lebesgue point of f. This means that each evaluation $f^\star(x_i)$ of the precise representative will equal the evaluation of any member of the equivalence class, almost surely (with respect to draws of x_i). This justifies the notation $f(x_i)$ used in the main text, for $f \in \mathrm{BV}(\Omega)$ and x_i drawn from a continuous probability distribution.

A.1.2 TV representation for piecewise constant functions

Here we provide result from which Proposition 1 will follow. First we give a more general definition of measure theoretic total variation, wherein the norm used to constrain the "test function" ϕ in the supremum is an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$,

$$\mathrm{TV}(f;\Omega,\|\cdot\|) = \sup\Big\{ \int_\Omega f(x)\,\mathrm{div}\,\phi(x)\,dx : \phi \in C_c^1(\Omega;\mathbb{R}^d), \\ \|\phi(x)\| \le 1 \text{ for all } x \in \Omega \Big\}. \tag{A.1}$$

Note that our earlier definition in (1.2) corresponds to the special case $\mathrm{TV}(f;\Omega,\|\cdot\|_2)$, that is, corresponds to choosing $\|\cdot\| = \|\cdot\|_2$ in (A.1). In the more general TV context, this special case is often called *isotropic* TV.

Proposition 8. Let $V_1,\dots,V_n$ be an open partition of Ω such that each V_i is semialgebraic. Let f be of the form

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

for arbitrary $\theta_1,\dots,\theta_n \in \mathbb{R}$. Then, for any norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$ (induced by the Euclidean inner product), we have

$$\mathrm{TV}(f;\Omega,\|\cdot\|) = \sum_{i,j=1}^n \bigg(\int_{\partial V_i \cap \partial V_j} \|n_i(t)\|_*\, d\mathcal{H}^{d-1}(t) \bigg) \cdot |\theta_i - \theta_j|,$$

where $n_i(t)$ is the measure theoretic unit outer normal for V_i at a boundary point $t \in \partial V_i$. In particular, in the isotropic case $\|\cdot\| = \|\cdot\|_2$,

$$\mathrm{TV}(f;\Omega,\|\cdot\|_2) = \sum_{i,j=1}^n \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j) \cdot |\theta_i - \theta_j|.$$

Proposition 8 is a special case of Proposition 3 with $d_2 = 1$ and $m = n$. The proof of Proposition 3 is given in Appendix B.1.

Remark 21. The condition that each V_i is semialgebraic may to weakened to what is called "polynomially bounded boundary measure." Namely, the proposition still holds if each map $r \mapsto \mathcal{H}^{d-1}(\partial V_i \cap B(0,r))$ is polynomially bounded (cf. Assumption 2.2 in Mikkelsen and Hansen, 2018). This is sufficient to guarantee a locally Lipschitz boundary (a prerequisite for the application of Gauss-Green) and to characterize the outer normals associated with the partition $V_1, \dots, V_n$.

A.2 Proofs for Section 2.3

A.2.1 Roadmap for the proof of Theorem 1

The proof of Theorem 1 consists of several parts, and we summarize them below. Some remarks on notation: throughout this section, we use σ_{Vor} for the constant c_d appearing in (2.18), and we abbreviate $\|\cdot\| = \|\cdot\|_2$. Also, we use $C^1(\Omega)$ and $C^2(\Omega)$ to denote the spaces of continuously differentiable and twice continuously differentiable functions, respectively, equipped with the L^∞ norm.

1. An edge $\{i,j\}$ in the Voronoi graph depends not only on x_i and x_j but also on all other design points $x_k, k \neq i,j$. In Lemma 5, we start by showing that the randomness due this dependence on $x_k, k \neq i,j$ is negligible,

$$\mathbb{E}\Big[\big(\mathrm{DTV}(f; w^{\mathrm{V}}) - U_{n,\mathrm{Vor}}(f)\big)^2\Big] \leq C\frac{\|f\|^2_{C^1(\Omega)}(\log n)^{(d+2)/d}}{n^{1/d}}, \tag{A.2}$$

   for a constant $C > 0$. The functional $U_{n,\mathrm{Vor}}(f)$ is an order-2 U-statistic,

$$U_{n,\mathrm{Vor}}(f) = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \big|f(x_i) - f(x_j)\big| H_{\mathrm{Vor}}(x_i, x_j),$$

   with kernel $H_{\mathrm{Vor}}(x,y)$ defined by

$$H_{\mathrm{Vor}}(x,y) = \mathbb{E}\big[\mathcal{H}^{d-1}(\partial V_i \cap \partial V_j)|x_i, x_j\big] = \int_{L\cap\Omega} \big(1 - p_x(z)\big)^{(n-2)}\, dz.$$

   Here $L = L_{xy}$ is the $(d-1)$-dimensional hyperplane $L = \{z : \|x-z\| = \|y-z\|\}$, and $p_x(z) = P(B(z, \|x-z\|))$. (Note that $p_x(z) = p_y(z)$ for all $z \in L$.)

2. We proceed to separately analyze the variance and bias of $U_{n,\mathrm{Vor}}(f)$. In Lemma 6, we establish that $U_{n,\mathrm{Vor}}(f)$ concentrates around its mean, giving the estimate, for a constant $C > 0$,

$$\mathrm{Var}\big[U_{n,\mathrm{Vor}}(f)\big] \leq C\frac{(\log n)^3}{n}\|f\|^2_{C^1(\Omega)}. \tag{A.3}$$

3. It remains to analyze the bias, the difference between the expectation of $U_{n,\mathrm{Vor}}(f)$ and continuum TV. Lemma 7 leverages the fact that the kernel $H_{\mathrm{Vor}}(x, y)$ is close to a spherically symmetric kernel—at least at points x, y sufficiently far from the boundary of Ω—to show that the expectation of the U-statistic $U_{n,\mathrm{Vor}}(f)$ is close to (an appropriately rescaled version of) the nonlocal functional

$$\mathrm{TV}_{\varepsilon,K}\big(f;\Omega,h\big) := \int_\Omega \int_\Omega |f(x)-f(y)| K_{\mathrm{Vor}}\bigg(\frac{\|y-x\|}{\varepsilon(x)}\bigg) h(x)\,dy\,dx, \tag{A.4}$$

for bandwidth $\varepsilon(x) = (np(x))^{-1/d}$, weight $h(x) = (p(x))^{(d+1)/d}$, and kernel $K_{\mathrm{Vor}}(t)$ defined in (A.10). Lemma 8 in turn shows that this nonlocal functional is close to (a scaling factor) times $\int_\Omega \|\nabla f\|$. Together, these lemmas imply that

$$\lim_{n\to\infty} \mathbb{E}[U_{n,\mathrm{Vor}}(f)] = \sigma_{\mathrm{Vor}} \int_\Omega \|\nabla f(x)\|\,dx. \tag{A.5}$$

Combining (A.2), (A.3), and (A.5) with Chebyshev's inequality implies the consistency result stated in (2.18). In the rest of this section, across Sections A.2.2–A.2.4, we state and prove the various lemmas referenced above.

A.2.2 Step 1: Voronoi TV approximates Voronoi U-statistic

Lemma 5 upper bounds the expected squared difference between Voronoi TV and the U-statistic $U_{n,\mathrm{Vor}}(f)$.

Lemma 5. *Suppose $x_{1:n}$ are sampled independently from a distribution P satisfiying A1. There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$ sufficiently large, and any $f \in C^1(\Omega)$,*

$$\mathbb{E}\Big[\big(\mathrm{DTV}(f;w^{\mathrm{V}}) - U_{n,\mathrm{Vor}}(f)\big)^2\Big] \le C\frac{\|f\|^2_{C^1(\Omega)}(\log n)^{(d+2)/d}}{n^{1/d}}.$$

Proof of Lemma 5. We begin by introducing some notation and basic inequalities used throughout this proof. Take $\varepsilon_0 = (\log n/n)^{1/d}$. Let $B_x(z) := B^o(z, \|x-z\|)$ denote the open ball centered at z of radius $\|x-z\|$, and note that by our assumptions on p, we have $p_x(z) := P(B_x(z))$. We will repeatedly use the estimates

$$p_x(z) \ge \frac{p_{\min}}{2d}\mu_d\|x-z\|^d,$$

and therefore for $c_1 = \frac{p_{\min}}{2d}\mu_d$,

$$(1-p_x(z))^n \le \exp(-c_1 n\|x-z\|^d).$$

It follows by Lemma 18 that for any constants $a, c > 0$, there exists a constant $C > 0$ depending only on a, c and d such that

$$\int_{L\cap\Omega} (1 - cp_x(z))^n \leq C\bigg(\frac{1\{\|x-y\| \leq C\varepsilon_0\}}{n^{(d-1)/d}} + \frac{1}{n^5}\bigg).$$

We will assume $n \geq 8$, so that the same estimate holds with respect to $n - 4 \geq n/2$. Finally for simplicity write $\Delta(x_i, x_j) := |f(x_i) - f(x_j)|\big(\mathcal{H}^{d-1}(\partial V_i \cap \partial V_j) - H_{\mathrm{Vor}}(x_i, x_j)\big)$.

We note immediately that, because $x_{1:n}$ are identically distributed, it follows from linearity of expectation that

$$\begin{aligned}
\mathbb{E}\Big[\big(\mathrm{DTV}_{n,\mathrm{Vor}}(f; w^{\mathrm{V}}) - U_{n,\mathrm{Vor}}(f)\big)^2\Big] &= \binom{n}{2}\mathbb{E}[\big(\Delta(x_1, x_2)\big)^2] \\
&\quad + \binom{n}{3}\mathbb{E}\big[\Delta(x_1, x_2)\Delta(x_1, x_3)\big] \\
&\quad + \binom{n}{4}\mathbb{E}\big[\Delta(x_1, x_2)\Delta(x_3, x_4)\big] \\
&=: \binom{n}{2}T_1 + \binom{n}{3}T_2 + \binom{n}{4}T_3.
\end{aligned}$$

We separately upper bound $|T_1|$ (which will make the main contribution to the overall upper bound) and $|T_2|$ and $|T_3|$ (which will be comparably negligible). In each case, the general idea is to use the fact that the fluctuations of the Voronoi edge weights $\mathcal{H}^{d-1}(\partial V_1 \cap \partial V_2)$ around the conditional expectation $H_{\mathrm{Vor}}(x_1, x_2)$ are small unless x_1 and x_2 are close together.

Upper bound on T_1. We begin by conditioning on x_1, x_2, and considering the conditional expectation

$$\mathbb{E}\big[(\Delta(x_1, x_2))^2 | x_1, x_2\big] = |f(x_1) - f(x_2)|^2 \mathrm{Var}(\mathcal{H}^{d-1}(\partial V_1 \cap \partial V_2)|x_1, x_2).$$

By Jensen's inequality,

$$\begin{aligned}
\mathrm{Var}(\mathcal{H}^{d-1}(\partial V_1 \cap \partial V_2)|x_1, x_2) &\leq \mathcal{H}^{d-1}(L \cap \Omega) \int_{L\cap\Omega} \mathrm{Var}\big(1\{P_n(B_{x_1}(z)) = 0\}|x_1\big)\, dz \\
&= \mathcal{H}^{d-1}(L \cap \Omega) \int_{L\cap\Omega} \big(1 - p_{x_1}(z)\big)^{(n-2)}\, dz \\
&\leq C\Big(\frac{1}{n^{(d-1)/d}} 1\{\|x_1 - x_2\| \leq C\varepsilon_0\} + \frac{1}{n^5}\Big).
\end{aligned}$$

Taking expectation over x_1 and x_2 gives

$$\begin{aligned}
T_1 &\le C\Big(\frac{\|f\|^2_{C^1(\Omega)}}{n^{(d-1)/d}}\int_\Omega\int_\Omega \|x-y\|^2 1\{\|x-y\|\le C\varepsilon_0\}\,dy\,dx + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big)\\
&\le C\Big(\frac{\|f\|^2_{C^1(\Omega)}\varepsilon_0^{(d+2)}}{n^{(d-1)/d}} + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big)\\
&= C\Big(\frac{\|f\|^2_{C^1(\Omega)}(\log n)^{(d+2)/d}}{n^{(2+1/d)}} + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big).
\end{aligned}$$

Upper bound on T_2. Again we begin by conditioning, this time on $x_{1:3}$, meaning we consider

$$\begin{aligned}
&\mathbb{E}\big[\Delta(x_1,x_2)\Delta(x_1,x_3)|x_{1:3}\big]\\
&\quad = |f(x_1)-f(x_2)||f(x_1)-f(x_3)|\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2),\mathcal{H}^{d-1}(\partial V_1\cap\partial V_3)|x_{1:3}\big].
\end{aligned}$$

We begin by focusing on this conditional covariance. Write $L=\{z\in\Omega:\|z-x_1\|=\|z-x_2\|\}$ and likewise $L'=\{z\in\Omega:\|z-x_1\|=\|z-x_3\|\}$. Exchanging covariance with integration gives

$$\begin{aligned}
&\Big|\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2),\mathcal{H}^{d-1}(\partial V_1\cap\partial V_3)|x_{1:3}\big]\Big|\\
&\le \int_L\int_{L'}\Big|\mathrm{Cov}\big[1\{P_n(B_{x_1}(z))=0\},1\{P_n(B_{x_1}(z'))=0\}|x_{1:3}\big]\Big|\,dz'\,dz\\
&\overset{(i)}{\le} \int_L\int_{L'}\Big(1-\frac{p_{x_1}(z)+p_{x_1}(z')}{2}\Big)^{(n-3)}\,dz\,dz'\\
&+\int_L\int_{L'}(1-p_{x_1}(z))^{(n-3)}(1-p_{x_1}(z'))^{(n-3)}\,dz'\,dz\\
&\le C\Big(\frac{1}{n^{(d-1)/d}}1\{\|x_1-x_2\|\le C\varepsilon_0\}+\frac{1}{n^5}\Big)\Big(\frac{1}{n^{(d-1)/d}}1\{\|x_1-x_3\|\le C\varepsilon_0\}+\frac{1}{n^5}\Big)\\
&\le C\Big(\frac{1}{n^{2(d-1)/d}}1\{\|x_1-x_2\|\le C\varepsilon_0\}1\{\|x_1-x_3\|\le C\varepsilon_0\}+\frac{1}{n^5}\Big).
\end{aligned}\tag{A.6}$$

The inequality (i) follows first from the standard fact that for positive random variables X and Y, $\big|\mathrm{Cov}[X,Y]\big|\le\mathbb{E}[XY]+\mathbb{E}[Y]\mathbb{E}[X]$, and second from the upper bound

$$\begin{aligned}
\mathbb{E}\Big[1\{P_n(B_{x_1}(z))=0\},1\{P_n(B_{x_1}(z'))=0\}\Big] &\le \Big(1-P\big(B_{x_1}(z)\cup B_{x_1}(z')\big)\Big)^{(n-3)}\\
&\le \left(1-\frac{P\big(B_{x_1}(z)\big)+P\big(B_{x_1}(z')\big)}{2}\right)^{(n-3)}.
\end{aligned}$$

Taking expectation over $x_{1:3}$, we have

$$\begin{aligned}T_2 &\le C\Big(\frac{\|f\|^2_{C^1(\Omega)}}{n^{2(d-1)/d}}\int_\Omega\int_\Omega\int_\Omega \|x-y\|\|x-z\|1\{\|x-y\|\le C\varepsilon_0\}1\{\|x-z\|\le C\varepsilon_0\}\,dz\,dy\,dx \\ &\qquad\qquad + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big) \\ &\le C\Big(\frac{\|f\|^2_{C^1(\Omega)}\varepsilon_0^{2(d+1)}}{n^{2(d-1)/d}} + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big) \\ &= C\Big(\frac{\|f\|^2_{C^1(\Omega)}(\log n)^{2(d+1)/d}}{n^4} + \frac{\|f\|^2_{L^\infty(\Omega)}}{n^5}\Big).\end{aligned}$$

Upper bound on T_3. Again we begin by conditioning, this time on $x_{1:4}$, so that

$$\begin{aligned}&\mathbb{E}\big[\Delta(x_1,x_2)\Delta(x_3,x_4)|x_{1:4}\big] \\ &\quad = |f(x_1)-f(x_2)||f(x_3)-f(x_4)|\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2),\mathcal{H}^{d-1}(\partial V_3\cap\partial V_4)|x_{1:4}\big],\end{aligned}$$

Write $L=\{z\in\Omega:\|z-x_1\|=\|z-x_2\|\}$ and likewise $L'=\{z\in\Omega:\|z-x_3\|=\|z-x_4\|\}$, we focus on the conditional covariance

$$\begin{aligned}&\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2),\mathcal{H}^{d-1}(\partial V_3\cap\partial V_4)|x_{1:4}\big] \\ &\quad = \int_L\int_{L'}\mathrm{Cov}\big[1\{P_n(B_{x_1}(z))=0\},1\{P_n(B_{x_3}(z'))=0\}|x_{1:4}\big]\,dz'\,dz\end{aligned}$$

We now show that this covariance is very small unless x_1 and x_3 are close. Specifically, suppose $\|x_1-x_3\|>\varepsilon_0$. Then either $\|z-x_1\|\ge\varepsilon_0/3$, or $\|z'-x_3\|\ge\varepsilon_0/3$, or $B_{x_1}(z)\cap B_{x_3}(z')=\emptyset$. In either of the first two cases, we have that

$$\begin{aligned}&\Big|\mathrm{Cov}\big[1\{P_n(B_{x_1}(z))=0\},1\{P_n(B_{x_3}(z'))=0\}|x_{1:4}\big]\Big| \\ &\quad\le 2\exp(-\frac{p_{\min}}{4d}(n-4)\|x_1-z\|^d)\exp(-\frac{p_{\min}}{4d}(n-4)\|x_3-z'\|^d)\} \\ &\quad\le 2\exp(-\frac{p_{\min}}{4d}(n-4)\varepsilon_0^d)\le\frac{C}{n^5}.\end{aligned}$$

In the third case, it follows that $P(B_{x_1}(z)\cup B_{x_3}(z'))=p_{x_1}(z)+p_{x_3}(z)$. Assume $x_3,x_4\notin B_{x_1}(z)$, and likewise $x_1,x_2\notin B_{x_3}(z')$, otherwise there is nothing to prove. We use the definition of covariance $\mathrm{Cov}[X,Y]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$ to obtain the

upper bound,

$$\begin{aligned}
&\Big|\mathrm{Cov}\big[1\{P_n(B_{x_1}(z))=0\}, 1\{P_n(B_{x_3}(z'))=0\}|x_{1:4}\big]\Big| \\
&\quad= \big|(1-(p_{x_1}(z)+p_{x_3}(z)))^{(n-4)} - (1-p_{x_1}(z))^{(n-4)}(1-p_{x_3}(z))^{(n-4)}\big| \\
&\quad= (1-p_{x_1}(z))^{(n-4)}(1-p_{x_3}(z))^{(n-4)}\bigg|\Big(1-\frac{p_{x_1}(z)p_{x_3}(z)}{(1-p_{x_1}(z))(1-p_{x_3}(z))}\Big)^{(n-4)} - 1\bigg| \\
&\quad\le (1-p_{x_1}(z))^{(n-4)}(1-p_{x_3}(z))^{(n-4)}p_{x_1}(z)p_{x_3}(z)n \\
&\quad\le p_{\max}^2\mu_d^2\exp(-\frac{p_{\min}}{4d}(n-4)\|x_1-z\|^d)\exp(-\frac{p_{\min}}{4d}(n-4)\|x_2-z\|^d)\|x_1-z\|^d\|x_3-z'\|^d n \\
&\quad\le C\exp(-\frac{p_{\min}}{4d}(n-4)\|x_1-z\|^d)\exp(-\frac{p_{\min}}{4d}(n-4)\|x_2-z\|^d)\varepsilon_0^{2d}n
\end{aligned}$$

Integrating over z, z', it follows that if $\|x_1-x_3\|>\varepsilon_0$, then

$$\begin{aligned}
&\Big|\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2), \mathcal{H}^{d-1}(\partial V_3\cap\partial V_4)|x_{1:4}\big]\Big| \\
&\qquad\le C\Big(\frac{\varepsilon_0^{2d}}{n^{(d-2)/d}}1\{\|x_1-x_2\|\le C\varepsilon_0\}1\{\|x_3-x_4\|\le C\varepsilon_0\}+\frac{1}{n^5}\Big).
\end{aligned}$$

Otherwise $\|x_1-x_3\|\le\varepsilon_0$, and using the same inequalities as in (A.6), we find that

$$\begin{aligned}
&\Big|\mathrm{Cov}\big[\mathcal{H}^{d-1}(\partial V_1\cap\partial V_2), \mathcal{H}^{d-1}(\partial V_3\cap\partial V_4)|x_{1:4}\big]\Big| \\
&\le C\Big(\frac{1}{n^{2(d-1)/d}}1\{\|x_1-x_2\|\le C\varepsilon_0\}1\{\|x_3-x_4\|\le C\varepsilon_0\}\{\|x_1-x_3\|\le\varepsilon_0\}+\frac{1}{n^5}\Big).
\end{aligned}$$

Taking expectation over $x_{1:4}$, we conclude that

$$\begin{aligned}
T_3 &\le C\bigg(\frac{\varepsilon_0^{2d}\|f\|_{C^1(\Omega)}^2}{n^{(d-2)/d}}\int_\Omega\int_\Omega\int_\Omega\int_\Omega \|x-y\|\|h-z\| \\
&\qquad\qquad\times 1\{\|x-y\|\le C\varepsilon_0\}1\{\|h-z\|\le C\varepsilon_0\}\,dh\,dz\,dy\,dx \\
&\qquad+\frac{\|f\|_{C^1(\Omega)}^2}{n^{2(d-1)/d}}\int_\Omega\int_\Omega\int_\Omega\int_\Omega \|x-y\|\|h-z\| \\
&\qquad\qquad\times 1\{\|x-y\|C\le\varepsilon_0\}1\{\|h-z\|\le C\varepsilon_0, \|x-h\|\le\varepsilon_0\}\,dh\,dz\,dy\,dx \\
&\qquad+\frac{\|f\|_{L^\infty(\Omega)}^2}{n^5}\bigg) \\
&\le C\Big(\frac{\|f\|_{C^1(\Omega)}^2\varepsilon_0^{4d+2}}{n^{(d-2)/d}}+\frac{\|f\|_{C^1(\Omega)}^2\varepsilon_0^{3d+2}}{n^{2(d-1)/d}}+\frac{\|f\|_{L^\infty(\Omega)}^2}{n^5}\Big) \\
&= C\Big(\frac{\|f\|_{C^1(\Omega)}^2(\log n)^{(4d+2)/d}}{n^5}+\frac{\|f\|_{L^\infty(\Omega)}^2}{n^5}\Big).
\end{aligned}$$

Combining our upper bounds on T_1-T_3 gives the claim of the lemma. □

A.2.3 Step 2: Variance of Voronoi U-statistic

Lemma 6 leverages classical theory regarding order-2 U-statistics to show that the Voronoi U-statistic $U_{n,\mathrm{Vor}}(f)$ concentrates around its expectation. This is closely related to an estimate provided in García Trillos et al. [2017], but not strictly implied by that result: it handles a specific kernel H_{Vor} that is not compactly supported, and functions f besides $f(x) = 1\{x \in A\}$ for some $A \subseteq \Omega$.

Lemma 6. *Suppose $x_{1:n}$ are sampled independently from a distribution P satisfiying A1. There exists a constant $C > 0$ such that for any $f \in C^1(\Omega)$,*

$$\mathrm{Var}\big[U_{n,\mathrm{Vor}}(f)\big] \le C\frac{(\log n)^3}{n}\|f\|^2_{C^1(\Omega)}. \tag{A.7}$$

Lemma 6 can be strengthened in several respects. Under the assumptions of the lemma, better bounds are available than (A.7) which do not depend on factors of $\log n$. Additionally, under weaker assumptions than $f \in C^1(\Omega)$, it is possible to obtain bounds which are looser than (A.7) but which still imply that $\mathrm{Var}\big[U_{n,\mathrm{Vor}}(f)\big] \to 0$ as $n \to \infty$. Neither of these are necessary to prove Theorem 1, and so we do not pursue them further.

Proof of Lemma 6. We will repeatedly use the following fact, which is a consequence of Lemma 18: there exists a constant $C > 0$ not depending on n such that for any $x, y \in \Omega$,

$$\begin{aligned} H_{\mathrm{Vor}}(x,y) &\le \int_{L\cap\Omega} \exp\big(-(p_{\min}/2d)\|x-z\|^d\big)\,dz \\ &\le C\Big(\frac{1}{n^{(d-1)/d}}1\{\|x-y\| \le C\varepsilon_0\} + \frac{1}{n^2}\Big). \end{aligned} \tag{A.8}$$

Now, we recall from Hoeffding's decomposition of U-statistics [Hoeffding, 1948] that the variance of $U_{n,\mathrm{Vor}}(f)$ can be written as

$$\mathrm{Var}[U_{n,\mathrm{Vor}}(f)] = \frac{1}{4}\Big(n(n-1)\mathrm{Var}[h(x_1,x_2)] + n(n-1)(n-2)\mathrm{Var}[h_1(x_1)]\Big) \tag{A.9}$$

where $h(x,y) = |f(x)-f(y)|H_{\mathrm{Vor}}(x,y)$ and $h_1(x) = \mathbb{E}[h(x_1,x_2)|x_1]$.

We now use (A.8) to upper bound the variance of h and h_1. For h, we have that

$$\begin{aligned} \mathrm{Var}[h(x_1,x_2)] &\le \mathbb{E}[h^2(x_1,x_2)] \\ &\le p^2_{\max}\|f\|^2_{C^1(\Omega)}\int_\Omega\int_\Omega \|y-x\|^2\big(H_{\mathrm{Vor}}(x,y)\big)^2\,dy\,dx \\ &\le C\|f\|^2_{C^1(\Omega)}\bigg(\frac{1}{n^{2(d-1)/d}}\int_\Omega\int_\Omega \|y-x\|^2 1\{\|x-y\| \le C\varepsilon_0\}\,dy\,dx + \frac{1}{n^4}\bigg) \\ &\le C\bigg(\varepsilon_0^{3d}\|f\|^2_{C^1(\Omega)} + \frac{\|f\|^2_{C^1(\Omega)}}{n^4}\bigg). \end{aligned}$$

For h_1, we have that for every $x \in \Omega$,

$$\begin{aligned}|h_1(x)| &\le \|f\|_{C^1(\Omega)} p_{\max} \int_\Omega \|y-x\| H_{\mathrm{Vor}}(y,x)\, dy \\ &\le C\|f\|_{C^1(\Omega)} \left(\frac{1}{n^{(d-1)/d}} \int_\Omega \|y-x\| 1\{\|y-x\| \le C\varepsilon_0\}\, dy + \frac{1}{n^2} \right) \\ &\le C\|f\|_{C^1(\Omega)} \left(\varepsilon_0^{2d} + \frac{1}{n^2} \right).\end{aligned}$$

Integrating over $x \in \Omega$, we conclude that

$$\mathrm{Var}[h_1(x_1)] \le \mathbb{E}[(h_1(x_1))^2] \le C\|f\|^2_{C^1(\Omega)} \left(\varepsilon_0^{4d} + \frac{1}{n^4} \right).$$

Plugging these estimates back into (A.9) gives the upper bound in (A.7). □

A.2.4 Step 3: Bias of Voronoi U-statistic

Under appropriate conditions, the expectation of $U_{n,\mathrm{Vor}}(f)$ is approximately equal to (an appropriately rescaled version of) the nonlocal functional (A.4) for bandwidth $\varepsilon_{(1)}(x) = (np(x))^{-1/d}$, weight $(p(x))^{(d+1)/d}$, and kernel

$$K_{\mathrm{Vor}}(t) = \int_0^\infty \exp\left(-\mu_d \left\{ \frac{t^2}{4} + s^2 \right\}^{d/2} \right) s^{d-2}\, ds. \tag{A.10}$$

Lemma 7. *Suppose $x_{1:n}$ are sampled independently from a distribution P satisfying A1. For any $f \in C^1(\Omega)$,*

$$\mathbb{E}\big[U_{n,\mathrm{Vor}}(f)\big] = n^{(d+1)/d} \frac{\eta_{d-2}}{2} \cdot \mathrm{TV}_{\varepsilon_{(1)}, K_{\mathrm{Vor}}}\big(f; \Omega, p^{(d+1)/d}\big) + O\left(\frac{(\log n)^{3+1/d}}{n^{1/d}} \|f\|_{C^1(\Omega)} \right).$$

Proof. We will use Lemma 17, which shows that at points $x, y \in \Omega$ sufficiently far from the boundary of Ω, the kernel $H_{\mathrm{Vor}}(x,y)$ is approximately equal to a spherical kernel. To invoke this lemma, we need to restrict our attention to points sufficiently far from the boundary. In particular, letting $h = h_n$ be defined as in Lemma 17, we conclude from (A.82) that

$$\begin{aligned}&\int_\Omega \int_\Omega |f(y) - f(x)| H_{\mathrm{Vor}}(x,y) p(y) p(x)\, dy\, dx = \\ &\qquad \int_{\Omega_h} \int_\Omega |f(y) - f(x)| H_{\mathrm{Vor}}(x,y) p(y) p(x)\, dy\, dx + O\left(\frac{h}{n^2} \|f\|_{C^1(\Omega)} \right),\end{aligned} \tag{A.11}$$

where we have used the assumption $f \in C^1(\Omega)$ and (A.82) to control the boundary term, since

$$\begin{aligned}
&\int_{\Omega\setminus\Omega_h}\int_{\Omega} |f(y)-f(x)| H_{\mathrm{Vor}}(x,y)p(y)p(x)\,dy\,dx \\
&\quad\le \frac{C_3 p_{\max}^2 \eta_{d-2}\|f\|_{C^1(\Omega)}}{n^{(d-1)/d}} \int_{\Omega\setminus\Omega_h}\int_{\Omega} \|y-x\| K_{\mathrm{Vor}}\biggl(\frac{\|y-x\|}{C_4 n^{1/d}}\biggr)\,dy\,dx \\
&\quad\overset{(i)}{\le} \frac{C_3 C_4^{(d+1)/d} p_{\max}^2 \eta_{d-2}\|f\|_{C^1(\Omega)}}{n^2} \int_{\Omega\setminus\Omega_h}\int_{\mathbb{R}^d} \|h\| K_{\mathrm{Vor}}(\|h\|)\,dh\,dx \\
&\quad\overset{(ii)}{\le} \frac{C_3 C_4^{(d+1)/d} p_{\max}^2 \eta_{d-2}\eta_{d-1}\|f\|_{C^1(\Omega)}}{n^2} \int_{\Omega\setminus\Omega_h}\int_0^{\infty} t^d K_{\mathrm{Vor}}(t)\,dt\,dx \\
&\quad\overset{(iii)}{\le} \frac{C\|f\|_{C^1(\Omega)}}{n^2}\mu(\Omega\setminus\Omega_h) \\
&\quad\le \frac{Ch\|f\|_{C^1(\Omega)}}{n^2},
\end{aligned} \tag{A.12}$$

where (i) follows by changing variables $h = (y-x)/C_3 n^{1/d}$, (ii) by converting to polar coordinates, and (iii) upon noticing that $\int_0^\infty t^d K_{\mathrm{Vor}}(t) < \infty$.

Returning to the first-order term in (A.11), we can use (A.81) to replace the integral with H_{Vor} by an integral with the Voronoi kernel K_{Vor}. Precisely,

$$\begin{aligned}
&\int_{\Omega_h}\int_{\Omega} |f(y)-f(x)| H_{\mathrm{Vor}}(x,y)p(y)p(x)\,dy\,dx \\
&\quad= \frac{\eta_{d-2}}{n^{(d-1)/d}} \int_{\Omega_h}\int_{\Omega} |f(y)-f(x)| K_{\mathrm{Vor}}\biggl(\frac{\|x-y\|}{\varepsilon_{(1)}}\biggr) p(y)\bigl(p(x)\bigr)^{1/d}\,dy\,dx \\
&\qquad + O\biggl(\frac{1}{n^3}\int_{\Omega}\int_{\Omega} |f(y)-f(x)|\,dy\,dx\biggr) \\
&\qquad + O\biggl(\frac{(\log n)^2}{n}\int_{\Omega}\int_{\Omega} |f(y)-f(x)| 1\Bigl\{\|x-y\| \le C(\log n/n)^{1/d}\Bigr\}\,dy\,dx\biggr) \\
&\quad= \frac{\eta_{d-2}}{n^{(d-1)/d}} \int_{\Omega_h}\int_{\Omega} |f(y)-f(x)| K_{\mathrm{Vor}}\biggl(\frac{\|x-y\|}{\varepsilon_{(1)}}\biggr) p(y)\bigl(p(x)\bigr)^{1/d}\,dy\,dx \\
&\qquad + O\biggl(\frac{\|f\|_{C^1(\Omega)}}{n^3} + \frac{(\log n)^{3+1/d}}{n^{2+1/d}}\|f\|_{C^1(\Omega)}\biggr) \\
&\quad= \frac{\eta_{d-2}}{n^{(d-1)/d}} \int_{\Omega}\int_{\Omega} |f(y)-f(x)| K_{\mathrm{Vor}}\biggl(\frac{\|x-y\|}{\varepsilon_{(1)}}\biggr) p(y)\bigl(p(x)\bigr)^{1/d}\,dy\,dx \\
&\qquad + O\biggl(\frac{\|f\|_{C^1(\Omega)}}{n^3} + \frac{(\log n)^{3+1/d}}{n^{2+1/d}}\|f\|_{C^1(\Omega)} + \frac{h\|f\|_{C^1(\Omega)}}{n^2}\biggr),
\end{aligned} \tag{A.13}$$

with the second equality following from the upper bound (A.39), and the third equality from exactly the same argument as in (A.12). Finally, we use the Lipschitz property of p to conclude that

$$\begin{aligned}&\int_\Omega \int_\Omega |f(y)-f(x)| K_{\mathrm{Vor}}\left(\frac{\|x-y\|}{\varepsilon_{(1)}}\right) p(y)\big(p(x)\big)^{1/d}\,dy\,dx \\ &\quad = \int_\Omega \int_\Omega |f(y)-f(x)| K_{\mathrm{Vor}}\left(\frac{\|x-y\|}{\varepsilon_{(1)}}\right) \big(p(x)\big)^{(d+1)/d}\,dy\,dx + O\left(\frac{\|f\|_{C^1(\Omega)}}{n^{(d+2)/2}}\right),\end{aligned} \tag{A.14}$$

since

$$\begin{aligned}&\int_\Omega \int_\Omega |f(y)-f(x)| K_{\mathrm{Vor}}\left(\frac{\|x-y\|}{\varepsilon_{(1)}}\right) |p(y)-p(x)| \big(p(x)\big)^{1/d}\,dy\,dx \\ &\quad \le C\|f\|_{C^1(\Omega)} p_{\max}^{1/d} \int_\Omega \int_\Omega \|y-x\|^2 K_{\mathrm{Vor}}\left(\frac{\|x-y\|}{\varepsilon_{(1)}}\right) dy\,dx \\ &\quad \le C\frac{\|f\|_{C^1(\Omega)} p_{\max}^{1/d}}{p_{\min}^{1/d} n^{(2+d)/d}} \int_\Omega \int_{\mathbb{R}^d} \|h\|^2 K_{\mathrm{Vor}}(\|h\|)\,dh\,dx \\ &\quad = C\frac{\|f\|_{C^1(\Omega)} p_{\max}^{1/d} \eta_{d-1}}{p_{\min}^{1/d} n^{(2+d)/d}} \int_\Omega \int_0^\infty t^{d+1} K_{\mathrm{Vor}}(t)\,dt\,dx \\ &\quad \le C\frac{\|f\|_{C^1(\Omega)}}{n^{(2+d)/d}},\end{aligned}$$

with the last inequality following since $\int_0^\infty t^{d+1} K_{\mathrm{Vor}}(t)\,dt = C < \infty$. Combining (A.11), (A.13) and (A.14) yields the final claim. □

Finally, Lemma 8 shows that the kernelized TV $\mathrm{TV}_{\varepsilon,K}(f;\Omega,h)$ converges to a continuum TV under appropriate conditions.

Assumption A2. The bandwidth $\varepsilon(x) = \bar{\varepsilon}_n g(x)$ for a sequence $\bar{\varepsilon}_n \to 0$ and a bounded function $g \in L^\infty(\Omega)$. The kernel function K satisfies $\int_0^\infty K(t) t^{d+1}\,dt < \infty$. The weight function $h \in L^\infty(\Omega)$.

Note that Assumption A1 implies that Assumption A2 is satisfied by bandwidth $\varepsilon_{(1)}$, kernel K_{Vor} and weight function $h = p^{(d+1)/d}$.

Lemma 8. *Assuming A2, for any $f \in C^2(\Omega)$,*

$$\lim_{n\to\infty} (\bar{\varepsilon}_n)^{-(d+1)} \mathrm{TV}_{\varepsilon,K}(f;\Omega,h) = \sigma_K \int_\Omega \|\nabla f(x)\| h(x) (g(x))^{d+1}\,dx \tag{A.15}$$

where

$$\sigma_K := \frac{2\eta_{d-2}}{(d-1)} \int_0^\infty K(t) t^d\,dt. \tag{A.16}$$

Proof. The proof of Lemma 8 follows closely the proof of some related results, e.g., Lemma 4.2 of García Trillos and Slepčev [2016]. We begin by summarizing the major steps.

1. We use a 2nd-order Taylor expansion to replace differencing by derivative inside the nonlocal TV.
2. Naturally, the nonlocal TV behaves rather differently than a local functional near the boundary of Ω. We show that the contribution of the integral near the boundary is negligible.
3. Finally, we reduce from a double integral to a single integral involving the norm $\|\nabla f\|$.

Step 1: Taylor expansion. Since $f \in C^2(\Omega)$ we have that

$$f(y) - f(x) = \nabla f(x)^\top (y - x) + O(\|f\|_{C^2(\Omega)} \|y - x\|^2).$$

Consequently,

$$\mathrm{TV}_{\varepsilon,K}(f; \Omega, h) = \int_\Omega \int_\Omega \Big(|\nabla f(x)^\top (y - x)| + O(\|f\|_{C^2(\Omega)}) \Big) K\left(\frac{\|y - x\|}{\varepsilon(x)} \right) h(x)\, dy\, dx.$$

We now upper bound the contribution of the $O(\|y - x\|^2)$-term. For each $x \in \Omega$,

$$\begin{aligned} &\int_\Omega \|y - x\| K\left(\frac{\|y - x\|^2}{\varepsilon(x)} \right) dy \\ &\qquad \le C|\varepsilon_n(x)|^{d+2} \int_{\mathbb{R}^d} \|z\|^2 K(\|z\|)\, dz \le C|\varepsilon_n(x)|^{d+2} \le C|\varepsilon_n(x)|^{d+2}, \end{aligned}$$

with the final inequality following from the assumption $\int_0^\infty t^{d+1} K(t)\, dt < \infty$. Integrating over Ω gives the upper bound

$$\int_\Omega \int_\Omega O(\|f\|_{C^2(\Omega)} \|y - x\|^2) K\left(\frac{\|y - x\|}{\varepsilon(x)} \right) h(x)\, dy\, dx = O(\|f\|_{C^2(\Omega)} \bar{\varepsilon}_n^{d+2}),$$

recalling that $h(x), g(x) \in L^\infty(\Omega)$.

Step 2: Contribution of boundary to nonlocal TV. Take $r = r_n$ to be any sequence such that $r_n/\bar{\varepsilon}_n \to \infty, r_n \to 0$. Breaking up the integrals in the definition of

nonlocal TV gives

$$\begin{aligned}
&\int_\Omega \int_\Omega |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\quad = \int_{\Omega_r} \int_{\mathbb{R}^d} |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\qquad - \int_{\Omega_r} \int_{\mathbb{R}^d \setminus \Omega} |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\qquad + \int_{\Omega \setminus \Omega_r} \int_{\Omega} |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\quad =: I_1 + I_2 + I_3.
\end{aligned}$$

Now we are going to show that I_2 and I_3 are negligible. For I_2, noting that $r/\varepsilon(x) \to \infty$ for all x, we have that for any $x \in \Omega_r$,

$$\begin{aligned}
&\int_{\mathbb{R}^d \setminus \Omega} |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy \\
&\qquad\qquad \le \|f\|_{C^1(\Omega)} \int_{\mathbb{R}^d \setminus \Omega} K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) \|y-x\|\, dy \\
&\qquad\qquad \le \|f\|_{C^1(\Omega)} (\varepsilon(x))^1 \int_{\mathbb{R}^d \setminus B(0, r/\varepsilon(x))} \|z\| K(\|z\|)\, dz \\
&\qquad\qquad \overset{(i)}{\le} C \|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1} \int_{r/\varepsilon(x))}^{\infty} t^{d+1} K(t)\, dt \\
&\qquad\qquad \overset{(ii)}{=} o(\|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1}),
\end{aligned}$$

where (i) follows from converting to polar coordinates and (ii) follows by the assumption $\int_0^\infty t^{d+1} K(t)\, dt < \infty$. Integrating over x yields $I_2 = o(\|f\|_{C^1(\Omega)} \bar{\varepsilon}_n^{d+1})$, since $h, g \in L^\infty(\Omega)$.

On the other hand for I_3, similar manipulations show that for every $x \in \Omega$,

$$\int_\Omega |\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) dy \le C \|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1}.$$

Noting that the tube $\Omega \setminus \Omega_r$ has volume at most Cr, we conclude that

$$I_3 \le C \|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1} \mu(\Omega \setminus \Omega_r) \le Cr \|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1} = o(\|f\|_{C^1(\Omega)} (\varepsilon(x))^{d+1}),$$

with the last inequality following since $r = o(1)$.

Step 3: Double integral to single integral. Now we proceed to reduce the double integral in I_1 to a single integral. Changing variables to $z = (y - x)/\varepsilon(x)$, converting to polar coordinates, and letting $w(x) = \nabla f(x)/\|\nabla f(x)\|$, we have that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|\nabla f(x)^\top (y-x)| K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) dy \\
&\qquad = (\varepsilon(x))^{d+1} \int_{\mathbb{R}^d} |\nabla f(x)^\top z| K(\|z\|)\, dz \\
&\qquad = (\varepsilon(x))^{d+1} \left(\int_{\mathbb{S}^{d-1}} |\nabla f(x)^\top \phi|\, d\mathcal{H}^{d-1}\right)\left(\int_0^\infty t^d K(t)\, dt\right) \\
&\qquad = (\varepsilon(x))^{d+1} \|\nabla f(x)\| \left(\int_{\mathbb{S}^{d-1}} |w(x)^\top \phi|\, d\mathcal{H}^{d-1}\right)\left(\int_0^\infty t^d K(t)\, dt\right) \\
&\qquad = (\varepsilon(x))^{d+1} \|\nabla f(x)\| \left(\int_{\mathbb{S}^{d-1}} |\phi_1|\, d\mathcal{H}^{d-1}\right)\left(\int_0^\infty t^d K(t)\, dt\right) \\
&\qquad = \sigma_K (\varepsilon(x))^{d+1} \|\nabla f(x)\|,
\end{aligned}$$

with the second to last equality following from the spherical symmetry of the integral, and the last equality by definition of σ_K. Integrating over $x \in \Omega_r$ gives

$$\begin{aligned}
I_1 &= \sigma_K \bar{\varepsilon}_n^{d+1} \int_{\Omega_r} \|\nabla f(x)\| h(x) (g(x))^{d+1}\, dx \\
&= \sigma_K \bar{\varepsilon}_n^{d+1} \int_{\Omega} \|\nabla f(x)\| h(x) (g(x))^{d+1}\, dx + o(\bar{\varepsilon}_n^{d+1} \|f\|_{C^1(\Omega)}),
\end{aligned}$$

with the second equality following from the same reasoning as was used in analyzing the integral I_3.

Putting the pieces together. We conclude that

$$\begin{aligned}
&(\bar{\varepsilon}_n)^{-(d+1)} \mathrm{TV}_{\varepsilon,K}(f;\Omega,h) \\
&\qquad = (\bar{\varepsilon}_n)^{-(d+1)} \int_\Omega \int_\Omega \Big(|\nabla f(x)^\top (y-x)|\Big) K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\qquad\qquad + O(\bar{\varepsilon}_n \|f\|_{C^2(\Omega)}) \\
&\qquad = (\bar{\varepsilon}_n)^{-(d+1)} \int_{\Omega_r} \int_{\mathbb{R}^d} \Big(|\nabla f(x)^\top (y-x)|\Big) K\left(\frac{\|y-x\|}{\varepsilon(x)}\right) h(x)\, dy\, dx \\
&\qquad\qquad + O(\bar{\varepsilon}_n \|f\|_{C^2(\Omega)}) + o(\|f\|_{C^1(\Omega)}) \\
&\qquad = \sigma_K \int_\Omega \int_\Omega \|\nabla f(x)\| h(x) (g(x))^{d+1}\, dx + O(\bar{\varepsilon}_n \|f\|_{C^2(\Omega)}) + o(\|f\|_{C^1(\Omega)}),
\end{aligned}$$

completing the proof of Lemma 8. □

A.3 Sensitivity analysis for Section 2.4

In Section 2.4, we chose the scale k, ε in the k-nearest neighbor and ε-neighborhood graphs to be such that their average degree would roughly match that of the Voronoi adjacency graph, and we remarked that mildly better results are attainable if one increases the connectivity of the graphs. Here, we present an analogous set of results to those found in Section 2.4, where the average degree of the k-nearest neighbor and ε-neighborhood graphs are roughly twice that of the graphs in Section 2.4. All other details of the experimental setup remain the same.

- In Figure A.1, the estimates of TV by the k-nearest neighbor and ε-neighborhood graphs approach their density-weighted limits more quickly than in Section 2.4, with slightly narrower variability bands.
- In Figure A.2, we see that ε-neighborhood TV denoising is now competitive with k-nearest neighbor TV denoising and the unweighted Voronoigram for the "low inside tube" setting. In the "high inside tube" and uniform sampling settings, the performance of k-nearest neighbor TV denoising improves slightly.

As previously remarked, the Voronoigram has no such auxiliary tuning parameter, so the weighted and unweighted Voronoigram results here are the same as in Section 2.4. We also note that with greater connectivity in the k-nearest neighbor and ε-neighborhood graphs comes greater computational burden in storing the graphs, as well as performing calculations with them. Therefore, it is advantageous to the practitioner to use the sparsest graph capable of achieving favorable performance.

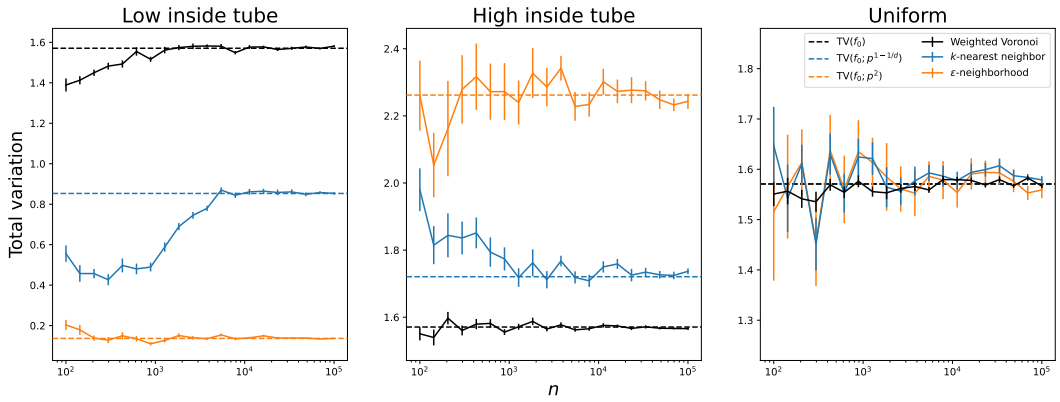

Figure A.1: *Results from the TV estimation experiment, with greater connectivity in the kNN and ε-neighborhood graphs. Compare these results to those in Figure 2.3.*

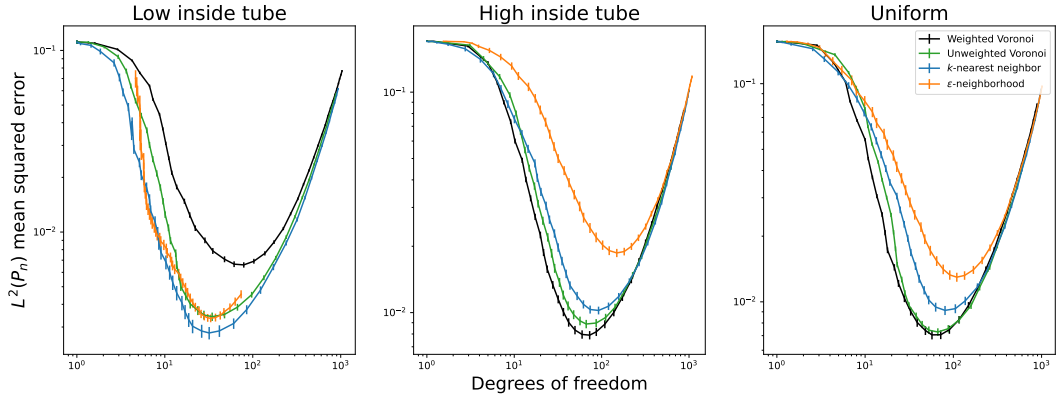

Figure A.2: *Results from the function estimation experiment, with greater connectivity in the kNN and ε-neighborhood graphs. Compare these results to those in Figure 2.5.*

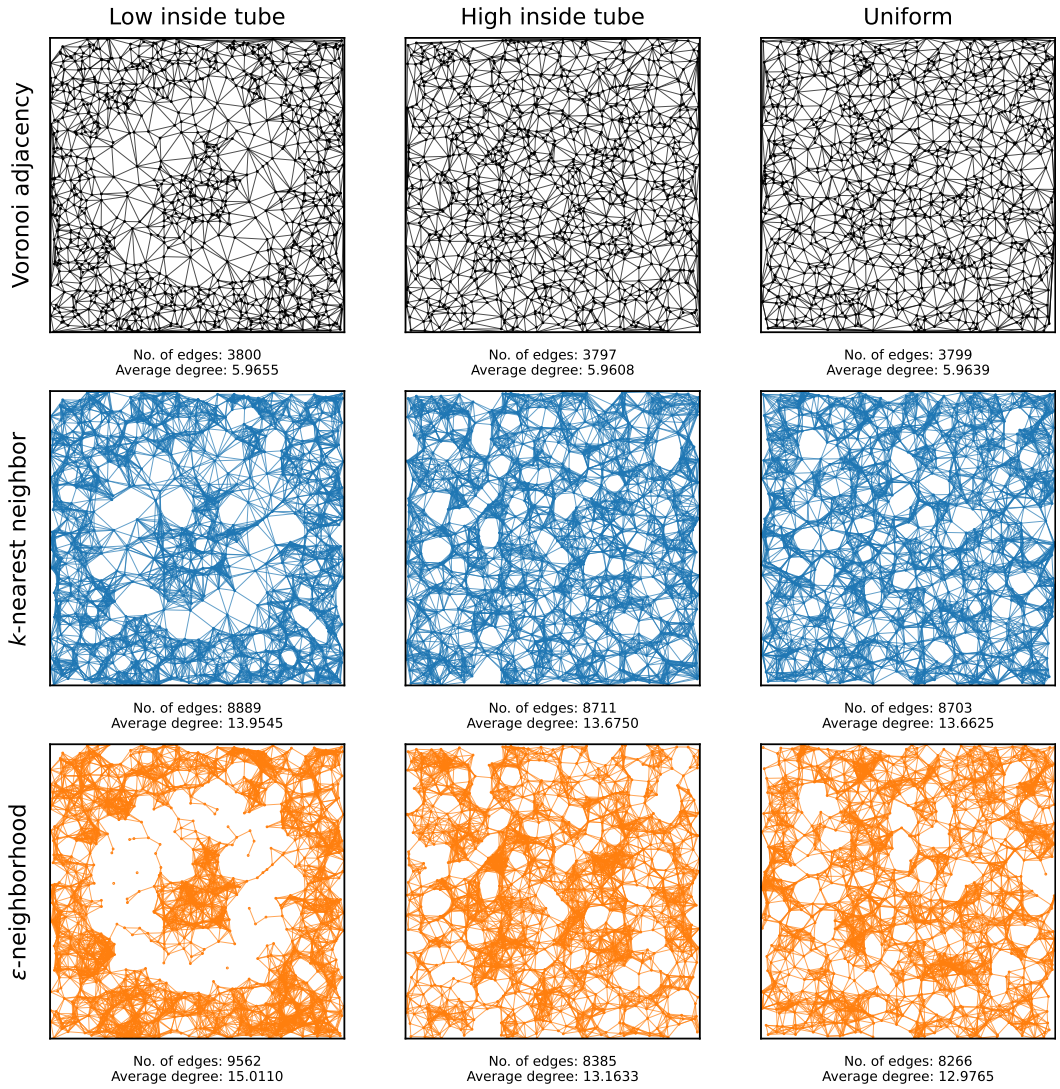

Figure A.3: *Visualization of the Voronoi, kNN, and ε-neighborhood graphs, with greater connectivity in the latter two graphs. (The Voronoi graph does not have such an auxiliary tuning parameter.) Compare these graphs to those in Figure 2.4.*

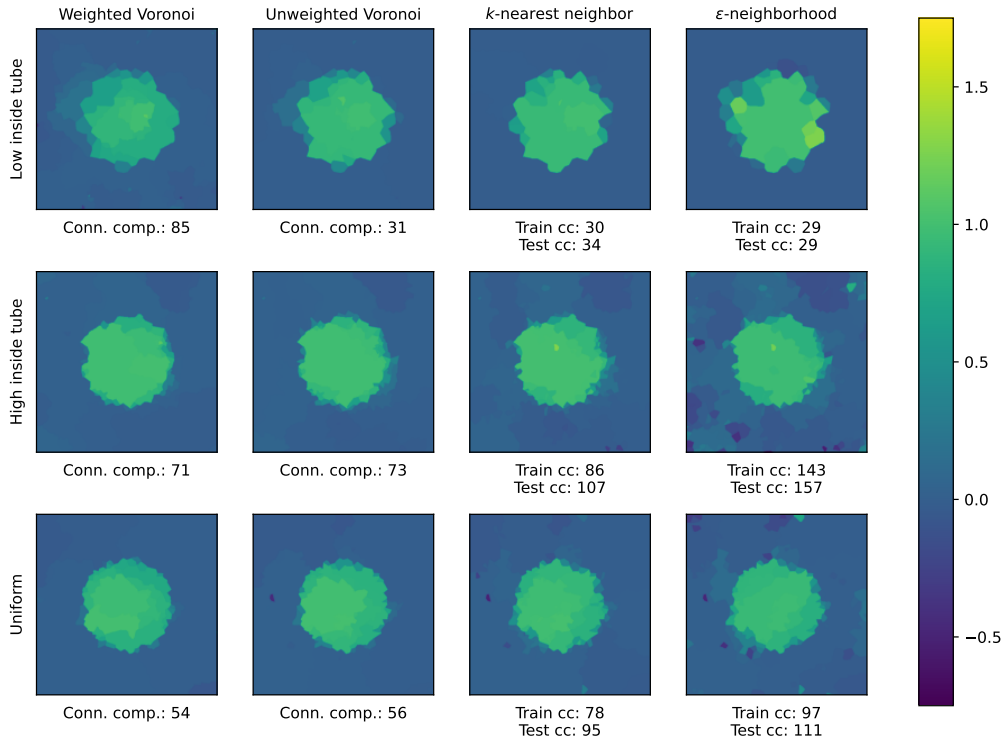

Figure A.4: *Extrapolants from graph TV denoising, with greater connectivity in the kNN and ε-neighborhood graphs. Compare these results to those in Figure 2.6.*

A.4 Proofs for Section 2.5

A.4.1 Proof of Theorem 2

From (2.29), in the discussion preceding Lemma 3, we have

$$\mathbb{E}\|\hat{f} - f_0\|_{L^2(P)}^2 \leq \mathbb{E}\Big[K_n \|\hat{f} - f_0\|_{L^2(P_n)}^2\Big] + 2\mathbb{E}\|\bar{f}_0 - f_0\|_{L^2(P)}^2, \tag{A.17}$$

where

$$K_n = 2p_{\max} n \cdot \left(\max_{i=1,\ldots,n} \mu(V_i)\right).$$

The second term is bounded by Lemma 3. We now outline the analysis of the first term. As in the $L^2(P_n)$ case we will decompose the error into the case where the design points are well-spaced and the case where they are not. This is formalized by the set $\mathscr{X} = \mathscr{X}_1 \cap \mathscr{X}_2$, where $\mathscr{X}_1, \mathscr{X}_2$ are defined in Appendix A.6. $x_{1:n}$ falls within this set with probability at least $1 - 3/n^4$, and notably on this set,

$$\max_i \mu(V_i) \leq C_1 \log n / n$$

for some $C_1 > 0$, since $\mathscr{X}_2$ is the set upon which the conclusion of Lemma 15 holds. We proceed by conditioning,

$$\begin{aligned}
&\mathbb{E}\Big[K_n\|\hat{f} - f_0\|^2_{L^2(P_n)}\Big] \\
&\quad = 2p_{\max}\Bigg(\mathbb{E}_x\left[\mathbb{E}_{z|x}\left[\max_i \mu(V_i)\|\hat{\theta} - \theta_0\|_2^2\right] 1\{x_{1:n} \in \mathscr{X}\}\right] \\
&\qquad\qquad + \mathbb{E}_x\left[\mathbb{E}_{z|x}\left[\max_i \mu(V_i)\|\hat{\theta} - \theta_0\|_2^2\right] 1\{x_{1:n} \notin \mathscr{X}\}\right]\Bigg).
\end{aligned} \tag{A.18}$$

Using the fact that $x_{1:n} \in \mathscr{X}$, the first term on the RHS of (A.18) may be bound,

$$\begin{aligned}
&\mathbb{E}_{z|x}\left[\max_i \mu(V_i)\|\hat{\theta} - \theta_0\|_2^2\right] 1\{x_{1:n} \in \mathscr{X}\} \\
&\qquad\qquad \leq C_1(\log n)\, \mathbb{E}_{z|x}\left[\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2\right] \cdot 1\{x_{1:n} \in \mathscr{X}\} \qquad \text{(A.19)} \\
&\qquad\qquad \leq C_2(\log n)\left(\frac{\lambda\|D\theta_0\|}{n} + \frac{\log^\alpha n}{n}\right), \qquad \text{(A.20)}
\end{aligned}$$

where the latter inequality is obtained by following the analysis of Lemma 1. For the second term on the RHS of (A.18), we apply the crude upper bound that $\mu(V_i) \leq \mu(\Omega) = 1$ for all $i = 1, \ldots, n$. Then apply (A.59) to obtain,

$$\begin{aligned}
&\mathbb{E}_{z|x}\left[\max_i \mu(V_i)\|\hat{\theta} - \theta_0\|_2^2\right] 1\{x_{1:n} \notin \mathscr{X}\} \\
&\qquad\qquad \leq \mathbb{E}_{z|x}\left[16\|z_{1:n}\|_2^2 + 2\lambda\|D\theta_0\|_1\right] 1\{x_{1:n} \notin \mathscr{X}\} \qquad \text{(A.21)} \\
&\qquad\qquad = (16n + 2\lambda\|D\theta_0\|_1)\, 1\{x_{1:n} \notin \mathscr{X}\}. \\
&\qquad\qquad \leq \left(16n + 4n^2\lambda\|\theta_0\|_\infty\|w\|_\infty\right) 1\{x_{1:n} \notin \mathscr{X}\}. \\
&\qquad\qquad \leq \left(16n + 4n^2\lambda\|\theta_0\|_\infty\right) 1\{x_{1:n} \notin \mathscr{X}\}, \qquad \text{(A.22)}
\end{aligned}$$

where we also use crude upper bounds on the discrete TV. Substitute (A.20) and (A.22) into (A.18) to obtain,

$$\begin{aligned}
\mathbb{E}\left[K_n\|\hat{f} - f_0\|^2_{L^2(P_n)}\right] &\leq C_3\left(\frac{(\log n)\lambda\mathbb{E}\|D\theta_0\|}{n} + \frac{(\log n)^{1+\alpha}}{n} + \lambda n^2\mathbb{P}\{x_{1:n} \notin \mathscr{X}\}\right) \\
&\leq C_4\left(\frac{(\log n)\lambda\mathbb{E}\|D\theta_0\|}{n} + \frac{(\log n)^{1+\alpha}}{n} + \frac{\lambda}{n^2}\right) \\
&\leq C_5\left(\frac{\sigma\tau_n(\log n)^{3/2+\alpha}\mathbb{E}\|D\theta_0\|}{n} + \frac{(\log n)^{1+\alpha}}{n}\right), \qquad \text{(A.23)}
\end{aligned}$$

where in the final line we have substituted in the value of $\lambda = c\sigma\tau_n(\log n)^{1/2+\alpha}$. Apply Lemma 2 to (A.23) and substitute back into (A.17) to obtain the claim. □

A.4.2 Proof of Theorem 3

To establish the lower bound in (2.24), we follow a classical approach, similar to that outlined in [del Álamo et al., 2021]: first we reduce the problem to estimating binary sequences, then we apply Assouad's lemma (Lemma 9). This results in a constrained maximization problem, which we analyze to establish the ultimate lower bound.

Step 1: Reduction to estimating binary sequences. We begin by associating functions f_θ with vertices of the hypercube $\Theta_S = \{0,1\}^S$, where $S \subseteq [m]^d$ for some $m \in \mathbb{N}$. To construct these functions f_θ, we partition Ω into cubes,

$$Q_i = \frac{1}{m}(i_1 - 1, i_1) \times \cdots \times \frac{1}{m}(i_d - 1, i_d), \quad \text{for } i \in [m]^d,$$

and for each $\theta \in \Theta_S$ take f_θ to be the piecewise constant function

$$f_\theta(x) := a \cdot \sum_{i \in S} \theta_i 1_{Q_i}(x), \tag{A.24}$$

where $1_{Q_i}(x) = 1(x \in Q_i)$ is the characteristic function of Q_i. Observe that for all $\theta \in \Theta_S$, letting $\epsilon := 1/m$,

$$\mathrm{TV}(f_\theta) \le 2da|S|\epsilon^{d-1}, \quad \text{and} \quad \|f_\theta\|_{L^\infty(\Omega)} \le a. \tag{A.25}$$

So long as the constraints in (A.25) are satisfied $\{f_\theta : \theta \in \Theta_S\} \subseteq \mathrm{BV}_\infty(L, M)$, and consequently

$$\inf_{\widehat{f}} \sup_{f_0 \in \mathrm{BV}_\infty(L,M)} \mathbb{E}_{f_0} \|\widehat{f} - f_0\|^2_{L^2(\Omega)} \ge \inf_{\widehat{f}} \max_{\theta \in \Theta_S} \mathbb{E}_\theta \|\widehat{f} - f_\theta\|^2_{L^2(\Omega)} \ge \frac{a^2\epsilon^d}{4} \inf_{\widehat{\theta}} \max_{\theta \in \Theta_S} \mathbb{E}_\theta \rho(\widehat{\theta}, \theta), \tag{A.26}$$

where $\rho(\theta, \theta') = \sum_{i \in S} |\theta_i - \theta'_i|$ is the Hamming distance between vertices $\theta, \theta' \in \Theta_S$. The second inequality in (A.26) is verified as follows: for a given $\widehat{f}$, letting

$$\widehat{\theta}_i = \begin{cases} 1, & \text{if } \fint_{Q_i} \widehat{f}(x)\, dx \ge a/2, \\ 0, & \text{otherwise,} \end{cases}$$

it follows that

$$\begin{aligned} \|\widehat{f} - f_\theta\|^2_{L^2(P)} &= \sum_{i \in [m]^d} \|\widehat{f} - f_\theta\|^2_{L^2(Q_i)} \\ &\ge \sum_{i \in S} \|\widehat{f} - f_\theta\|^2_{L^2(Q_i)} \\ &\ge \frac{a^2\epsilon^d}{4} \sum_{i \in S} 1\{\widehat{\theta}_i \ne \theta_i\}. \end{aligned}$$

Step 2: Application of Assouad's lemma. Given a measurable space $(\mathcal{Z}, \mathcal{A})$, and a set of probability measures $\mathcal{M} = \{\mu_\theta : \theta \in \Theta_S\}$ on $(\mathcal{Z}, \mathcal{A})$, Assouad's lemma lower bounds the minimax risk over Θ_S, when loss is measured by the Hamming distance $\rho(\widehat{\theta}, \theta) := \sum_{i \in S} |\widehat{\theta}_i - \theta_i|$. We use a form of Assouad's lemma given in Tsybakov [2009].

Lemma 9 (Lemma 2.12 of Tsybakov [2009])**.** *Suppose that for each $\theta, \theta' \in \Theta_S$: $\rho(\theta, \theta') = 1$, we have that $\mathrm{KL}(\mu_\theta, \mu_{\theta'}) \leq \alpha < \infty$. It follows that*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_S} \mathbb{E}_\theta \rho(\widehat{\theta}, \theta) \geq \frac{|S|}{2} \max\left(\frac{1}{2}\exp(-\alpha), (1 - \sqrt{\alpha/2})\right).$$

To apply Assouad's lemma in our context, we take $\mathcal{Z} = (\Omega \times \mathbb{R})^{\otimes n}$, and associate each $\theta \in \Theta_S$ with the measure $\mu_\theta^{(n)}$, the n-times product of measure $\mu_\theta = \mathrm{Unif}(\Omega) \times N(f_\theta(x), 1)$. We now upper bound the KL divergence $\mathrm{KL}(\mu_\theta^{(n)}, \mu_{\theta'}^{(n)})$ when $\rho(\theta, \theta') = 1$; letting $i \in S$ be the single index at which $\theta_i \neq \theta'_i$,

$$\begin{aligned}
\mathrm{KL}(\mu_\theta, \mu_{\theta'}) &= \int_{\Omega \times \mathbb{R}} \log\left(\frac{\phi(y - f_\theta(x))}{\phi(y - f_{\theta'}(x))}\right) \phi(y - f_\theta(x))\, dy\, dx \\
&= \int_{Q_i \times \mathbb{R}} \log\left(\frac{\phi(y - a\theta_i)}{\phi(y - a\theta'_i)}\right) \phi(y - a\theta_i)\, dy\, dx \\
&= \epsilon^d \int_{\mathbb{R}} \log\left(\frac{\phi(y - a\theta_i)}{\phi(y - a\theta'_i)}\right) \phi(y - a\theta_i)\, dy \\
&= \frac{\epsilon^d a^2}{2},
\end{aligned}$$

and it follows that $\mathrm{KL}(\mu_\theta^{(n)}, \mu_{\theta'}^{(n)}) \leq n\epsilon^d a^2/2$. Consequently, so long as (A.25) is satisfied and

$$\frac{n\epsilon^d a^2}{2} \leq 1,$$

we may apply Lemma 9, and deduce from (A.26) that

$$\inf_{\widehat{f}} \sup_{f_0 \in \mathrm{BV}_\infty(L,M)} \mathbb{E}_{f_0} \|\widehat{f} - f_0\|^2_{L^2(\Omega)} \geq \frac{a^2\epsilon^d}{4} \inf_{\widehat{\theta}} \max_{\theta \in \Theta_S} \mathbb{E}_\theta \rho(\widehat{\theta}, \theta) \geq \frac{a^2\epsilon^d |S|}{16 \exp(1)}. \quad \text{(A.27)}$$

Step 3: Lower bound. The upshot of Steps 1 and 2 is that the solution to the following constrained maximization problem yields a lower bound on the minimax

risk: letting $s = |S|$,

$$\begin{aligned} \text{maximize} \quad & \frac{a^2 \epsilon^d s}{16 \exp(1)}, \\ \text{subject to} \quad & 1 \leq s \leq \epsilon^{-d}, \\ & a s \epsilon^{d-1} \leq \frac{L}{2d}, \\ & a \leq M, \\ & \frac{n a^2 \epsilon^d}{2} \leq 1. \end{aligned}$$

Setting $a = M, \epsilon = (\frac{2}{a^2 n})^{1/d}$, and $s = \frac{L}{2da} \epsilon^{-(d-1)}$ is feasible for this problem if $2dM(\frac{M^2 n}{2})^{-\frac{(d-1)}{d}} \leq L \leq 2dM(\frac{M^2 n}{2})^{1/d}$, and implies that the optimal value is at least $\frac{2^{1/d}}{32 \exp(1) d} LM(M^2 n)^{-1/d}$. This implies the claim (2.24) upon suitable choices of constants. □

A.4.3 Proof of Lemma 1

In this proof, write $\theta_0 := (f_0(x_1), \ldots, f_0(x_n))$ and $\mathbb{E}_{z|x}[\cdot] = \mathbb{E}[\cdot | x_{1:n}]$. We will use D to represent the modified edge incidence operator with either clipped edge weights or unit weights; the following analysis, which uses the scaling factor τ_n, applies to both. Let

$$\mathscr{X} = \mathscr{X}_1 \cap \mathscr{X}_2, \tag{A.28}$$

with $\mathscr{X}_1, \mathscr{X}_2$ as in Section A.6. By the law of iterated expectation,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\right] = \mathbb{E}_x\left[\mathbb{E}_{z|x}\left[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\right] \cdot 1\{x_{1:n} \in \mathscr{X}\}\right] \\ + \mathbb{E}_x\left[\mathbb{E}_{z|x}\left[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\right] \cdot 1\{x_{1:n} \notin \mathscr{X}\}\right]. \end{aligned} \tag{A.29}$$

We now upper bound each term on the right hand side separately.

For the first term, we will proceed by comparing the penalty operator D to the averaging operator (A.63) and surrogate operator T corresponding to the graph (A.64). By construction $x_{1:n} \in \mathscr{X}$ implies, for (ξ_k, u_k) the kth singular value/left singular

vector of T, that

$$\begin{aligned}\lambda &\geq C_1 \sigma \tau_n (\log n)^{1/2+\alpha} \\ &\geq \max\Bigg\{ 8 \max_{\ell} |\mathcal{C}_\ell|^{1/2} \Phi_1(D, T, A) \cdot \sigma \sqrt{\log 2n^4 \cdot \sum_{k=2}^{n} \frac{\|u_k\|_\infty^2}{\xi_k^2}}, \\ &\qquad\qquad \Phi_2(D, T, A) \cdot \sigma \sqrt{2 \log n} \Bigg\},\end{aligned}$$

where the latter inequality follows from combining (A.60), (A.61) with (A.65), (A.66) in the clipped weights case, or (A.67), (A.68) in the unit weights case, for an appropriately chosen C_1. We may therefore apply Theorem 8 with D, T, and A, which gives

$$\mathbb{E}_{z|x}\bigg[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\bigg] \cdot 1\{x_{1:n} \in \mathscr{X}\} \leq C\bigg(\frac{\lambda \|D\theta_0\|_1}{n} + \frac{\log^\alpha n}{n}\bigg), \tag{A.30}$$

On the other hand, to upper bound the second term in (A.29) we use (A.59),

$$\begin{aligned}\mathbb{E}_{z|x}\Big[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\Big] \cdot 1\{x_{z|x} \notin \mathscr{X}\} &\leq \mathbb{E}_{z|x}\Big[\frac{16\|z_{1:n}\|_2^2}{n} + \frac{2\lambda\|D\theta_0\|_1}{n}\Big] 1\{x_{1:n} \notin \mathscr{X}\} \\ &\leq \Big(16 + \frac{2\lambda\|D\theta_0\|_1}{n}\Big) 1\{x_{1:n} \notin \mathscr{X}\}.\end{aligned} \tag{A.31}$$

Substituting (A.30) and (A.31) into (A.29), we conclude that

$$\begin{aligned}\mathbb{E}\bigg[\frac{1}{n}\|\widehat{\theta} - \theta_0\|_2^2\bigg] &\leq C\Big(\frac{\lambda \mathbb{E}\|D\theta_0\|_1}{n} + \frac{\log^\alpha n}{n} + \mathbb{P}(x_{1:n} \notin \mathscr{X})\Big) \\ &\leq C\Big(\frac{\lambda \mathbb{E}\|D\theta_0\|_1}{n} + \frac{\log^\alpha n}{n}\Big) \\ &= C\Big(\frac{\sigma\tau_n (\log n)^{1/2+\alpha} \mathbb{E}\|D\theta_0\|_1}{n} + \frac{\log^\alpha n}{n}\Big),\end{aligned} \tag{A.32}$$

with the second inequality following from Lemma 13, and the equality from the choice of $\lambda = C_1 \sigma \tau_n (\log n)^{1/2+\alpha}$. □

A.4.4 Proof of Lemma 2

We prove the claim (2.27) separately for the unit weights and clipped weights case (recall that they differ by a scaling factor $\bar{\tau}_n$. We will subsequently abbreviate $f := f_0$ and use the notation $\mathrm{DTV}(\,\cdot\,; w^{\varepsilon \leftarrow r})$ to denote the ε-neighborhood graph TV, having set $\varepsilon = r$.

Unit weights

Our goal is to upper bound

$$\mathbb{E}\Big[\,\mathrm{DTV}\left(f(x_{1:n});\,\check{w}^{\mathrm{V}}\right)\Big] = n(n-1)\mathbb{E}\Big[|f(x_1)-f(x_2)|1\{\mathcal{H}^{d-1}(\bar{V}_1\cap\bar{V}_2)>0\}\Big].$$

By conditioning, we can rewrite the expectation above as

$$p_{\max}^2\int_\Omega\int_\Omega |f(y)-f(x)|\mathbb{P}_{x_{3:n}}\{\mathcal{H}^{d-1}(\bar{V}_x\cap\bar{V}_y)>0\}\,dy\,dx, \tag{A.33}$$

where $V_x=\{z:\|z-x\|_2<\|z-x_i\|\ \forall i=2,3,\dots,n\}$, and likewise for V_y. Note that V_x and V_y are random subsets of $\mathbb{R}^d$.

We now give an upper bound on the probability that the random cells $\bar{V}_x$ and $\bar{V}_y$ intersect on a set of positive Hausdorff measure, by relating the problem to uniform concentration of the empirical mass of balls in $\mathbb{R}^d$. The upper bound will be crude, in that it may depend on suboptimal multiplicative constants, but sufficient for our purposes. Define $r(V_x) := \sup\{\|z-x\| : z\in V_x\}$. Observe that if $\|y-x\| > r(V_x)+r(V_y)$, then $\bar{V}_x\cap\bar{V}_y=\emptyset$, since for any $z\in V_x$, by the triangle inequality

$$\{\|z-y\|\geq\|y-x\|-\|z-x\|>r(V_y)\}\Longrightarrow\{z\notin V_y\};$$

therefore

$$\{\mathcal{H}^{d-1}(\bar{V}_x\cap\bar{V}_y)>0\}\Longrightarrow\{\|y-x\|\leq r(V_x)+r(V_y)\}.$$

Now, choose $z\in V_x$ for which $\|z-x\|=r(V_x)$. Observe that the ball $B(z,r(V_x)/2)$ must have empirical mass 0, i.e., $B(z,r(V_x)/2)\cap\{x_3,\dots,x_n\}=\emptyset$ (indeed, this same fact must hold for any $r<r(V_x)$). Therefore,

$$\mathbb{P}_{x_{3:n}}\{r(V_x)\geq t\}\leq\mathbb{P}_{x_{3:n}}\Big\{\exists z: B(z,t/2)\cap\{x_3,\dots,x_n\}=\emptyset\Big\}.$$

It follows from Lemma 14 that if $t_{n,\delta}=c\big(\frac{1}{n}(d\log n+\log(1/\delta)\big)^{1/d}<t_0$, where t_0 is a constant not depending on n,δ, then

$$\mathbb{P}_{x_{3:n}}\{\exists z: B(z,t_{n,\delta}/2)\cap\{x_3,\dots,x_n\}=\emptyset\}\leq\delta.$$

Summarizing this reasoning, we have

$$\begin{aligned}
\mathbb{P}_{x_{3:n}}\{\mathcal{H}^{d-1}(\bar{V}_x\cap\bar{V}_y)>0\} &\leq \mathbb{P}_{x_{3:n}}\Big\{\|y-x\|\leq r(V_x)+r(V_y)\Big\}\\
&\leq \mathbb{P}_{x_{3:n}}\Big\{\|y-x\|\leq 2r(V_x)\Big\}+\mathbb{P}_{x_{3:n}}\Big\{\|y-x\|\leq 2r(V_y)\Big\}\\
&\leq \mathbb{P}_{x_{3:n}}\Big\{\exists z:|B(z,\|x-y\|/4)\cap\{x_3,\dots,x_n\}|=\emptyset\Big\}\\
&\qquad+\mathbb{P}_{x_{3:n}}\Big\{\exists z:|B(z,\|x-y\|/4)\cap\{x_3,\dots,x_n\}|=\emptyset\Big\}\\
&\leq\begin{cases}2, & \text{if } \|x-y\|_2\leq 2t_{n,\delta},\\ 2\delta, & \text{otherwise.}\end{cases}
\end{aligned}$$

Setting $\delta_n = n^{-(d+1)/d}$ and plugging this back into (A.33), we conclude that if $t_{n,\delta_n} < t_0$, then

$$\mathbb{E}\Big[\operatorname{DTV}\Big(f(x_{1:n});\, \breve{w}^{\mathrm{V}}\Big)\Big] \leq 2n(n-1)\int_\Omega\int_\Omega |f(y)-f(x)|\Big(1\{\|x-y\| \leq 2t_{n,\delta_n}\} + 2\delta_n\Big)\,dy\,dx \tag{A.34}$$

$$\leq 2\mathbb{E}[\operatorname{DTV}(f; w^{\varepsilon \leftarrow t_{n,\delta}})] + 2n^{1-1/d}\int_\Omega\int_\Omega |f(y)-f(x)|\,dy\,dx. \tag{A.35}$$

Note that since $\lim_{n\to\infty} t_{n,\delta_n} = 0$, the condition $t_{n,\delta_n} < t_0$ will automatically be satisfied for all n sufficiently large.

We now conclude the proof by upper bounding each term in (A.35). The first term refers to the expected ε-neighborhood graph total variation of f when $\varepsilon = t_{n,\delta_n}$, and by (A.39) satisfies

$$\mathbb{E}[\operatorname{DTV}_{n,t_{n,\delta}}(f)] \leq Cn^2(t_{n,\delta_n})^{d+1}\operatorname{TV}(f;\Omega) \leq Cn^{1-1/d}(\log n)^{(d+1)/d}\operatorname{TV}(f;\Omega).$$

The second term above can be upper bounded using a Poincaré inequality for $\mathrm{BV}(\Omega)$ functions, i.e.,

$$\int_\Omega\int_\Omega |f(y)-f(x)|\,dy\,dx \leq 2\int_\Omega |f(x)-\bar{f}(x)|\,dx \leq C\operatorname{TV}(f;\Omega).$$

Plugging these upper bounds back into (A.35) yields the claimed result (2.27) in the unit weights case. □

Clipped weights

We now show (2.27) using clipped weights. Our goal is to upper bound

$$\mathbb{E}\Big[\operatorname{DTV}\Big(f(x_{1:n});\, \tilde{w}^{\mathrm{V}}\Big)\Big] = n(n-1)\mathbb{E}\Big[|f(x_1)-f(x_2)| \max\{c_0 n^{-(d-1)/d}1\{\mathcal{H}^{d-1}(\bar{V}_1\cap\bar{V}_2)>0\}, \mathcal{H}^{d-1}(\bar{V}_1\cap\bar{V}_2)\}\Big].$$

By conditioning, we may rewrite the expectation above as

$$p_{\max}^2\int_\Omega\int_\Omega |f(y)-f(x)| \times \mathbb{E}_{x_{3:n}}\big[\max\{c_0 n^{-(d-1)/d}1\{\mathcal{H}^{d-1}(\bar{V}_x\cap\bar{V}_y)>0\}, \mathcal{H}^{d-1}(\bar{V}_x\cap\bar{V}_y)\}\big]\,dy\,dx, \tag{A.36}$$

where $V_x = \{z : \|z - x\|_2 < \|z - x_i\| \; \forall i = 2, 3, \ldots, n\}$, and likewise for V_y. Note that V_x and V_y are random subsets of $\mathbb{R}^d$. We now focus on controlling the inner expectation of (A.36). Upper bound the maximum of two positive functions with their sum to obtain,

$$\mathbb{E}_{3:n}\Big[\max\{c_0 n^{-(d-1)/d} 1\{\mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y) > 0\}, \mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y)\}\Big] \\ \leq c_0 n^{-(d-1)/d} \mathbb{P}\{\mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y) > 0\} + \mathbb{E}\left[\mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y)\right]. \tag{A.37}$$

We recognize the first term on the RHS of (A.37) as having already been analyzed in the unit weights case; we now focus on the second term. The latter "Voronoi kernel" term may be rewritten,

$$\mathbb{E}_{x_{3:n}}\left[\mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y)\right] = \int_{L \cap \Omega} (1 - p_x(z))^{n-2} dz,$$

where $L = \{z : \|x - z\| = \|y - z\|\}$ and $p_x(z) = P(B(z, \|x - z\|))$. Observe by Assumption A1 that $p_x(z) \geq p_{\min} \mu_d \|x - z\|^d$, and therefore

$$\int_{L \cap \Omega} (1 - p_x(z))^{n-2} \leq \exp(-cn\|x - z\|^d),$$

for some $c > 0$. Apply Lemma A.16 with $a = 2$ to therefore bound,

$$\mathbb{E}_{x_{3:n}}\left[\mathcal{H}^{d-1}(\bar{V}_x \cap \bar{V}_y)\right] \leq C_1 \left(\frac{1\{\|x - y\| \leq C_2 (\log n / n)^{1/d}\}}{n^{(d-1)/d}} + \frac{1}{n^2} \right), \tag{A.38}$$

for constants $C_1, C_2 > 0$. Substitute (A.38) into (A.37) and (A.36) to obtain,

$$\begin{aligned}
&\mathbb{E}\Big[\,\mathrm{DTV}\left(f(x_{1:n});\, \tilde{w}\right)\Big] \\
&\quad \leq p_{\max}^2 n^2 \int_\Omega \int_\Omega |f(y) - f(x)| \Bigg(c_0 n^{-(d-1)/d} \mathbb{P}_{3:n}\{\mathcal{H}(\bar{V}_x \cap \bar{V}_y) > 0\} \\
&\qquad\qquad + C_1 \frac{1\{\|x - y\| \leq C_2 (\log n / n)^{1/d}\}}{n^{(d-1)/d}} + \frac{C_1}{n^2} \Bigg)\, dy\, dx \\
&\quad \leq p_{\max}^2 c_0 n^{-(d-1)/d} \mathbb{E}\left[\mathrm{DTV}(f(x_{1:n}); \check{w}^{\mathrm{V}})\right] \\
&\qquad\qquad + p_{\max}^2 C_1 n^{-(d-1)/d} \mathbb{E}\left[\mathrm{DTV}(f(x_{1:n}); w^{\varepsilon \leftarrow C_2 (\log n/n)^{1/d}})\right] \\
&\qquad\qquad + p_{\max}^2 C_1 \int_\Omega \int_\Omega |f(y) - f(x)|\, dy\, dx \\
&\quad = T_1 + T_2 + T_3.
\end{aligned}$$

We bound each of the terms above in turn. The first term appeals to (2.27) in the unit weights case, which we have already proved.

$$\begin{aligned} T_1 &= p_{\max}^2 c_0 n^{-(d-1)/d} \mathbb{E}\left[\mathrm{DTV}(f(x_{1:n}); \check{w}^{\mathrm{V}})\right] \\ &\leq C_3 n^{-(d-1)/d} n^{(d-1)/d} (\log n)^{1+1/d} \,\mathrm{TV}(f) \\ &= C_3 (\log n)^{1+1/d} \,\mathrm{TV}(f). \end{aligned}$$

The second term refers to the expected ε-neighborhood graph total variation of f when $\varepsilon = C_2(\log n/n)^{1/d}$, which by (A.39) satisfies,

$$\begin{aligned} T_2 &= p_{\max}^2 C_1 n^{-(d-1)/d} \mathbb{E}\left[\mathrm{DTV}\left(f(x_{1:n}); w^{\varepsilon \leftarrow C_2(\log n/n)^{1/d}}\right)\right] \\ &\leq C_4 n^{-(d-1)/d} n^2 (\log n/n)^{(d+1)/d} \,\mathrm{TV}(f) \\ &\leq C_4 (\log n)^{1+1/d} \,\mathrm{TV}(f). \end{aligned}$$

The third term can be controlled via the Poincaré inequality,

$$\begin{aligned} T_3 &= p_{\max}^2 C_1 \int_\Omega \int_\Omega |f(y) - \bar{f} + \bar{f} - f(x)| \, dy \, dx \\ &\leq C_5 \int_\Omega |f(x) - \bar{f}| \, dx \\ &\leq C_5 \,\mathrm{TV}(f), \end{aligned}$$

where $\bar{f} := \fint_\Omega f$. □

ε-neighborhood and kNN expected discrete TV

Lemma 10. *Under Assumption A1, there exist constants $c, C_1, C_2 > 0$ such that for all sufficiently large n and $f_0 \in \mathrm{BV}(\Omega)$,*

- *The ε-neighborhood graph total variation, for any $\varepsilon > 0$, satisfies*

$$\mathbb{E}\Big[\,\mathrm{DTV}\left(f_0(x_{1:n};\ w^{\varepsilon})\right)\Big] \leq C_1 n^2 \varepsilon^{d+1} \,\mathrm{TV}(f_0). \tag{A.39}$$

- *The k-nearest neighbors graph total variation, for any $k \in \mathbb{N}$, satisfies*

$$\mathbb{E}\Big[\,\mathrm{DTV}\left(f_0(x_{1:n};\ w^{k})\right)\Big] \leq C_2 \left(n^{1-1/d} k^{(d+1)/d} + n^2 \exp(-ck)\right) \mathrm{TV}(f_0). \tag{A.40}$$

Proof.

ε-neighborhood expected discrete TV. This follows the proof of Lemma 1 in Green et al. [2021a], with two adaptations to move from Sobolev $H^2(\Omega)$ to the $\mathrm{BV}(\Omega)$: we deal in absolute differences rather than squared differences, and an approximation argument is invoked at the end to account for the existence of non-weakly differentiable functions in $\mathrm{BV}(\Omega)$.

Begin by rewriting,

$$\begin{aligned}\mathbb{E}\left[\sum_{i,j=1}^{n}|f(x_i)-f(x_j)|\cdot 1\{\|x_i-x_j\|\le\varepsilon\}\right]\\ =\frac{n(n-1)}{2}\mathbb{E}\left[|f(X')-f(X)|K\left(\frac{\|X'-X\|}{\varepsilon}\right)\right],\end{aligned}\tag{A.41}$$

where X and X' are random variables independently drawn from P following Assumption A1 and $K(t)=1\{t\le 1\}$. Now, take Ω' to be an arbitrary bounded open set such that $B(x,c_0)\subseteq\Omega'$ for all $x\in\Omega$.

For the remainder of this proof, we assume that (i) $f\in BV(\Omega')$ and (ii) $\|f\|_{BV(\Omega')}\le C'\|f\|_{BV(\Omega)}$ for some constant C' independent of f. These conditions are guaranteed by the Extension Theorem (Evans, 2010; Section 5.4 Theorem 1), which promises an extension operator $E:W^{1,p}(\Omega)\to W^{1,p}(\Omega)$ (take $p=1$ and the BV case is established through an approximation argument). We also assume that $f\in C^\infty(\Omega)$, which is addressed through via an approximation argument at the end. Since $f\in C^\infty(\Omega)$, we may rewrite a difference in terms of an integrated derivative:

$$f(x')-f(x)=\int_0^1\nabla f(x+t(x'-x))^\top(x'-x)dx.\tag{A.42}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[|f(X')-f(X)|K\left(\frac{\|X'-X\|}{\varepsilon}\right)\right]\\ \le p_{\max}^2\int_\Omega\int_\Omega|f(x')-f(x)|K\left(\frac{\|x'-x\|}{\varepsilon}\right)dx'dx,\end{aligned}\tag{A.43}$$

and the final step is to bound the double integral. We have

$$\begin{aligned}
&\int_\Omega \int_\Omega |f(x') - f(x)| K\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' dx \\
&\quad \overset{(i)}{=} \int_\Omega \int_\Omega \Big| \int_0^1 \nabla f(x + t(x' - x))^\top (x' - x) dt \Big| K\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' dx \\
&\quad \overset{(ii)}{\leq} \int_\Omega \int_\Omega \int_0^1 |\nabla f(x + t(x' - x))^\top (x' - x)| K\left(\frac{\|x' - x\|}{\varepsilon}\right) dt dx' dx \\
&\quad \overset{(iii)}{=} \int_\Omega \int_{B(0,1)} \int_0^1 |\nabla f(x + t\varepsilon z)^\top (\varepsilon z)| K(\|z\|) \varepsilon^d dt dz dx \\
&\quad = \varepsilon^{d+1} \int_\Omega \int_{B(0,1)} \int_0^1 |\nabla f(x + t\varepsilon z)^\top z| K(\|z\|) dt dz dx \\
&\quad \overset{(iv)}{\leq} \varepsilon^{d+1} \int_{\Omega'} \int_{B(0,1)} \int_0^1 |\nabla f(\tilde{x})^\top z| K(\|z\|) dt dz d\tilde{x},
\end{aligned}$$

where we obtain *(i)* by the fundamental theorem of calculus; *(ii)* by Jensen's inequality; *(iii)* by setting $z = (x' - x)/\varepsilon$, and *(iv)* by setting $\tilde{x} = x + t\varepsilon z$.

Next, we apply the Cauchy-Schwarz to $|\nabla f(\tilde{x})^\top z|$ to obtain,

$$\begin{aligned}
\int_{B(0,1)} |\nabla f(\tilde{x})^\top z| K(\|z\|) dz &\leq \int_{B(0,1)} \|\nabla f(\tilde{x})\| \|z\| K(\|z\|) dz \\
&= \|\nabla f(\tilde{x})\| \int_{(B(0,1)} \|z\| K(\|z\|) dz \\
&= C_d \|\nabla f(\tilde{x})\|
\end{aligned}$$

Substituting back in to the previous derivation, we obtain

$$\begin{aligned}
\int_\Omega \int_\Omega |f(x') - f(x)| K\left(\frac{\|x' - x\|}{\varepsilon}\right) dx' dx &\leq C_d \varepsilon^{d+1} \int_{\Omega'} \int_0^1 \|\nabla f(\tilde{x})\|_1 dt d\tilde{x} \\
&= C_d \varepsilon^{d+1} \|Df\|(\Omega') \\
&\leq C_d C' \varepsilon^{d+1} \|Df\|(\Omega)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{2} \sum_{i,j=1}^n |f(x_i) - f(x_j)| \cdot 1\{\|x_i - x_j\| \leq \varepsilon\}\right] &\leq \frac{n(n-1)}{2} p_{\max}^2 C_d C' \varepsilon^{d+1} \|Df\|(\Omega) \\
&\leq C_2 n^2 \varepsilon^{d+1} \|Df\|(\Omega)
\end{aligned}$$

Finally, we provide an approximation argument to justify the assumption that $f \in C^1(\Omega')$. For a function $f \in \mathrm{BV}(\Omega')$, we may construct a sequence of functions $f_k \in C^\infty(\Omega')$ via mollification such that $f_k \to f$ μ-a.e. (specifically, at all Lebesgue points) and $\|Df_k\|(\Omega') \to \|Df\|(\Omega')$ as $k \to \infty$ (Evans and Gariepy, 2015; Theorems 4.1 & 5.3). Via an application of Fatou's lemma, we find that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}\sum_{i,j=1}^n |f(x_i) - f(x_j)| \cdot 1\{\|x_i - x_j\| \le \varepsilon\}\right] \\
&\qquad = \mathbb{E}\left[\frac{1}{2}\sum_{i,j=1}^n |\lim_{k\to\infty} f_k(x_i) - f_k(x_j)| \cdot 1\{\|x_i - x_j\| \le \varepsilon\}\right] \\
&\qquad = \mathbb{E}\left[\liminf_{k\to\infty} \frac{1}{2}\sum_{i,j=1}^n |f_k(x_i) - f_k(x_j)| \cdot 1\{\|x_i - x_j\| \le \varepsilon\}\right] \\
&\qquad \le \liminf_{k\to\infty} \mathbb{E}\left[\frac{1}{2}\sum_{i,j=1}^n |f_k(x_i) - f_k(x_j)| \cdot 1\{\|x_i - x_j\| \le \varepsilon\}\right] \\
&\qquad \le \liminf_{k\to\infty} Cn^2\varepsilon^{d+1}\|Df_k\|(\Omega) \\
&\qquad = Cn^2\varepsilon^{d+1}\|Df\|(\Omega)
\end{aligned}$$

***k*-nearest neighbors expected discrete TV.** Let $\varepsilon_k(x) := \|x - x_{(k)}(x)\|_2$ and $\varepsilon_k(x,y) = \max\{\varepsilon_k(x), \varepsilon_k(y)\}$ be data-dependent radii. Notice that

$$\mathrm{DTV}_{n,k}(f) = \frac{1}{2}\sum_{i,j=1}^n |f(x_i) - f(x_j)| \cdot 1\big\{\|x_i - x_j\| \le \varepsilon_k(x_i, x_j)\big\}.$$

By linearity of expectation and conditioning, the expected k-nearest neighbor TV can be written as a double integral,

$$\begin{aligned}
&\mathbb{E}[\mathrm{DTV}(f; w^k)] \\
&\quad = n(n-1)\mathbb{E}\Big[|f(x_i) - f(x_j)|\, 1\big\{\|x_i - x_j\| \le \varepsilon_k(x_i, x_j)\big\}\Big] \\
&\quad = n(n-1)\mathbb{E}\Big[\mathbb{E}\Big[|f(x_i) - f(x_j)|\, 1\big\{\|x_i - x_j\| \le \varepsilon_k(x_i, x_j)\big\}|x_i, x_j\Big]\Big] \\
&\quad \le n(n-1)\int_\Omega\int_\Omega |f(y) - f(x)|\, \mathbb{P}\big\{\|x - y\| \le \varepsilon_k(x,y)\big\}\, dx\, dy \\
&\quad \le n(n-1)\int_\Omega\int_\Omega |f(y) - f(x)| \left(\mathbb{P}\big\{\|x - y\| \le \varepsilon_k(x)\big\} + \mathbb{P}\big\{\|x - y\| \le \varepsilon_k(x)\big\}\right) dx\, dy
\end{aligned}$$

(The first inequality above is nearly an equality for large n, and the second inequality follows by a union bound.)

We now derive an upper bound $\mathbb{P}\big\{\|x-y\| \le \varepsilon_k(x)\big\}$. First, observe that the event $\|x-y\| \le \varepsilon_k(x)$ is equivalent to $|B(x, \|y-x\|) \cap x_{1:n}| < k$. Suppose $\|y-x\| \ge C(k/n)^{1/d}$ for $C = (\frac{2d}{p_{\min}\mu_d})^{1/d}$. Then

$$p_k(x,y) := P(B(x, \|y-x\|)) \ge \frac{p_{\min}}{2d}\mu_d \|y-x\|^d \ge \frac{2k}{n},$$

and applying standard concentration bounds (Bernstein's inequality) to the tails of a binomial distribution, it follows that

$$\begin{aligned}
&\mathbb{P}\bigg\{ |B(x,\|y-x\|) \cap x_{1:n}| < k \bigg\} \\
&\qquad = \mathbb{P}\bigg\{ |B(x, \|y-x\|) \cap x_{1:n}| - np_k(x,y) < k - np_k(x,y) \bigg\} \\
&\qquad \le \exp\bigg(-\frac{c(np_k(x,y)-k)^2}{np_k(x,y) + |np_k(x,y)-k|} \bigg) \\
&\qquad \le \exp(-ck).
\end{aligned}$$

Otherwise if $\|y-x\| < C(k/n)^{1/d}$, we use the trivial upper bound 1 on the probability of an event. To summarize, we have shown

$$\mathbb{P}\big(\|x-y\| \le \varepsilon_k(x)\big) \le \begin{cases} 1, & \text{if } \|x-y\| < C(k/n)^{1/d}, \\ \exp(-ck), & \text{otherwise.} \end{cases}$$

It follows from (A.40) that

$$\begin{aligned}
\mathbb{E}[\mathrm{DTV}(f; w^k)] &\le 2n^2 \int_\Omega \int_\Omega |f(y) - f(x)| \, \Big(\big(1\{\|x-y\| < C(k/n)^{1/d}\}\big) \\
&\qquad\qquad + \exp(-ck) \Big) \, dx \, dy \\
&\le C\big(\mathbb{E}[\mathrm{DTV}(f; w^{\varepsilon \leftarrow C(k/n)^{1/d}})] + n^2 \exp(-ck) \mathrm{TV}(f, \Omega)\big); \qquad \text{(A.44)}
\end{aligned}$$

the first term on the right hand side of the second inequality is the expected ε-neighborhood graph TV of f, with radius $C(k/n)^{1/d}$, while the second term is obtained from the Poincaré inequality

$$\int_\Omega \int_\Omega |f(y) - f(x)| \, dy \, dx = \int_\Omega \int_\Omega \Big| f(y) - \bar{f} + \bar{f} - f(x) \Big| \, dy \, dx \le C\big(\mathrm{TV}(f; \Omega)\big), \tag{A.45}$$

where $\bar{f} = \fint_\Omega f(x) \, dx$ is the average of f over Ω. The claimed upper bound (A.40) follows from applying inequality (A.39), with $\varepsilon = C(k/n)^{1/d}$, to (A.44). □

A.4.5 Proof of Lemma 3

Recall that $\|g\|_{L^2(P)} \le p_{\max}\|g\|_{L^2(\mu)}$ for any $g \in L^2(\mu)$ and note that $\|\bar{f}_0\|_{L^\infty(\mu)} \le M$ with probability one. By Hölder's inequality,

$$\begin{aligned}\mathbb{E}\|\bar{f}_0 - f_0\|_{L^2(\mu)}^2 &\le \mathbb{E}\Big[\|\bar{f}_0 - f_0\|_{L^1(\mu)} \cdot \|\bar{f}_0 - f_0\|_{L^\infty(\mu)}\Big] \\ &\le 2M\,\mathbb{E}\|\bar{f}_0 - f_0\|_{L^1(\mu)},\end{aligned} \tag{A.46}$$

and the problem is reduced to upper bounding the expected $L^1(\mu)$ loss of $\bar{f}_0$. By Fubini's Theorem we may exchange expectation with integral, giving

$$\begin{aligned}\mathbb{E}\|\bar{f}_0 - f_0\|_{L^1(\mu)} &= \int_\Omega \mathbb{E}|\bar{f}_0(x) - f_0(x)|\,dx \\ &= \int_\Omega\int_\Omega |f_0(y) - f_0(x)| p_x^{(1)}(y)\,dy\,dx,\end{aligned} \tag{A.47}$$

where $p_x^{(1)}(\cdot)$ is the density of $x_{(1)}(x)$. We now give a closed form expression for this density, before proceeding to lower bound (A.47).

Closed-form expression for $p_x^{(1)}$. Suppose P satisfies Assumption A1. For any $y \in \Omega$ and $0 < r < \operatorname{dist}(y, \partial\Omega)$, we have

$$\begin{aligned}\mathbb{P}\big\{x_{(1)}(x) \in B(y,r)\big\} &\le n\,\mathbb{P}\big\{x_1 \in B(y,r)\big\}\big(\mathbb{P}\{x_2 \notin B(x, \|y-x\|)\}\big)^{(n-1)} \\ &\le np_{\max}\mu\big(B(y,r)\big)\Big(1 - P\big(B(x,\|y-x\|)\big)\Big)^{(n-1)}.\end{aligned}$$

Taking limits as $r \to 0$ gives

$$p_x^{(1)}(y) = \lim_{r\to 0}\frac{\mathbb{P}\big\{x_{(1)}(x) \in B(y,r)\big\}}{\mu(B(y,r))} = np_{\max}\Big(1 - P\big(B(x,\|y-x\|)\big)\Big)^{(n-1)}.$$

Upper bound on (A.47)**.** There exists a constant C_d such that for all $x, y \in \Omega$,

$$P\big(B(x,\|y-x\|)\big) \ge \frac{p_{\min}}{C_d}\mu(B(x,\|y-x\|)) = \frac{p_{\min}\mu_d}{C_d}\|y-x\|^d.$$

This implies an upper bound on the density of $x_{(1)}(x)$,

$$\begin{aligned}p_x^{(1)}(y) &\le n\left(1 - \frac{p_{\min}\mu_d}{C_d}\|y-x\|^d\right)^{(n-1)} \\ &\le n\exp\left(-\frac{p_{\min}\mu_d}{C_d}\Big(\frac{\|y-x\|}{n^{-1/d}}\Big)^d\right),\end{aligned}$$

where we have used the inequality $(1-x)^n \leq \exp(-nx)$ for $|x| \leq 1$. Using the inequality, valid for all monotone non-increasing functions $g : [0,\infty) \to [0,\infty)$, that $g(t) \leq 1\{t \leq t_0\}g(0) + g(t_0)$, we further conclude that

$$p_x^{(1)}(y) \leq n1\{\|y-x\| \leq \varepsilon_n^{(1)}\} + \frac{1}{n},$$

for $\varepsilon_n^{(1)} := (\frac{2C_d}{p_{\min}\mu_d}(\log n/n))^{1/d}$. Plugging back into (A.47), we see that the expected $L^1(\mu)$ error is upper bounded by the expected discrete TV of a neighborhood graph with particular kernel and radius, plus a remainder term. Specifically,

$$\begin{aligned}
\mathbb{E}\|\bar{f}_0 - f_0\|_{L^1(\mu)} &\leq n\int_\Omega\int_\Omega |f_0(y)-f_0(x)|1\{\|y-x\| \leq \varepsilon_n^{(1)}\}\,dy\,dx \\
&\qquad + \frac{1}{n}\int_\Omega\int_\Omega |f_0(y)-f_0(x)|\,dy\,dx \\
&\leq n\int_\Omega\int_\Omega |f_0(y)-f_0(x)|1\{\|y-x\| \leq \varepsilon_n^{(1)}\}\,dy\,dx + \frac{C\ \mathrm{TV}(f_0;\Omega)}{n} \qquad \text{(A.48)} \\
&= \frac{1}{n}\mathrm{E}[\mathrm{DTV}(f_0; w^{\varepsilon\leftarrow\varepsilon_n^{(1)}}))] + \frac{C\ \mathrm{TV}(f_0;\Omega)}{n},
\end{aligned}$$

where (A.48) above follows from the Poincaré inequality (A.45). We can therefore apply (A.39), which upper bounds the expected ε-neighborhood graph TV, and conclude that

$$\mathbb{E}\|\bar{f}_0 - f_0\|_{L^1(\mu)} \leq C\left(\frac{(\log n)^{1+1/d}}{n^{1/d}} + \frac{1}{n}\right)\mathrm{TV}(f_0;\Omega) \leq C\left(\frac{L(\log n)^{1+1/d}}{n^{1/d}}\right).$$

Inserting this upper bound into (A.46) completes the proof of Lemma 3. □

A.4.6 Proof of Theorem 4

The analysis of the ε-neighborhood and kNN TV denoising estimators proceeds identically, so we consider them together. Henceforth let D denote the penalty operator for either estimator and $\hat{f}$ denote their 1NN extrapolants. Follow the proof of Theorem 2 (given in Appendix A.4.1) to decompose the $L^2(P)$ error for some $C > 0$,

$$\mathbb{E}\left[\|\hat{f} - f_0\|_{L^2(P)}^2\right] \leq C\left(\frac{\lambda \log n\ \mathbb{E}\|D\theta_0\|}{n} + \frac{(\log n)^{1+\alpha}}{n} + \frac{LM(\log n)^{1+1/d}}{n^{1/d}}\right), \qquad \text{(A.49)}$$

where we have applied Lemma 3 which controls the 1NN extrapolation error. Lemma 10 provides that under the standard assumptions, there exist constants $C_1, C_1' > 0$ such that for all sufficiently large n and $\theta_0 = f_0(x_{1:n})$, $f_0 \in \mathrm{BV}(\Omega)$,

- setting $\varepsilon = c_1(\log^\alpha n/n)^{1/d}$,

$$\mathbb{E}\|D^\varepsilon\theta_0\|_1 \leq C_1 n^{(d-1)/d}(\log n)^{\alpha+\alpha/d}\,\mathrm{TV}(f_0); \tag{A.50}$$

- setting $k = c_1'(\log n)^3$,

$$\mathbb{E}\|D^k\theta_0\|_1 \leq C_1' n^{(d-1)/d}(\log n)^{3+3/d}\,\mathrm{TV}(f_0). \tag{A.51}$$

Take these values of ε, k and $\lambda = c\sigma(\log n)^{1/2-\alpha}$, $c = c_2, c_2'$, and substitute (A.50), (A.51) into (A.49) to obtain the claim. □

Note that the $L^2(P_n)$ in-sample error may be obtained similarly, beginning with an analysis identical to that of Lemma 1 to obtain the preliminary upper bound,

$$\mathbb{E}\left[\|\hat{f} - f_0\|_{L^2(P_n)}^2\right] \leq C\left(\frac{\lambda\,\mathbb{E}\|D\theta_0\|}{n} + \frac{(\log n)^{1+\alpha}}{n}\right).$$

A.5 Analysis of graph TV denoising

In this section, we review tools for analyzing graph total variation denoising. Suppose an unknown $\theta_0 \in \mathbb{R}^n$ and observations $y_1, \ldots, y_n$,

$$y_i = \theta_{0i} + z_i, \qquad i = 1, \ldots, n, \tag{A.52}$$

where $z_i \sim \mathcal{N}(0, \sigma^2)$. The graph total variation denoising estimator $\hat{\theta}$ associated with a graph $G = (V, E)$, $|V| = n$, is given by

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y_{1:n} - \theta\|_2^2 + \lambda\|D\theta\|_1, \tag{A.53}$$

where $D \in \mathbb{R}^{m\times n}$ is the edge incidence matrix of G.

The initial analysis of graph total variation denoising was performed by Hutter and Rigollet [2016] for the two-dimensional grid. Sadhanala et al. [2016] subsequently generalized the analysis to d-dimensional lattices, and Wang et al. [2016] provided tools for the analysis of general graphs. These techniques rely on direct analysis of properties of graph G and the penalty D in induces, which is tractable when G has a known and regular properties (e.g., it is a lattice graph).

Unfortunately, direct analysis on D may not always be feasible. It may be possible, however, to compare the operator D to a *surrogate operator* whose properties we analyze instead. For our purposes, we compare D to a linear operator which first takes averages on a partition, and then computes differences across cells of the partition. Comparison to this type of surrogate operator was used by Padilla et al. [2020] to bound the risk of graph total variation denoising in probability; the following theorem provides an analogous risk bound in expectation. We note that elements of the "surrogate operator analysis" are also found in Padilla et al. [2018].

Theorem 8. *Suppose we observe data according to model* (A.52) *and compute the graph TV denoising estimator* $\hat{\theta}$ *of* (A.53). *Let* $A \in \mathbb{R}^{n \times n}$ *denote an averaging operator over* $\bar{N}$ *groups of the form,*

$$A = \begin{bmatrix} n_1^{-1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top & 0 & \dots & 0 \\ 0 & n_2^{-1}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_{\bar{N}}^{-1}\mathbf{1}_{n_{\bar{N}}}\mathbf{1}_{n_{\bar{N}}}^\top \end{bmatrix},$$

with $M := \max_j n_j$, *and let* $\bar{A} \in \mathbb{R}^{\bar{N} \times n}$ *be the same matrix with redundant rows removed. Further let* $T \in \mathbb{R}^{\bar{m} \times \bar{N}}$ *be a surrogate penalty operator, with singular value decomposition* $T = U\Sigma V^\top$, *such that*

$$\|T\bar{A}\theta\|_1 \leq \Phi_1(D, T, A)\|D\theta\|_1, \tag{A.54}$$

$$\|(I - A)\theta\|_1 \leq \Phi_2(D, T, A)\|D\theta\|_1, \tag{A.55}$$

for quantities $\Phi_1(D, T, A), \Phi_2(D, T, A)$ *that may depend on* n, *for all* $\theta \in \mathbb{R}^n$. *If the penalty parameter*

$$\lambda > \max\left\{ 8M^{1/2}\Phi_1(D, T, A) \cdot \sigma\sqrt{\log(2n^4)\sum_{k=2}^{\bar{N}} \frac{\|u_k\|_\infty^2}{\xi_k^2}}, \Phi_2(D, T, A) \cdot \sigma\sqrt{2\log(n)} \right\} \tag{A.56}$$

where u_k *is the* k*th column of* U *and* ξ_k *the* k*th diagonal entry of* Σ, *then there exists a constant* $C > 0$ *such that*

$$\mathbb{E}\left[\frac{1}{n}\|\hat{\theta} - \theta_0\|_2^2\right] \leq C\left(\frac{\lambda\|D\theta_0\|_1}{n} + \frac{M\text{nullity}(T)}{n}\right). \tag{A.57}$$

Proof. We follow the approach of Padilla et al. [2020], with adaptations to provide a bound in expectation rather than in probability. From the basic inequality,

$$\|\hat{\theta} - \theta_0\|_2^2 \leq 2\langle z_{1:n}, \hat{\theta} - \theta_0\rangle + \lambda(\|D\theta_0\|_1 - \|D\hat{\theta}\|_1),$$

where $z_{1:n} \in \mathbb{R}^n$ is the vector of error terms z_i, $i = 1, \dots, n$. We provide two deterministic bounds under the "good case" that the error term falls into the set,

$$\mathcal{Z}_\lambda = \left\{ z_{1:n} : \max\left\{ M^{1/2}\Phi_1(D, T, A) \sup_{\bar{A}\theta \in \text{row}(T): \|T\bar{A}\theta\|_1 \leq 1} |\langle \overline{z_{1:n}}, \bar{A}\theta\rangle|, \right.\right. \\ \left.\left. \Phi_2(D, T, A)\|z_{1:n}\|_\infty \right\} \leq \frac{\lambda}{8} \right\}, \tag{A.58}$$

where $\overline{z_{1:n}} \in \mathbb{R}^{\bar{N}}$ has entries $\overline{z_{1:n}}_j = n_j^{1/2}(\bar{A}z_{1:n})_j$, and under the "bad case" that $z_{1:n} \notin \mathcal{Z}_\lambda$.

Upper bound in the "good case". Decompose the first term on the RHS

$$\begin{aligned}\langle z_{1:n}, \hat{\theta} - \theta_0 \rangle &= \langle z_{1:n}, \hat{\theta} - A\hat{\theta} \rangle + \langle z_{1:n}, A\theta_0 - \theta_0 \rangle + \langle z_{1:n}, A(\theta_0 - \hat{\theta}) \rangle \\ &\leq \langle z_{1:n}, A(\theta_0 - \hat{\theta}) \rangle + \|z_{1:n}\|_\infty (\|(I-A)\hat{\theta}\|_1 + \|(I-A)\theta_0\|_1) \\ &\leq \langle z_{1:n}, A(\theta_0 - \hat{\theta}) \rangle + \|z_{1:n}\|_\infty \Phi(D,T,A)(\|D\theta_0\|_1 + \|D\hat{\theta}\|_1),\end{aligned}$$

where the final inequality follows from (A.55). Observe that we may rewrite, for any $\theta \in \mathbb{R}^n$,

$$\begin{aligned}\langle z_{1:n}, A\theta \rangle &= \sum_{j=1}^{\bar{N}} \sum_{i=1}^{n_j} z_{1:n(\sum_{k=1}^{j-1} n_k)+i} (\bar{A}\theta)_j \\ &\stackrel{d}{=} \sum_{j=1}^{\bar{N}} n_j^{1/2} z_{1:nj} (\bar{A}\theta)_j \\ \Rightarrow \langle z_{1:n}, A\theta \rangle &\leq M^{1/2} |\langle \overline{z_{1:n}}, \bar{A}\theta \rangle| \\ &\leq M^{1/2} \left| \langle \mathrm{proj}_V(\overline{z_{1:n}}), \bar{A}\theta \rangle + \langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}\theta \rangle \right| \\ &\leq M^{1/2} \left| \|\mathrm{proj}_V(\overline{z_{1:n}})\|_2 \|\bar{A}\theta\|_2 + \langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}\theta \rangle \right| \\ &\leq M^{1/2} \left(\|\mathrm{proj}_V(\overline{z_{1:n}})\|_2 \|\theta\|_2 + |\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}\theta \rangle| \right|,\end{aligned}$$

where $\overline{z_{1:n}} \in \mathbb{R}^{\bar{N}}$ has independent $\mathcal{N}(0, \sigma^2)$ entries and $V = \mathrm{null}(T)$. Substitute back in to obtain

$$\begin{aligned}\|\hat{\theta} - \theta_0\|_2^2 \leq 2M^{1/2} (&\|\mathrm{proj}_V(\overline{z_{1:n}})\|_2 \|\hat{\theta} - \theta_0\|_2 + |\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}(\hat{\theta} - \theta_0) \rangle|) \\ &+ 2\|z_{1:n}\|_\infty \Phi(D,T,A)(\|D\theta_0\|_1 + \|D\hat{\theta}\|_1) + \lambda(\|D\theta_0\|_1 - \|D\hat{\theta}\|_1),\end{aligned}$$

and consequently

$$\begin{aligned}\|\hat{\theta} - \theta_0\|_2 (&\|\hat{\theta} - \theta_0\|_2 - 2M^{1/2} \|\mathrm{proj}_V(\overline{z_{1:n}})\|_2) \\ &\leq 2M^{1/2} |\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}(\hat{\theta} - \theta_0) \rangle| \\ &\quad + 2\|z_{1:n}\|_\infty \Phi(D,T,A)(\|D\theta_0\|_1 + \|D\hat{\theta}\|_1) + \lambda(\|D\theta_0\|_1 - \|D\hat{\theta}\|_1).\end{aligned}$$

Case 1. $\|\hat{\theta} - \theta_0\|_2 \leq 4M^{1/2} \|\mathrm{proj}_V(\overline{z_{1:n}})\|_2$.
Case 2. $\|\hat{\theta} - \theta_0\|_2 > 4M^{1/2} \|\mathrm{proj}_V(\overline{z_{1:n}})\|_2$. Then,

$$\begin{aligned}\|\hat{\theta} - \theta_0\|_2^2 \leq &4M^{1/2} |\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}(\hat{\theta} - \theta_0) \rangle| \\ &+ 4\|z_{1:n}\|_\infty \Phi(D,T,A)(\|D\theta_0\|_1 + \|D\hat{\theta}\|_1) + \lambda(\|D\theta_0\|_1 - \|D\hat{\theta}\|_1).\end{aligned}$$

We then bound

$$\begin{aligned}
&|\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \bar{A}(\hat{\theta}-\theta_0)\rangle| \\
&\qquad = \left| \left\langle \mathrm{proj}_{V^\perp}(\overline{z_{1:n}}), \frac{\bar{A}(\hat{\theta}-\theta_0)}{\|T\bar{A}(\hat{\theta}-\theta_0)\|_1} \right\rangle \|T\bar{A}(\hat{\theta}-\theta_0)\|_1 \right| \\
&\qquad \le \sup_{\bar{A}\theta \in V^\perp : \|T\bar{A}\theta\|_1 \le 1} |\langle \overline{z_{1:n}}, \bar{A}\theta\rangle| \|T\bar{A}(\hat{\theta}-\theta_0)\|_1 \\
&\qquad \le \sup_{\bar{A}\theta \in V^\perp : \|T\bar{A}\theta\|_1 \le 1} |\langle \overline{z_{1:n}}, \bar{A}\theta\rangle| \Phi(D,T,A)(\|D\hat{\theta}\|_1 + \|D\theta_0\|_1),
\end{aligned}$$

where the last inequality follows by (A.55). Conditioning on $z_{1:n} \in \mathcal{Z}_\lambda$, we find that under Case 2,

$$\begin{aligned}
\|\hat{\theta}-\theta_0\|_2^2 &\le \frac{\lambda}{2}(\|D\hat{\theta}\|_1 + \|D\theta_0\|_1) + \frac{\lambda}{2}(\|D\hat{\theta}\|_1 + \|D\theta_0\|_1) + \lambda(\|D\theta_0\|_1 - \|D\hat{\theta}\|_1) \\
&\le 2\lambda \|D\theta_0\|_1.
\end{aligned}$$

Therefore, conditioning on $z_{1:n} \in \mathcal{Z}_\lambda$ and combining Case 1 and Case 2, we obtain that

$$\|\hat{\theta}-\theta_0\|_2^2 \le 16M \|\mathrm{proj}_V(\overline{z_{1:n}})\|_2^2 + 2\lambda\|D\theta_0\|_1.$$

Upper bound in the "bad case". On the "bad event" that $z_{1:n} \notin \mathcal{Z}_\lambda$, we apply Hölder directly to the basic inequality to bound

$$\|\hat{\theta}-\theta_0\|_2^2 \le 2\|z_{1:n}\|_2 \|\hat{\theta}-\theta_0\|_2 + \lambda\|D\theta_0\|_1,$$

and rearrange to obtain

$$\|\hat{\theta}-\theta_0\|_2^2 \le 16\|z_{1:n}\|_2^2 + 2\lambda\|D\theta_0\|_1. \tag{A.59}$$

Combining the "good case" and "bad case" upper bounds.

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\|\hat{\theta}-\theta_0\|_2^2 &= \frac{1}{n}\mathbb{E}\left[\|\hat{\theta}-\theta_0\|_2^2 \mathbf{1}\{z_{1:n} \in \mathcal{Z}_\lambda\} + \|\hat{\theta}-\theta_0\|_2^2 \mathbf{1}\{z_{1:n} \notin \mathcal{Z}_\lambda\}\right] \\
&\le \frac{1}{n}\Bigg[\mathbb{E}\left[16M\|\mathrm{proj}_V(\overline{z_{1:n}})\|_2^2 + 2\lambda\|D\theta_0\|_1\right] \\
&\qquad\qquad + \mathbb{E}\left[(16\|z_{1:n}\|_2^2 + 2\lambda\|D\theta_0\|_1)\mathbf{1}\{z_{1:n} \notin \mathcal{Z}_\lambda\}\right]\Bigg] \\
&\le \frac{1}{n}\left[16M\dim(V) + 4\lambda\|D\theta_0\|_1 + \sqrt{\mathbb{E}[\|z_{1:n}\|_2^4]} \cdot \mathbb{P}[z_{1:n} \notin \mathcal{Z}_\lambda]\right] \\
&\le \frac{1}{n}\left[16M\dim(V) + 4\lambda\|D\theta_0\|_1 + \sqrt{3}n \cdot \mathbb{P}[z_{1:n} \notin \mathcal{Z}_\lambda]\right]
\end{aligned}$$

It remains to bound the probability of the bad case,

$$\begin{aligned}\mathbb{P}\{z_{1:n} \notin \mathcal{Z}_\lambda\} &\leq \mathbb{P}\left\{M^{1/2} \sup_{\bar{A}\theta \in V^\perp : \|T\bar{A}\theta\|_1 \leq 1} |\langle \overline{z_{1:n}}, \bar{A}\theta \rangle| \geq \lambda/8\Phi(D,T,A)\right\} \\ &\quad + \mathbb{P}\left\{\|z_{1:n}\|_\infty \geq \lambda/8\Phi(D,T,A)\right\} \\ &\leq \mathbb{P}\{M^{1/2}\Phi(D,T,A)\|(T^+)^\top \overline{z_{1:n}}\|_\infty \geq \lambda/8\} + \mathbb{P}\{\Phi(D,T,A)\|z_{1:n}\|_\infty \geq \lambda/8\}.\end{aligned}$$

Standard results on the maxima of Gaussians provide that

$$\begin{aligned}\mathbb{P}\Bigg\{M^{1/2}\Phi_1(D,T,A)\|(T^+)^\top \overline{z_{1:n}}\|_\infty \geq \\ M^{1/2}\Phi_1(D,T,A)\cdot\sigma\sqrt{\log(2n^2/\delta)\cdot\sum_{k=2}^{\bar{N}}\frac{\|u_k\|_\infty^2}{\xi_k^2}}\Bigg\} \leq \delta, \\ \mathbb{P}\left\{\Phi_2(D,T,A)\|z_{1:n}\|_\infty \geq \Phi_2(D,T,A)\cdot\sigma\sqrt{\log(2n^2/\delta)}\right\} \leq \delta.\end{aligned}$$

Recalling the choice of penalty parameter,

$$\begin{aligned}\lambda > \max\Bigg\{8M^{1/2}\Phi_1(D,T,A)\cdot\sigma\sqrt{\log(2n^4)\sum_{k=2}^{\bar{N}}\frac{\|u_k\|_\infty^2}{\xi_k^2}}, \\ \Phi_2(D,T,A)\cdot\sigma\sqrt{2\log(n)}\Bigg\},\end{aligned}$$

we conclude that

$$\mathbb{P}\{z_{1:n} \notin \mathcal{Z}_\lambda\} \leq \frac{2}{n^2},$$

completing the proof. □

We now state a well-known result controlling certain functionals of the lattice difference operator. These quantities have been analyzed by others studying graph total variation denoising on lattices, e.g., Hutter and Rigollet [2016] and Sadhanala et al. [2017].

Lemma 11. *Let T be the edge incidence operator of the d-dimensional lattice graph N elements per direction. Denote $n = N^d$. The left singular vectors of T satisfy an incoherence condition,*

$$\|u_j\|_\infty \leq \frac{C_d}{\sqrt{n}}, \qquad j = 1, \ldots, n,$$

for some $C_d > 0$, and its singular values satisfy an asymptotic scaling,

$$c_d(j/n)^{1/d} \le \xi_j \le C_d(j/n)^{1/d}, \qquad j = 2, \dots, n,$$

for some $0 < c_d < C_d$. Consequently,

$$\sum_{j=2}^{n} \frac{\|u_j\|_\infty^2}{\xi_j^2} = C_d \begin{cases} \log n & d = 2, \\ 1 & d > 2. \end{cases} \tag{A.60}$$

A.6 Embeddings for random graphs

We begin by providing a result that controls the number of sample points that fall into each cell of a lattice mesh.

Lemma 12. *Suppose $x_1, \dots, x_n$ are sampled from a distribution P supported on $(0,1)^d$ with density p such that $0 < p_{\min} < p(x) < p_{\max} < 1$ for all $x \in (0,1)^d$. Form a partition of $(0,1)^d$ using an equally spaced mesh with $N = C_1(p_{\min} n/\log^\alpha n)^{1/d}$, $\alpha > 1$, along each dimension. Let $\mathcal{C}_\ell$ denote the ℓth cell of the mesh, and let $|\mathcal{C}_\ell|$ denote its empirical content. Then for all $x_{1:n} \in \mathscr{X}_1$, with $\mathbb{P}\{x_{1:n} \in \mathscr{X}_1\} \ge 1 - 2/n^4$,*

$$\max_\ell |\mathcal{C}_\ell| \le C_3 \log^\alpha n, \tag{A.61}$$

$$\min_\ell |\mathcal{C}_\ell| \ge c_4 \log^\alpha n, \tag{A.62}$$

for n sufficiently large, where $C_3, c_4 > 0$ depend only on $p_{\min}, p_{\max}, d$.

Proof. From standard concentration bounds (e.g., Von Luxburg et al., 2014; Proposition 27) on a random variable $m \sim \mathrm{Bin}(n, p)$, for all $\delta \in (0, 1]$,

$$\mathbb{P}\{m \ge (1+\delta)np\} \le \exp\{-\frac{1}{3}\delta^2 np\},$$
$$\mathbb{P}\{m \le (1-\delta)np\} \le \exp\{-\frac{1}{3}\delta^2 np\}.$$

Apply these bounds with $p = \mathbb{P}\{x \in \mathcal{C}_\ell\}$ to obtain that

$$\mathbb{P}\left\{\max_\ell |\mathcal{C}_\ell| \ge (1+\delta)C_1^d \frac{p_{\max}}{p_{\min}} \log^\alpha n\right\} \le N^d \exp\left\{-\frac{1}{3}\delta^2 C_1^d \log^\alpha n\right\},$$
$$\mathbb{P}\left\{\min_\ell |\mathcal{C}_\ell| \le (1-\delta)C_1^d \log^\alpha n\right\} \le N^d \exp\left\{-\frac{1}{3}\delta^2 C_1^d \log^\alpha n\right\},$$

for all $\delta \in (0,1)$. Setting the RHS to $1/n^4$,

$$\frac{C_1^d p_{\min} n}{\log^\alpha n} \exp\{-\frac{1}{3}\delta^2 C_1^{-d} \log^\alpha n\} \le \frac{1}{n^4}$$
$$\log(C_1^d p_{\min}) - \log(\log^\alpha n) - \frac{1}{3}\delta^2 C_1^{-d} \log^\alpha n \le -5 \log n;$$

it follows that

$$\begin{aligned}\frac{1}{3}\delta^2 C_1^{-d}\log^\alpha n &\geq 5\log n + \log(C_1^d p_{\min}) - \log(\log^\alpha n)\\ \delta^2 &\geq 3C_1^d\left(5\log^{1-\alpha} n + \frac{\log(C_1^d p_{\min})}{\log^\alpha n} - \frac{\log(\log^\alpha n)}{\log^\alpha n}\right)\\ \delta &\geq C_2\log^{(1-\alpha)/2} n,\end{aligned}$$

for some $C_2 > 0$ for all n sufficiently large. Therefore deduce that

$$\begin{aligned}\mathbb{P}\left\{\max_\ell |\mathcal{C}_\ell| \geq C_1^d\frac{p_{\max}}{p_{\min}}\log^\alpha n + C_1^d C_2\frac{p_{\max}}{p_{\min}}\log^{(1+\alpha)/2} n\right\} &\leq \frac{1}{n^4},\\ \mathbb{P}\left\{\min_\ell |\mathcal{C}_\ell| \leq C_1^d\log^\alpha n - C_1^d C_2\log^{(1+\alpha)/2} n\right\} &\leq \frac{1}{n^4}.\end{aligned}$$

Recall that $\alpha > 1$ by assumption, and choose $C_3, c_4 > 0$ with n sufficiently large to obtain the claim. □

The following lemma establishes embeddings from certain random graphs into a coarser lattice graph.

Lemma 13. *Partition the domain $(0,1)^d$ using an equally spaced mesh with $N = C_1(p_{\min}n/\log^\alpha n)^{1/d}$ elements per direction. Suppose that $x_{1:n} \in \mathscr{X}_1$, with $x_{1:n}$ re-indexed such that*

$$\begin{aligned}x_1,\dots,x_{|\mathcal{C}_1|} &\in \mathcal{C}_1,\\ x_{|\mathcal{C}_1|+1},\dots,x_{|\mathcal{C}_1|+|\mathcal{C}_2|} &\in \mathcal{C}_2,\\ &\vdots\\ x_{\sum_{\ell=1}^{N^d-1}|\mathcal{C}_\ell|+1},\dots,x_{N^d} &\in \mathcal{C}_{N^d}.\end{aligned}$$

Consider the averaging operator

$$A = \begin{bmatrix} |\mathcal{C}_1|^{-1}\mathbf{1}_{|\mathcal{C}_1|}\mathbf{1}_{|\mathcal{C}_1|}^\top & 0 & \dots & 0\\ 0 & |\mathcal{C}_2|^{-1}\mathbf{1}_{|\mathcal{C}_2|}\mathbf{1}_{|\mathcal{C}_2|}^\top & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & |\mathcal{C}_{N^d}|^{-1}\mathbf{1}_{|\mathcal{C}_{N^d}|}\mathbf{1}_{|\mathcal{C}_{N^d}|}^\top \end{bmatrix}, \tag{A.63}$$

and the lattice difference operator T based on the graph

$$G_T = (\{1,\dots,N^d\}, E_T), \tag{A.64}$$

where $(i,j) \in E_T$ if the midpoints of $\mathcal{C}_i, \mathcal{C}_j$ are $1/N$ apart. Also, let $\bar{A} \in \mathbb{R}^{N^d\times n}$ be the matrix obtained by dropping the redundant rows of A.

- *Build the Voronoi graph from $x_{1:n}$, and let $\tilde{D}^{\mathrm{V}}$ denote the edge incidence operator with edge set E^{V} and edge weights $\tilde{w}_{ij}^{\mathrm{V}} = \max\{c_0 n^{-(d-1)/d}, w_{ij}^{\mathrm{V}}\}$ for each i, j. Further condition on the set $\mathscr{X}_2$ such that the result of Lemma 14 holds with probability $1 - 1/n^4$ (equivalently, the set that the result of Lemma 15 holds with probability $1 - 1/n^4$). Then there exists a constant $C_6 > 0$ such that for all $\theta \in \mathbb{R}^n$,*

$$\|T\bar{A}\theta\|_1 \leq C_6 n^{(d-1)/d} \|\tilde{D}^{\mathrm{V}}\theta\|_1. \tag{A.65}$$

$$\|(I - A)\theta\|_1 \leq C_6 (\log n)^{\alpha} n^{(d-1)/d} \|\tilde{D}^{\mathrm{V}}\theta\|_1, \tag{A.66}$$

- *Build the Voronoi graph from $x_{1:n}$, and let $\check{D}^{\mathrm{V}}$ denote the edge incidence operator with edge set E^{V} and edge weights $\check{w}_{ij}^{\mathrm{V}} = 1$ for each i, j such that $w_{i,j}^{\mathrm{V}} > 0$. Further condition on the set $\mathscr{X}_2$ such that the result of Lemma 14 holds with probability $1 - 1/n^4$. Then there exists a constant $C_7 > 0$ such that for all $\theta \in \mathbb{R}^n$,*

$$\|T\bar{A}\theta\|_1 \leq C_7 \|\check{D}^{\mathrm{V}}\theta\|_1. \tag{A.67}$$

$$\|(I - A)\theta\|_1 \leq C_7 (\log n)^{\alpha} \|\check{D}^{\mathrm{V}}\theta\|_1, \tag{A.68}$$

- *Build the ε-neighborhood graph from $x_{1:n}$, with $\varepsilon \geq 2\sqrt{d}/N$. Then with the constant c_4 from Lemma 12, it holds that for all $\theta \in \mathbb{R}^n$,*

$$\|T\bar{A}\theta\|_1 \leq \frac{1}{c_4^2 \log^{2\alpha} n} \|D^{\varepsilon}\theta\|_1. \tag{A.69}$$

$$\|(I - A)\theta\|_1 \leq \frac{2}{c_4 \log^{\alpha} n} \|D^{\varepsilon}\theta\|_1, \tag{A.70}$$

- *Build the k-nearest neighbors graph from $x_{1:n}$, with $k \geq C_5 \log^3 n$. Further condition on the set $\mathscr{X}_2$ such that the result of Lemma 14 holds with probability $1 - 1/n^4$. Then with the constant c_4 from Lemma 12, it holds for all $\theta \in \mathbb{R}^n$,*

$$\|T\bar{A}\theta\|_1 \leq \frac{1}{c_4^2 \log^{2\alpha} n} \|D^{k}\theta\|_1. \tag{A.71}$$

$$\|(I - A)\theta\|_1 \leq \frac{2}{c_4 \log^{\alpha} n} \|D^{k}\theta\|_1, \tag{A.72}$$

Proof. **ε-neighborhood graph.** First, we prove (A.69) and (A.70). For the former,

observe that

$$\begin{aligned}\|T\bar{A}\theta\|_1 &= \sum_{(k,\ell)\in E_T}\left||\mathcal{C}_k|^{-1}\sum_{i\in\mathcal{C}_k}\theta_i - |\mathcal{C}_\ell|^{-1}\sum_{j\in\mathcal{C}_\ell}\theta_j\right| \\ &\le \sum_{(k,\ell)\in E_T}\frac{1}{|\mathcal{C}_k||\mathcal{C}_\ell|}\sum_{i\in|\mathcal{C}_k|,j\in|\mathcal{C}_\ell|}|\theta_i-\theta_j| \\ &\le \frac{1}{c_4^2\log^{2\alpha}n}\sum_{(k,\ell)\in E_T}\sum_{i\in\mathcal{C}_k,j\in\mathcal{C}_\ell}|\theta_i-\theta_j| \\ &\le \frac{1}{c_4^2\log^{2\alpha}n}\|D^\varepsilon\theta\|_1,\end{aligned}$$

as $\varepsilon = 2\sqrt{d}/N$. For the latter, similarly deduce that

$$\begin{aligned}\|(I-A)\theta\|_1 &= \sum_{i=1}^n\left|\theta_i - |\mathcal{C}(i)|^{-1}\sum_{j\in\mathcal{C}(i)}\theta_j\right| \\ &\le \sum_{i=1}^n|\mathcal{C}(i)|^{-1}\left|\sum_{j\in\mathcal{C}(i)}\theta_j-\theta_i\right| \\ &\le \sum_{i=1}^n|\mathcal{C}(i)|^{-1}\sum_{j\in\mathcal{C}(i)}|\theta_i-\theta_j| \\ &= \sum_{\ell=1}^{N^d}|\mathcal{C}_\ell|^{-1}\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_\ell}|\theta_i-\theta_j| \\ &\le \frac{2}{c_4\log^\alpha n}\sum_{\ell=1}^{N^d}\sum_{i<j\in\mathcal{C}_\ell}|\theta_i-\theta_j| \\ &\le \frac{2}{c_4\log^\alpha n}\|D^\varepsilon\theta\|_1.\end{aligned}$$

***k*-nearest neighbors graph.** Recall that we have conditioned on the set $\mathscr{X}_2$ such that the result of Lemma 14 holds. In particular, (A.75) gives that

$$\min_{i=1,\dots,n}\varepsilon_k(x_i) \ge C\left(\frac{k}{n}\right)^{1/d},$$

where $\varepsilon_k(x_i) := \|x_i - x_{(k)}(x_i)\|_2$. The results (A.71) and (A.72) then follow by observing that on the event $\mathscr{X}_2$, the k-nearest neighbors graph with $k \geq C_5 \log^3 n$ dominates the ε-neighborhood graph with $\varepsilon = 2\sqrt{d}/N$.

Voronoi adjacency graph. We will prove the results (A.67) and (A.68) by providing a graph comparison inequality between the ε-neighborhood graph with $\varepsilon = 2\sqrt{d}/N$ and the Voronoi adjacency graph. The results (A.65), (A.66) follow from the inequality $\|\check{D}^{\mathrm{V}}\theta\|_1 \leq c_0^{-1} n^{(d-1)/d} \|\tilde{D}^{\mathrm{V}}\theta\|_1$ for all $\theta \in \mathbb{R}^n$.

Intuition and outline. The central goal of this proof is to show that

$$\|D^{\varepsilon}\theta\|_1 \leq C(n)\|\check{D}^{\mathrm{V}}\theta\|_1,$$

for all $\theta \in \mathbb{R}^n$, where $C(n)$ is at most polylogarithmic in n. This will be accomplished by

(i) verifying that for any $\{x_i, x_j\} \in E^{\varepsilon}$, there exists a path $\{x_i, x_{k_1}\}, \{x_{k_1}, x_{k_2}\}, \ldots, \{x_{k_{ij}}, x_j\} \in E^{\mathrm{V}}$, and

(ii) showing that if one uses the shortest path in the Voronoi adjacency graph G^{V} to connect each $\{x_i, x_j\} \in E^{\varepsilon}$, then no one edge is used more than $C_9 \log^{2\alpha} n$ times, where C_9 is a positive constant and $\alpha > 1$ may be chosen.

Step (i). Consider x_i, x_j such that $\{x_i, x_j\} \in E^{\varepsilon}$. We will show the existence of a path between x_i and x_j in G^{V} and also characterize some properties of the path for step (ii).

By definition, $\|x_i - x_j\| \leq \varepsilon$. Denote

$$x_{ij} := \frac{x_i + x_j}{2},$$
$$r_{ij} := \|x_i - x_{ij}\|.$$

Consider the subgraph $G^{ij} = (V^{ij}, E^{ij})$, where

$$V^{ij} := \{V_k : V_k \cap B(x_{ij}, r_{ij}) \neq \emptyset\},$$
$$E^{ij} := \{\{V_k, V_\ell\} : V_k, V_\ell \in V^{ij}, \mathcal{H}^{d-1}(\partial V_k \cap \partial V_\ell) > 0\},$$

where $B(x_{ij}, r_{ij})$ is the closed ball centered at x_{ij} with radius r_{ij}. By construction, $x_i, x_j \in V^{ij}$, and by Lemma 16, G^{ij} is connected. Therefore a path between x_i and x_j exists in the graph G^{ij} (one can use, e.g., breadth-first search or Dijkstra's algorithm to find such a path).

Step (ii). For any $\{x_i, x_j\} \in E^{\varepsilon} \setminus E^{\mathrm{V}}$, we create a path in G^{V} as prescribed in step (i). With these paths created, we upper bound the number of times any edge in E^{V} is

used. We do so by uniformly bounding above the number of times a vertex x_k appears in these paths (and since each edge involves two vertices, this immediately yields an upper bound on the number of times an edge appears in these paths). We split this into two substeps:

(a) first, we derive a necessary condition for x_k to appear in the path between x_i and x_j;

(b) then, we will upper bound the number of possible pairs x_i, x_j such that this necessary condition is satisfied.

Step (ii a). For x_k to appear in the path between x_i and x_j as designed in step (i), it is necessary for $V_k \in V^{ij}$. Consider $x \in V_k \cap B(x_{ij}, r_{ij})$. Since x belongs to the Voronoi cell V_k,

$$\|x - x_k\| < \min\{\|x - x_i\|, \|x - x_j\|\},$$

but since x also lies in $B(x_{ij}, r_{ij})$,

$$\|x - x_{ij}\| < r_{ij}.$$

It follows that,

$$\begin{aligned}\|x_k - x_{ij}\| &\leq \|x - x_k\| + \|x - x_{ij}\| \\ &\leq \|x - x_i\| + \|x - x_{ij}\| \\ &\leq \|x - x_{ij}\| + \|x_i - x_{ij}\| + \|x - x_{ij}\| \\ &\leq 3r_{ij},\end{aligned}$$

thus if $V_k \in V^{ij}$, then it is necessary for $x_k \in B(x_{ij}, 3r_{ij})$.

Step (ii b). Recalling $\varepsilon = 2\sqrt{d}/N$, where $N = C_1(p_{\min} n/\log^\alpha n)^{1/d}$, we have a uniform upper bound of

$$\max_{\{x_i,x_j\}\in E^\varepsilon} r_{ij} \leq C_8 \left(\frac{\log^\alpha n}{n}\right)^{1/d},$$

for some $C_8 > 0$. Thus, we conclude that for an edge of x_k to be involved in a path between x_i and x_j, it is necessary for

$$x_{ij} \in B(x_k, 3C_8(\log n/n)^{1/d}),$$

or more loosely,

$$x_i, x_j \in B(x_k, 4C_8(\log n/n)^{1/d}),$$

recalling that $r_{ij} = \|x_{ij} - x_i\| = \|x_{ij} - x_j\|$ and the uniform upper bound on r_{ij}. Therefore, the number of paths in which any x_k may appear is bounded above,

$$(nP_n(\cdot, 4C_8(\log^\alpha n/n)^{1/d}))^2 \leq C_9 \log^{2\alpha} n,$$

where the final inequality is obtained by (A.74). □

A.7 Auxiliary lemmas and proofs

A.7.1 Useful concentration results

The following is an immediate consequence of the well-known fact that the set of balls B in $\mathbb{R}^d$ has VC dimension $d+1$ (e.g., Lemma 16 of Chaudhuri and Dasgupta, 2010).

Lemma 14. *Suppose $x_1, \ldots, x_n$ are drawn from P satisfying Assumption A1. There exist constants C_1-C_5 depending only on d, $p_{\min}$, and $p_{\max}$ such that the following statements hold: with probability at least $1-\delta$, for any $z \in \Omega$,*

$$\{|B(z,r) \cap \{x_1, \ldots, x_n\}| = 0\} \Longrightarrow \left\{ r < C_1 \left(\frac{\log n + \log(1/\delta)}{n} \right)^{1/d} \right\}, \tag{A.73}$$

and

$$\left\{ r < C_2 \Big(\frac{k - C_3(d \log n + \log(1/\delta) + \sqrt{k(d \log n + \log(1/\delta))})}{n} \Big)^{1/d} \right\} \Longrightarrow \{|B(z,r) \cap \{x_1, \ldots, x_n\}| < k\}. \tag{A.74}$$

In particular, if $k \geq C_4 (\log(1/\delta))^2 \log n$, then

$$\{|B(z,r) \cap \{x_1, \ldots, x_n\}| \geq k\} \Longrightarrow \left\{ r \geq C_5 \left(\frac{k}{n} \right)^{1/d} \right\}. \tag{A.75}$$

A.7.2 Properties of the Voronoi diagram

High probability control of cell geometry

The following lemma shows that with high probability, no Voronoi cell is very large. Let $r(V_i) := \max\{\|x - x_i\| : x \in V_i\}$ be the radius of the Voronoi cell V_i.

Lemma 15. *Suppose $x_1, \ldots, x_n$ are drawn from P satisfying Assumption A1. There exist constants C_1 and C_2 such that the following statement holds: for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\max_{i=1,\ldots,n} r(V_i) \leq C_1 \left(\frac{\log n + \log(1/\delta)}{n} \right)^{1/d}, \tag{A.76}$$

and

$$\max_{i=1,\ldots,n} \mu(V_i) \leq C_2 \left(\frac{\log n + \log(1/\delta)}{n} \right). \tag{A.77}$$

Proof. If $x \in V_i$, then $|B(x, \frac{1}{2}\|x - x_i\|) \cap \{x_1, \ldots, x_n\}| = 0$. (Note that the same holds true if $\frac{1}{2}$ is replaced with any $a \in [0, 1)$). Taking x to be such that $\|x - x_i\| = r(V_i)$, it follows by Lemma 14 that

$$\frac{1}{2} r(V_i) = \frac{1}{2}\|x - x_i\| \leq C\left(\frac{\log n + \log(1/\delta)}{n}\right)^{1/d},$$

with probability at least $1 - \delta$. Multiplying both sides by 2 and taking a maximum over $i = 1, \ldots, n$ gives (A.76). The upper bound (A.77) on the maximum Lebesgue measure of V_i follows immediately, since $V_i \subseteq B(x, r(V_i))$. □

Connectedness of the Voronoi adjacency graph

The following lemma relates graph theoretic connectedness to a kind of topological connectedness that excludes connectedness using sets of $\mathcal{H}^{d-1}$-measure zero.

Lemma 16. *Let $\Omega \subset \mathbb{R}^d$ be open such that there does* not *exist any set $S \subsetneq \Omega$ with $\mathcal{H}^{d-1}(S) = 0$ such that $\Omega \backslash S$ is disconnected. Let $\{V_1, \ldots, V_m\}$ denote an open polyhedral partition of Ω. Then the graph $G = (\{V_1, \ldots, V_m\}, E)$, where*

$$E = \{\{V_i, V_j\} : \mathcal{H}^{d-1}(\partial V_i \cap \partial V_j) > 0\},$$

is connected.

Proof. Assume by way of contradiction that G is disconnected. Therefore there exists sets of vertices $\mathcal{C}_1, \mathcal{C}_2$ such that

$$\mathcal{H}^{d-1}(\bar{V}_i \cap \bar{V}_j) = 0, \tag{A.78}$$

for all $V_i \in \mathcal{C}_1$, $V_j \in \mathcal{C}_2$. Next, define

$$\Omega_1 := (\cup_{V_i \in \mathcal{C}_1} \bar{V}_i)^\circ,$$
$$\Omega_2 := (\cup_{V_j \in \mathcal{C}_2} \bar{V}_j)^\circ,$$

such that $\{\Omega_1, \Omega_2\}$ constitutes an open partition of Ω. Let

$$\begin{aligned} S &:= \Omega \setminus (\Omega_1 \cup \Omega_2) && \text{(A.79)} \\ &= \Omega \cap ((\partial\Omega_1 \cap \partial\Omega_2) \cup ((\Omega_1^c)^\circ \cap (\Omega_2^c)^\circ)) \\ &= \Omega \cap ((\partial\Omega_1 \cap \partial\Omega_2) \cup (\Omega_2 \cap \Omega_1)) \\ &= \Omega \cap \partial\Omega_1 \cap \partial\Omega_2. && \text{(A.80)} \end{aligned}$$

From (A.78) and (A.80) we see that $\mathcal{H}^{d-1}(S) = 0$. On the other hand, (A.79) yields that $\Omega \setminus S = \Omega_1 \cup \Omega_2$ is disconnected (Ω_1, Ω_2 are open and disjoint). □

Analysis of the Voronoi kernel

Recall that in the proof of Theorem 1, we compare Voronoi TV to a U-statistic involving the kernel function

$$H_{\mathrm{Vor}}(x,y) = \mathbb{E}[\mathcal{H}^{d-1}(\partial V_{x_1} \cap \partial V_{x_2}) | x_1 = x, x_2 = y] = \int_{L\cap\Omega} (1 - p_x(z))^{(n-2)}\, dz.$$

The following lemma shows that this kernel function is close to a spherically symmetric kernel.

Lemma 17. *Suppose $x_1, \dots, x_n$ are sampled from distribution P satisfying A1. There exist constants C_1-$C_4 > 0$ such that for $h = h_n = C_1(3\log n/n)^{1/d}$, the following statements hold.*

- *For any $x, y \in \Omega_h$,*

$$\begin{aligned} H_{\mathrm{Vor}}(x,y) = {} & \frac{\eta_{d-2}}{(np(x))^{\frac{d-1}{d}}} K_{\mathrm{Vor}}\left(\frac{\|y-x\|}{\varepsilon_{(1)}}\right) \\ & + O\left(\frac{1}{n^3} + \frac{(\log n)^2}{n} 1\{\|x-y\| \le C_2(\log n/n)^{1/d}\}\right) \end{aligned} \tag{A.81}$$

- *For any $x, y \in \Omega$,*

$$H_{\mathrm{Vor}}(x,y) \le \frac{C_3}{n^{(d-1)/d}} K_{\mathrm{Vor}}\left(\frac{\|y-x\|}{C_4 n^{1/d}}\right). \tag{A.82}$$

Proof of (A.81). We now replace the integral above with one involving an exponential function that can be more easily evaluated. Then we evaluate this latter integral.

Step 1: Reduction to easier integral. Let $\Omega_x = \{z \in \Omega : \mathrm{dist}(z, \partial\Omega) > \|z - x\|\}$. (Note that $L \cap \Omega_x = L \cap \Omega_y$.) Separate the integral into two parts,

$$\begin{aligned} & \int_{L\cap\Omega} \Big(1 - p_x(z)\Big)^{(n-2)}\, dz \\ & \qquad = \int_{L\cap\Omega_x} \Big(1 - p_x(z)\Big)^{(n-2)}\, dz + \int_{L\cap(\Omega\setminus\Omega_x)} \Big(1 - p_x(z)\Big)^{(n-2)}\, dz. \end{aligned}$$

We start by showing that the second term above is negligible for $x, y \in \Omega_h$. For any $z \in \Omega \setminus \Omega_x$, it follows by the triangle inequality that

$$\mathrm{dist}(x, \partial\Omega) \le \|x - z\| + \mathrm{dist}(z, \Omega) \le 2\|x - z\|.$$

Since $x \in \Omega_h$, it follows that $p_x(z) \geq (p_{\min}/2d)\|z-x\|^d \geq (p_{\min}/2^{d+1}d)(\mathrm{dist}(x, \partial\Omega))^d \geq (p_{\min}/2^{d+1}d)h^d$. Integrating over $z \in \Omega \setminus \Omega_x$ implies an upper bound on the second term,

$$\begin{aligned}\int_{L\cap(\Omega\setminus\Omega_x)} \Big(1 - p_x(z)\Big)^{(n-2)}\, dz &\leq \int_{L\cap(\Omega\setminus\Omega_x)} \exp(-(n-2)p_x(z))\, dz \\ &= O(\exp(-(p_{\min}/2^{d+2}d)nh^d)) \\ &= O(\frac{1}{n^3}),\end{aligned}$$

with the last line following upon choosing $C_1 \geq (p_{\min}/2^{d+2}d)^{-1/d}$ in the definition of h.

On the other hand, if $z \in \Omega_x$ then $B(z, \|z-x\|) \subset \Omega$. Consequently, letting $\widetilde{p}_x(z) := p(x)\mu_d\|z-x\|^d$, it follows by the Lipschitz property of p that

$$|p_x(z) - \widetilde{p}_x(z)| \leq \int_{B(z,\|z-x\|)} |p(z) - p(x)|\, dz \leq C\mu_d\|z-x\|^{d+1},$$

and

$$|\exp\big(-np_x(z)\big) - \exp\big(-n\widetilde{p}_x(z)\big)| \leq C\mu_d\|z-x\|^{d+1}n.$$

Additionally recall that $\exp(-np) \geq (1-p)^n \geq \exp(-np)(1-np^2)$ for any $|p| < 1$. Combining these facts, we conclude that

$$\begin{aligned}&\int_{L\cap\Omega_x} \Big(1 - p_x(z)\Big)^n\, dz \\ &\quad= \int_{L\cap\Omega_x} \exp(-np_x(z))\big(1 + O(np_x(z)^2)\big)\, dz \\ &\quad= \int_{L\cap\Omega_x} \exp(-n\widetilde{p}_x(z))\Big(1 + O(n\|z-x\|^{2d}) + O(n\|z-x\|^{d+1})\Big)\, dz \\ &\quad\overset{(i)}{=} \int_{L\cap\Omega_x} \exp(-np(x)\mu_d\|x-z\|^d)\, dz \\ &\qquad\qquad + O\Big(\frac{1}{n^3} + \frac{1}{n}1\big\{\|x-y\| \leq C_2(\log n/n)^{1/d}\big\}\Big) \\ &\quad\overset{(ii)}{=} \int_L \exp\big(-np(x)\mu_d\|x-z\|^d\big)\, dz \\ &\qquad\qquad + O\Big(\frac{1}{n^3} + \frac{(\log n)^2}{n}1\big\{\|x-y\| \leq C_2(\log n/n)^{1/d}\big\}\Big).\end{aligned} \tag{A.83}$$

We prove the last two equalities, which control the remainder terms, after completing our analysis of the leading order term.

Step 2: Leading order term. Let $r = \|x-y\|/2$. Due to rotational symmetry, we may as well take $x = re_1, y = -re_1$, in which case the integral becomes

$$\begin{aligned}\int_L \exp\bigl(-np(x)\mu_d\|x-z\|^d\bigr)\,dz &= \int_{\{0\}\times\mathbb{R}^{d-1}} \exp\bigl(-np(x)\mu_d\|re_1-z\|^d\bigr)\,dz \\ &= \int_{\mathbb{R}^{d-1}} \exp\bigl(-np(x)\mu_d(r^2+\|z\|^2)^{d/2}\bigr)\,dz,\end{aligned}$$

with the latter equality following from the Pythagorean theorem. Converting to polar coordinates, we see that

$$\begin{aligned}&\int_{\mathbb{R}^{d-1}} \exp\bigl(-np(x)\mu_d(r^2+\|z\|^2)^{d/2}\bigr)\,dz \\ &\qquad= \int_0^\infty \int_{\mathbb{S}^{d-2}} \exp\bigl(-np(x)\mu_d(r^2+t^2)^{d/2}\bigr)t^{d-2}\,d\theta\,dt \\ &\qquad= \eta_{d-2}\int_0^\infty \exp\bigl(-np(x)\mu_d(r^2+t^2)^{d/2}\bigr)t^{d-2}\,dt \\ &\qquad= \frac{\eta_{d-2}}{(np(x))^{\frac{d-1}{d}}}\int_0^\infty \exp\Bigl(-\mu_d\bigl\{\bigl(r(np(x))^{1/d}\bigr)^2+s^2\bigr\}^{d/2}\Bigr)s^{d-2}\,ds, \\ &\qquad= \frac{\eta_{d-2}}{(np(x))^{\frac{d-1}{d}}}K_{\mathrm{Vor}}\biggl(\frac{\|y-x\|}{\varepsilon_{(1)}}\biggr),\end{aligned}$$

with the second to last equality following by substituting $s = t/(np(x))^{-1/d}$.

Controlling remainder terms. We complete the proof of (A.81) by establishing *(i)* and *(ii)* in (A.83).

Proof of *(i)*. Take $\varepsilon_0 = (4\log n/\mu_d p_{\min} n)^{1/d}$, and note that if $\|z-x\| \geq \varepsilon_0$ then $\exp\bigl(-\mu_d np(x)\|z-x\|^d\bigr) \leq \frac{1}{n^4}$. Recalling the definition of $\widetilde{p}_x(z)$, we have

$$\begin{aligned}&n\int_{L\cap\Omega_x} \exp\bigl(-n\widetilde{p}_x(z)\bigr)\|z-x\|^{d+1}\,dz \\ &= n\int_{L\cap\Omega_x} \exp\bigl(-\mu_d np(x)\|z-x\|^d\bigr)\|z-x\|^{d+1}\,dz \\ &\leq n\int_{L\cap B(x,\varepsilon_0)} \exp\bigl(-\mu_d np_{\min}\|z-x\|^d\bigr)\|z-x\|^{d+1}\,dz + \frac{\mathcal{H}^{d-1}(L\cap\Omega)}{n^3} \qquad \text{(A.84)} \\ &\leq n\varepsilon_0^{d+1}\int_{L\cap B(x,\varepsilon_0)} \exp\bigl(-\mu_d np_{\min}\|z-x\|^d\bigr)\,dz + \frac{\mathcal{H}^{d-1}(L\cap\Omega)}{n^3} \\ &\leq n\varepsilon_0^{d+1}\mathcal{H}^{d-1}(L\cap B(x,\varepsilon_0)) + \frac{\mathcal{H}^{d-1}(L\cap\Omega)}{n^3}.\end{aligned}$$

For any x, y we have $\mathcal{H}^{d-1}(L \cap B(x, \varepsilon_0)) \le \mu_{d-1}\varepsilon_0^{d-1}$. If additionally $\|x-y\|/2 > \varepsilon_0$ then $L \cap B(x, \varepsilon_0) = \emptyset$, and so $\mathcal{H}^{d-1}(L \cap B(x, \varepsilon_0)) = 0$. Compactly, these estimates can be written as

$$\mathcal{H}^{d-1}(L \cap B(x, \varepsilon_0)) \le \mu_{d-1} 1\{\|x-y\| \le 2\varepsilon_0\}\varepsilon_0^{d-1}.$$

Plugging this back into (A.84), we conclude that

$$\begin{aligned} n \int_{L \cap \Omega_x} \exp\bigl(-n\widetilde{p}_x(z)\bigr)\|z-x\|^{d+1}\,dz & \\ &\le n\varepsilon_0^{2d} 1\{\|x-y\| \le 2\varepsilon_0\} + \frac{\mathcal{H}^{d-1}(L \cap \Omega)}{n^3} \\ &\le C\Big(\frac{(\log n)^2}{n} 1\{\|x-y\| \le C_2(\log n/n)^{1/d}\} + \frac{1}{n^3}\Big), \end{aligned}$$

for $C_2 = 2(4/(p_{\min}\mu_d))^{1/d}$. This is precisely the claim.

Proof of *(ii)*. Recall the fact established previously, that if $z \in L \setminus \Omega_x$ then $\|z-x\| \ge h/2$. Therefore,

$$\begin{aligned} \int_{L \setminus \Omega_x} & \exp\bigl(-n\widetilde{p}_x(z)\bigr)\,dz \\ &\le \int_{L \setminus \Omega_x} \exp\bigl(-\mu_d n p_{\min}\|z-x\|^d\bigr)\,dz \\ &\le \int_{L \setminus B(x,2)} \exp\bigl(-\mu_d n p_{\min}\|z-x\|^d\bigr)\,dz \\ &\qquad + \int_{(L \cap B(x,2) \setminus \Omega_x)} \exp\bigl(-\mu_d p_{\min} n (h/2)^d\bigr)\,dz \\ &\le \int_{L \setminus B(x,2)} \exp\bigl(-\mu_d n p_{\min}\|z-x\|^d\bigr)\,dz + \frac{\mathcal{H}^{d-1}(L \cap B(x,2))}{n^3}, \end{aligned}$$

with the last inequality following upon choosing $C_1 \ge 2/(\mu_d p_{\min})^{1/d}$ in the definition of h. The remaining integral is exponentially small in n, proving the upper bound *(ii)*. □

Proof of (A.82). Note immediately that

$$H_{\mathrm{Vor}}(x,y) \le \int_{L \cap \Omega} \exp(-np_x(z))\,dz \le \int_L \exp(-n\mu_d p_{\min}\|x-z\|^d/2d)\,dz.$$

We have already analyzed this integral in the proof of (A.81), with the analysis implying that

$$\int_L \exp(-n\mu_d p_{\min}\|x-z\|^d/2d)\,dz = \frac{\eta_{d-2}(2d)^{\frac{d-1}{d}}}{(np_{\min})^{(d-1)/d}} K_{\mathrm{Vor}}\biggl(\frac{\|y-x\|}{(2dn/p_{\min})^{1/d}}\biggr).$$

This is exactly (A.82) with $C_3 = \eta_{d-2}(2d/p_{\min})^{(d-1)/d}$ and $C_4 = (2d/p_{\min})^{1/d}$. □

Compact kernel approximation

The kernel function $H_{\mathrm{Vor}}(x, y)$ is not compactly supported, and in our analysis it will frequently be convenient to approximate it by a compactly supported kernel. The following lemma does the trick. Let $\varepsilon_0 := (\log n/n)^{1/d}$.

Lemma 18. *Let $x, y \in \Omega$, and $L = \{z : \|x - z\| = \|y - x\|\}$. For any $a, c > 0$, there exists a constant $C > 0$ depending only on a, c and d such that*

$$\int_{L\cap\Omega} \exp(-cn\|x - z\|^d)\, dz \le C\left(\frac{1\{\|x - y\| \le C\varepsilon_0\}}{n^{(d-1)/d}} + \frac{1}{n^a}\right) \tag{A.85}$$

where $\varepsilon_0 := (\log n/n)^{1/d}$.

Proof. Let $\widetilde{\varepsilon}_0 = C_1\varepsilon_0$ for $C_1 = (a/c)^{1/d}$. The key is that if $\|x - z\| \ge \widetilde{\varepsilon}_0$, then

$$\exp(-cn\|x - z\|^d) \le \frac{1}{n^a}.$$

Now suppose $\|y - x\| > 2\widetilde{\varepsilon}_0$. Then $\|x - z\| \ge \widetilde{\varepsilon}_0$ for all $z \in L$, and

$$\int_{L\cap\Omega} \exp(-cn\|x - z\|^d)\, dz \le \frac{\mathcal{H}^{d-1}(L \cap \Omega)}{n^a}.$$

It follows that

$$\begin{aligned}
&\int_{L\cap\Omega} \exp(-cn\|x - z\|^d)\, dz \\
&\quad\le 1\{\|y - x\| \le 2\widetilde{\varepsilon}_0\} \int_{L\cap\Omega} \exp(-cn\|x - z\|^d)\, dz + \frac{\mathcal{H}^{d-1}(L \cap \Omega)}{n^a} \\
&\quad\le 1\{\|y - x\| \le 2\widetilde{\varepsilon}_0\} \int_{B_{d-1}((x+y)/2,\widetilde{\varepsilon}_0)} \exp(-cn\|x - z\|^d)\, dz + 2\frac{\mathcal{H}^{d-1}(L \cap \Omega)}{n^a} \\
&\quad\le \frac{1\{\|y - x\| \le 2\widetilde{\varepsilon}_0\}}{n^{(d-1)/d}} \int_{\mathbb{R}^{d-1}} \exp(-\|z\|^d)\, dz + 2\frac{\mathcal{H}^{d-1}(L \cap \Omega)}{n^a} \\
&\quad\le C_2\left(\frac{1\{\|y - x\| \le 2\widetilde{\varepsilon}_0\}}{n^{(d-1)/d}} + \frac{1}{n^a}\right).
\end{aligned}$$

for $C_2 = \max\{\int_{\mathbb{R}^{d-1}} \exp(-\|z\|^d)\, dz, 2\mathcal{H}^{d-1}(L \cap \Omega)\}$. Equation (A.85) follows upon taking $C = \max\{2C_1, C_2\}$. □

Appendix B

Supplement to Chapter 3

B.1 Proofs for Section 3.2

B.1.1 Proof of Proposition 3

Proof. For the matrix-valued $\phi : \mathbb{R}^d \to \mathbb{R}^{d_1 \times d_2}$, introduce the notation $\phi_k : \mathbb{R}^d_1 \to \mathbb{R}^d_1$ to denote a column of its value.

$$
\begin{aligned}
&\mathrm{TV}(g;\Omega, \|\cdot\|) \\
&\quad = \sup \left\{ \sum_{k=1}^{d_2} \int_\Omega f_k(x) \operatorname{div} \phi_k(x) dx : \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall x \right\} \\
&\quad = \sup \left\{ \sum_{k=1}^{d_2} \sum_{i=1}^{n} \int_{V_i} \gamma_{ik} \operatorname{div}(\phi_k(x) dx : \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall x \right\} \\
&\quad = \sup \Bigg\{ \sum_{k=1}^{d_2} \sum_{i=1}^{n} \gamma_{ik} \int_{\partial V_i} \langle \phi_k(t), n_i(t) \rangle d\mathcal{H}^{d-1}(t) : \\
&\qquad\qquad \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall x \Bigg\} \qquad \text{(B.1)} \\
&\quad = \sup \Bigg\{ \sum_{k=1}^{d_2} \sum_{i \sim j} \Bigg(\gamma_{ik} \int_{\partial V_i \cap \partial V_j} \langle \phi_k(t), n_i(t) \rangle d\mathcal{H}^{d-1}(t) \\
&\qquad\qquad + \gamma_{jk} \int_{\partial V_i \cap \partial V_j} \langle \phi_k(t), n_j(t) \rangle d\mathcal{H}^{d-1}(t) \Bigg) \\
&\qquad + \sum_{i : \bar{V}_i \cap \partial\Omega \neq \emptyset} \gamma_{ik} \underbrace{\int_{\partial V_i \cap \partial\Omega} \langle \phi_k(t), n_i(t) \rangle d\mathcal{H}^{d-1}(t)}_{=0;\ (\phi_k \text{ compactly supported})} : \qquad \text{(B.2)} \\
&\qquad\qquad \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall t \Bigg\} \\
&\quad = \sup \Bigg\{ \sum_{k=1}^{d_2} \sum_{i \sim j} \int_{\partial V_i \cap \partial V_j} \langle \phi_k(t), (\gamma_{ik} - \gamma_{jk}) n_i(t) \rangle d\mathcal{H}^{d-1}(t) : \\
&\qquad\qquad \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall t \Bigg\} \\
&\quad = \sup \Bigg\{ \sum_{i \sim j} \int_{\partial V_i \cap \partial V_j} \langle \phi(t), (\gamma_i - \gamma_j)^\top \otimes n_i(t) \rangle d\mathcal{H}^{d-1}(t) : \\
&\qquad\qquad \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \le 1 \ \forall t \Bigg\}
\end{aligned}
$$

where we obtain (B.1) by applying the Gauss-Green Theorem [Evans and Gariepy, 2015, Theorem 5.16]; and (B.2) by observing that when the boundaries of exactly two $V_i \neq V_j$ intersect, they have opposing outer normals, and when the boundaries of three or more $V_i \neq V_j \neq V_k \neq \cdots$ intersect, the outer normal vector is zero [Mikkelsen and Hansen, 2018, Lemma A.2(b)]. Apply Hölder's inequality to obtain an upper bound,

$$\begin{aligned}
&\mathrm{TV}(g;\Omega, \|\cdot\|) \\
&\quad \leq \sup \left\{ \sum_{i \sim j} \int_{\partial V_i \cap \partial V_j} \|\phi(t)\|_* \|(\gamma_i - \gamma_j)^\top \otimes n_i(t)\| d\mathcal{H}^{d-1}(t) : \right. \\
&\qquad\qquad \left. \phi \in C_c^1(\Omega; \mathbb{R}^{d_1 \times d_2}), \|\phi\|_* \leq 1 \ \forall t \right\} \\
&\quad = \sum_{i \sim j} \int_{\partial V_i \cap \partial V_j} \|(\gamma_i - \gamma_j)^\top \otimes n_i(t)\| d\mathcal{H}^{d-1}(t),
\end{aligned}$$

where recall $\|\cdot\|, \|\cdot\|_*$ are dual norms. Finally, we obtain a matching lower bound via a mollification argument. The target of our approximating sequence will be a pointwise duality map with respect to $\|\cdot\|$. Define the function $\phi_0 : \cup_{i \sim j} \partial V_i \cap \partial V_j \to \mathbb{R}^{d_1 \times d_2}$ by

$$\phi_0(t) \in \{g / \|g\|_* : g \in F(n_i(t)), t \in \partial V_i \cap V_j\},$$

and its piecewise constant extension to Ω, $\tilde{\phi} : \Omega \to \mathbb{R}^{d_1 \times d_2}$, by

$$\tilde{\phi}(x) = \phi_0 \left(t \in \operatorname*{argmin}_s \{ \|x - s\|_2 : s \in \cup_{i \sim j} \partial V_i \cap \partial V_j \} \right),$$

where recall for a Banach space E and its continuous dual $E*$, $F : E \to P(E^*)$ is the duality map defined by

$$F(x_0) = \{ \|f_0\|_{E^*} = \|x_0\|_E \text{and} \langle f_0, x_0 \rangle_{(E, E^*)} = \|x_0\|_E^2 \},$$

and moreover when E^* is strictly convex, the duality map is singleton-valued [Brezis, 2011]. Observe that $\tilde{\phi} \in L^p_{\mathrm{loc}}(\Omega)$, $1 \leq p < \infty$, and thus there exitss an approximating sequence $\{\tilde{\phi}_k\}$, $\tilde{\phi}_k \in C_c^\infty(\Omega; \mathbb{R}^{d_1 \times d_2})$ such that $\lim_k \tilde{\phi}_k \to \tilde{\phi}_k$ μ-a.e. We invoke Fatou's

lemma and properties of the duality map to obtain a matching lower bound,

$$
\begin{aligned}
&\mathrm{TV}(g;\Omega,\|\cdot\|) \\
&\quad= \sup\left\{ \sum_{i\sim j} \int_{\partial V_i\cap\partial V_j} \langle \phi(t), (\gamma_i-\gamma_j)^\top \otimes n_i(t)\rangle d\mathcal{H}^{d-1}(t) : \right. \\
&\qquad\qquad \left. \phi\in C_c^1(\Omega;\mathbb{R}^{d_1\times d_2}), \|\phi\|_*\le 1\ \forall t \right\} \\
&\quad\ge \sum_{i\sim j} \liminf_{k\to\infty} \int_{\partial V_i\cap\partial V_j} \langle \tilde\phi_k(t), (\gamma_i-\gamma_j)^\top \otimes n_i(t)\rangle d\mathcal{H}^{d-1}(t) \\
&\quad\ge \sum_{i\sim j} \int_{\partial V_i\cap\partial V_j} \langle \liminf_{k\to\infty} \tilde\phi_k(t), (\gamma_i-\gamma_j)^\top \otimes n_i(t)\rangle d\mathcal{H}^{d-1}(t) \\
&\quad= \sum_{i\sim j} \int_{\partial V_i\cap\partial V_j} \langle \tilde\phi(t), (\gamma_i-\gamma_j)^\top \otimes n_i(t)\rangle d\mathcal{H}^{d-1}(t) \\
&\quad= \sum_{i\sim j} \int_{\partial V_i\cap\partial V_j} \|(\gamma_i-\gamma_j)^\top \otimes n_i(t)\| d\mathcal{H}^{d-1}(t).
\end{aligned}
$$

This gives (3.5). Subsequently, (3.6) is obtained by observing that $\|x^\top \otimes y\|_{q,p} = \|x\|_p\|y\|_q$ and recalling that $n_i(t)$ has unit length in the Euclidean norm. □

B.1.2 Proof of Proposition 5

The following proof refers to a technical lemma given at the end of this subsection.

Proof. We begin with the discrete penalty

$$\sum_{\{i,j\}\in E^{\mathcal{DT}}} w_{ij}^{\mathcal{DT}} \cdot \|G_{s_i}\theta - G_{s_j}\theta\|_2,$$

which is known to be equivalent to the continuous-time penalty of (3.3) for $f\in\mathcal{F}_n^{\mathcal{DT}}$ by the gradient variation representation of that function class. Each summand of the above display equation is the ℓ_2-norm of a gradient difference.

- First, the gradient difference is computable as the difference of linear transformations of the parameter vector θ.
- Second, the ℓ_2-norm of the gradient difference is computable by taking the absolute value of an inner product between the gradient difference and a fixed vector which depends only on $\mathcal{DT}$ and not on θ; this is due to Lemma 19.

With these two facts in hand, it immediately follows that the discrete penalty may be written as the ℓ_1-norm of a linear transformation D of the parameter vector θ. □

This technical lemma shows that for a piecewise linear function on a triangulation (i.e., a partition of $\Omega \subset \mathbb{R}^d$ where each element is a d-simplex), the difference in gradients between two adjoining triangles is a scalar multiple of the normal vector to the face joining those triangles.

Lemma 19. *Let $f \in \mathcal{F}_n^{\mathcal{DT}}$. Let s_i and s_j share a face, i.e., $\mathcal{H}^{d-1}(\partial s_i \cap \partial s_j) > 0$, and $g_i := g_i(\theta) = \nabla f|_{s_i}$. Then*

$$\|g_i - g_j\|_2 = |\langle g_i - g_j, \eta_{ij} \rangle| \tag{B.3}$$

where η_{ij} is such that $\langle \eta_{ij}, u - v \rangle = 0$, $u, v \in \partial s_i \cap \partial s_j, u \neq v$, $\|\eta_{ij}\|_2 = 1$, i.e., the unit-length normal vector to the face joining s_i, s_j.

Proof. For any matrix A, we denote by A_i the ith row of A and by $A_{i:j}$ the submatrix induced by taking the ith through jth rows of A (endpoints inclusive). Similarly, for any vector θ, θ_i denotes the ith entry of θ and $\theta_{i:j}$ the ith through jth entries. We denote by $X_{1:d} \in \mathbb{R}^{d \times d}$ the matrix which takes the d shared points as its rows.

The key insight in this proof is that the continuity of f constrains the gradients in two neighboring simplices to lie in a one-dimensional affine space. This affine space is parameterized by $\gamma + c\beta$, γ, β fixed in $\mathbb{R}^{d+1}$ and c varying over $\mathbb{R}$. The gradient proper occupies the latter d components of $\gamma + c\beta$, while we call the full $(d+1)$-dimensional vector (which includes an "offset" term in the first component) the "extended gradient".

It turns out that $\beta_{2:(d+1)}$ is equal to η, the vector normal to the hyperplane coinciding with the face shared by the two neighboring simplices. Having shown this, it follows that the gradient difference takes the form $(c - c')\eta$.

First, we show that extended gradient for two neighboring simplices takes the form $\gamma + c\beta$, $\gamma, \beta \in \mathbb{R}^{d+1}$, $c \in \mathbb{R}$. Write the two linear systems that define the gradient for each of the two neighboring simplices, calling them s, s'. Recall that because they are neighboring, these simplices share d vertices and hence (by continuity) match in value

on these d vertices. They only differ in location and value on the final, $(d+1)$st vertex.

$$\underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{d1} & \cdots & x_{dd} \\ 1 & x_{(d+1)1} & \cdots & x_{(d+1)d} \end{bmatrix}}_{:=A} \begin{bmatrix} b \\ g_1 \\ \vdots \\ g_d \end{bmatrix} = \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \\ \theta_{d+1} \end{bmatrix}}_{:=\theta}$$

$$\underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{d1} & \cdots & x_{dd} \\ 1 & x'_{(d+1)1} & \cdots & x'_{(d+1)d} \end{bmatrix}}_{:=A'} \begin{bmatrix} b' \\ g'_1 \\ \vdots \\ g'_d \end{bmatrix} = \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \\ \theta'_{d+1} \end{bmatrix}}_{:=\theta'}$$

Recall that $A_{1:d} = A'_{1:d}$ and $\theta_{1:d} = \theta'_{1:d}$ by definition. We also observe that $A_{1:d}$ must be rank deficient and, assuming it has rank d, possesses a one-dimensional nullspace. Hence, the linear system consisting of the first d rows has an infinite number of solutions lying in an affine space of dimension one, which we may write as $\{\gamma + c\beta : c \in \mathbb{R}\}$.

The inclusion of A_{d+1} (or A'_{d+1}), a $(d+1)$st row linearly independent to the first d, provides the necessary constraint make the solution unique. Suppose we solve the subsystem given by the first d rows by taking the minimum norm solution, calling this $\gamma \in \mathbb{R}^{d+1}$. Further let $\beta \in \mathbb{R}^{d+1}$ denote the unit vector spanning the nullspace of $A_{1:d}$. By definition, $\langle \beta, A_i \rangle = 0$ for $i = 1, \ldots, d$. Therefore, for the problems with $d+1$ constraints, we may directly obtain "updated" solutions $\tilde{\gamma} = \gamma + \alpha\beta$, $\tilde{\gamma}' = \gamma + \alpha'\beta$, where we define

$$\alpha := \frac{\theta_{d+1} - \langle A_{d+1}, \gamma \rangle}{\langle A_{d+1}, \beta \rangle}$$
$$\alpha' := \frac{\theta'_{d+1} - \langle A'_{d+1}, \gamma \rangle}{\langle A_{d+1}, \beta \rangle}$$

which satisfy the first d constraints (because we have only added a perturbation lying along their nullspace) and also satisfies the $(d+1)$st constraint (by construction).

It remains for us to show that the normal vector $\eta \in \operatorname{span}(\beta_{2:(d+1)})$, β spanning the nullspace of $A_{1:d}$. Recall that $\eta \perp H(X_{1:d})$, the hyperplane spanned by the d points shared between the two neighboring simplices. Every element in $H(X_{1:d})$ may be written as $c\eta + \phi$, with $\phi \in \perp \{\eta\}$ and $c = \min_{c'}\{c'\eta \in H(X_{1:d})\}$, and it follows that for any element $x' \in H(X_{1:d})$, $\langle x', \eta \rangle = c$.

Because $\eta \in \mathbb{R}^d$, it is determined in d constraints as such,

$$X_{1:d}\eta = c1 \Leftrightarrow X_{1:d}\eta - c1 = 0,$$

and we observe that up to scaling, this problem is equivalent to determining the nullspace of

$$\begin{bmatrix} 1 & X_{1:d} \end{bmatrix} \begin{bmatrix} -c \\ \eta \end{bmatrix} = 0.$$

Finally, take $\beta = \begin{bmatrix} -c \\ \eta \end{bmatrix}$. Hence, $\eta \in \operatorname{span}(\beta_{2:(d+1)})$. □

B.1.3 Proof of Proposition 6

Proof. We investigate the nullity of $D_{\mathcal{A}^c}$ via its equivalent characterization as the dimension of the solution set to the system of linear equations

$$D_{\mathcal{A}^c}\theta = 0. \tag{B.4}$$

Block structure. First, we group the $|\mathcal{A}^c|$ equations in (B.4) into blocks based by the connected components, giving the system

$$\begin{cases} D_{\mathcal{B}_1}\theta & = 0 \\ D_{\mathcal{B}_2}\theta & = 0 \\ & \vdots \\ D_{\mathcal{B}_{\tilde{m}}}\theta & = 0 \end{cases} \tag{B.5}$$

where $\mathcal{B}_j$ contains index i if the inactive facet $i \in \mathcal{A}^c$ belongs to the jth connected component (it is easy to check that each inactive facet belongs to only one connected component).

Change of variables. We now introduce a change of variables for (B.5). Letting $n_j := |x_{1:n} \cap \bar{\mathcal{C}}_j|$, $j = 1, \ldots, \tilde{m}$, and $\tilde{n} := \sum_{j=1}^{\tilde{m}} n_j$, introduce the block vector $\beta \in \mathbb{R}^{\tilde{n}}$, where the jth block is indexed β_{ij} for $i : x_i \in \bar{\mathcal{C}}_j$. β serves as a re-indexed form of θ, where θ_i is duplicated into multiple blocks if x_i lies in multiple connected components (this occurs when x_i lies on the boundary between connected components). Introduce $\tilde{D}_{\mathcal{B}_j} \in \mathbb{R}^{|\mathcal{B}_j| \times \tilde{n}}$ as a version of $D_{\mathcal{B}_j}$ subject to analogous re-indexing and duplication along its columns, but with nonzero entries only in columns $1 + \sum_{j'<j} n_{j'}$ through $\sum_{j' \le j} n_{j'}$, i.e., corresponding to the jth block of β.

Consider now the system of equations

$$\begin{cases} \tilde{D}_{\mathcal{B}_1}\beta & = 0 \\ \tilde{D}_{\mathcal{B}_2}\beta & = 0 \\ & \vdots \\ \tilde{D}_{\mathcal{B}_{\tilde{m}}}\beta & = 0 \\ \beta_{ij_i} - \beta_{i\ell} & = 0 \qquad x_i \in \partial\mathcal{C}_\ell, \ell > j_i \end{cases} \tag{B.6}$$

where $j_i = \min\{j : x_i \in \bar{\mathcal{C}}_j\}$. It is clear that the systems (B.4) and (B.6) have the same solution set after identifying θ_i with β_{ij_i}. Therefore we may investigate the solution set dimension of the latter to determine the nullity of $D_{\mathcal{A}^c}$.

Piecewise affine structure conditions. Consider the jth block of the linear system (B.6)

$$\tilde{D}_{\mathcal{B}_j}\beta = 0. \tag{B.7}$$

Observing that (B.7) is a homogeneous system of linear equations and that $\tilde{D}_{\mathcal{B}_j}$ has nonzero entries only in columns $1 + \sum_{j'<j} n_{j'}$ through $\sum_{j'\leq j} n_{j'}$, we conclude that its solution is of the form

$$\Big(\prod_{j'<j} \mathbb{R}^{n_{j'}}\Big) \times S_j \times \Big(\prod_{j'>j} \mathbb{R}^{n_{j'}}\Big),$$

where S_j is a subspace of $\mathbb{R}^{n_j}$. In particular, recalling that $\tilde{D}_{\mathcal{B}_j}\beta = 0$ constrains $(x_i, \beta_{ij}), i : x_i \in \bar{\mathcal{C}}_j$ to lie on a hyperplane in dimension $\mathbb{R}^{d+1}$, we conclude that S_j is a subspace of dimension $d+1$. Applying this argument to blocks $j = 1, \ldots, \tilde{m}$ and intersecting their solution sets, we find that the solution set to the system of equations

$$\begin{cases} \tilde{D}_{\mathcal{B}_1}\beta & = 0 \\ \tilde{D}_{\mathcal{B}_2}\beta & = 0 \\ & \vdots \\ \tilde{D}_{\mathcal{B}_{\tilde{m}}}\beta & = 0 \end{cases}$$

is

$$S_1 \times S_2 \times \cdots \times S_{\tilde{m}},$$

the outer product of $\tilde{m}$ subspaces, each having dimension $d+1$, i.e., the solution set has dimension $\tilde{m}(d+1)$. This solution set corresponds the functions which are piecewise affine on $\mathcal{T}(x_{1:n}; y_{1:n}, \lambda)$ *without* satisfying continuity across the pieces.

Continuity conditions. We now consider the full system (B.6), which includes the continuity conditions. The solution set to each constraint

$$\beta_{ij_i} - \beta_{i\ell} = 0$$

is a $(\tilde{n}-1)$-dimensional subspace of $\mathbb{R}^{\tilde{n}}$. In order to show the claim that (B.6) has a solution set of dimension

$$\tilde{m}(d+1) - \sum_{i=1}^{n} \Big(\sum_{j=1}^{\tilde{m}} 1\{x_i \in \bar{\mathcal{C}}_j\} - 1 \Big),$$

we must demonstrate two conditions:

(a) there are $\sum_{i=1}^{n}(\sum_{j=1}^{\tilde{m}} 1\{x_i \in \bar{\mathcal{C}}_j\} - 1)$ continuity conditions, and

(b) each of the equations $\beta_{ij_i} - \beta_{i\ell} = 0$ is linearly independent of all other equations in (B.6), i.e., that this constraint is not implied by any of the other constraints in the system.

Condition (a) is easily checked by examining the construction of β. To check condition (b), we first interpret linear dependence of $\beta_{ij_i} - \beta_{i\ell} = 0$ on the other equations in (B.6) as stating that the other equations in (B.6) require

$$f_{\bar{\mathcal{C}}_{j_i}}(x_i) = f_{\bar{\mathcal{C}}_\ell}(x_i),$$

where $f_{\bar{\mathcal{C}}_j}$ is the affine function on $\mathcal{C}_j$ determined by (B.7), extended to take its limiting values on $\partial\mathcal{C}_j$. Due to the affine nature of $f_{\bar{\mathcal{C}}_{j_i}}, f_{\bar{\mathcal{C}}_\ell}$, this "continuity" of f between $\mathcal{C}_{j_i}$ and $\mathcal{C}_\ell$ at x_i can only happen

(i) at x_i and nowhere else along $\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell$, or

(ii) along all of $\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell$.

Continuity only at x_i. For (i) to be implied by the other equations in (B.6), $\beta_{ij_i} - \beta_{i\ell} = 0$ must be implied by the other continuity conditions at the same vertex, i.e., by the constraints $\beta_{ij_i} - \beta_{ik} = 0$, $k \neq \ell$. This is not possible by construction of the constraints (which enforce equality across $\beta_{i\cdot}$ through a hub-and-spoke model).

Continuity along all of $\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell$. For (ii) to be implied by the other equations in (B.6), we must have $|\{\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell \cap x_{1:n}\}| \geq d+1$, since (ii) requires that d vertices (other than x_i) lie on $\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell$, enforcing continuity on all of $\partial\mathcal{C}_{j_i} \cap \partial\mathcal{C}_\ell$. This possibility is ruled out by Proposition 4. □

B.2 Proofs for Section 3.3

B.2.1 Proof of Theorem 5

The subsequent subsections incrementally develop the result (3.25), with proofs of technical lemmas deferred to the Appendix.

Basic inequality & analysis overview

Abbreviate $J(\cdot) := \mathrm{DGV}(\cdot; \mathcal{T}, w)$. We begin by deriving a basic inequality,

$$\begin{aligned} &\|y_{1:n} - \hat{f}^{\mathcal{T}}(x_{1:n})\|_2^2 + \lambda J(\hat{f}^{\mathcal{T}}) \leq \|y_{1:n} - f(x_{1:n})\|_2^2 + \lambda J(f) \\ &\qquad \Rightarrow \|(f - \hat{f}^{\mathcal{T}})(x_{1:n})\|_2^2 \leq 2\langle z_{1:n}, (\hat{f} - f)(x_{1:n})\rangle_2 + \lambda(J(f) - J(\hat{f})), \end{aligned}$$

for any $f \in \mathcal{F}_n^{\mathcal{T}}$, and in particular

$$\|(f_0 - \hat{f}^{\mathcal{T}})(x_{1:n})\|_2^2 \leq 2\langle z_{1:n}, (\hat{f} - f_0)(x_{1:n})\rangle_2 + \lambda(J(f) - J(\hat{f})). \tag{B.8}$$

Our principal concern now is to control the quantity

$$\langle z_{1:n}, f(x_{1:n})\rangle = \sum_{i=1}^{n} f(x_i) z_i, \tag{B.9}$$

$f \in \mathcal{F}_n^{\mathcal{T}}$, in terms of $J(f)$.

Elements of the analysis. The analysis of (B.9) involves the following elements:

- The function f, which is CPWL on a triangulation, will be compared to its "projection" $\Pi_\Gamma f$ onto the set of function which are piecewise linear an a lattice-based partition of Ω. This incurs a truncation error measuring the distance of $\Pi_\Gamma f$ to f, which is shown to be controlled by the discrete gradient variation of f.
- The comparison of f to $\Pi_\Gamma f$ reduces the analysis of (B.9) into the Gaussian complexity of the second-order discrete-difference operator on a lattice. This complexity class (and its higher-order generalizations) has been rigorously studied by Sadhanala et al. [2021].
- We find that the complexity of $\Pi_\Gamma f$, measured in terms of the lattice-based second-difference operator, scales with the discrete gradient variation of f, measured using second-order discrete-differences on the triangulation.
- This analysis allows us to use the graph trend filtering framework of Wang et al. [2016] and the spectral analysis of lattice-based discrete-difference operators of Sadhanala et al. [2021] to control the penalized triogram complexity class. The comparison of penalty operators, and in particular the comparison to a lattice-based penalty operators, is inspired by the surrogate operator analysis of Padilla et al. [2018, 2020].

Lattice-based discretization

Consider a grid-based partition of $\Omega = (0,1)^d$ with the following properties.

- The grid Γ will be instantiated with N^d cells γ_i, where $i \in [N]^d$. $\Gamma = \{\gamma_i : i \in [N]^d\}$ is an open partition of Ω.
- Assume that N is chosen such that $\gamma \cap \bar{\mathcal{T}} \neq \emptyset$, $\gamma \in \Gamma$.
- The sidelength of each γ is $1/N$.
- Define the adjacency graph $G_\Gamma = (\Gamma, E_\Gamma)$, where $\gamma_i, \gamma_j \in E_\Gamma$ iff $\|i-j\|_1 = 1$. Coincidentally, this condition is equivalent to $\mathcal{H}^{d-1}(\partial\gamma_i \cap \partial\gamma_j) > 0$.
- Define a knot t_γ associated with each cell. t_γ is the "lower-left" corner of the cell γ; i.e., for γ_i, $t_{\Gamma_i} := (i-1)/N$.
- For a grid partition Γ of Ω, we introduce a space of piecewise linear functions on Γ,

$$\mathcal{F}_\Gamma = \{f(x) := \Big\{ \sum_{\gamma\in\Gamma} 1\{x \in \gamma\} \cdot \big(\alpha_\gamma + \beta_\gamma(x - t_\gamma)\big) : \alpha_\gamma \in \mathbb{R}, \beta_\gamma \in \mathbb{R}^d \Big\}. \quad \text{(B.10)}$$

  Note carefully that unlike $\mathcal{F}_n^{\mathcal{T}}$, the functions in $\mathcal{F}_\Gamma$ are *not* required to satisfy continuity.

- We also define a "projection" operation from $\Pi_\Gamma : \mathcal{F}_n^{\mathcal{T}} \to \mathcal{F}_\Gamma$,

$$\Pi_\Gamma f(x) = \sum_{\gamma\in\Gamma} 1\{x \in \gamma\} \cdot \big(f(t_\gamma) + \nabla f(t_\gamma)^\top (x - t_\gamma)\big). \quad \text{(B.11)}$$

  That is, the $\alpha_\gamma, \beta_\gamma$ associated with $\Pi_\Gamma f$ are the function and gradient evaluations at t_γ. If $t_\gamma \notin \bar{\mathcal{T}}$, take the linear function from any simplex s with nonempty intersection with γ (this is guaranteed by assumption), extend it to all of γ, and define α_γ, β_γ accordingly.

Remainder term

First, we split the Gaussian process term (B.9) into a "main term" and a "remainder term", corresponding to different subspaces of $\mathbb{R}^{|\Gamma|}$. The main term will be controlled via certain spectral functionals of the lattice-based difference operators. The remainder term will correspond to a parametric rate in the subcritical regime, and in the supercritical regime the size of the remainder subspace will be carefully tuned to obtain the desired rate.

Begin by rewriting the Gaussian process term

$$\langle f(x_{1:n}), z_{1:n} \rangle = \langle \tilde{f}, \epsilon \rangle,$$

where $\tilde{f} \in \mathbb{R}^{|\Gamma|}$ has elements

$$\tilde{f}_\gamma := \Big(\sum_{x_i \in \gamma} \big(f(x_i)\big)^2 \Big)^{1/2}$$

and $\epsilon \in \mathbb{R}^{|\Gamma|}$ has independent elements $\epsilon_\gamma \sim N(0, \sigma^2)$. For an index set $\mathcal{S} \subset [N]^d$, let $P_\mathcal{S}$ denote the projector onto some dimension-$|\mathcal{S}|$ subspace of $\mathbb{R}^{|\Gamma|}$ corresponding to $\mathcal{S}$. With this notation in hand, decompose the Gaussian process term into a remainder term and a main term,

$$\begin{aligned} \langle f(x_{1:n}), z_{1:n} \rangle &\stackrel{d}{=} \langle \tilde{f}, \epsilon \rangle \\ &= \langle \tilde{f}, (I - P_\mathcal{S})\epsilon \rangle + \langle \tilde{f}, P_\mathcal{S}\epsilon \rangle \\ &\leq \|f(x_{1:n})\|_2 \|(I - P_\mathcal{S})\epsilon\|_2 + \langle \tilde{f}, P_\mathcal{S}\epsilon \rangle. \end{aligned}$$

After substitution back into the basic inequality and solving a quadratic inequality in $\|f(x_{1:n})\|_2$, the remainder term will be $O_\mathbb{P}(1 - |\mathcal{S}|/n)$ in average error. We now turn our attention to the main term $\langle \tilde{f}, P_\mathcal{S}\epsilon \rangle$ and will return to the remainder term at the end.

Approximation error

We now focus on the main term, $\langle \tilde{f}, P_\mathcal{S}\epsilon \rangle$. Because $P_\mathcal{S}$ is a projector, each entry

$$\tilde{\epsilon}_\gamma =: (P_\mathcal{S}\epsilon)_\gamma \sim N(0, \nu_\gamma^2),$$

where $\nu_\gamma^2 \leq \sigma^2$. Note importantly that $\tilde{\epsilon}_\gamma, \epsilon_{\gamma'}$ are not necessarily independent, but this will not pose a problem in our analysis. The pointwise products are distributed

$$\begin{aligned} \tilde{f}_\gamma \cdot \tilde{\epsilon}_\gamma &= \Big(\sum_{x_i \in \gamma} \big(f(x_i)\big)^2 \Big)^{1/2} \tilde{\epsilon}_\gamma \\ &\stackrel{d}{=} N\left(0, \nu_\gamma^2 \sum_{x_i \in \gamma} \big(f(x_i)\big)^2\right) \\ &\stackrel{d}{=} \sum_{x_i \in \gamma} \big(f(x_i)\big) \tilde{\epsilon}_{\gamma,i}, \end{aligned}$$

where $\tilde{\epsilon}_{\gamma,i} \sim N(0, \nu_\gamma^2)$ with independence across i. We may then decompose and rewrite

$$\begin{aligned}
\langle \tilde{f}, P_{\mathcal{S}}\epsilon \rangle &\overset{d}{=} \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \big(f(x_i)\big)\tilde{\epsilon}_{\gamma,i}, \\
&= \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \big(f(x_i) - \Pi_\Gamma f(x_i) + \Pi_\Gamma f(x_i)\big)\tilde{\epsilon}_{\gamma,i}, \\
&= \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \big(f(x_i) - \Pi_\Gamma f(x_i)\big)\tilde{\epsilon}_{\gamma,i} + \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i} \\
&\leq \|(f - \Pi_\Gamma f)(x_{1:n})\|_1 \|z_{1:n}\|_\infty + \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i}
\end{aligned}$$

to obtain an approximation term and a lattice-based Gaussian complexity term. The approximation term is controlled using the following result.

Lemma 20. *The approximation error between a CPWL $f \in \mathcal{F}_n^{\mathcal{T}}$ and its lattice-based approximation $\Pi_\Gamma f$ may be bound above in terms of its gradient variation,*

$$\|(f - \Pi_\Gamma f)(x_{1:n})\|_1 \leq \frac{\sqrt{d} n_{\mathcal{T}}(\Gamma) n_\Gamma(\mathcal{T}) n_x(\Gamma)}{N w_{\min}} \cdot \mathrm{DGV}(f; \mathcal{T}, w), \tag{B.12}$$

where $w_{\min} := \min_{(s_i,s_j)\in E_{\mathcal{T}}} w_{ij}$.

The proof is deferred to Appendix B.2.2.

Lattice-based Gaussian process

We now study the lattice-based Gaussian process term.

Equivalent representation. Begin by representing $\Pi_\Gamma f$ explicitly in terms of the piecewise intercepts and slopes, i.e.,

$$\sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i} = \sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma} \big(\alpha_\gamma + \beta_\gamma^\top (x_i - t_\gamma)\big)\tilde{\epsilon}_{\gamma,i}.$$

We now consider the Gaussian processes involving intercepts and the slopes separately and derive representations of these Gaussian processes that are equivalent in distribution. For the intercepts, we observe directly that due to independence in i, we have

$$\sum_{x_i\in\gamma} \alpha_\gamma \tilde{\epsilon}_{\gamma,i} \overset{d}{=} \alpha_\gamma \cdot \sqrt{n_x(\gamma)}\tilde{\epsilon}_\gamma.$$

For the slopes, we decompose by dimension,

$$\begin{aligned}\sum_{x_i\in\gamma}\beta_\gamma^\top(t_\gamma - x_i)\tilde{\epsilon}_{\gamma,i} &= \sum_{x_i\in\gamma}\sum_{j=1}^{d}\beta_{\gamma,j}(t_\gamma - x_i)_j\tilde{\epsilon}_{\gamma,i}\\ &= \sum_{j=1}^{d}\beta_{\gamma,j}\sum_{x_i\in\gamma}(t_\gamma - x_i)_j\tilde{\epsilon}_{\gamma,i}\\ &\overset{d}{=} \sum_{j=1}^{d}\beta_{\gamma,j}N\Big(0,\nu_\gamma^2\sum_{x_i\in\gamma}(t_\gamma - x_i)_j^2\Big)\\ &\overset{d}{=} \sum_{j=1}^{d}\beta_{\gamma,j}\cdot\Big(\sum_{x_i\in\gamma}(t_\gamma - x_i)_j^2\Big)^{1/2}\tilde{\epsilon}_\gamma.\end{aligned}$$

Conclude that

$$\sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma}\Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i} = \langle\alpha, SP_{\mathcal{S}}\epsilon\rangle + \sum_{j=1}^{d}\langle\beta_j, S_jP_{\mathcal{S}}\epsilon\rangle, \tag{B.13}$$

where $\alpha,\beta_j\in\mathbb{R}^{|\Gamma|}$, $j=1,\dots,d$ collect the intercepts and directional derivatives into vectors, and S, S_j are diagonal scaling matrices with entries

$$\begin{aligned}S_\gamma &= \sqrt{n_x(\gamma)},\\ (S_j)_\gamma &= \Big(\sum_{x_i\in\gamma}(t_\gamma - x_i)_j^2\Big)^{1/2}.\end{aligned}$$

Upper bound in terms of lattice-based complexity. Following the notation of Lemma 23, for each m, define the index set

$$\mathcal{S}_r^{(m)} := \{i\in[N]^d : \|(i-m-1)\|_2\geq r\},$$

for some $r\in[1,\sqrt{d}N]$, and set

$$\mathcal{S} = \mathcal{S}_r^{(2)}.$$

Taking the singular value decomposition $D_\Gamma^{(2)} = U\Xi V^\top$ with the singular vectors and values indexed on the lattice $[N]^d$, note that

$$\operatorname{span}(\{V_i : i\in\mathcal{S}_r^{(2)}\})\subset\operatorname{span}(\{V_i : i\in\mathcal{S}_1^{(2)}\}) = R^{(2)}\subset R^{(1)}.$$

We now treat the intercepts and slopes separately. For the intercepts, follow the analysis of Wang et al. [2016] and Sadhanala et al. [2021] to upper bound,

$$\begin{aligned}\langle \alpha, SP_{\mathcal{S}_r^{(2)}}\epsilon\rangle &= \epsilon^\top P_{\mathcal{S}_r^{(2)}} S P_{R^{(2)}}\alpha \\ &= \epsilon^\top P_{\mathcal{S}_r^{(2)}} S (D_\Gamma^{(2)})^+ D_\Gamma^{(2)}\alpha \\ &\le \|((D_\Gamma^{(2)})^+)^\top S P_{\mathcal{S}_r^{(2)}}\epsilon\|_\infty \|D_\Gamma^{(2)}\alpha\|_1.\end{aligned}$$

For the slopes, we approach coordinate-wise to obtain a similar upper bound,

$$\begin{aligned}\langle \beta_j, S_j P_{\mathcal{S}_r^{(2)}}\epsilon\rangle &= \epsilon^\top P_{\mathcal{S}_r^{(2)}} S_j P_{R^{(1)}}\beta_j \\ &= \epsilon^\top P_{\mathcal{S}_r^{(2)}} S_j (D_\Gamma^{(1)})^+ D_\Gamma^{(1)}\beta_j \\ &\le \|((D_\Gamma^{(1)})^+)^\top S_j P_{\mathcal{S}_r^{(2)}}\epsilon\|_\infty \|D_\Gamma^{(1)}\beta_j\|_1 \\ &\le \|((D_\Gamma^{(1)})^+)^\top S_j P_{\mathcal{S}_1^{(1)}}\epsilon\|_\infty \|D_\Gamma^{(1)}\beta_j\|_1,\end{aligned}$$

noting in the final line that $\mathcal{S}_r^{(2)} \subset \mathcal{S}_1^{(1)}$.

Control via spectral functionals. Apply Lemma 24 to upper bound

$$\begin{aligned}\|((D_\Gamma^{(2)})^+)^\top S P_{\mathcal{S}_r^{(2)}}\epsilon\|_\infty &\le \sigma \frac{\mu_2 \|S\|_2}{\sqrt{|\Gamma|}} \sqrt{2(\log|\Gamma| + \delta)\sum_{i\in\mathcal{S}_r^{(2)}} \frac{1}{\left(\xi_i^{(2)}\right)^2}} \\ &\le \sigma \frac{\mu_2 \sqrt{n_x(\Gamma)}}{\sqrt{|\Gamma|}} \sqrt{2(\log|\Gamma| + \delta)\sum_{i\in\mathcal{S}_r^{(2)}} \frac{1}{\left(\xi_i^{(2)}\right)^2}}\end{aligned}$$

with probability at least $1 - 2\exp(-\delta)$, where $\xi_i^{(m)}$ are the singular values of $D_\Gamma^{(m)}$. Similarly, for each $j \in [d]$, we may upper bound

$$\begin{aligned}\|((D_\Gamma^{(1)})^+)^\top S_j P_{\mathcal{S}^{(1)}}\epsilon\|_\infty &\le \sigma \frac{\mu_1 \max_{\gamma\in\Gamma}\sqrt{\sum_{x_i\in\gamma}(t_\gamma - x_i)_j^2}}{\sqrt{|\Gamma|}} \sqrt{2(\log|\Gamma| + \delta)\sum_{i\in\mathcal{S}^{(1)}} \frac{1}{\left(\xi_i^{(1)}\right)^2}} \\ &\le \sigma \frac{\mu_1 \sqrt{n_x(\Gamma)}}{N\sqrt{|\Gamma|}} \sqrt{2(\log|\Gamma| + \delta)\sum_{i\in\mathcal{S}^{(1)}} \frac{1}{\left(\xi_i^{(1)}\right)^2}}\end{aligned}$$

with probability at least $1 - 2\exp(-\delta)$. We conclude that (B.13) is bound above by

$$\sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma}\Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i}$$
$$\leq \sigma\mu\sqrt{n_x(\Gamma)}\sqrt{2(\log|\Gamma|+\delta)}$$
$$\times\left(\sqrt{|\Gamma|^{-1}\sum_{i\in\mathcal{S}_r^{(2)}}\frac{1}{\left(\xi_i^{(2)}\right)^2}}\cdot\|D_\Gamma^{(2)}\alpha\|_1+\sqrt{|\Gamma|^{-1}\sum_{i\in\mathcal{S}^{(1)}}\frac{1}{\left(\xi_i^{(1)}\right)^2}}\cdot\sum_{j=1}^{d}\|D_\Gamma^{(1)}\beta_j\|_1\right).$$

Lattice-based constraint sets

The following two results show that when $\Pi_\Gamma f$ is obtained from a function $f \in \mathcal{F}_n^{\mathcal{T}}$ satisfying bounded discrete gradient variation, the intercepts of $\Pi_\Gamma f$ satisfy an integrated second-difference constraint, and the slopes satisfy an integrated first-difference contraint. Their proofs are deferred to Appendices B.2.2 and B.2.2.

Lemma 21. *Suppose a CPWL $f \in \mathcal{F}_n^{\mathcal{T}}$ and its projection $\Pi_\Gamma f$ onto PWL functions on Γ. Letting $\gamma \sim \gamma'$ denote the neighbor relationship in G_Γ,*

$$\sum_{\gamma\sim\gamma'\sim\gamma''}|\alpha_{\gamma''}-2\alpha_{\gamma'}+\alpha_\gamma|\leq\frac{3^d n_\Gamma(\mathcal{T})\left(1+4n_{\mathcal{T}}(\Gamma)\sqrt{d}\right)}{Nw_{\min}}\cdot\mathrm{DGV}(f;\mathcal{T},w),\qquad\text{(B.14)}$$

where $w_{\min} := \min_{(s_i,s_j)\in E_{\mathcal{T}}} w_{ij}$.

Lemma 22. *Suppose a CPWL $f \in \mathcal{F}_n^{\mathcal{T}}$ and its projection $\Pi_\Gamma f$ onto PWL functions on Γ. Letting $\gamma \sim \gamma'$ denote the neighbor relationship in G_Γ,*

$$\sum_{\gamma\sim\gamma'}\|\beta_{\gamma'}-\beta_\gamma\|_2\leq\frac{3^d n_\Gamma(\mathcal{T})}{w_{\min}}\cdot\mathrm{DGV}(f;\mathcal{T},w),\qquad\text{(B.15)}$$

where $w_{\min} := \min_{(s_i,s_j)\in E_{\mathcal{T}}} w_{ij}$.

Putting it all together

The foregoing analysis has established that the triogram Gaussian process term (B.9) may be upper bound by a remainder term, an approximation term, and a lattice-based complexity term,

$$\langle f(x_{1:n}),z_{1:n}\rangle\leq\|f(x_{1:n})\|_2\|(I-P_{\mathcal{S}_r^{(2)}})\epsilon\|_2$$
$$+\|(f-\Pi_\Gamma f)(x_{1:n})\|_1\|z_{1:n}\|_\infty+\sum_{\gamma\in\Gamma}\sum_{x_i\in\gamma}\Pi_\Gamma f(x_i)\tilde{\epsilon}_{\gamma,i}.$$

Lemma 20 bounds the approximation error,

$$\|(f - \Pi_\Gamma f)(x_{1:n})\|_1 \leq \frac{\tilde{C}_a(n)}{N w_{\min}} \cdot \mathrm{DGV}(f; \mathcal{T}, w)$$

with probability at least $1 - 4\exp(-\delta)$, where

$$\tilde{C}_1(n) := \sqrt{d} n_{\mathcal{T}}(\Gamma) n_\Gamma(\mathcal{T}) n_x(\Gamma).$$

The results of Sections B.2.1 and B.2.1 reveal that (B.13) is bound above by

$$\sum_{\gamma \in \Gamma} \sum_{x_i \in \gamma} \Pi_\Gamma f(x_i) \tilde{\epsilon}_{\gamma,i} \leq \frac{\tilde{C}_2(n,\delta)}{N w_{\min}} \cdot \big(\Sigma^{(1)} + \Sigma_r^{(2)}\big) \cdot \mathrm{DGV}(f; \mathcal{T}, w),$$

where we use the abbreviations

$$\tilde{C}_2(n,\delta) := \sigma \cdot 4 \cdot 3^d \, \mu \, n_\Gamma(\mathcal{T}) \, n_{\mathcal{T}}(\Gamma) \sqrt{2 d n_x(\Gamma)(\log|\Gamma| + \delta)}$$

and

$$\Sigma^{(1)} := \sqrt{|\Gamma|^{-1} \sum_{i \in \mathcal{S}^{(1)}} \frac{1}{\big(\xi_i^{(1)}\big)^2}},$$
$$\Sigma_r^{(2)} := \sqrt{|\Gamma|^{-1} \sum_{i \in \mathcal{S}_r^{(2)}} \frac{1}{\big(\xi_i^{(2)}\big)^2}}.$$

This yields an upper bound on (B.9) of

$$\langle f(x_{1:n}), z_{1:n} \rangle \leq \|f(x_{1:n})\|_2 \|(I - P_{\mathcal{S}_r^{(2)}})\epsilon\|_2 + \frac{\tilde{C}_1(n)\|z_{1:n}\|_\infty + \tilde{C}_2(n,\delta)\big(\Sigma^{(1)} + \Sigma_r^{(2)}\big)}{N w_{\min}} \cdot \mathrm{DGV}(f; \mathcal{T}, w),$$

with probability at least $1 - 4\exp(-\delta)$. Substitute this back into the basic inequality (B.8) with $f = \hat{f}^{\mathcal{T}} - f_0$ to obtain an upper bound at the same probability of

$$\begin{aligned}\|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 \leq{}& 4\|(I - P_{\mathcal{S}_r^{(2)}})\epsilon\|_2^2 \\ &+ \frac{4\tilde{C}_1(n)\|z_{1:n}\|_\infty + 4\tilde{C}_2(n,\delta)\big(\Sigma^{(1)} + \Sigma_r^{(2)}\big)}{N w_{\min}} \\ &\quad \times \mathrm{DGV}(\hat{f}^{\mathcal{T}} - f_0; \mathcal{T}, w) \\ &+ \lambda\big(\mathrm{DGV}(f_0; \mathcal{T}, w) - \mathrm{DGV}(\hat{f}^{\mathcal{T}}; \mathcal{T}, w)\big),\end{aligned}$$

after solving a quadratic inequality of the form $ax^2 - bx - c \leq 0$ in $x = \|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2$. When the tuning parameter is set such that

$$\lambda \geq \frac{4\tilde{C}_1(n)\|z_{1:n}\|_\infty + 4\tilde{C}_2(n,\delta)\big(\Sigma^{(1)} + \Sigma_r^{(2)}\big)}{Nw_{\min}}, \tag{B.16}$$

the penalized triogram $\hat{f}^{\mathcal{T}}$ satisfies

$$\begin{aligned} &n^{-1}\|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 \\ &\quad\leq \frac{4\big((r+2)^d + \delta_1\big)}{n} \\ &\qquad + \frac{4\sigma\tilde{C}_1(n)\sqrt{\log(2n)} + 4\tilde{C}_2(n,\delta_2)\big(\Sigma^{(1)} + \Sigma_r^{(2)}\big)}{nNw_{\min}}\,\mathrm{DGV}(f_0;\mathcal{T},w) \end{aligned} \tag{B.17}$$

with probability at least $1 - \exp(-\delta_1/8) - 4\exp(-\delta_2) - 1/n$.

In the following results, we choose

$$N \asymp \left(\frac{n}{\log^{\alpha_\Gamma} n}\right)^{1/d}$$

for a user-chosen $\alpha_\Gamma > 1$.

Subcritical regime: $d > 4$. Set $r = 1$. Lemma 23 provides that there exists a constant $c > 0$ depending only on d such that

$$\Sigma^{(1)} \asymp \Sigma_1^{(2)} \leq c.$$

This yields a rate of convergence of

$$\begin{aligned} &n^{-1}\|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 \\ &\quad\leq \frac{4\big(3^d + \delta_1\big)}{n} \\ &\qquad + \frac{4(\log n)^{\alpha_\Gamma/d}\big(\sigma\tilde{C}_1(n)\sqrt{\log(2n)} + c\tilde{C}_2(n,\delta_2)\big)}{n^{1+1/d}w_{\min}}\,\mathrm{DGV}(f_0;\mathcal{T},w) \end{aligned} \tag{B.18}$$

with probability at least $1 - \exp(-\delta_1/8) - 4\exp(-\delta_2) - 1/n$. Suppose the choice of N and $\mathcal{T}$ are such that

$$\tilde{C}_1(n), \tilde{C}_2(n,\delta_2) = \tilde{O}(1)$$

and

$$\frac{\mathrm{DGV}(f_0;\mathcal{T},w)}{n^{1/d}w_{\min}} = \tilde{O}\big(n^{1-2/d}\big)$$

are both satisfied. Then the penalized triogram with $\mathcal{T}$, w satisfies

$$n^{-1}\|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 = \tilde{O}_{\mathbb{P}}\big(n^{-2/d}\big).$$

Critical boundary: $d = 4$. Set $r = 1$. Lemma 23 provides that there exists a constant $c > 0$ depending only on d such that

$$\Sigma^{(1)} \leq c$$

and

$$\Sigma_1^{(2)} \leq c \log(N^d).$$

Arguments exactly following the case $d > 4$ confirm that when $d = 4$, the penalized triogram satisfies

$$n^{-1} \|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 = \tilde{O}_{\mathbb{P}}\big(n^{-2/d}\big).$$

Supercritical regime: $d = 2, 3$. First, note from Lemma 23 that for the first-difference operator,

$$\Sigma^{(1)} \leq c \begin{cases} \log(N^d) & d = 2, \\ 1 & d = 3, \end{cases}$$

for a constant $c > 0$ depending only on d. This will contribute a lower-order term to the rate compared with $\Sigma_r^{(2)}$, and so we safely ignore it, along with the approximation error term. For the second-difference operator, Lemma 23 prescribes a bound of

$$\Sigma_r^{(2)} \leq c\sqrt{N^{4-d} r^{d-4}}$$

when $r \in [1, \sqrt{d}N]$. Following the analysis of Sadhanala et al. [2021], we choose r to balance the remainder term with the Gaussian complexity term, i.e., balance

$$(r+2)^d \quad \text{with} \quad c(\log n)^{\frac{\alpha_\Gamma(d-2)}{2d}} \tilde{C}_2(n, \delta_2)\big(n^{\frac{4-d}{2d}} r^{\frac{d-4}{2}}\big) \frac{\mathrm{DGV}(f_0; \mathcal{T}, w)}{n^{1/d} w_{\min}}.$$

Choose

$$(r+2)^d \asymp \left((\log n)^{\frac{\alpha_\Gamma(d-2)}{2d}} \tilde{C}_2(n, \delta_2) \frac{\mathrm{DGV}(f_0; \mathcal{T}, w)}{n^{1/d} w_{\min}}\right)^{\frac{2d}{4+d}} n^{\frac{4-d}{4+d}}.$$

Abbreviating

$$\tilde{C}_3(n, \delta_2) = \left((\log n)^{\frac{\alpha_\Gamma(d-2)}{2d}} \tilde{C}_2(n, \delta_2)\right)^{\frac{2d}{4+d}},$$

it follows that for some $C_4 > 0$ depending only on d, the average squared error of the penalized triogram is upper bound by

$$n^{-1} \|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 \leq C_4 \left(\frac{\delta_1}{n} + \tilde{C}_3(n, \delta_2) \left(\frac{\mathrm{DGV}(f_0; \mathcal{T}, w)}{n^{1/d} w_{\min}}\right)^{\frac{2d}{4+d}} n^{\frac{4-d}{4+d} - 1}\right) \quad \text{(B.19)}$$

with probability at least $1 - \exp(-\delta_1/8) - 4\exp(-\delta_2)$. Suppose the choice of N and $\mathcal{T}$ are such that

$$C_3(n, \delta_2) = \tilde{O}(1)$$

and

$$\frac{\mathrm{DGV}(f_0; \mathcal{T}, w)}{n^{1/d} w_{\min}} = \tilde{O}\big(n^{1-2/d}\big)$$

are both satisfied. Then the penalized triogram satisfies

$$n^{-1}\|(\hat{f}^{\mathcal{T}} - f_0)(x_{1:n})\|_2^2 = \tilde{O}_{\mathbb{P}}\big(n^{-\frac{4}{4+d}}\big).$$

B.2.2 Technical lemmas for the proof of Theorem 5

Introduce the following notation for the purposes of the forthcoming proofs.

- For $\gamma \neq \gamma'$, we define a path $\pi_{\mathcal{T}}(\gamma, \gamma') = \{s_1, \ldots, s_\ell\}$, where $t_\gamma \in s_1$, $t_{\gamma'} \in s_\ell$ and $(s_i, s_{i+1}) \in E_{\mathcal{T}}$ for $i = 1, \ldots, \ell - 1$. By convention, we take the shortest path.
- $\pi_{\mathcal{T}}(\Gamma) := \{\pi_{\mathcal{T}}(\gamma, \gamma') : (\gamma, \gamma') \in E_\Gamma\}$ is the collection of all paths between adjacent grid cells.
- Note that the maximum path length in $\pi_{\mathcal{T}}(\Gamma)$ is bound above by $2n_{\mathcal{T}}(\Gamma)$, since in the shortest each simplex will be visited at most once.
- We will also overload the path notation to have $\pi(s, s') = \{s_1, \ldots, s_\ell\}$, $s_1 = s$, $s_\ell = s'$, and $(s_i, s_{i+1}) \in E_{\mathcal{T}}$ for $i = 1, \ldots, \ell - 1$. Again, we take the shortest path by convention.

Proof of Lemma 20

Proof. Fix a cell γ and a point $x \in \gamma$. Let s_γ whose linear function matches that of γ (i.e., s such that $t_\gamma \in s$), and let s_x be the triangle to which x belongs (if there are several, then pick one arbitrarily).

Case 1. $s_\gamma = s_x$. Then $f(x) = \Pi_\Gamma f(x)$ and we are done.

Case 2. There exists a path $\pi(s_\gamma, s_x) = \{s_0, \ldots, s_\ell\}$, $\ell \geq 1$, such that $s_0 = s_\gamma$, $s_\ell = s_x$, and $s_i \cap \gamma \neq \emptyset$ for all $s_i \in \pi(s_\gamma, s_x)$. Consider points $p_0, \ldots, p_{k+1}$ such that $p_0 = t_\gamma$, $p_{k+1} = x$, and $p_{i+1} \in s_{i+1} \cap s_i \cap \gamma$ for each $i = 0, \ldots, k - 1$. Use the CPWL structure

of f on $\mathcal{T}$ to write,

$$\begin{aligned} f(x) &= f(p_{k+1}) \\ &= f(p_0) + \sum_{j=0}^{\ell} f(p_{j+1}) - f(p_j) \\ &= f(p_0) + \sum_{j=0}^{\ell} \beta_{s_j}^\top (p_{j+1} - p_j). \end{aligned}$$

Recalling the definition of $\Pi_\Gamma f$ on each cell γ and that $p_0 = t_\gamma$, $s_0 = s_\gamma$, we may express

$$\begin{aligned} \Pi_\Gamma f(x) &= f(p_0) + \beta_\gamma^\top (x - p_0) \\ &= f(p_0) + \sum_{j=0}^{\ell} \beta_\gamma^\top (p_{j+1} - p_j). \end{aligned}$$

Therefore, at $x \in \gamma$, the approximation error may be expressed,

$$\begin{aligned} |f(x) - \Pi_\Gamma f(x)| &= \Big| \sum_{j=0}^{\ell} (\beta_{s_j} - \beta_{s_0})^\top (p_{j+1} - p_j) \Big| \\ &\overset{(i)}{\leq} \sum_{j=0}^{\ell} \|\beta_{s_j} - \beta_{s_0}\| \|p_{j+1} - p_j\| \\ &\overset{(ii)}{\leq} \sum_{j=0}^{\ell} \sum_{i=0}^{j-1} \|\beta_{s_{i+1}} - \beta_{s_i}\| \|p_{j+1} - p_j\| \\ &\overset{(iii)}{\leq} \sqrt{d}/N \cdot \sum_{j=0}^{\ell} \sum_{i=0}^{j-1} \|\beta_{s_{i+1}} - \beta_{s_i}\| \\ &\overset{(iv)}{\leq} \frac{\sqrt{d} n_{\mathcal{T}}(\gamma)}{N} \cdot \sum_{j=0}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\|, \end{aligned}$$

where *(i)* uses Cauchy-Schwarz and the triangle inequality; *(ii)* uses a telescoping sum and triangle inequality again; *(iii)* uses an upper bound on the diameter of any cell γ; and *(iv)* upper bounds the length of the path $\pi(s_\gamma, s_x)$ by the number of simplices that intersect cell γ (since in the shortest path, each simplex would be visited at most once).

Apply this argument to all $\gamma \in \Gamma$ and sum over x_i, $i = 1, \dots, n$, to obtain

$$\begin{aligned}\sum_{i=1}^{n} \left| f(x_i) - \Pi_\Gamma f(x_i) \right| &\leq \frac{\sqrt{d} n_{\mathcal{T}}(\Gamma)}{N} \sum_{i=1}^{n} \sum_{j=0}^{\ell(x_i)-1} \|\beta_{s_{j+1}^{(x_i)}} - \beta_{s_j^{(x_i)}}\| \\ &\overset{(i)}{\leq} \frac{\sqrt{n} n_{\mathcal{T}}(\Gamma) n_\Gamma(\mathcal{T}) n_x(\Gamma)}{N} \sum_{(s,s') \in E_{\mathcal{T}}} \|\beta_s - \beta_{s'}\| \\ &\overset{(ii)}{\leq} \frac{\sqrt{d} n_{\mathcal{T}}(\Gamma) n_\Gamma(\mathcal{T}) n_x(\Gamma)}{N w_{\min}} \cdot \sum_{(s_i,s_j) \in E_{\mathcal{T}}} w_{ij} \cdot \left\| \nabla f|_{s_i} - \nabla f|_{s_j} \right\|,\end{aligned}$$

where *(i)* uses a crude upper bound on the number of paths a simplex s may appear in (the number of cells it appears in times the number of sample points in each cell); and *(ii)* re-expresses the result in terms of a (weighted) gradient variation of the original CPWL $f \in \mathcal{F}_n^{\mathcal{T}}$. □

Proof of Lemma 21

Proof. Consider three consecutive grid cells $\gamma = \gamma_j$, $\gamma' = \gamma_{j+e_i}$, $\gamma'' = \gamma_{j+2e_i}$, for some $i = 1, \dots, d$. We will analyze the term

$$|\alpha_{\gamma''} - 2\alpha_{\gamma'} + \alpha_\gamma| = |(\alpha_{\gamma''} - \alpha_{\gamma'}) - (\alpha_{\gamma'} - \alpha_\gamma)|$$

by analyzing each of the first differences. We begin with $\alpha_{\gamma'} - \alpha_\gamma$. Recall the convention that s_γ denotes the simplex whose linear function matches that of γ (i.e., s such that $t_\gamma \in s$), and similarly for $s_{\gamma'}$.

<u>*Case 1.*</u> $s_\gamma = s_{\gamma'}$. We may represent the intercept first-difference by

$$\alpha_{\gamma'} - \alpha_\gamma = \beta_\gamma^\top \big(t_{\gamma'} - t_\gamma\big).$$

<u>*Case 2.*</u> $s_\gamma \neq s_{\gamma'}$. In this case, we must work a little harder. Consider the path $\pi(\gamma, \gamma') = \{s_1, \dots, s_\ell\}$, $\ell \geq 2$, with $s_1 = s_\gamma$, $s_\ell = s_{\gamma'}$, and $\partial s_{i+1} \cap \partial s_i \cap (\gamma \cup \gamma') \neq \emptyset$, $i = 1, \dots, \ell - 1$. Associate to this path the points $p_1, \dots, p_{\ell+1}$, where $p_1 = t_\gamma$, $p_{\ell+1} = t_{\gamma'}$, and $p_j \in \partial s_j \cap \partial s_{j-1} \cap (\gamma \cup \gamma')$, $j = 2, \dots, \ell$.

Form the telescoping sum,

$$\alpha_{\gamma'} - \alpha_\gamma = f(p_{\ell+1}) - f(p_1) = \sum_{j=1}^{\ell} \beta_{s_j}^\top (p_{j+1} - p_j),$$

and separately note that

$$\beta_\gamma^\top (t'_\gamma - t_\gamma) = \beta_\gamma^\top (p_{\ell+1} - p_1) = \sum_{j=1}^{\ell} \beta_\gamma^\top (p_{j+1} - p_j).$$

Add and subtract β_γ within the telescoping sum to obtain that

$$\alpha_{\gamma'} - \alpha_\gamma = \beta_\gamma^\top (t_{\gamma'} - t_\gamma) + \sum_{j=1}^{\ell} (\beta_{s_j} - \beta_\gamma)^\top (p_{j+1} - p_j).$$

Similarly, we may reason that

$$\alpha_{\gamma''} - \alpha_{\gamma'} = \beta_{\gamma'}^\top (t_{\gamma''} - t_{\gamma'}) + \sum_{j=1}^{\ell'} (\beta_{s'_j} - \beta_{\gamma'})^\top (p'_{j+1} - p'_j).$$

Take a difference of terms to obtain,

$$\begin{aligned}
&(\alpha_{\gamma''} - \alpha_{\gamma'}) - (\alpha_{\gamma'} - \alpha_\gamma) \\
&\quad = \beta_{\gamma'}^\top (t_{\gamma''} - t_{\gamma'}) - \beta_\gamma^\top (t_{\gamma'} - t_\gamma) \\
&\qquad\qquad + \sum_{j=1}^{\ell'} (\beta_{s'_j} - \beta_{\gamma'})^\top (p'_{j+1} - p'_j) - \sum_{j=1}^{\ell} (\beta_{s_j} - \beta_\gamma)^\top (p_{j+1} - p_j) \\
&\quad = \frac{e_i}{N}^\top (\beta_{\gamma'} - \beta_\gamma) + \sum_{j=1}^{\ell'} (\beta_{s'_j} - \beta_{\gamma'})^\top (p'_{j+1} - p'_j) - \sum_{j=1}^{\ell} (\beta_{s_j} - \beta_\gamma)^\top (p_{j+1} - p_j),
\end{aligned}$$

where in the final line we recall that $t_{\gamma''} - t_{\gamma'} = t_{\gamma'} - t_\gamma = e_i/N$. Apply the triangle inequality to deduce that

$$\begin{aligned}
&|(\alpha_{\gamma''} - \alpha_{\gamma'}) - (\alpha_{\gamma'} - \alpha_\gamma)| \\
&\quad \le \frac{1}{N} \|\beta_{\gamma'} - \beta_\gamma\| + \sum_{j=1}^{\ell'} \|\beta_{s'_j} - \beta_{\gamma'}\| \|p'_{j+1} - p'_j\| + \sum_{j=1}^{\ell} \|\beta_{s_j} - \beta_\gamma\| \|p_{j+1} - p_j\| \\
&\quad \overset{(i)}{\le} \frac{1}{N} \|\beta_{\gamma'} - \beta_\gamma\| + \frac{\sqrt{d}}{N} \Big(\sum_{j=1}^{\ell'} \|\beta_{s'_j} - \beta_{\gamma'}\| + \sum_{j=1}^{\ell} \|\beta_{s_j} - \beta_\gamma\| \Big) \\
&\quad \overset{(ii)}{\le} \frac{1}{N} \|\beta_{\gamma'} - \beta_\gamma\| + \frac{\sqrt{d}}{N} \Big(\sum_{j=1}^{\ell'-1} \sum_{i=1}^{j} \|\beta_{s'_{i+1}} - \beta_{s'_i}\| + \sum_{j=1}^{\ell-1} \sum_{i=1}^{j} \|\beta_{s_{i+1}} - \beta_{s_i}\| \Big) \\
&\quad \overset{(iii)}{\le} \frac{1}{N} \|\beta_{\gamma'} - \beta_\gamma\| + \frac{2n_{\mathcal{T}}(\Gamma)\sqrt{d}}{N} \Big(\sum_{j=1}^{\ell'-1} \|\beta_{s'_{j+1}} - \beta_{s'_j}\| + \sum_{j=1}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\| \Big) \\
&\quad \overset{(iv)}{\le} \frac{1}{N} \sum_{j=1}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\| + \frac{2n_{\mathcal{T}}(\Gamma)\sqrt{d}}{N} \Big(\sum_{j=1}^{\ell'-1} \|\beta_{s'_{j+1}} - \beta_{s'_j}\| + \sum_{j=1}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\| \Big),
\end{aligned}$$

where we obtain *(i)* from an upper bound on the cell diameter; *(ii)* from a telescoping sum (and recalling that $\beta_\gamma = \beta_{s_1}$ and similarly for γ'); *(iii)* from uniformly upper bounding the path length by twice the number of simplices appearing in any grid cell; and *(iv)* from another telescoping sum.

Sum over all $\gamma = \gamma_j$, $\gamma' = \gamma_{j+e_i}$, $\gamma'' = \gamma_{j+2e_i}$, $j = 1, \dots, N-2$, $i = 1, \dots, d$, to obtain,

$$\begin{aligned}
&\sum_{\gamma \sim \gamma' \sim \gamma''} |(\alpha_{\gamma''} - \alpha_{\gamma'}) - (\alpha_{\gamma'} - \alpha_\gamma)| \\
&\qquad \le 3^d n_\Gamma(\mathcal{T}) \Big(\frac{1}{N} \sum_{(s,s') \in E_\mathcal{T}} \|\beta_{s'} - \beta_s\| + \frac{4 n_\mathcal{T}(\Gamma)\sqrt{d}}{N} \sum_{(s,s') \in E_\mathcal{T}} \|\beta_{s'} - \beta_s\| \Big) \\
&\qquad \le \frac{3^d n_\Gamma(\mathcal{T}) \big(1 + 4 n_\mathcal{T}(\Gamma)\sqrt{d}\big)}{N w_{\min}} \cdot \sum_{(s_i,s_j) \in E_\mathcal{T}} w_{ij} \cdot \big\| \nabla f|_{s_i} - \nabla f|_{s_j} \big\|.
\end{aligned}$$

□

Proof of Lemma 22

Proof. Consider $(\gamma, \gamma') \in E_\Gamma$. Let s_γ denote the simplex whose linear function matches that of γ (i.e., s such that $t_\gamma \in s$), and similarly for $s_{\gamma'}$.

<u>*Case 1.*</u> $s_\gamma = s_{\gamma'}$. Then $\beta_\gamma = \beta_{\gamma'}$ and we are done.

<u>*Case 2.*</u> $s_\gamma \neq s_{\gamma'}$. There exists a path from s_γ to $s_{\gamma'}$, i.e., $\pi(\gamma, \gamma') = \{s_1, \dots, s_\ell\}$, $\ell \ge 2$, with $s_1 = s_\gamma$, $s_\ell = s_{\gamma'}$, and $\partial s_{i+1} \cap \partial s_i \cap (\gamma \cup \gamma') \neq \emptyset$, $i = 1, \dots, \ell - 1$. Expand the difference in coefficients as a telescoping sum and apply the triangle inequality to upper bound,

$$\|\beta_{\gamma'} - \beta_\gamma\| \le \sum_{j=1}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\|.$$

Summing over all neighboring grid cells,

$$\begin{aligned}
\sum_{\gamma \sim \gamma'} \|\beta_{\gamma'} - \beta_\gamma\| &\le \sum_{\gamma \sim \gamma'} \sum_{j=1}^{\ell-1} \|\beta_{s_{j+1}} - \beta_{s_j}\| \\
&\overset{(i)}{\le} 3^d n_\Gamma(\mathcal{T}) \sum_{(s,s') \in E_\mathcal{T}} \|\beta_{s'} - \beta_s\| \\
&\le \frac{3^d n_\Gamma(\mathcal{T})}{w_{\min}} \cdot \sum_{(s_i,s_j) \in E_\mathcal{T}} w_{ij} \cdot \big\| \nabla f|_{s_i} - \nabla f|_{s_j} \big\|,
\end{aligned}$$

where *(i)* is obtained by observing that there are at most 3^d paths emerging from each grid cell, in which each simplex may participate at most once in the shortest path. □

Spectral functionals of the lattice-based discrete-difference operator

The following lemma summarizes results of Sadhanala et al. [2021] controlling certain spectral functionals of the lattice-based discrete-difference operator.

Lemma 23 (Sadhanala et al., 2021; Lemmas 1 and 6)**.** *Let Γ be a d-dimensional lattice with N vertices along each dimension, for a total of $|\Gamma| = N^d$ vertices. Denote by $D_\Gamma^{(m)}$ the mth-order discrete-difference operator on Γ. Let its singular value decomposition be $D_\Gamma^{(m)} = U\Xi V^\top$, with lattice-based indices $i = (i_1, \dots, i_d) \in [N]^d$, where the indices $[m]^d$ corresponding to $\mathrm{null}(D_\Gamma^{(m)})$ are left unused.*

- *There exists a subset $\mathcal{S}^{(m)} = [N]^d \setminus [m+1]^d$ of the left singular vectors satisfying incoherence, i.e.,*

$$\|u_i\|_\infty \leq \frac{\mu}{\sqrt{N^d}}, \qquad i \in \mathcal{S}^{(m)}, \tag{B.20}$$

  for some $\mu \geq 1$ that depends only on m, d.

- *The singular values ξ_i corresponding to these left singular vectors satisfies a certain decay structure. In the case $2m \leq d$,*

$$\sum_{i \in \mathcal{S}^{(m)}} \frac{1}{\xi_i^2} \leq c \begin{cases} N^d & 2m < d \\ N^d \log(N^d) & 2m = d, \end{cases} \tag{B.21}$$

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

- *In the case $2m > d$, for $r \in [1, \sqrt{d}N]$, the singular values in the subset $\mathcal{S}_r^{(m)} = \{i : \|(i - m - 1)_+\|_2 \geq r\} \subset \mathcal{S}$ obey the decay structure*

$$\sum_{i \in \mathcal{S}_r^{(m)}} \frac{1}{\xi_i^2} \leq c N^{2m} r^{d-2m}, \tag{B.22}$$

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

Error control via incoherence

Lemma 24. *Let $D \in \mathbb{R}^{r \times n}$ have rank q, with singular values $\xi_1 \leq \cdots \leq \xi_q$ corresponding to left singular vectors $u_1, \dots, u_q \in \mathbb{R}^r$. Denote the full singular value decomposition $D = U\Xi V^\top$. Assume an index set $\mathcal{S} \subset [q]$ upon which the vectors $u_i, i \in \mathcal{S}$ are incoherent, meaning that for a constant $\mu \geq 1$,*

$$\|u_i\|_\infty \leq \frac{\mu}{\sqrt{n}}, \qquad i \in \mathcal{S}.$$

Furthermore, let $S \in \mathbb{R}^{n\times n}$ be a diagonal scaling matrix. For $\epsilon \sim N(0, \sigma^2 I_n)$,

$$\|(D^+)^\top S P_{\mathcal{S}} \epsilon\|_\infty \leq \sigma \frac{\mu \|S\|_2}{\sqrt{n}} \sqrt{2(\log r + \delta) \sum_{i\in\mathcal{S}} \frac{1}{\xi_i^2}} \tag{B.23}$$

with probability at least $1 - 2\exp(-\delta)$.

Proof. We follow the proof of Wang et al. [2016, Theorem 6]. Define

$$g_j = P_{\mathcal{S}} S D^+ e_j, \qquad j \in [r],$$

with e_j the jth canonical basis vector. Observe for a fixed j that

$$\begin{aligned} g_j &= V_{\mathcal{S}} V_{\mathcal{S}}^\top S V \Xi^{-1} U^\top e_j \\ &= \tilde{V}_{\mathcal{S}} (S\Xi^{-1}) U^\top e_j, \end{aligned}$$

where $\tilde{V}_{\mathcal{S}} \in \mathbb{R}^{n\times q}$ has ith column V_i for $i \in \mathcal{S}$ and 0 for $i \notin \mathcal{S}$. Therefore for each $j \in [r]$,

$$\begin{aligned} \|g_j\|_2^2 &= \frac{\mu^2}{n} \sum_{i\in\mathcal{S}} \frac{S_{ii}^2}{\xi_i^2} \\ &\leq \frac{\mu^2 \|S\|_2^2}{n} \sum_{i\in\mathcal{S}} \frac{1}{\xi_i^2}. \end{aligned}$$

Conclude via a union bound that

$$\begin{aligned} \|(D^+)^\top S P_{\mathcal{S}} \epsilon\|_\infty &= \max_{j\in[r]} \left| g_j^\top \epsilon \right| \\ &\leq \sigma \frac{\mu \|S\|_2}{\sqrt{n}} \sqrt{2(\log r + \delta) \sum_{i\in\mathcal{S}} \frac{1}{\xi_i^2}} \end{aligned}$$

with probability at least $1 - 2\exp(-\delta)$. □

B.2.3 Proofs and technical lemmas for Section 3.3.2

Empirical content of grid cells

The following lemma appears previously as Lemma 12. We reproduce it, with notation matching our setting, for completeness.

Lemma 25. *Suppose $x_1, \dots, x_n$ are sampled from a distribution P supported on $(0,1)^d$ with density p such that $0 < p_{\min} < p(x) < p_{\max} < 1$ for all $x \in (0,1)^d$. Form a partition of $(0,1)^d$ using an equally spaced mesh with $N = C_1(p_{\min} n / \log^{\alpha_\Gamma} n)^{1/d}$, $\alpha_\Gamma > 1$, along each dimension. Let γ_ℓ denote the ℓth cell of the mesh, and let $|\gamma_\ell|$ denote its empirical content. Then for all $x_{1:n} \in \mathscr{X}_1$, with $\mathbb{P}\{x_{1:n} \in \mathscr{X}_1\} \geq 1 - 2/n^4$,*

$$\max_\ell |\gamma_\ell| \leq C_3 \log^{\alpha_\Gamma} n, \tag{B.24}$$

$$\min_\ell |\gamma_\ell| \geq c_4 \log^{\alpha_\Gamma} n, \tag{B.25}$$

for n sufficiently large, where $C_3, c_4 > 0$ depend only on $p_{\min}, p_{\max}, d$.

Geometric properties of Delaunay simplices

Lemma 26. *Sample $x_1, \dots, x_n$ from a distribution P following Assumption A1. Form the Delaunay triangulation $\mathcal{DT}(x_{1:n})$ and consider the subtriangulation $\widetilde{\mathcal{DT}}(x_{1:n})$. There exists a constant C_5 depending only on $p_{\min}$, d, $\alpha \geq 1$ such that for all sufficiently large n,*

$$\max_{s \in \widetilde{\mathcal{DT}}} r(s) \leq C_5 (\log n / n)^{1/d}, \tag{B.26}$$

with probability at least $1 - n^{-\alpha}$.

The following proof is an adaptation of Bern et al. [1991, Theorem 1], which considered the minus-sampling setting.

Proof. Let $\mathcal{C}(s)$ denote the circumsphere of a simplex $s \in \widetilde{\mathcal{DT}}$. (Although it is called a circum*sphere*, we will actually take $\mathcal{C}(s) := B(s)^\circ$, where $B(s)$ is the smallest Euclidean ball containing s.) We will provide an upper bound on $r(\mathcal{C}(s))$, $s \in \widetilde{\mathcal{DT}}$, leveraging a defining property of the Delaunay triangulation: $x_{1:n} \cap \mathcal{C}(s) = 0$ for all $s \in \mathcal{DT}$ [1]. In other words, we will bound the probability that there exists a set of $d+1$ points with empty circumsphere of a certain radius or larger. This can be done via the union bound,

$$\begin{aligned}
\mathbb{P}\big\{\exists x_{1:(d+1)} \subset \tilde{x}_{1:n} : |x_{1:n} \cap \mathcal{C}(x_{1:(d+1)})| = 0, & r(\mathcal{C}(x_{1:(d+1)})) \geq \varepsilon\big\} \\
&\leq \binom{n}{k+1} (1 - p_{\min} \varepsilon^d)^{(n-d-1)} \\
&\leq C n^{d+1} \exp\big(- p_{\min}(n-d-1)\varepsilon^d \big),
\end{aligned}$$

[1] The lack of a tilde here is intentional; this defining condition holds for all members of the triangulation $\mathcal{DT}$, not just for the subtriangulation $\widetilde{\mathcal{DT}}$. However, the probabilistic upper bound on the radius only holds for simplices in the subtriangulation.

for sufficiently large n and some fixed constant C. In the first inequality, we have used the crucial condition that $x_{1:(d+1)} \in \tilde{x}_{1:n}$ to ensure that we can bound the probability mass of $\mathcal{C}(x_{1:(d+1)})$ from below, so long as $\varepsilon \leq C'(p_{\min}^{-1}(\alpha+d+1)\log n/n)^{1/d}$. Choose $\varepsilon = C'(p_{\min}^{-1}(\alpha+d+1)\log n/n)^{1/d}$ to obtain the claim. □

Corollary 2. *Sample $x_1, \ldots, x_n$ from a distribution P following Assumption A1. Form the Delaunay triangulation $\mathcal{DT}(x_{1:n})$ and consider the subtriangulation $\widetilde{\mathcal{DT}}(x_{1:n})$. There exists a constants C_6, C_7 depending only on $p_{\min}$, d, $\alpha \geq 1$ such that the following statements hold: with probability at least $1 - n^{-\alpha}$, for any simplex $s \in \widetilde{\mathcal{DT}}(x_{1:n})$, the volume of s is upper bounded*

$$\mu(s) \leq C_6 \log n/n, \tag{B.27}$$

and the surface area of any facet F_i, $i = 1, \ldots, d+1$ of s is upper bounded

$$\mathcal{H}^{d-1}(F_i) \leq C_7(\log n/n)^{(d-1)/d}. \tag{B.28}$$

Overlap between Delaunay simplices and grid cells

Lemma 27. *Sample $x_1, \ldots, x_n$ from a distribution P following Assumption A1. Form the Delaunay triangulation $\mathcal{DT}(x_{1:n})$ and consider the subtriangulation $\widetilde{\mathcal{DT}}(x_{1:n})$. Form a grid open partition Γ of Ω with N^d cells γ_i, each with sidelength $1/N$. On the same event $\mathscr{X}_2$ upon which the conclusion of Lemma 26 holds, there exists a constant C_6 such that the number of grid cells γ overlapping any simplex s is upper bounded,*

$$\max_{s \in \widetilde{\mathcal{DT}}} |\{\gamma \in \Gamma : s \cap \gamma \neq \emptyset\}| \leq \max\left\{ \left\lceil \frac{C_6 \log n/n}{(1/N)^d} \right\rceil, 2^d \right\}. \tag{B.29}$$

Proof. The result follows from a simple volume packing argument. □

Lemma 28. *Sample $x_1, \ldots, x_n$ from a distribution P following Assumption A1. Form the Delaunay triangulation $\mathcal{DT}(x_{1:n})$ and consider the subtriangulation $\widetilde{\mathcal{DT}}(x_{1:n})$. Form a grid open partition Γ of Ω with N^d cells γ_i, each with sidelength $1/N$. Take $N = C_1(p_{\min} n/\log^{\alpha_\Gamma} n)^{1/n}$, $\alpha_\Gamma > 1$. On the same event $\mathscr{X}_1$ upon which the conclusion of Lemma 25 holds, there exists a constant C_8 depending only on $p_{\min}, p_{\max}, d$ such that the number of simplices s overlapping any cell γ is upper bounded,*

$$\max_{\gamma \in \Gamma} |\{s \in \widetilde{\mathcal{DT}} : s \cap \gamma \neq \emptyset\}| \leq C_8(\log n)^{\alpha_\Gamma(d+1)}. \tag{B.30}$$

Proof. Fix a $\gamma_i \in \Gamma$. We will first prove that any simplex s which has nonempty intersection with γ_i must be entirely contained in

$$\Gamma_i := \cup\{\bar{\gamma}_j : \|\gamma_j - \gamma_i\|_\infty \leq 3\}.$$

Suppose not. Then there exists $s \in \widetilde{\mathcal{DT}}$ such that $x, x' \in s$ with $x \in \Omega \setminus \Gamma_i$ and $x' \in \gamma_i$. By construction,

$$\|x - x'\| \geq 3/N.$$

Consider $B(x, x')$, the smallest ball with x, x' on its surface. There exists $B' \in B(x, x')$ such that $\operatorname{diam}(B') = 3/N$ and $B' \subset \Gamma_i \setminus \gamma_i$. It then follows that there exists $\gamma_j \subset B'$, $1 \leq \|j - i\|_\infty \leq 3$. On the event $\mathscr{X}_1$, γ_j contains at least $c_4 \log^{\alpha_\Gamma} n$ sample points, contradicting the claim that $s \in \widetilde{\mathcal{DT}}$.

We now proceed to demonstrate the claim, using a crude upper bound based on the maximum number of sample points on each grid cell under the event $\mathscr{X}_1$.

$$\begin{aligned}
|\{s : s \cap \gamma_i \neq \emptyset\}| &\leq \binom{|\{x_{1:n} \cap \Gamma_i\}|}{d+1} \\
&\leq C\big(|\{x_{1:n} \cap \Gamma_i\}|\big)^{d+1} \\
&\leq C(5^d C_4 \log^{\alpha_\Gamma} n)^{d+1} \\
&\leq C_8 (\log n)^{\alpha_\Gamma (d+1)}.
\end{aligned}$$

This upper bound holds simultaneously for all γ_i. □

B.2.4 Proof of Theorem 6

Proof. First, apply Lemmas 25, 26, 27, 28 to enforce probabilistic control of the necessary functionals of $\widetilde{\mathcal{DT}}$, and then apply the analysis of Section 3.3.1 to $\hat{f}^{\widetilde{\mathcal{DT}}}$ on $\tilde{\Omega}$ in order to obtain the rate

$$\frac{1}{\tilde{n}} \sum_{x_i \in \tilde{x}_{1:n}} \big(\hat{f}^{\widetilde{\mathcal{DT}}}(x_i) - f_0(x_i)\big)^2 = \begin{cases} \tilde{O}_{\mathbb{P}}\big(L\tilde{n}^{-\frac{4}{4+d}}\big) & d < 4, \\ \tilde{O}_{\mathbb{P}}\big(L\tilde{n}^{-\frac{2}{d}}\big) & d \geq 4, \end{cases}$$

We now show that on the set $\mathscr{X}_1$ on which the conclusion of Lemma 25 holds, $\tilde{n}$ scales asymptotically like n. Take $\alpha_\Gamma = \alpha > 1$ for the user-chosen parameters in Lemmas 25 and 26. For $x_{1:n} \in \mathscr{X}_1$,

$$\begin{aligned}
\tilde{n} &:= |x_{1:n} \cap \tilde{\Omega}| \\
&= n - \sum_{\gamma : \bar{\gamma} \cap \partial\Omega \neq \emptyset} |x_{1:n} \cap \gamma| \\
&\geq n - 2dN^{d-1} \max_{\gamma \in \Gamma} |x_{1:n} \cap \gamma| \\
&\geq n - 2dC_4 N^{d-1} \log^{\alpha} n \\
&\geq n - C_5 (\log n)^{\alpha/d} n^{1-1/d} \\
&\geq c_6 n,
\end{aligned}$$

for some $c_6 > 0$ and n sufficiently large. □

B.2.5 Proof of Theorem 7

Proof. The proof proceeds in three steps: first, we reduce the problem to estimating binary sequences; next, we apply Assouad's lemma; and finally, we optimize the resulting constrained maximization problem to obtain the lower bounds. In the supercritical regime, optimizing all the free parameters in the construction provides a valid lower bound. In the subcritical regime, optimizing all the free parameters yields a vacuous constant-order lower bound (reflecting the impossibility of estimation when $d \geq 4$), and so we restrict our attention to estimating functions in $\mathrm{BGV}(L)$ which are additionally essentially bounded, i.e., functions in $\mathrm{BGV}_\infty(L, M)$.

Step 1: Reduction to estimating binary sequences. We begin by associating functions f_θ with the vertices of a hypercube $\Theta_S = \{0,1\}^S$, where $S \subseteq [m]^d$ for some $m \in \{1, 2, \dots\}$. The functions f_θ are constructed as follows: we partition Ω into cubes,

$$Q_i = \frac{1}{m}(i_1 - 1, i_1) \times \cdots \times \frac{1}{m}(i_d - 1, i - d), \qquad i \in [m]^d,$$

and for each $\theta \in \Theta_S$ take f_θ to be the continuous piecewise linear function,

$$f_\theta(x) = \sum_{i \in S} \theta g(x; a, Q_i), \tag{B.31}$$

where the function $g(x; a, Q_i)$ is given by

$$g(x; a, Q_i) = 2ma \cdot d(x, \partial Q_i) \cdot 1\{x \in Q_i\}.$$

A direct calculation reveals that for all $\theta \in \Theta_S$, $\epsilon := 1/m$,

$$\mathrm{TV}(\nabla f_\theta) \leq C_d a |S| \epsilon^{d-2} \tag{B.32}$$

for a constant C_d that depends only on dimension. So long as the gradient variation (B.32) is bound above by L, we may reduce

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BGV}(L)} \mathbb{E}_{f_0} \|\hat{f} - f_0\|^2_{L^2(\Omega)} \geq \inf_{\hat{f}} \sup_{\theta \in \Theta_S} \mathbb{E}_\theta \|\hat{f} - f_\theta\|^2_{L^2(\Omega)} \geq \frac{a^2 \epsilon^d}{64(d+2)^2} \mathbb{E}_\theta \rho(\hat{\theta}, \theta), \tag{B.33}$$

where $\rho(\theta, \theta') = \sum_{i \in S} |\theta_i - \theta'_i|$ is the Hamming distances between vertices $\theta, \theta' \in \Theta_S$. To obtain the latter inequality in (B.33), first for a given $\hat{f}$, take

$$\hat{\theta}_i = \begin{cases} 1 & \fint_{Q_i} \hat{f} \geq \frac{a}{8(d+2)}, \\ 0 & \text{otherwise}, \end{cases}$$

and calculate

$$\begin{aligned}\|\hat{f} - f_\theta\|^2_{L^2(\Omega)} &= \sum_{i\in[m]^d} \|\hat{f} - f_\theta\|^2_{L^2(Q_i)} \\ &\geq \sum_{i\in S} \|\hat{f} - f_\theta\|^2_{L^2(Q_i)} \\ &\geq \frac{a^2\epsilon^d}{64(d+2)^2} \sum_{i\in S} 1\{\hat{\theta}_i \neq \theta_i\}.\end{aligned}$$

The final inequality above is obtained by considering two cases. In the first case, $\hat{\theta}_i = 0$, $\theta_i = 1$, and we find that

$$\begin{aligned}\|\hat{f} - f_\theta\|^2_{L^2(Q_i)} &= \int_{Q_i} (\hat{f}(x) - g(x; a, Q_i))^2 dx \\ &= \int_{Q_i} (\hat{f}(x) - \fint_{Q_i} \hat{f})^2 dx + \int_{Q_i} (\fint_{Q_i} \hat{f} - g(x; a, Q_i))^2 dx \\ &\quad + \int_{Q_i} (\hat{f}(x) - \fint_{Q_i} \hat{f})(\fint_{Q_i} \hat{f} - g(x; a, Q_i)) dx \\ &\geq \int_{Q_i} (\fint_{Q_i} \hat{f} - g(x; a, Q_i))^2 dx \\ &= \int_{Q_i} (g(x; a, Q_i))^2 dx - 2(\fint_{Q_i} \hat{f}) \int_{Q_i} g(x; a, Q_i) dx + (\fint_{Q_i} \hat{f})^2 \epsilon^d \\ &\geq \frac{a^2\epsilon^d}{4(d+1)(d+2)},\end{aligned}$$

upon taking $\fint_{Q_i} \hat{f} < \frac{a}{8(d+2)}$. In the second case, $\hat{\theta}_i = 1$, $\theta_i = 0$, and we find that

$$\begin{aligned}\|\hat{f} - f_\theta\|^2_{L^2(Q_i)} &= \int_{Q_i} (\hat{f}(x))^2 dx \\ &= \int_{Q_i} (\hat{f}(x) - \fint_{Q_i} \hat{f})^2 dx + \int_{Q_i} (\fint_{Q_i} \hat{f})^2 dx + (\fint_{Q_i} \hat{f}) \int_{Q_i} (\hat{f}(x) - \fint_{Q_i} \hat{f}) dx \\ &\geq \int_{Q_i} (\fint_{Q_i} \hat{f})^2 dx \\ &\geq \frac{a^2\epsilon^d}{64(d+2)^2},\end{aligned}$$

upon taking $\fint_{Q_i} \hat{f} \geq \frac{a}{8(d+2)}$.

Step 2: Application of Assouad's lemma. Given a measurable space $(\mathcal{Z}, \mathcal{A})$ and a set of probability measures $\mathcal{M} = \{\mu_\theta : \theta \in \Theta_S\}$, Assouad's lemma lower bounds the minimax risk over Θ_S, when loss is measured by the Hamming distance $\rho(\hat{\theta})$.

Lemma 29 (Tsybakov, 2009; Lemma 2.12)**.** *Suppose that for each $\theta, \theta' \in \Theta_S : \rho(\theta, \theta') = 1$, we have that $\mathrm{KL}(\mu_\theta, \mu_{\theta'}) \leq \alpha < \infty$. It follows that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_S} \mathbb{E}_\theta \rho(\hat{\theta}, \theta) \geq \frac{|S|}{2} \max\left(\frac{1}{2}\exp(-\alpha), \left(1 - \sqrt{\alpha/2}\right)\right).$$

We take $\mathcal{Z} = (\Omega \times \mathbb{R})^{\otimes n}$ and associate each $\theta \in \Theta_S$ with the measure $\mu_\theta^{(n)}$, the n-times product of measure $\mu_\theta = \mathrm{Unif}(\Omega) \times N(f_\theta(x), 1)$. We now upper bound the Kullback-Leibler divergence $\mathrm{KL}(\mu_\theta^{(n)}, \mu_{\theta'}^{(n)})$ when $\rho(\theta, \theta') = 1$. Letting $i \in S$ be the single index at which $\theta_i \neq \theta_i'$ and ϕ be the density for a standard normal random variable, we may calculate

$$\begin{aligned}
\mathrm{KL}\left(\mu_\theta^{(n)}, \mu_{\theta'}^{(n)}\right) &= \int_\Omega \int_{\mathbb{R}} \log\left(\frac{\phi\big(y - f_\theta(x)\big)}{\phi\big(y - f_{\theta'}(x)\big)}\right) \phi\big(y - f_\theta(x)\big)\, dy\, dx \\
&= \int_{Q_i} \int_{\mathbb{R}} \log\left(\frac{\phi\big(y - f_\theta(x)\big)}{\phi\big(y - f_{\theta'}(x)\big)}\right) \phi\big(y - f_\theta(x)\big)\, dy\, dx \\
&= \int_{Q_i} \int_{\mathbb{R}} \left(\frac{(\theta_i' - \theta_i)\big(g(x; a, Q_i)\big)^2 + 2(\theta_i - \theta_i')g(x; a, Q_i)y}{2}\right) \\
&\qquad\qquad \times \phi\big(y - f_\theta(x)\big)\, dy\, dx \\
&= \frac{1}{2}\int_{Q_i} \big(g(x; a, Q_i)\big)^2\, dx \\
&= \frac{a^2\epsilon^d}{2(d+1)(d+2)},
\end{aligned}$$

and conclude that

$$\mathrm{KL}\left(\mu_\theta^{(n)}, \mu_{\theta'}^{(n)}\right) \leq \frac{n\epsilon^d a^2}{2(d+1)(d+2)}.$$

It follows that as long as a, S, ϵ have been chosen such that $f_\theta \in \mathrm{BGV}(L), \theta \in \Theta_S$ and

$$\frac{n\epsilon^d a^2}{2(d+1)(d+2)} \leq 1,$$

we may apply Lemma 29 to conclude that

$$\inf_{\hat{f}} \sup_{f_0 \in \mathrm{BGV}(L)} \mathbb{E}_{f_0} \|\hat{f} - f_0\|_{L^2(\Omega)}^2 \geq \frac{a^2\epsilon^d}{64(d+2)^2} \inf_{\hat{\theta}} \max_{\theta \in \Theta_S} \mathbb{E}_\theta \rho(\hat{\theta}, \theta) \geq \frac{a^2\epsilon^d |S|}{256(d+2)^2 \exp(1)}. \tag{B.34}$$

Step 3a: Supercritical lower bound. The minimax risk over $\mathrm{BGV}(L)$ may be lower bounded by a solution to the following constrained maximization problem: letting $s = |S|$,

$$\begin{aligned} \text{maximize} \quad & \frac{a^2 \epsilon^d s}{256(d+2)^2 \exp(1)}, \\ \text{subject to} \quad & 1 \le s \le \epsilon^{-d}, \\ & as\epsilon^{d-2} \le \frac{L}{C_d}, \\ & \frac{n\epsilon^d a^2}{2(d+1)(d+2)} \le 1. \end{aligned}$$

Set $\epsilon = (\frac{2(d+1)(d+2)}{a^2 n})^{1/d}$ and $s = (\frac{L}{C_d a})\epsilon^{-(d-2)}$ to consider the problem

$$\begin{aligned} \text{maximize} \quad & \frac{\big(2(d+1)(d+2)\big)^{2/d}}{256(d+2)^2 \exp(1) C_d} L a^{1-4/d} n^{-2/d}, \\ \text{subject to} \quad & 1 \le \left(\frac{1}{C_d \big(2(d+1)(d+2)\big)^{\frac{d-2}{d}}} \right) L a^{\frac{d-4}{d}} n^{\frac{d-2}{d}} \le \frac{a^2 n}{2(d+1)(d+2)}. \end{aligned} \tag{B.35}$$

When $d < 4$, the lower bound may be improved by taking $a > 0$ small, subject to the constraints in (B.35). The first constraint is trivially satisfied for $a > 0$ small, and the second constraint reveals the lower bound

$$a \ge \left(\frac{(2(d+1)(d+2))^{2/d}}{C_d} \right)^{\frac{d}{4+d}} L^{\frac{d}{4+d}} n^{\frac{-2}{4+d}},$$

which when substituted into the criterion of (B.35) yields (3.31).

Step 3b: Subcritical lower bound. Returning to the constrained maximization problem (B.35), we see that when $d > 4$, the criterion is maximized by taking a large, again trivially satisfying the first constraint. The second constraint is also satisfied by taking a large, implying an arbitrarily large lower bound on the estimation error. As a result, we instead pursue a lower bound on the estimation error over $\mathrm{BGV}_\infty(L, M) := \mathrm{BGV}(L) \cap L^\infty(M)$. Arguments identical to the foregoing imply the amended constrained maximization

$$\begin{aligned} \text{maximize} \quad & \frac{\big(2(d+1)(d+2)\big)^{2/d}}{256(d+2)^2 \exp(1) C_d} L a^{1-4/d} n^{-2/d}, \\ \text{subject to} \quad & 1 \le \left(\frac{1}{C_d \big(2(d+1)(d+2)\big)^{\frac{d-2}{d}}} \right) L a^{\frac{d-4}{d}} n^{\frac{d-2}{d}} \le \frac{a^2 n}{2(d+1)(d+2)}, \\ & a \le M. \end{aligned} \tag{B.36}$$

Take $a = M$ to obtain (3.32). □

B.2.6 Proofs and technical lemmas for Section 3.3.4

Proof of Lemma 4

In this subsection, re-notate $\tilde{n}, \tilde{x}_{1:n}, \tilde{\Omega}$ to $n, x_{1:n}, \Omega$ respectively; and $\tilde{\mathcal{T}}$ to $\widetilde{\mathcal{DT}}$.

Proof. Let $\mathscr{C}^{d+2}(x_{1:n})$ denote all size-$(d+2)$ combinations of points from $x_1, \dots, x_n$. For each element $(x_1, \dots, x_{d+2})$ in $\mathscr{C}^{d+2}(x_{1:n})$, assume without loss of generality that the points are ordered such that the simplices $s_1 = (x_1, x_3, \dots, x_{d+2})$, $s_2 = (x_2, x_3, \dots, x_{d+2})$ are the Delaunay simplices of $(x_1, \dots, x_{d+2})$. First, upper bound the discrete gradient variation (unweighted) via Lemma 30 and a Hölder inequality,

$$\begin{aligned}
&\sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \left\|\hat{g}(s_1) - \hat{g}(s_2)\right\|_2 1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}\} \\
&\quad\leq \max_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \left(\lambda_{\min}^{-1/2}(M_1^\top M_1) + \lambda_{\min}^{-1/2}(M_2^\top M_2)\right) \\
&\qquad\qquad \times 1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}\} \\
&\qquad\qquad \times \sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\| \, 1\{x_2, \dots, x_{d+2} \in B(x_1, r^*)\}.
\end{aligned}$$

The first term we recognize as $\lambda_{\min}^{-1/2}(\tilde{\mathcal{T}})$. We now analyze the second term, take its expectation, and apply Markov's inequality to obtain the result.

$$\begin{aligned}
&\sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\| \, 1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}\} \\
&= \sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\| \big(1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}, r(s_1) \wedge r(s_2) \le r\} \\
&\qquad\qquad + 1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}, r(s_1) \wedge r(s_2) > r\}\big)
\end{aligned}$$

Choose $r = r^* := C(p_{\min}^{-1}(\alpha + d + 1)\log n/n)^{1/d}$ for a fixed constant $C > 0$ as in Lemma 26. The latter term is zero with probability at least $1 - n^{-\alpha}$, so we focus on the former term.

$$\begin{aligned}
&\sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\| \, 1\{s_1 \in \tilde{\mathcal{T}}, s_2 \in \tilde{\mathcal{T}}, (s_1, s_2) \in \tilde{E}_{\mathcal{T}}, r(s_1) \wedge r(s_2) \le r^*\} \\
&\qquad\leq \sum_{(x_1,\dots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\| \, 1\{r(s_1) \wedge r(s_2) \le r^*\}.
\end{aligned}$$

Take an expectation to compute

$$\begin{aligned}
&\mathbb{E}_{x_{1:n}}\Bigg[\sum_{(x_1,\ldots,x_{d+2})\in\mathscr{C}^{d+2}(x_{1:n})} \|R_{x_{2:(d+2)}}\|\ 1\{x_2,\ldots,x_{d+2}\in B(x_1,r^*)\}\Bigg] \\
&\quad= n^{d+2}\mathbb{E}_{x_{1:(d+2)}}\mathbb{E}\big[\|R_{x_{2:(d+2)}}\|\ 1\{x_2,\ldots,x_{d+2}\in B(x_1,r^*)\}\big] \\
&\quad\le n^{d+2}p_{\max}^{d+2}\int_\Omega\int_\Omega\cdots\int_\Omega \|R_{x_{2:(d+2)}}\|\ 1\{x_2,\ldots,x_{d+2}\in B(x_1,r^*)\}dx_{d+2}\ldots dx_2dx_1 \\
&\quad= n^{d+2}p_{\max}^{d+2}\int_\Omega\int_{B(x_1,r^*)}\cdots\int_{B(x_1,r^*)} \|R_{x_{2:(d+2)}}\|\ dx_{d+2}\ldots dx_2dx_1 \\
&\quad= n^{d+2}p_{\max}^{d+2}\mu(B(\cdot,r^*))^{d+1} \\
&\qquad\qquad\times\int_\Omega\int_{B(x_1,r^*)}\cdots\int_{B(x_1,r^*)} \|R_{x_{2:(d+2)}}\|\mu(B(\cdot,r^*))^{-d-1}dx_{d+2}\ldots dx_2dx_1 \\
&\quad= \frac{2^{d^2+d}p_{\max}^{d+2}}{p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n\int_\Omega \mathbb{E}_{\substack{x_2,\ldots,x_{d+2}\\ \sim\mathrm{Unif}(B(x_1,r^*))}}\big[\|R_{x_{2:(d+2)}}\|\big]dx_1. \qquad \text{(B.37)}
\end{aligned}$$

We now analyze the inner expectation.

$$\begin{aligned}
&\mathbb{E}_{\substack{x_2,\ldots,x_{d+2}\\ \sim\mathrm{Unif}(B(x_1,r^*))}}\big[\|R_{x_{2:(d+2)}}\|\big] \\
&= \mathbb{E}_{\substack{x_2,\ldots,x_{d+2}\\ \sim\mathrm{Unif}(B(x_1,r^*))}}\big[\sum_{i=2}^{d+2}|R_{x_i}|\big] \\
&= (d+1)\mathbb{E}_{x\sim\mathrm{Unif}(B(x_1,r^*))}\Big[\big|\sum_{|\beta|=2}(x-x_1)^\beta R_\beta(x)\big|\Big] \\
&\le (d+1)\sum_{|\beta|=2}\mathbb{E}_{x\sim\mathrm{Unif}(B(x_1,r^*))}\Big[\|x-x_1\|_2^2|R_\beta(x)|\Big] \\
&\le (d+1)(r^*)^2\sum_{|\beta|=2}(\mu_d(r^*)^d)^{-1}\int_{B(x_1,r^*)}\left|\int_0^1(1-t)D^\beta f(x_1+t(x-x_1))dt\right|dx \\
&\le \mu_d^{-1}(d+1)(r^*)^{2-d}\sum_{|\beta|=2}\int_0^1\int_{B(x_1,r^*)}|D^\beta f(x_1+t(x-x_1))|dxdt \\
&= (\star)
\end{aligned}$$

Employ a change of variables: $t=t$, $z=x_1+t(x-x_1)$, and rename z back to x to obtain

$$\begin{aligned}
(\star) &\le \mu_d^{-1}(d+1)(r^*)^{2-d}\int_0^1\frac{1}{t^d}\int_{B(x_1,tr^*)}\sum_{|\beta|=2}|D^\beta f(x)|dxdt \\
&\le \mu_d^{-1}(d^2+d)(r^*)^{2-d}\int_0^1\frac{1}{t^d}\int_{B(x_1,tr^*)}\|Hf(x)\|_F dxdt. \qquad \text{(B.38)}
\end{aligned}$$

Finally, we substitute (B.38) back into (B.37).

$$\begin{aligned}
\text{(B.37)} &= \frac{2^{d^2+d}(d^2+d)p_{\max}^{d+2}}{\mu_d p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n(r^*)^{2-d}\int_\Omega\int_0^1\frac{1}{t^d}\int_{B(x_1,tr^*)}\|Hf(x)\|_F dx dt dx_1 \\
&= \frac{2^{d^2+d}(d^2+d)p_{\max}^{d+2}}{\mu_d p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n(r^*)^{2-d}\int_0^1\frac{1}{t^d}\int_\Omega\int_{B(x,tr^*)}\|Hf(x)\|_F dx_1 dx dt \\
&= \frac{2^{d^2+d}(d^2+d)p_{\max}^{d+2}}{p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n(r^*)^{2}\int_0^1\int_\Omega\|Hf(x)\|_F dx dt \\
&= \frac{2^{d^2+d}(d^2+d)p_{\max}^{d+2}}{p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n(r^*)^{2}\,\mathrm{TV}(\nabla f;\Omega) \\
&= \frac{2^{d^2+d}(d^2+d)p_{\max}^{d+2}}{p_{\min}^{d+1}}\alpha^{d+1}(\log n)^{d+1}n(r^*)^{2}\,\mathrm{TV}(\nabla f;\Omega).
\end{aligned}$$

Apply Markov's inequality and the weights w_{ij} to obtain the claim. □

Normed discrete second difference

Lemma 30. *Given a twice-differentiable function $f:\Omega\to\mathbb{R}$ with a convex domain $\Omega\subseteq\mathbb{R}^d$ and points $x_1,\dots,x_{d+2}\in\Omega$, form the simplices*

$$\begin{aligned}
s_1 &= (x_1, x_3, \dots, x_{d+2}),\\
s_2 &= (x_2, x_3, \dots, x_{d+2}),
\end{aligned}$$

and the approximate gradients $\hat{g}(s_1),\hat{g}(s_2)$ by linearly interpolating f on those simplices. Then the normed discrete second difference satisfies

$$\|\hat{g}(s_1)-\hat{g}(s_2)\| \le \left(\frac{1}{\sqrt{\lambda_{\min}(M_1^\top M_1)}}+\frac{1}{\sqrt{\lambda_{\min}(M_2^\top M_2)}}\right)\|R_{x_{2:(d+2)}}\|, \tag{B.39}$$

where $M_1, M_2\in\mathbb{R}^{(d+1)\times(d+1)}$ are given by

$$M_1 = \begin{bmatrix} 1 & 0_{1\times d}\\ 1 & x_3-x_1\\ \vdots & \vdots\\ 1 & x_{d+2}-x_1\end{bmatrix} \qquad M_2 = \begin{bmatrix} 1 & 0_{1\times d}\\ 1 & x_3-x_2\\ \vdots & \vdots\\ 1 & x_{d+2}-x_2\end{bmatrix},$$

and the remainder term $R_{x_{2:(d+2)}}$ is given by

$$R_{x_{2:(d+2)}} = \begin{bmatrix} \sum_{|\beta|=2}(x_2-x_1)^\beta R_\beta(x_2)\\ \vdots\\ \sum_{|\beta|=2}(x_{d+2}-x_1)^\beta R_\beta(x_{d+2})\end{bmatrix}$$

with

$$R_\beta(x) = \int_0^1 (1-t) D^\beta f(x_1 + t(x - x_1))dt.$$

Proof. Consider the second-order multivariate Taylor expansion of f about $x = x_1$,

$$\begin{aligned} f(x_1) &= f(x_1) \\ f(x_2) &= f(x_1) + \nabla f(x_1)^\top (x_2 - x_1) + \sum_{|\beta|=2} (x_2 - x_1)^\beta R_\beta(x_1) \\ &\vdots \\ f(x_{d+2}) &= f(x_1) + \nabla f(x_1)^\top (x_{d+2} - x_1) + \sum_{|\beta|=2} (x_{d+2} - x_1)^\beta R_\beta(x_{d+2}). \end{aligned} \tag{B.40}$$

The first and latter d equations in (B.40) may be rewritten,

$$f(x_{1,3:(d+2)}) = M_1 \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + R_{x_{1,3:(d+2)}}.$$

Therefore the approximate gradient $\hat{g}(s_1)$ can be related to the gradient at x_1,

$$\hat{g}(s_1) = \nabla f(x_1) + (M_1^{-1} R_{x_{1,3:(d+2)}})_{2:(d+1)} \tag{B.41}$$

For the approximate gradient $\hat{g}(s_2)$, first consider the latter $d+1$ equations in (B.40) as

$$\begin{aligned} f(x_{2:(d+2)}) &= \begin{bmatrix} 1 & x_2 - x_1 \\ \vdots & \vdots \\ 1 & x_{d+2} - x_1 \end{bmatrix} \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + R_{x_{2:(d+2)}} \\ &= \begin{bmatrix} 1 & 0_{1\times d} \\ 1 & x_3 - x_2 \\ \vdots & \vdots \\ 1 & x_{d+2} - x_2 \end{bmatrix} \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + \begin{bmatrix} 0 & x_2 - x_1 \\ 0 & x_2 - x_1 \\ \vdots & \vdots \\ 0 & x_2 - x_1 \end{bmatrix} \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + R_{x_{2:(d+2)}}. \end{aligned}$$

Applying M_2^{-1} to both sides then gives

$$\begin{aligned} M_2^{-1} f(x_{2:(d+2)}) &= \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + M_2^{-1} \begin{bmatrix} 0 & x_2 - x_1 \\ 0 & x_2 - x_1 \\ \vdots & \vdots \\ 0 & x_2 - x_1 \end{bmatrix} \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + M_2^{-1} R_{x_{2:(d+2)}} \\ &= \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + \begin{bmatrix} 0 & x_2 - x_1 \\ 0 & 0_{1\times d} \\ \vdots & \vdots \\ 0 & 0_{1\times d} \end{bmatrix} \begin{bmatrix} f(x_1) \\ \nabla f(x_1) \end{bmatrix} + M_2^{-1} R_{x_{2:(d+2)}}, \end{aligned}$$

since the all-ones vector is the first column of M_2. This yields, for the approximate gradient $\hat{g}(s_2)$,

$$\hat{g}(s_2) = \nabla f(x_1) + (M_2^{-1} R_{x_{2:(d+2)}})_{2:(d+1)}. \tag{B.42}$$

Finally, combine (B.41) and (B.42) to obtain a bound on the normed discrete second difference,

$$\begin{aligned}
\|\hat{g}(s_1) - \hat{g}(s_2)\| &= \|(M_1^{-1} R_{x_{1,3:(d+2)}} - M_2^{-1} R_{x_{2:(d+2)}})_{2:(d+1)}\| \\
&\leq \|M_1^{-1} R_{x_{1,3:(d+2)}}\| + \|M_2^{-1} R_{x_{2:(d+2)}}\| \\
&\leq \left(\frac{1}{\sqrt{\lambda_{\min}(M_1^\top M_1)}} + \frac{1}{\sqrt{\lambda_{\min}(M_2^\top M_2)}} \right) \|R_{x_{2:(d+2)}}\|.
\end{aligned}$$

□

Minimum eigenvalue of the Gram matrix

We begin by recalling two useful results. The first is a lower bound on the minimum eigenvalue of a Wishart matrix.

Theorem 9 (Edelman, 1988; Lemma 8.2)**.** *Suppose $Y \in \mathbb{R}^{d \times d}$ has columns $Y_i \sim N(0, I_d)$. As $\delta_1 \to 0$, the minimum eigenvalue of $Y^\top Y$ satisfies*

$$\mathbb{P}\big\{ \lambda_{\min}(Y^\top Y) < \delta_1 \big\} \sim \sqrt{d\delta_1}. \tag{B.43}$$

The second useful result we recall is a quantitative form of Sylvester's law of inertia.

Theorem 10 (Ostrowski, 1959)**.** *Let A be a Hermitian matrix of order n and B be an arbitrary real nonsingular matrix of the same order. Denote the eigenvalues of A by*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n,$$

and those of $B^* AB$ by

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_n.$$

Then

$$\rho_i = \xi_i \lambda_i, \qquad i = 1, \ldots, n, \tag{B.44}$$

where ξ_i lie between the largest eigenvalue ν_1 and the smallest eigenvalue ν_n of the positive definite matrix $B^* B$.

We now state a few useful lemmas.

Lemma 31. *The random vector $X_i \sim \mathrm{Unif}(B_d(0,1))$ drawn from the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ factorizes into independent scale and orientation*

components,

$$\begin{aligned} &X_i = R_i\theta_i, \\ &R_i \text{ has density } f(r) = dr^{d-1}, \\ &\theta_i \sim \mathrm{Unif}(S^{d-1}), \\ &R_i \perp\!\!\!\perp \theta_i. \end{aligned} \tag{B.45}$$

Lemma 32. *The random vector $Y_i \sim N(0, I_d)$ drawn from an isotropic Gaussian distribution factorizes into independent scale and orientation components,*

$$\begin{aligned} &Y_i = S_i\varphi_i, \\ &S_i \sim \chi_d, \\ &\varphi_i \sim \mathrm{Unif}(S^{d-1}), \\ &S_i \perp\!\!\!\perp \varphi_i. \end{aligned} \tag{B.46}$$

Lemma 33. *There exists a scaling function $g : [0, +\infty) \to [0, +\infty)$ such that for $X_i \sim \mathrm{Unif}(B_d(0, 1))$ and $Y_i \sim N(0, I_d)$,*

$$Y_i \cdot g(\|Y_i\|) \overset{d}{=} X_i. \tag{B.47}$$

Proof. Lemmas 31 and 32 allow us to factorize X_i and Y_i into independent scale and orientation components, as in (B.45) and (B.46), respectively. Rewrite

$$Y_i \cdot g(\|Y_i\|) = g(S_i) \cdot S_i \cdot \varphi_i,$$

and observe that $\varphi_i \overset{d}{=} \theta_i$ by construction. It only remains to verify the existence of $g : [0, +\infty) \to [0, +\infty)$ such that

$$g(S_i) \cdot S_i \overset{d}{=} R_i.$$

We construct such a function directly, by taking

$$g(s) = \frac{\mathrm{Quantile}_R(F_S(s))}{s},$$

where $\mathrm{Quantile}_R$ is the quantile function for the distribution with density function $f(r) = dr^{d-1}$ and F_S is the cumulative distribution function for the chi distribution with d degrees of freedom. □

Corollary 3. *Suppose $Y \in \mathbb{R}^{d\times d}$ is a random matrix whose columns Y_i are drawn independently from the isotropic Gaussian distribution $N(0, I_d)$. Then there exists a diagonal matrix D_Y (depending on Y) such that $Y D_Y$ has columns $(Y D_Y)_i$ drawn independently from the uniform distribution on the unit Euclidean ball $\mathrm{Unif}(B_d(0, 1))$.*

Lemma 34. *Suppose a matrix $X \in \mathbb{R}^{d\times d}$ with columns $X_i \sim \mathrm{Unif}(B_d(0,1))$ independently drawn from the uniform distribution on the unit Euclidean ball. There exist constants C_1, C_2 depending only on d such that for sufficiently small $\delta_1 \in (0,1)$ and any $\delta_2 \in (0,1)$, the minimum eigenvalue of its Gram matrix satisfies*

$$\lambda_{\min}(X^\top X) \geq (\delta_1\delta_2)^2 \tag{B.48}$$

with probability at least $1 - C_1\delta_1 - d\delta_2^{d/2} - C_2\exp\{-\delta_2^{-1}\}$.

Proof. Recall from Corollary 3 that under a column scaling,

$$X \overset{d}{=} YD_Y,$$

where the matrix $Y \in \mathbb{R}^{d\times d}$ has columns drawn independently from the isotropic Gaussian $N(0, I_d)$ (and D_Y is a diagonal matrix that depends on Y). Applying Ostrowski's theorem (B.44), we lower bound

$$\begin{aligned}\lambda_{\min}(X^\top X) &= \lambda_{\min}(D_Y^\top Y^\top Y D_Y)\\ &\geq \big(\min_i (D_Y)_{ii}\big)^2 \lambda_{\min}(Y^\top Y).\end{aligned} \tag{B.49}$$

We now lower bound (B.49) by $\delta_1\delta_2$ in probability. Decompose via the union bound,

$$\begin{aligned}&\mathbb{P}\bigg(\Big\{\big(\min_i (D_Y)_{ii}\big)^2 < \delta_2\Big\} \cup \Big\{\lambda_{\min}(Y^\top Y) < \delta_1\Big\}\bigg)\\ &\qquad\leq \mathbb{P}\Big\{\lambda_{\min}(Y^\top Y) < \delta_1\Big\} + \mathbb{P}\Big\{\big(\min_i (D_Y)_{ii}\big)^2 < \delta_2\Big\}\\ &\qquad=: p_0 + p_1.\end{aligned}$$

For p_0, we quantify from (B.43),

$$p_0 \leq C_3\sqrt{d\delta_1},$$

for sufficiently small δ_1. For p_1, we can apply the union bound, denoting $S \sim \chi_d$,

$$\begin{aligned}p_1 = \mathbb{P}\Big\{\big(\min_i (D_Y)_{ii}\big)^2 < \delta_2\Big\} &\leq d\cdot\mathbb{P}\{g(S) < \sqrt{\delta_2}\}\\ &= d\cdot\mathbb{P}\Big\{\mathrm{Quantile}_R(F_S(S))/S < \sqrt{\delta_2}\Big\}\\ &= d\cdot\mathbb{P}\Big\{(F_S(S))^{1/d}/S < \sqrt{\delta_2}\Big\}.\end{aligned}$$

Finally, condition to obtain the bound

$$\begin{aligned}
\mathbb{P}\left\{\frac{\big(F_S(S)\big)^{1/d}}{S} < \sqrt{\delta_2}\right\} &= \mathbb{P}\left\{\frac{\big(F_S(S)\big)^{1/d}}{S} < \sqrt{\delta_2}, S < \delta_3\right\} \\
&\qquad + \mathbb{P}\left\{\frac{\big(F_S(S)\big)^{1/d}}{S} < \sqrt{\delta_2}, S \geq \delta_3\right\} \\
&\leq \mathbb{P}\left\{\frac{\big(F_S(S)\big)^{1/d}}{S} < \sqrt{\delta_2} | S < \delta_3\right\} + \mathbb{P}\{S \geq \delta_3\} \\
&\leq \mathbb{P}\big\{F_S(S) < (\sqrt{\delta_2}\delta_3)^d | S < \delta_3\big\}\mathbb{P}\{S < \delta_3\} + \mathbb{P}\{S \geq \delta_3\} \\
&\leq \mathbb{P}\big\{F_S(S) < (\sqrt{\delta_2}\delta_3)^d\big\} + \mathbb{P}\{S \geq \delta_3\} \\
&\overset{(i)}{\leq} (\sqrt{\delta_2}\delta_3)^d + \exp\big\{-(\delta_3^2 - 2\delta d + d^2)/2\sigma^2\big\} \\
&\leq (\sqrt{\delta_2}\delta_3)^d + C_4 \exp\{-\delta_3^2\} \\
&\overset{(ii)}{\leq} \delta_2^{d/4} + C_4 \exp\{-\delta_2^{-1/2}\},
\end{aligned}$$

where in *(i)* we use the fact that a chi-distributed random variable with d degrees of freedom is sub-Gaussian with mean d and variance proxy σ^2, and in *(ii)* we set $\delta_3 = \delta_2^{-1/4}$. □

Lemma 35. *Suppose a matrix $\tilde{X} \in \mathbb{R}^{(d+1)\times(d+1)}$ of the form*

$$\tilde{X} = \begin{bmatrix} 1 & 1_{1\times d} \\ 0_{d\times 1} & X \end{bmatrix},$$

where $X \in \mathbb{R}^{d\times d}$ is positive definite and whose smallest singular value is $\sigma_{\min}(X) = \lambda^{1/2}$. Then the smallest singular value of $\tilde{X}$ satisfies

$$\sigma_{\min}(\tilde{X}) \geq \left(\frac{\lambda}{d+2}\right)^{1/2} \tag{B.50}$$

for sufficiently small $\lambda > 0$.

Proof. Let $v^* \in S^{d-1}$ such that $\|Xv^*\|_2^2 = \lambda$. Using the variational form of the singular

value,

$$\begin{aligned}
\sigma_{\min}^2(\tilde{X}) &= \inf_{\tilde{u}\in S^d} \|\tilde{X}\tilde{u}\|_2^2 \\
&= \inf_{\tilde{v}\in \mathrm{null}(1_{d+1}), \tilde{w}\in \mathrm{null}(\{0\}\times\mathbb{R}^d), \|\tilde{v}+\tilde{w}\|_2=1} \|\tilde{X}(\tilde{v}+\tilde{w})\|_2^2 \\
&= \inf_{\tilde{v}\in \mathrm{null}(1_{d+1}), w\in\mathbb{R}, \|\tilde{v}+we_1\|_2=1} \left\| \begin{bmatrix} 0 \\ X\tilde{v}_{2:(d+1)} \end{bmatrix} + \begin{bmatrix} w \\ 0_{d\times 1} \end{bmatrix} \right\|_2^2 \\
&= \inf_{\tilde{v}\in \mathrm{null}(1_{d+1}), w\in\mathbb{R}, \|\tilde{v}+we_1\|_2=1} \|X\tilde{v}_{2:(d+1)}\|_2^2 + w^2 \\
&\overset{(i)}{=} \inf_{\alpha^2+(w-\alpha 1^\top v^*)^2=1, \alpha\in[0,1]} \alpha^2\lambda + w^2 \\
&= \inf_{\alpha\in[0,1]} \alpha^2\lambda + \left((1-\alpha^2)^{1/2} + \alpha 1^\top v^*\right)^2 \\
&\overset{(ii)}{\geq} \inf_{\alpha\in[0,1]} \alpha^2\lambda + (1-\alpha+\alpha 1^\top v^*)^2 \\
&= \frac{\lambda}{\lambda + (1^\top v^* - 1)^2} \\
&\overset{(iii)}{\geq} \frac{\lambda}{1 + (1^\top v^* - 1)^2} \\
&\overset{(iv)}{\geq} \frac{\lambda}{d+2},
\end{aligned}$$

where in *(i)* we minimize $\|\tilde{X}\tilde{v}_{2:(d+1)}\|_2^2$ by taking $\tilde{v}_{2:(d+1)} = \alpha v^*$ for some $\alpha \in (0,1)$ and observe that for $\tilde{v} \in \mathrm{null}(1_{d+1})$, $\tilde{v}_1 = -1^\top \tilde{v}_{2:(d+1)}$; in *(ii)* we use the inequality $\sqrt{1-\alpha^2} \geq 1-\alpha$ for $\alpha \in [0,1]$; *(iii)* holds for $\lambda > 0$ sufficiently small; and *(iv)* follows from the upper bound $|1^\top v^*| \leq \|v^*\|_1 \leq \sqrt{d}$ for a vector $v^* \in S^{d-1}$. □

Lemma 36. *Suppose a matrix $M \in \mathbb{R}^{(d+1)\times(d+1)}$ of the form*

$$M = \begin{bmatrix} 1 & 0_{1\times d} \\ 1_{d\times 1} & X \end{bmatrix},$$

where $X \in \mathbb{R}^{d\times d}$ has columns $X_i \sim \mathrm{Unif}(B_d(0,r))$ independently drawn from the uniform distribution on the unit Euclidean ball. There exist constants C_1, C_2 depending only on d such that for sufficiently small $\delta_1 \in (0,1)$ and any $\delta_2 \in (0,1)$, the minimum eigenvalue of the Gram matrix $M^\top M$ satisfies

$$\lambda_{\min}(M^\top M) \geq \frac{(r\delta_1\delta_2)^2}{d+2} \tag{B.51}$$

with probability at least $1 - C_1\delta_1 - d\delta_2^{d/2} - C_2\exp\{-\delta_2^{-1}\}$.

Proof. The result follows from Lemma 34 with X/r and Lemma 35. □

B.3 Sensitivity analysis for Section 3.3

In this section, we additionally report results from the experiments of Section 3.4.1 using the smaller sample sizes of $n = 500, 1000$. We find that the conclusions of Section 3.4.1 remain unchanged. The figures here correspond to Figures 3.9 and 3.10.

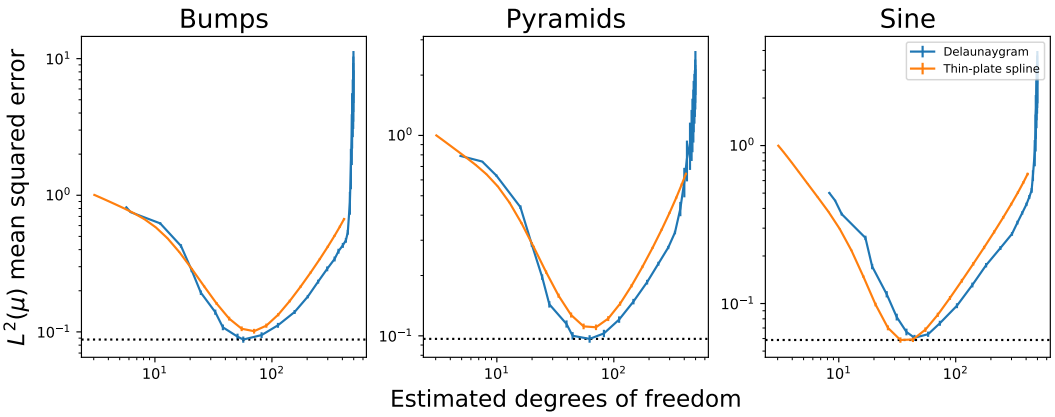

Figure B.1: *MSE for* $n = 500$. *Compare these results to those in Figure 3.9.*

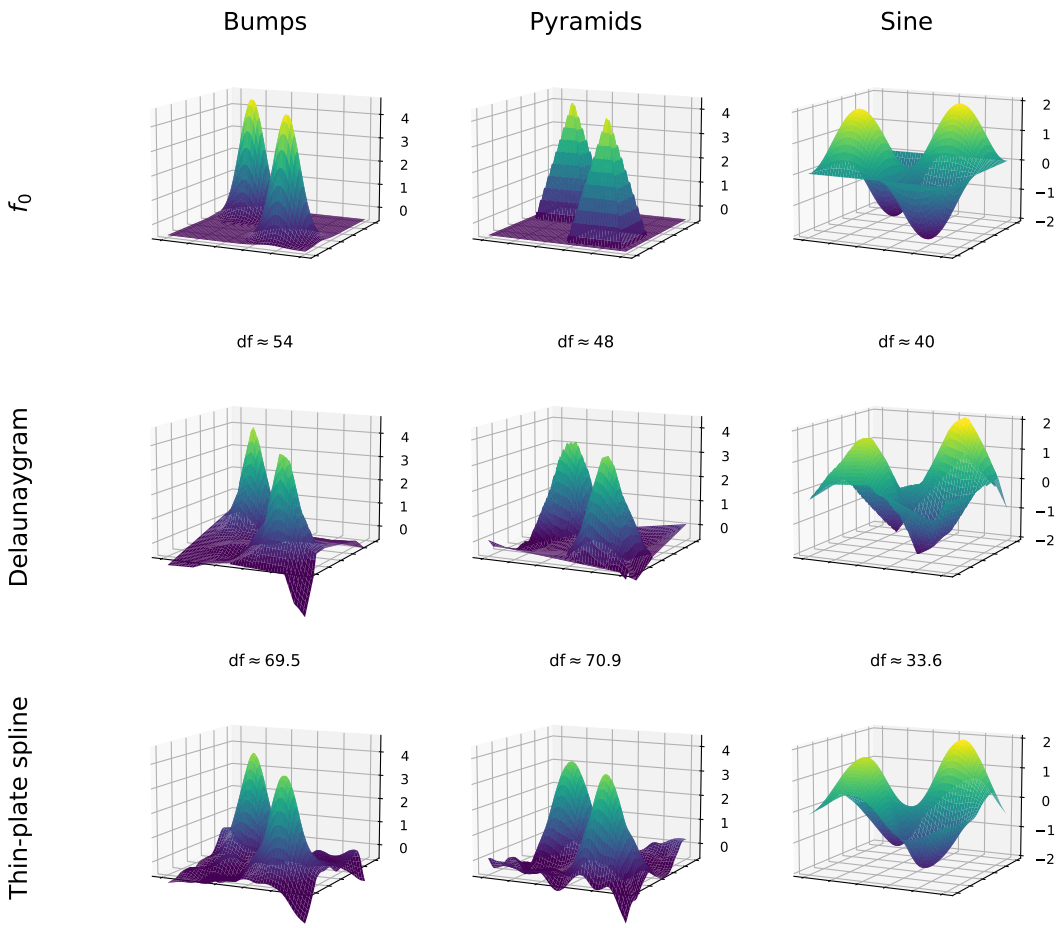

Figure B.2: *Predictions for* $n = 500$. *Compare these plots to those in Figure 3.10.*

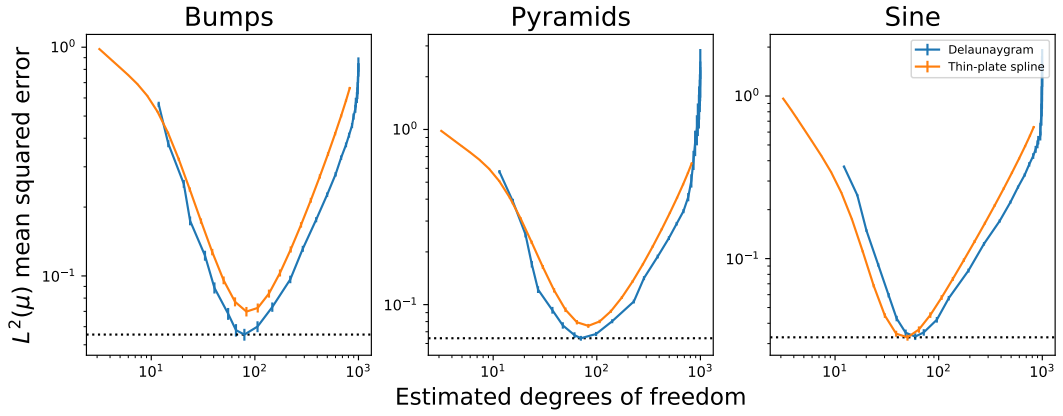

Figure B.3: *MSE for* $n = 1000$. *Compare these results to those in Figure 3.9.*

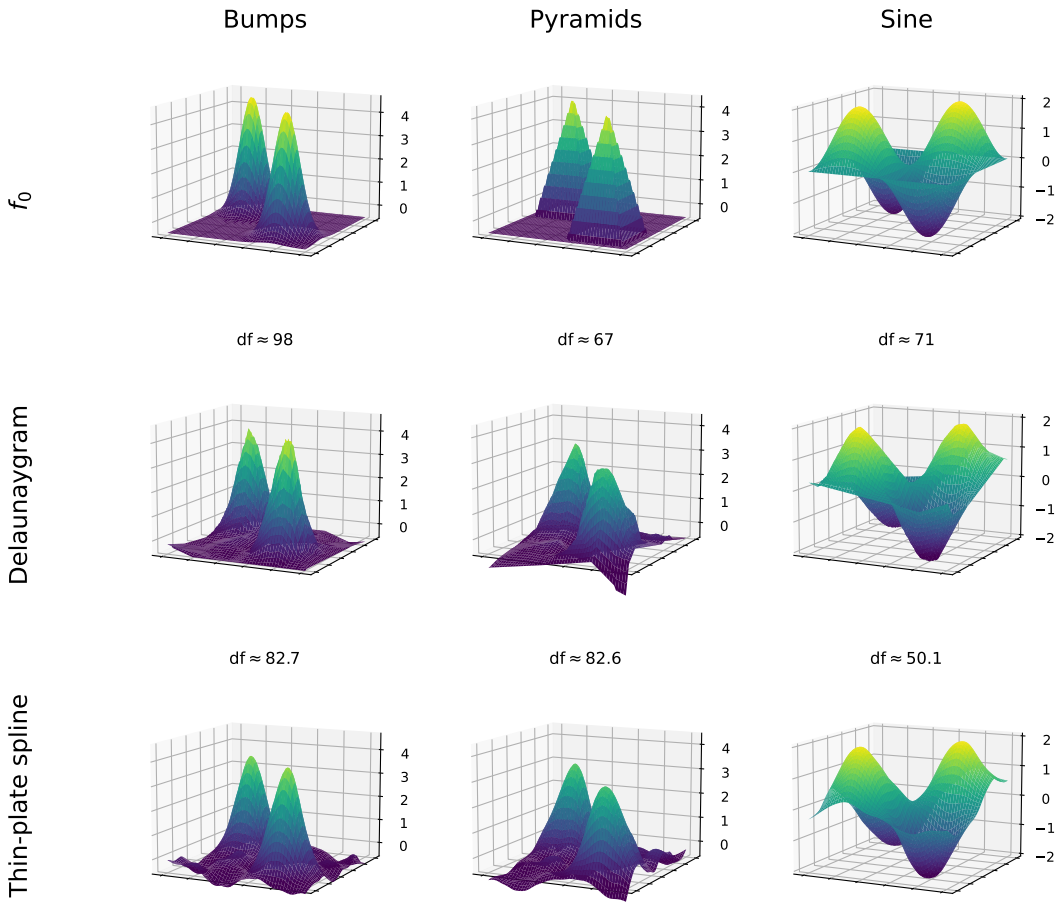

Figure B.4: *Predictions for* $n = 1000$. *Compare these plots to those in Figure 3.10.*