Adaptive Non-Parametric Regression With the K-NN Fused Lasso

Oscar Hernan Madrid Padilla[1]
omadrid@berkeley.edu

James Sharpnack[2]
jsharpna@ucdavis.edu

Yanzhen Chen [5]
imyanzhen@ust.hk

Daniela Witten[3] [4]
dwitten@uw.edu

[1] Department of Statistics, UCLA
[2] Department of Statistics, UC Davis
[3] Department of Statistics, University of Washington
[4] Department of Biostatistics, University of Washington
[5] Department of ISOM, Hong Kong University of Science and Technology

July 7, 2019

Abstract

The fused lasso, also known as total-variation denoising, is a locally-adaptive function estimator over a regular grid of design points. In this paper, we extend the fused lasso to settings in which the points do not occur on a regular grid, leading to an approach for non-parametric regression. This approach, which we call the *K-nearest neighbors (K-NN) fused lasso*, involves (i) computing the K-NN graph of the design points; and (ii) performing the fused lasso over this K-NN graph. We show that this procedure has a number of theoretical advantages over competing approaches: specifically, it inherits *local adaptivity* from its connection to the fused lasso, and it inherits *manifold adaptivity* from its connection to the K-NN approach. We show that excellent results are obtained in a simulation study and on an application to flu data. For completeness, we also study an estimator that makes use of an ϵ-graph rather than a K-NN graph, and contrast this with the K-NN fused lasso.

Keywords: non-parametric regression, local adaptivity, manifold adaptivity, total variation, fused lasso

1 Introduction

In this paper, we consider the non-parametric regression setting in which we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$, of the pair of random variables $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a metric space with metric $d_{\mathcal{X}}$. We suppose that the model

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

holds, where $f_0 : \mathcal{X} \to \mathbb{R}$ is an unknown function that we wish to estimate. This problem arises in many settings, including demographic applications (Petersen et al., 2016a; Sadhanala and Tibshirani, 2017), environmental data analysis (Hengl et al., 2007), image processing (Rudin et al., 1992), and causal inference (Wager and Athey, 2018).

A substantial body of work has considered estimating the function f_0 in (1) at the observations $X = x_1, \ldots, x_n$ (i.e. denoising) as well as at other values of the random variable X (i.e. prediction). This includes

seminal papers by Breiman et al. (1984), Duchon (1977), and Friedman (1991), as well as more recent work by Petersen et al. (2016b), Petersen et al. (2016a), and Sadhanala and Tibshirani (2017). A number of previous papers have focused in particular on manifold adaptivity, i.e, adapting to the dimensionality of the data; these include work on local polynomial regression by Bickel and Li (2007) and Cheng and Wu (2013), K-NN regression by Kpotufe (2011), Gaussian processes by Yang et al. (2015) and Yang and Dunson (2016), and tree–based estimators such as those in Kpotufe (2009) and Kpotufe and Dasgupta (2012). We refer the reader to Györfi et al. (2006) for a very detailed survey of other classical non-parametric regression methods. The vast majority of this work performs well in function classes with variation controlled uniformly throughout the domain, such as Lipschitz and L_2 Sobolev classes. Donoho and Johnstone (1998) and Härdle et al. (2012) generalize this setting by considering functions of bounded variation and Besov classes. In this work, we focus on piecewise Lipschitz and bounded variation functions, as these classes can have functions with non-smooth regions as well as smooth regions (Wang et al., 2016).

Recently, interest has focused on so-called *trend filtering* (Kim et al., 2009), which seeks to estimate $f_0(\cdot)$ under the assumption that its discrete derivatives are sparse, in a setting in which we have access to an unweighted graph that quantifies the pairwise relationships between the n observations. In particular, the fused lasso, also known as zeroth-order trend filtering or total variation denoising (Mammen and van de Geer, 1997; Rudin et al., 1992; Tibshirani et al., 2005; Wang et al., 2016), solves the optimization problem

$$\text{minimize}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{(i,j) \in E} |\theta_i - \theta_j| \right\}, \tag{2}$$

where λ is a non-negative tuning parameter, and where $(i, j) \in E$ if and only if there is an edge between the ith and jth observations in the underlying graph. Then, $\hat{f}(x_i) = \hat{\theta}_i$. Computational aspects of the fused lasso have been studied extensively in the case of chain graphs (Johnson, 2013; Barbero and Sra, 2014; Davies and Kovac, 2001) as well as general graphs (Chambolle and Darbon, 2009; Chambolle and Pock, 2011; Landrieu and Obozinski, 2015; Hoefling, 2010; Tibshirani and Taylor, 2011; Chambolle and Darbon, 2009). Furthermore, the fused lasso is known to have excellent theoretical properties. In one dimension, Mammen and van de Geer (1997) and Tibshirani (2014) showed that the fused lasso attains nearly minimax rates in mean squared error (MSE) for estimating functions of bounded variation. More recently, also in one dimension, Lin et al. (2017) and Guntuboyina et al. (2017) independently proved that the fused lasso is nearly minimax under the assumption that f_0 is piecewise constant. In grid graphs, Hutter and Rigollet (2016), Sadhanala et al. (2016), and Sadhanala et al. (2017) proved minimax results for the fused lasso when estimating signals of interest in applications of image denoising. In more general graph structures, Padilla et al. (2018) showed that the fused lasso is consistent for denoising problems, provided that the underlying signal has total variation along the graph that divided by n goes to zero. Other graph models that have been studied in the literature include tree graphs in Padilla et al. (2018) and Ortelli and van de Geer (2018), and star and Erdős-Rényi graphs in Hutter and Rigollet (2016).

In this paper, we extend the utility of the fused lasso approach by combining it with the *K-nearest neighbors* (K-NN) procedure. K-NN has been well-studied from a theoretical (Stone, 1977; Chaudhuri and Dasgupta, 2014; Von Luxburg et al., 2014; Alamgir et al., 2014), methodological (Dasgupta, 2012; Kontorovich et al., 2016; Singh and Póczos, 2016; Dasgupta and Kpotufe, 2014), and algorithmic (Friedman et al., 1977; Dasgupta and Sinha, 2013; Zhang et al., 2012) perspective. One key feature of K-NN methods is that they automatically have a finer resolution in regions with a higher density of design points; this is particularly consequential when the underlying density is highly non-uniform. We study the extreme case in which the data are supported over multiple manifolds of mixed intrinsic dimension. An estimator that adapts to this setting is said to achieve *manifold adaptivity*.

In this paper, we exploit recent theoretical developments in the fused lasso and the K-NN procedure in order to obtain a single approach that inherits the advantages of both methods. In greater detail, we extend

the fused lasso to the general non-parametric setting of (1), by performing a two-step procedure.

Step 1. We construct a K-nearest-neighbor (K-NN) graph, by placing an edge between each observation and the K observations to which it is closest, in terms of the metric $d_{\mathcal{X}}$.

Step 2. We apply the fused lasso to this K-NN graph.

The resulting *K-NN fused lasso* (K-NN-FL) estimator appeared in the context of image processing in Elmoataz et al. (2008) and Ferradans et al. (2014), and more recently in an application of graph trend filtering in Wang et al. (2016). We are the first to study its theoretical properties. We also consider a variant obtained by replacing the K-NN graph in Step 1 with an ϵ-nearest-neighbor (ϵ-NN) graph, which contains an edge between x_i and x_j only if $d_{\mathcal{X}}(x_i, x_j) < \epsilon$.

The main contributions of this paper are as follows:

Local adaptivity. We show that provided that f_0 has bounded variation, along with an additional condition that generalizes piecewise Lipschitz continuity (Assumption 5), then the mean squared errors of both the K-NN-FL estimator and the ϵ-NN-FL estimator scale like $n^{-1/d}$, ignoring logarithmic factors; here, $d > 1$ is the dimension of $\mathcal{X}$. In fact, this matches the minimax rate for estimating a two-dimensional Lipschitz function (Györfi et al., 2006), but over a much wider function class.

Manifold adaptivity. Suppose that the covariates are i.i.d. samples from a mixture model $\sum_{l=1}^{\ell} \pi_l^* p_l$, where $p_1, \ldots, p_\ell$ are unknown bounded densities, and the weights $\pi_l^* \in [0,1]$ satisfy $\sum_{l=1}^{\ell} \pi_l^* = 1$. Suppose further that for $l = 1, \ldots, \ell$, the support $\mathcal{X}_l$ of p_l is homeomorphic (see Assumption 3) to $[0,1]^{d_l} = [0,1] \times [0,1] \times \ldots \times [0,1]$, where $d_l > 1$ is the intrinsic dimension of $\mathcal{X}_l$. We show that under mild conditions, if the restriction of f_0 to $\mathcal{X}_l$ is a function of bounded variation, then the K-NN-FL estimator attains the rate $\sum_{l=1}^{\ell} \pi_l^* (\pi_l^* n)^{-1/d_l}$. We can obtain intuition for this rate by noticing that $\pi_l^* n$ is the expected number of samples from the lth component, and hence $(\pi_l^* n)^{-1/d_l}$ is the expected rate for the lth component. Therefore, our rate is the weighted average of the expected rates for the different components.

2 Methodology

2.1 The K-Nearest-Neighbor and ϵ-Nearest-Neighbor Fused Lasso

Both the K-NN-FL and ϵ-NN-FL approaches are simple two-step procedures. The first step involves constructing a graph on the n observations. The K-NN graph, $G_K = (V, E_K)$, has vertex set $V = \{1, \ldots, n\}$, and its edge set E_K contains the pair (i, j) if and only if x_i is among the K-nearest neighbors (with respect to the metric $d_{\mathcal{X}}$) of x_j, or vice versa. By contrast, the ϵ-graph, $G_\epsilon = (V, E_\epsilon)$, contains the edge (i, j) in E_ϵ if and only if $d_{\mathcal{X}}(x_i, x_j) < \epsilon$.

After constructing the graph, we apply the fused lasso to $y = \begin{pmatrix} y_1 \ldots y_n \end{pmatrix}^T$ over the graph G (either G_K or G_ϵ). We can re-write the fused lasso optimization problem (2) as

$$\hat{\theta} \;=\; \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 \;+\; \lambda \left\| \nabla_G \theta \right\|_1 \right\}, \tag{3}$$

where $\lambda > 0$ is a tuning parameter, and ∇_G is an oriented incidence matrix of G; each row of ∇_G corresponds to an edge in G. For instance, if the kth edge in G connects the ith and jth observations, then

$$(\nabla_G)_{k,l} = \begin{cases} 1 & \text{if } l = i \\ -1 & \text{if } l = j \\ 0 & \text{otherwise} \end{cases},$$

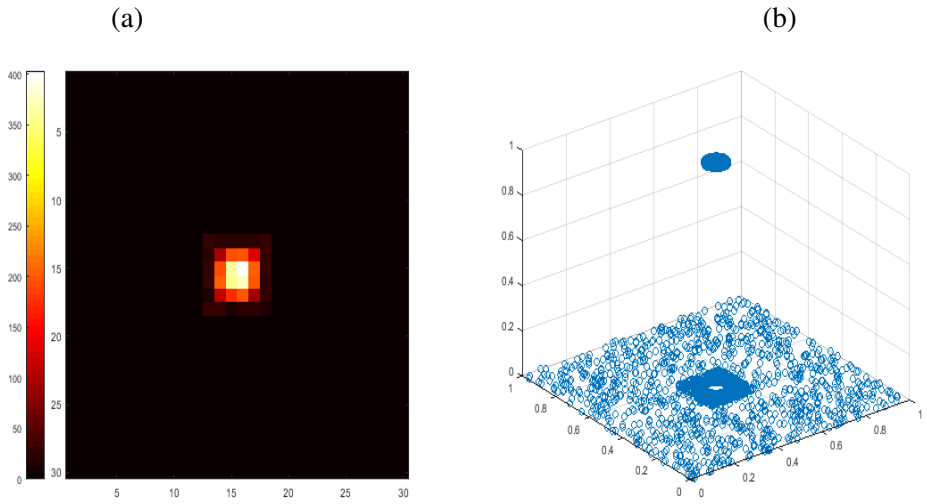

Figure 1: *(a):* A heatmap of $n = 5000$ draws from (5). *(b):* $n = 5000$ samples generated as in (1), with $\varepsilon_i \overset{\text{ind}}{\sim} N(0, 0.5)$, X has probability density function as in (5), and f_0 is given in (6). The vertical axis corresponds to $f_0(x_i)$, and the other two axes display the two covariates.

and so $(\nabla_G \theta)_k = \theta_i - \theta_j$. This definition of ∇_G implicitly assumes an ordering of the nodes and edges, which may be chosen arbitrarily without loss of generality. In this paper, we mostly focus on the setting where $G = G_K$ is the K-NN graph. We also include an analysis of the ϵ-graph, which results from using $G = G_\epsilon$, as a point of contrast.

Given the estimator $\hat{\theta}$ defined in (3), we predict the response at a new observation $x \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$ according to

$$\hat{f}(x) = \frac{1}{\sum_{j=1}^n k(x_j, x)} \sum_{i=1}^n \hat{\theta}_i \, k(x_i, x). \tag{4}$$

In the case of K-NN-FL, we take $k(x_i, x) = \mathbf{1}_{\{x_i \in \mathcal{N}_K(x)\}}$, where $\mathcal{N}_K(x)$ is the set of K nearest neighbors of x in the training data. In the case of ϵ-NN-FL, we take $k(x_i, x) = \mathbf{1}_{\{d_{\mathcal{X}}(x_i, x) < \epsilon\}}$. (Given a set A, $\mathbf{1}_A(x)$ is the indicator function that takes on a value of 1 if $x \in A$, and 0 otherwise.) Note that for the ϵ-NN-FL estimator, the prediction rule in (4) might not be well-defined if all the training points are farther than ϵ from x. When that is the case, we set $\hat{f}(x)$ to equal the fitted value of the nearest training point.

We construct the K-NN and ϵ-NN graphs using standard `Matlab` functions such as `knnsearch` and `bsxfun`; this results in a computational complexity of $O(n^2)$. We solve the fused lasso with the parametric max-flow algorithm from Chambolle and Darbon (2009), for which software is available from the authors' website, `http://www.cmap.polytechnique.fr/~antonin/software/`; it is in practice much faster than its worst-case complexity of $O(m\, n^2)$, where m is the number of edges in the graph (Boykov and Kolmogorov, 2004; Chambolle and Darbon, 2009).

In ϵ-NN and K-NN, the values of ϵ and K directly affect the sparsity of the graphs, and hence the computational performance of the ϵ-NN-FL and K-NN-FL estimators. Corollary 3.23 in Miller et al. (1997) provides an upper bound on the maximum degree of arbitrary K-NN graphs in $\mathbb{R}^d$.

2.2 Example

To illustrate the main advantages of K-NN-FL, we construct a simple example. We refer to the ability to adapt to the local smoothness of the regression function as local adaptivity, and the ability to adapt to the density of the design points as manifold adaptivity. The performance gains of K-NN-FL are most pronounced when these two effects happen in concert, namely, when the regression function is less smooth where design points are denser. These properties are manifested in the following example.

We generate $X \in \mathbb{R}^2$ according to the probability density function

$$p(x) \;=\; \frac{1}{5}\,\mathbf{1}_{\{[0,1]^2\setminus[0.4,0.6]^2\}}(x) \;+\; \frac{16}{25}\,\mathbf{1}_{\{[0.45,0.55]^2\}}(x) \;+\; \frac{4}{25}\,\mathbf{1}_{\{[0.4,0.6]^2\setminus[0.45,0.55]^2\}}(x). \tag{5}$$

Thus, p concentrates 64% of its mass in the small interval $[0.45, 0.55]^2$, and 80% in $[0.4, 0.6]^2$. The left-hand panel of Figure 1 displays a heatmap of $n = 5000$ observations drawn from (5).

We define $f_0 : \mathbb{R}^2 \to \mathbb{R}$ in (1) to be the piecewise constant function

$$f_0(x) \;=\; \mathbf{1}_{\left\{\left\|x-\frac{1}{2}(1,1)^T\right\|_2^2 \le \frac{2}{1000}\right\}}(x). \tag{6}$$

We then generate $\{(x_i, y_i)\}_{i=1}^n$ with $n = 5000$ from (1); the regression function is displayed in the right-hand panel of Figure 1. This simulation study has the following characteristics: (a) the function f_0 in (6) is not Lipschitz, but does have low total variation, and (b) the probability density function p is non-uniform with higher density in the region where f_0 is less smooth.

We compared the following methods in this example:

1. K-NN-FL, with the number of neighbors set to $K = 5$, and the tuning parameter λ chosen to minimize the average MSE over 100 Monte Carlo replicates.

2. CART (Breiman et al., 1984), with the complexity parameter chosen to minimize the average MSE over 100 Monte Carlo replicates.

3. K-NN regression (see e.g. Stone, 1977), with the number of neighbors K set to minimize the average MSE over 100 Monte Carlo replicates.

The estimated regression functions resulting from these three approaches are displayed in Figure 2. We see that K-NN-FL can adapt to low-density and high-density regions of the distribution of covariates, as well as to the local structure of the regression function. By contrast, CART displays some artifacts due to the binary splits that make up the decision tree, and K-NN regression undersmoothes in large areas of the domain.

In practice, we anticipate that K-NN-FL will outperform its competitors when the data are highly-concentrated around a low-dimensional manifold and the regression function is non-smooth in that region, as in the above example. In our theoretical analysis, we will consider the special case in which the data lie precisely on a low-dimensional manifold, or a mixture of low-dimensional manifolds.

3 Local Adaptivity of K-NN-FL and ϵ-NN-FL

3.1 Assumptions

We assume that, in (1), the elements of $\varepsilon \;=\; (\varepsilon_1, \ldots, \varepsilon_n)^T$ are independent and identically-distributed mean-zero sub-Gaussian random variables,

$$E(\varepsilon_i) \;=\; 0, \qquad \mathrm{pr}\left(|\varepsilon_i| \;>\; t\right) \;\le\; C\,\exp\left\{-t^2/(2\sigma^2)\right\}, \quad i = 1, \ldots, n, \quad \text{for all } t > 0, \tag{7}$$

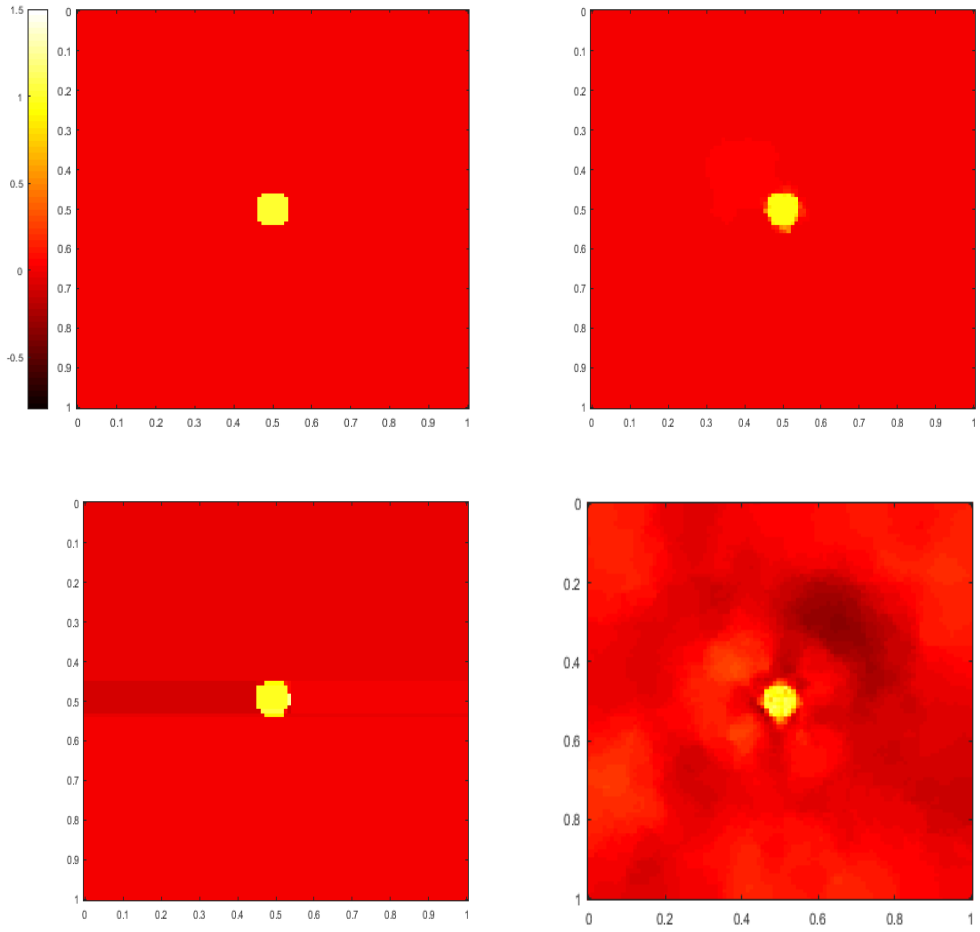

Figure 2: *Top Left:* The function f_0 in (6), evaluated at an evenly-spaced grid of size 100×100 in $[0,1]^2$. *Top Right:* The estimate of f_0 obtained via K-NN-FL. *Bottom Left:* The estimate of f_0 obtained via CART. *Bottom Right:* The estimate of f_0 obtained via K-NN regression.

for some positive constants σ and C. Furthermore, we assume that ε is independent of X.

In addition, for a set $A \subset \mathcal{A}$ with $(\mathcal{A}, d_{\mathcal{A}})$ a metric space, we write $B_\epsilon(A) = \{a : \text{exists } a' \in A, \text{ with } d_{\mathcal{A}}(a, a') \leq \epsilon\}$. We let ∂A denote the boundary of the set A. Moreover, the MSE of $\hat{\theta}$ is defined as $\|\hat{\theta} - \theta^*\|_n^2 = n^{-1} \sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|x\|_2 = (x_1^2 + \cdots + x_d^2)^{1/2}$. For $s \in \mathbb{N}$, we set $\mathbf{1}_s = (1, \ldots, 1)^T \in \mathbb{R}^s$. In the covariate space $\mathcal{X}$, we consider the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, induced by the metric $d_{\mathcal{X}}$. We let μ be a measure on $\mathcal{B}(\mathcal{X})$. We complement the model in (1) by assuming that the covariates satisfy $x_i \overset{\text{ind}}{\sim} p(x)$. Thus, p is the probability density function associated with the distribution of x_i, with respect to the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. Note that $\mathcal{X}$ can be a manifold of dimension d in a space of much higher dimension.

We begin by stating assumptions on the distribution of the covariates $p(\cdot)$, and on the metric space $(\mathcal{X}, d_{\mathcal{X}})$. In the theoretical results in Section 3 of Györfi et al. (2006), it is assumed that p is the probability density function of the uniform distribution in $[0,1]^d$. In this section, we will require only that p is bounded above and below. This condition appeared in the framework for studying K-NN graphs in Von Luxburg et al. (2014), and in the work on density quantization by Alamgir et al. (2014).

Assumption 1. *The density p satisfies $0 < p_{\min} < p(x) < p_{\max}$, for all $x \in \mathcal{X}$, where $p_{\min}, p_{\max} \in \mathbb{R}$.*

Although we do not require that $\mathcal{X}$ be a Euclidean space, we do require that balls in $\mathcal{X}$ have volume (with respect to μ) that behaves similarly to the Lebesgue measure of balls in $\mathbb{R}^d$. This is expressed in the next assumption, which appeared as part of the definition of a valid region (Definition 2) in Von Luxburg et al. (2014).

Assumption 2. *The base measure μ in $\mathcal{X}$ satisfies*

$$c_{1,d} r^d \leq \mu\{B_r(x)\} \leq c_{2,d} r^d, \quad \text{for all } x \in \mathcal{X},$$

for all $0 < r < r_0$, where r_0, $c_{1,d}$, and $c_{2,d}$ are positive constants, and $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

Next, we make an assumption about the topology of the space $\mathcal{X}$. We require that the space have no holes, and is topologically equivalent to $[0,1]^d$, in the sense that there exists a continuous bijection between $\mathcal{X}$ and $[0,1]^d$.

Assumption 3. *There exists a homeomorphism (a continuous bijection with a continuous inverse) $h : \mathcal{X} \to [0,1]^d$, such that*

$$L_{\min} d_{\mathcal{X}}(x, x') \leq \|h(x) - h(x')\|_2 \leq L_{\max} d_{\mathcal{X}}(x, x'), \quad \text{for all } x, x' \in \mathcal{X},$$

for some positive constants $L_{\min}, L_{\max}$, where $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

Note that Assumptions 2 and 3 immediately hold if $\mathcal{X} = [0,1]^d$, with $d_{\mathcal{X}}$ the Euclidean distance, h the identity mapping in $[0,1]^d$, and μ the Lebesgue measure in $[0,1]^d$. A metric space $(\mathcal{X}, d_{\mathcal{X}})$ that satisfies Assumption 3 is a special case of a differential manifold; the intuition is that the space $\mathcal{X}$ is a chart of the atlas for said differential manifold.

In Assumptions 2 and 3, we assume $d > 1$, since local adaptivity in non-parametric regression is well understood in one dimension. For instance, see Tibshirani (2014),Wang et al. (2016), Guntuboyina et al. (2017), and references therein.

We now proceed to state conditions on the regression function f_0 defined in (1). The first assumption simply requires bounded variation of the composition of the regression function with the homeomorphism h from Assumption 3.

Assumption 4. *The function $g_0 = f_0 \circ h^{-1}$ has bounded variation, i.e. $g_0 \in BV\{(0,1)^d\}$, and g_0 is also bounded. Here $(0,1)^d$ is the interior of $[0,1]^d$, and $BV\{(0,1)^d\}$ is the class of functions in $(0,1)^d$ with bounded variation. We refer the reader to Section S2 in the Supplementary Material for the explicit construction of the $BV\{(0,1)^d\}$ class. The function h was defined in Assumption 3.*

It is worth mentioning that if $\mathcal{X} = [0,1]^d$ and $h(\cdot)$ is the identity function in $[0,1]^d$, then Assumption 4 simply states that f_0 has bounded variation. However, in order to allow for more general scenarios, the condition is stated in terms of the function g_0 which has domain in the unit box, whereas the domain of f_0 is the more general set $\mathcal{X}$.

We now recall the definition of a piecewise Lipschitz function, which induces a much larger class than the set of Lipschitz functions, as it allows for discontinuities.

Definition 1. Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. We say that a bounded function $g : [0,1]^d \to \mathbb{R}$ is *piecewise Lipschitz* if there exists a set $\mathcal{S} \subset (0,1)^d$ that has the following properties:

1. The set $\mathcal{S}$ has Lebesgue measure zero.

2. For some constants $C_{\mathcal{S}}, \epsilon_0 > 0$, we have that $\mu(h^{-1}\{B_\epsilon(\mathcal{S}) \cup ([0,1]^d \backslash \Omega_\epsilon)\}) \leq C_{\mathcal{S}}\, \epsilon$ for all $0 < \epsilon < \epsilon_0$.

3. There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, then $|g(z) - g(z')| \leq L_0 \|z - z'\|_2$.

Roughly speaking, Definition 1 says that g is piecewise Lipschitz if there exists a small set $\mathcal{S}$ that partitions $[0,1]^d$ in such a way that g is Lipschitz within each connected component of the partition. Theorem 2.2.1 in Ziemer (2012) implies that if g is piecewise Lipschitz, then g has bounded variation on any open set within a connected component.

Theorem 1 will require Assumption 5, which is a milder assumption on g_0 than piecewise Lipschitz continuity (Definition 1). We now present some notation that is needed in order to introduce Assumption 5.

For $\epsilon > 0$ small enough, we denote by $\mathcal{P}_\epsilon$ a rectangular partition of $(0,1)^d$ induced by $0, \epsilon, 2\epsilon, \ldots, \epsilon(\lfloor 1/\epsilon \rfloor - 1), 1$, so that all the elements of $\mathcal{P}_\epsilon$ have volume of order ϵ^d. We define $\Omega_{2\epsilon} := [0,1]^d \backslash B_{2\epsilon}(\partial[0,1]^d)$. Then, for a set $\mathcal{S} \subset (0,1)^d$, we define

$$\mathcal{P}_{\epsilon,\mathcal{S}} := \{A \cap \Omega_{2\epsilon} \backslash B_{2\,\epsilon}(\mathcal{S}) \ : \ A \in \mathcal{P}_\epsilon, \ \ A \cap \Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S}) \neq \emptyset\};$$

this is the partition induced in $\Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S})$ by the grid $\mathcal{P}_\epsilon$.

For a function g with domain $[0,1]^d$, we define

$$S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}}) := \sum_{A \in \mathcal{P}_{\epsilon,\mathcal{S}}} \sup_{z_A \in A} \frac{1}{\epsilon} \int_{B_\epsilon(z_A)} |g(z_A) - g(z)|\, dz. \tag{8}$$

If g is piecewise Lipschitz, then $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ is bounded; see Section S3.1 in the Supplementary Material.

Next, we define

$$S_2\,(g, \mathcal{P}_{\epsilon,\mathcal{S}}) := \sum_{A \in \mathcal{P}_{\epsilon,\mathcal{S}}} \sup_{z_A \in A} T(g, z_A)\ \epsilon^d, \tag{9}$$

where

$$T(g, z_A) \ = \ \sup_{z \in B_\epsilon(z_A)} \sum_{l=1}^{d} \left| \int_{\|z'\|_2 \leq \epsilon} \frac{\partial \psi(z'/\epsilon)}{\partial z_l} \left\{ \frac{g(z_A - z') - g(z - z')}{\|z - z_A\|_2\, \epsilon^d} \right\} dz' \right|, \tag{10}$$

and ψ is a test function (see Section S1 in the Supplementary Material). Thus, (9) is the summation, over evenly-sized rectangles of volume ϵ that intersect $\Omega_{2\,\epsilon} \backslash B_{2\,\epsilon}(\mathcal{S})$, of the supremum of the function in (10). The latter, for a function g, can be thought as the average Lipschitz constant near z_A — see the expression inside curly braces in (10) — weighted by the derivative of a test function. The scaling factor ϵ^d in (10) arises because the integral is taken over a set of measure proportional to ϵ^d.

As with $S_1\,(g, \mathcal{P}_{\epsilon,\mathcal{S}})$, one can verify that if g is a piecewise Lipschitz function, then $S_2\,(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ is bounded.

We now make use of (8) and (9) in order to state our next condition on $g_0 = f_0 \circ h^{-1}$. This next condition is milder than assuming that g_0 is piecewise Lipschitz, see Definition 1.

Assumption 5. *Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. There exists a set $\mathcal{S} \subset (0,1)^d$ that satisfies the following:*

1. *The set S has Lebesgue measure zero.*

2. *For some constants $C_{\mathcal{S}}, \epsilon_0 > 0$, we have that $\mu\{h^{-1}[B_\epsilon(\mathcal{S}) \cup \{(0,1)^d \backslash \Omega_\epsilon\}]\} \leq C_{\mathcal{S}}\, \epsilon$ for all $0 < \epsilon < \epsilon_0$.*

3. *The summations $S_1(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})$ and $S_2(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})$ are bounded:*

$$\sup_{0<\epsilon<\epsilon_0} \max\{S_1(g_0, \mathcal{P}_{\epsilon,\mathcal{S}}), S_2(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})\} < \infty.$$

Finally, we refer the reader to Section S3 in the Supplementary Material for a discussion on Assumptions 4–5. In particular, we present an example illustrating that the class of piecewise Lipschitz functions is, in general, different from the class of functions for which Assumptions 4–5 hold. However, both classes contain the class of Lipschitz functions, where the latter amounts to $\mathcal{S} = \emptyset$ in Definition 1.

3.2 Results

Letting $\theta_i^* = f_0(x_i)$, we express the MSEs of K-NN-FL and ϵ-NN-FL in terms of the total variation of θ^* with respect to the K-NN and ϵ-NN graphs.

Theorem 1. *Let $K \asymp \log^{1+2r} n$ for some $r > 0$. Then under Assumptions 1–3, with an appropriate choice of the tuning parameter λ, the K-NN-FL estimator $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\text{pr}}\left(\frac{\log^{1+2r} n}{n} + \frac{\log^{1.5+r} n}{n}\|\nabla_{G_K}\theta^*\|_1\right).$$

This upper bound also holds for the ϵ-NN-FL estimator with $\epsilon \asymp \log^{(1+2r)/d} n/n^{1/d}$, if we replace $\|\nabla_{G_K}\theta^*\|_1$ with $\|\nabla_{G_\epsilon}\theta^*\|_1$, and with an appropriate choice of λ.

Clearly, the upper bound in Theorem 1 is a function of $\|\nabla_{G_K}\theta^*\|_1$ or $\|\nabla_{G_\epsilon}\theta^*\|_1$, for the K-NN or ϵ-NN graph, respectively. For the grid graph considered in Sadhanala et al. (2016), $\|\nabla_G\theta^*\|_1 \asymp n^{1-1/d}$, leading to the rate $n^{-1/d}$. However, for a general graph, there is no a priori reason to expect that $\|\nabla_G\theta^*\|_1 \asymp n^{1-1/d}$. Notably, our next result shows that $\|\nabla_G\theta^*\|_1 \asymp n^{1-1/d}$ for $G \in \{G_K, G_\epsilon\}$, under the assumptions discussed in Section 3.1.

Theorem 2. *Under Assumptions 1–5, or Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, if $K \asymp \log^{1+2r} n$ for some $r > 0$, then for an appropriate choice of the tuning parameter λ, the K-NN-FL estimator defined in (3) satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\text{pr}}\left(\frac{\log^{\alpha} n}{n^{1/d}}\right), \tag{11}$$

with $\alpha = 3r + 5/2 + (2r+1)/d$. Moreover, under Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, then $\hat{f}$ defined in (4) with the K-NN-FL estimator satisfies

$$E_{X\sim p}\left\{\left|f_0(X) - \hat{f}(X)\right|^2\right\} = O_{\text{pr}}\left\{\frac{\log^{\alpha} n}{n^{1/d}}\right\}. \tag{12}$$

Furthermore, under the same assumptions, (11) and (12) hold for the ϵ-NN-FL estimator with $\epsilon \asymp \log^{(1+2r)/d} n/n^{1/d}$.

Theorem 2 indicates that under Assumptions 1–5 or Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, both the K-NN-FL and ϵ-NN-FL estimators attain the convergence rate $n^{-1/d}$, ignoring logarithmic terms. Importantly, Theorem 3.2 from Györfi et al. (2006) shows that in the two-dimensional setting, this rate is actually minimax for estimation of Lipschitz continuous functions, when the design points are uniformly drawn from $[0,1]^2$. Thus, when $d = 2$ both K-NN-FL and ϵ-NN-FL are minimax for estimating functions in the class implied by Assumptions 1–5, and also in the class of piecewise Lipschitz functions implied by Assumptions 1–3 and Definition 1. In higher dimensions ($d > 2$), by the lower bound

in Proposition 2 from Castro et al. (2005), we can conclude that K-NN-FL and ϵ-NN-FL attain nearly minimax rates for estimating piecewise Lipschitz functions, whereas it is unknown if the same is true under Assumptions 1–5. Notably, a different method, similar in spirit to CART, was introduced in Appendix E of Castro et al. (2005). Castro et al. (2005) showed that this approach is also nearly minimax for estimating elements in the class of piecewise Lipschitz functions, although is unclear whether a computationally feasible implementation of their algorithm is available.

We see from Theorem 2 that both ϵ-NN-FL and K-NN-FL are locally adaptive, in the sense that they can adapt to the form of the function f_0. Specifically, these estimators do not require knowledge of the set $\mathcal{S}$ in Assumption 5 or Definition 1. This is similar in spirit to the one-dimensional fused lasso, which does not require knowledge of the breakpoints when estimating a piecewise Lipschitz function.

However, there is an important difference between the applicability of Theorem 2 for K-NN-FL and ϵ-NN-FL. In order to attain the rate in Theorem 2, ϵ-NN-FL requires knowledge of the dimension d, since this quantity appears in the rate of decay of ϵ. But in practice, the value of d might not be clear: for instance, suppose that $\mathcal{X} = [0,1]^2 \times \{0\}$; this is a subset of $[0,1]^3$, but it is homeomorphic to $[0,1]^2$, so that $d = 2$. If d is unknown, then it can be challenging to choose ϵ for ϵ-NN-FL. By contrast, the choice of K in K-NN-FL only involves the sample size n. Consequently, local adaptivity of K-NN-FL may be much easier to achieve in practice.

4 Manifold Adaptivity of K-NN-FL

In this section, we allow the observations $\{(x_i, y_i)\}_{i=1}^n$ to be drawn from a mixture distribution, in which each mixture component satisfies the assumptions in Section 3. Under these assumptions, we show that the K-NN-FL estimator can still achieve a desirable rate.

We assume

$$\begin{array}{lcll} y_i & = & \theta_i^* + \varepsilon_i, & i = 1, \dots, n, \\ \theta_i^* & = & f_{0,z_i}(x_i), & \\ x_i & \sim & p_{z_i}(x), & \\ \text{pr}(z_i = l) & \sim & \pi_l^*, & \text{for } l = 1, \dots, \ell. \end{array} \tag{13}$$

where ε satisfies (7), $\pi_l^* \in [0,1]$ with $\sum_{l=1}^{\ell} \pi_l^* = 1$, p_l is a density with support $\mathcal{X}_l \subset \mathcal{X}$, $f_{0,l} : \mathcal{X}_l \to \mathbb{R}$, and $\{\mathcal{X}_l\}_{l=1,\dots,\ell}$ is a collection of subsets of $\mathcal{X}$. For simplicity, we will assume that $\mathcal{X} \subset \mathbb{R}^d$ for some $d > 1$, and $d_{\mathcal{X}}$ is the Euclidean distance. In (13), the observed data is $\{(x_i, y_i)\}_{i=1}^n$. The remaining ingredients in (13) are either latent or unknown.

We further assume that each set $\mathcal{X}_l$ is homeomorphic to a Euclidean box of dimension depending on l, as follows:

Assumption 6. *For $l = 1, \dots, \ell$, the set $\mathcal{X}_l$ satisfies Assumptions 1–3 with metric given by $d_{\mathcal{X}}$, with dimension $d_l \in \mathbb{N} \backslash \{0, 1\}$, and with μ equal to some measure μ_l. In addition:*

1. *There exists a positive constant $\tilde{c}_l$ such that the set $\partial X_l := \bigcup_{l' \neq l} \mathcal{X}_{l'} \cap \mathcal{X}_l$ satisfies*

$$\mu_l \left\{ B_\epsilon \left(\partial X_l \right) \bigcap X_l \right\} \leq \tilde{c}_l \epsilon, \tag{14}$$

   for any small enough $\epsilon > 0$.

2. *There exists a positive constant r_l such that for any $x \in \mathcal{X}_l$, either*

$$\inf_{x'' \in \partial X_l} d_{\mathcal{X}}(x, x'') < d_{\mathcal{X}}(x, x') \quad \text{for all} \quad x' \in \mathcal{X} \backslash \mathcal{X}_l, \tag{15}$$

   or $B_\epsilon(x) \subset \mathcal{X}_l$ for all $\epsilon < r_l$.

The constraints implied by Assumption 6 are very natural. Inequality (14) states that the intersections of the manifolds $\mathcal{X}_1, \ldots, \mathcal{X}_\ell$ are small. To put this in perspective, if the extrinsic space ($\mathcal{X}$) were $[0,1]^d$ with the Lebesgue measure, then balls of radius of ϵ would have measure ϵ^d which is less than ϵ for all $d > 1$, and the set $B_\epsilon(\partial[0,1]^d) \cap [0,1]^d$ would have measure that scales like ϵ, which is the same scaling appearing (14). Furthermore, (15) holds if $\mathcal{X}_1, \ldots, \mathcal{X}_\ell$ are compact and convex subsets of $\mathbb{R}^d$ whose interiors are disjoint.

We are now ready to extend Theorem 2 to the framework described in this section.

Theorem 3. *Suppose that the data are generated as in (13), and Assumption 6 holds. Suppose also that the functions $f_{0,1}, \ldots, f_{0,\ell}$ either satisfy Assumptions 4–5 or are piecewise Lipschitz in the domain $\mathcal{X}_l$. Then for an appropriate choice of the tuning parameter λ, the K-NN-FL estimator defined in (3) satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\text{pr}}\left\{ \text{poly}(\log n) \sum_{l=1}^{\ell} \frac{\pi_l^*}{(\pi_l^* n)^{1/d_l}} \right\},$$

provided that $n \min\{\pi_l^* : l \in [\ell]\} \geq c_0 n^{r_0}$, and $K \asymp \log^{1+2r} n$ for some constants $c_0, r_0, r > 0$, where $\text{poly}(\cdot)$ is a polynomial function. Here, the π_l^*'s are allowed to change with n.

Notice that when $d_l = d$ for all $l \in [\ell]$ in Theorem 3, then we obtain, ignoring logarithmic factors, the rate $n^{-1/d}$ which is minimax when the functions $f_{0,l}$ are piecewise Lipschitz. The rate is also minimax when $d = 2$ and the functions $f_{0,l}$ satisfy Assumptions 4–5. In addition, our rates can be compared with the existing literature on manifold adaptivity. Specifically, when $d = 2$, the rate $n^{-1/2}$ is attained by local polynomial regression (Bickel and Li, 2007) and Gaussian process regression (Yang and Dunson, 2016) for the class of differentiable functions with bounded partial derivatives, and by K-NN regression for Lipschitz functions (Kpotufe, 2011). In higher dimensions, Bickel and Li (2007), Yang and Dunson (2016) and Kpotufe (2011) attain better rates than $n^{-1/d}$ on smaller classes of functions that do not allow for discontinuities.

Finally, we refer the reader to Section S11 of the Supplementary Material for an example suggesting that the ϵ-NN-FL estimator may not be manifold adaptive.

5 Experiments

Throughout this section, we take $d_{\mathcal{X}}$ to be Euclidean distance.

5.1 Simulated Data

In this section, we compare the following approaches:

- ϵ-NN-FL, with ϵ held fixed, and λ treated as a tuning parameter.
- K-NN-FL, with K held fixed, and λ treated as a tuning parameter.
- CART (Breiman et al., 1984), implemented in the `R` package `rpart`, with the complexity parameter treated as a tuning parameter.
- MARS (Friedman, 1991), implemented in the `R` package `earth`, with the penalty parameter treated as a tuning parameter.
- Random forests (Breiman, 2001), implemented in the `R` package `randomForest`, with the number of trees fixed at 800, and with the minimum size of each terminal node treated as a tuning parameter.

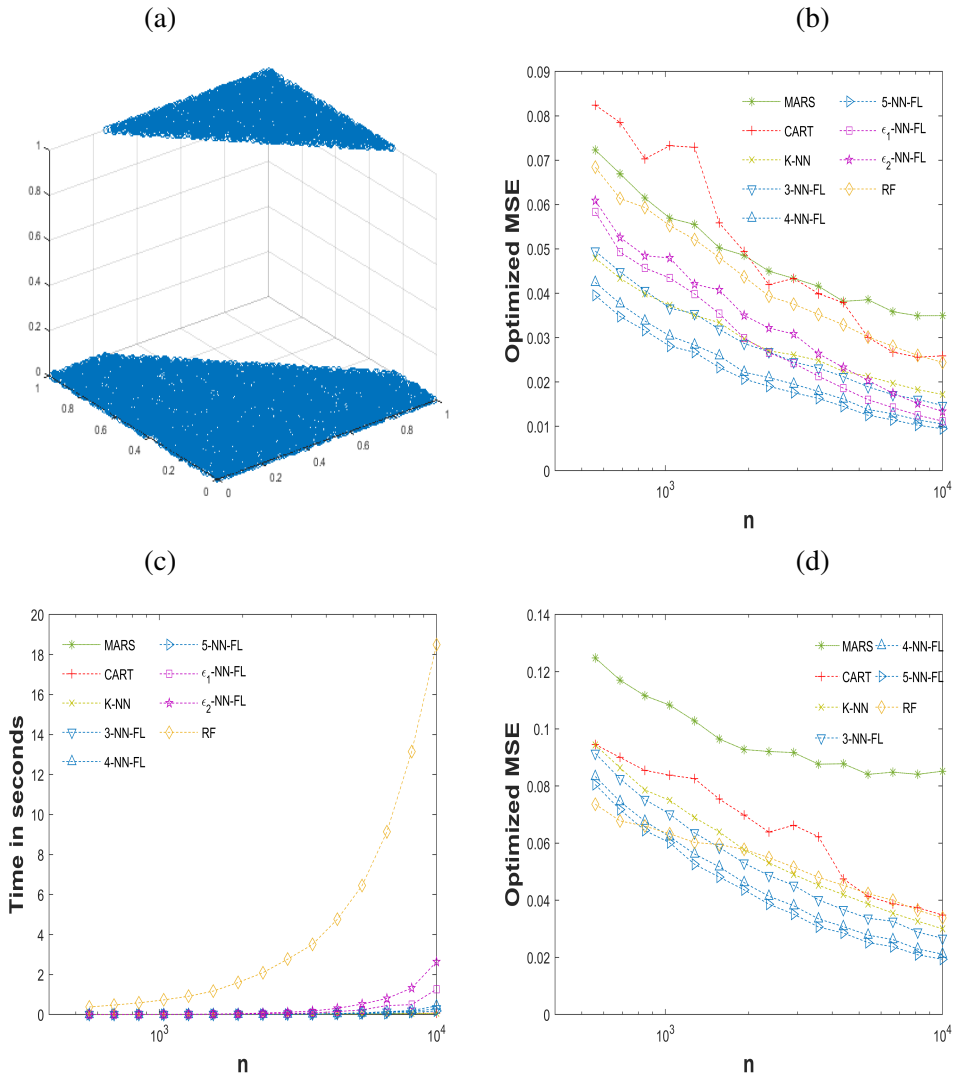

Figure 3: *(a):* A scatterplot of data generated under Scenario 1. The vertical axis displays $f_0(x_i)$, while the other two axes display the two covariates. *(b):* Optimized MSE (averaged over 150 Monte Carlo simulations) of competing methods under Scenario 1. Here $\epsilon_1 = \frac{3}{4}(\log n/n)^{1/2}$ and $\epsilon_2 = (\log n/n)^{1/2}$. *(c):* Computational time (in seconds) for Scenario 1, averaged over 150 Monte Carlo simulations. *(d):* Optimized MSE (averaged over 150 Monte Carlo simulations) of competing methods under Scenario 2.

- K-NN regression (e.g. Stone, 1977), implemented in `Matlab` using the function `knnsearch`, with K treated as a tuning parameter.

We evaluate each method's performance using the MSE, as defined in Section 3.1. Specifically, we apply each method to 150 Monte Carlo data sets with a range of tuning parameter values. For each method, we then identify the tuning parameter value that leads to the smallest average MSE over the 150 data sets. We refer to this smallest average MSE as the *optimized MSE* in what follows.

In our first two scenarios we consider $d = 2$ covariates, and let the sample size n vary.

Scenario 1. The function $f_0 : [0,1]^2 \to \mathbb{R}$ is piecewise constant,

$$f_0(x) = \mathbf{1}_{\{t \in \mathbb{R}^2 \,:\, \|t - \frac{3}{4}(1,1)^T\|_2 < \|t - \frac{1}{2}(1,1)^T\|_2\}}(x). \tag{16}$$

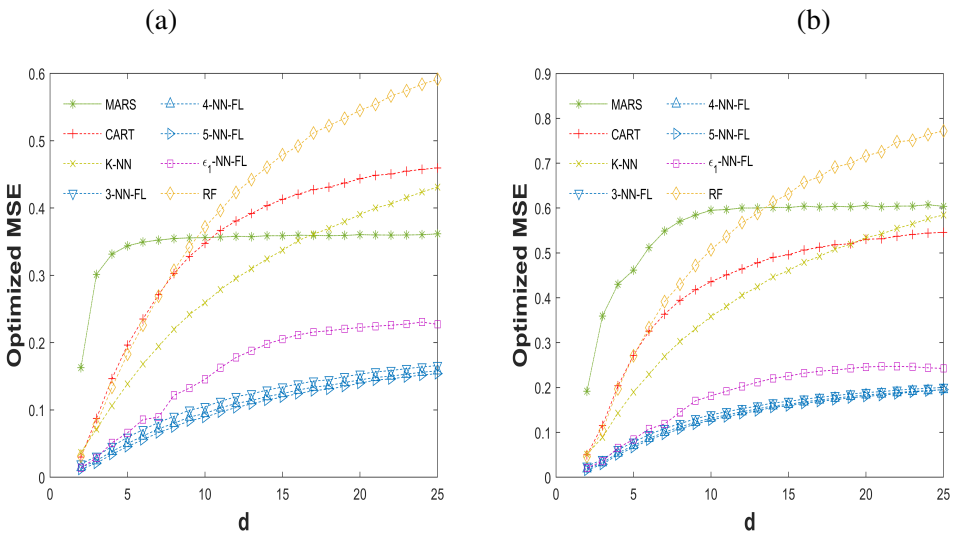

Figure 4: *(a)* Optimized MSE, averaged over 150 Monte Carlo simulations, for Scenario 3. *(b):* Optimized MSE, averaged over 150 Monte Carlo simulations, for Scenario 4. In both Scenario 3 and Scenario 4, ϵ_1 is chosen to be the largest value such that the total number of edges in the graph G_{ϵ_1} is at most 50000.

The covariates are drawn from a uniform distribution on $[0,1]^2$. The data are generated as in (1) with $N(0,1)$ errors.

Scenario 2. The function $f_0 : [0,1]^2 \to \mathbb{R}$ is as in (6), with generative density for X as in (5). The data are generated as in (1) with $N(0,1)$ errors.

Data generated under Scenario 1 are displayed in Figure 3(a). Data generated under Scenario 2 are displayed in Figure 1(b).

Figure 3(b) and Figure 3(d) display the optimized MSE, as a function of the sample size, for Scenarios 1 and 2, respectively. K-NN-FL gives the best results in both scenarios. ϵ-NN-FL performs a bit worse than K-NN-FL in Scenario 1, and very poorly in Scenario 2 (results not shown).

Timing results for all approaches under Scenario 1 are reported in Figure 3(c). For all methods, the times reported are averaged over a range of tuning parameter values. For instance, for K-NN-FL, we fix K and compute the time for different choices of λ; we then report the average of those times.

For our next two scenarios, we consider $n = 8000$ and values of d in $\{2, \ldots, 25\}$.

Scenario 3. The function $f_0 : [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 1 & \text{if } \left\|x - \frac{1}{4}\mathbf{1}_d\right\|_2 < \left\|x - \frac{3}{4}\mathbf{1}_d\right\|_2 \\ -1 & \text{otherwise,} \end{cases},$$

and the density p is uniform in $[0,1]^d$. The data are generated as in (1) with $\varepsilon_i \overset{\text{ind}}{\sim} N(0, 0.3)$.

Scenario 4. The function $f_0 : [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 2 & \text{if } \|x - q_1\|_2 < \min\{\|x - q_2\|_2, \|x - q_3\|_2, \|x - q_4\|_2\} \\ 1 & \text{if } \|x - q_2\|_2 < \min\{\|x - q_1\|_2, \|x - q_3\|_2, \|x - q_4\|_2\} \\ 0 & \text{if } \|x - q_3\|_2 < \min\{\|x - q_1\|_2, \|x - q_2\|_2, \|x - q_4\|_2\} \\ -1 & \text{otherwise} \end{cases},$$

where $q_1 = \left(\frac{1}{4}\mathbf{1}^T_{\lfloor d/2 \rfloor}, \frac{1}{2}\mathbf{1}^T_{d-\lfloor d/2 \rfloor}\right)^T$, $q_2 = \left(\frac{1}{2}\mathbf{1}^T_{\lfloor d/2 \rfloor}, \frac{1}{4}\mathbf{1}^T_{d-\lfloor d/2 \rfloor}\right)^T$, $q_3 = \left(\frac{3}{4}\mathbf{1}^T_{\lfloor d/2 \rfloor}, \frac{1}{2}\mathbf{1}^T_{d-\lfloor d/2 \rfloor}\right)^T$ and

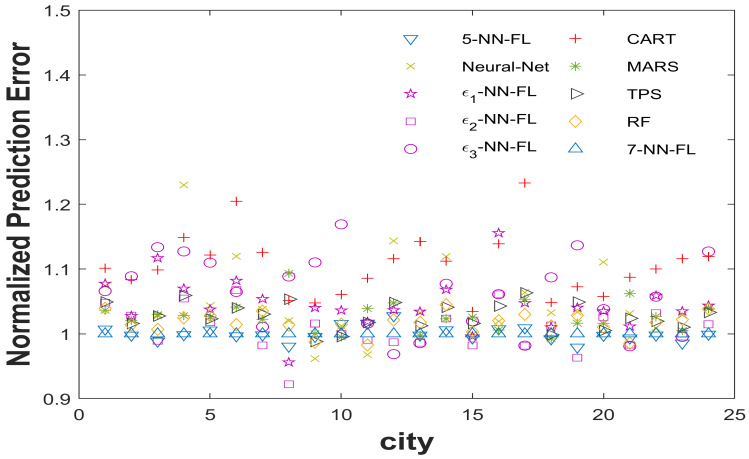

Figure 5: Results for the flu data. "Normalized Prediction Error" was obtained by dividing each method's test set prediction error by the test set prediction error of K-NN-FL, with $K = 7$.

$q_4 = \left(\frac{1}{2}\mathbf{1}^T_{\lfloor d/2 \rfloor}, \frac{3}{4}\mathbf{1}^T_{d-\lfloor d/2 \rfloor}\right)^T$. Once again, the generative density for X is uniform in $[0,1]^d$. The data are generated as in (1) with $\varepsilon_i \overset{\text{ind}}{\sim} N(0, 0.3)$.

Optimized MSE for each approach is displayed in Figure 4. When d is small, most methods perform well; however, as d increases, the competing methods quickly deteriorate, whereas K-NN-FL continues to perform well.

5.2 Flu Data

Our data consists of flu activity and atmospheric conditions between January 1, 2003 and December 31, 2009 across different cities in Texas. Our data-use agreement does not permit dissemination of the flu activity, which comes from medical records. The atmospheric conditions, which include temperature and air quality (particulate matter), can be obtained directly from CDC Wonder (`http://wonder.cdc.gov/`). Using the number of flu-related doctor's office visits as the dependent variable, we fit a separate non-parametric regression model to each of 24 cities; each day was treated as a separate observation, so that the number of samples is $n = 2556$ in each city. Five independent variables are included in the regression: maximum and average observed concentration of particulate matter, maximum and minimum temperature, and day of the year. All variables are scaled to lie in $[0, 1]$. We performed 50 75%/25% splits of the data into a training set and a test set. All models were fit on the training data, with 5-fold cross-validation to select tuning parameter values. Then prediction performance was evaluated on the test set.

We apply K-NN-FL with $K \in \{5, 7\}$, and ϵ-NN-FL with $\epsilon = j/n^{1/d}$ for $j \in \{1, 2, 3\}$, which is motivated by Theorem 2, and with larger choices of ϵ leading to worse performance. We also fit neural networks (Hagan et al., 1996; implemented in `Matlab` using the functions `newfit` and `train`), thin plate splines (Duchon, 1977; implemented using the `R` package `fields`), and MARS, CART, and random forests, as described in Section 5.1.

Average test set prediction error (across the 50 test sets) is displayed in Figure 5. We see that K-NN-FL and ϵ-NN-FL have the best performances. In particular, K-NN-FL performs best in 13 out the 24 cities, and second best in 6 cities. In 8 of the 24 cities, ϵ-NN-FL performs best.

We contend that K-NN-FL achieves superior performance because it adapts to heterogeneity in the density of design points, p (manifold adaptivity), and heterogeneity in the smoothness of the regression

function, f_0 (local adaptivity). In our theoretical results, we have substantiated this contention through prediction error rate bounds for a large class of regression functions of heterogeneous smoothness and a large class of underlying measures with heterogeneous intrinsic dimensionality. Our experiments demonstrate that these theoretical advantages translate into practical performance gains.

A Notation

Throughout, given $m \in \mathbb{N}$, we denote by $[m]$ the set $\{1, \ldots, m\}$. For $a \in A$, $A \subset \mathcal{A}$, with $(\mathcal{A}, d_{\mathcal{A}})$ a metric space, we write

$$d_{\mathcal{A}}(a, A) \;=\; \inf_{b \in A} d_{\mathcal{A}}(a, b).$$

In the case of the space $\mathbb{R}^d$, we will use the notation $\mathrm{dist}(\cdot, \cdot)$ for the metric induced by the norm $\|\cdot\|_\infty$, and we will write $B_\epsilon(x)$ for the ball $B_\epsilon(x, \|\cdot\|_2)$. We use the notation $\|\cdot\|_n$ for the rescaling of the $\|\cdot\|_2$ norm, such that for $x \in \mathbb{R}^n$,

$$\|x\|_n^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Furthermore, we write

$$\Omega_\epsilon \;=\; [0,1]^d \backslash B_\epsilon(\partial[0,1]^d). \tag{17}$$

Thus, Ω_ϵ is the set of points in the interior of $[0,1]^d$ such that balls of radius ϵ with center in such points are also contained in $[0,1]^d$.

Given an open set $\Omega \subset \mathbb{R}^d$, as is usual in mathematical analysis, we denote by $C_c^1(\Omega, \mathbb{R}^d)$ the set of functions $\phi : \Omega \to \mathbb{R}^d$ that have compact support and continuous first derivative. We also write $C^\infty(\Omega)$ for the class of functions $\phi : \Omega \to \mathbb{R}$ which have derivatives of all orders. The function $\psi : \mathbb{R}^d \to \mathbb{R}$ given as

$$\psi(z) \;=\; \begin{cases} C_1 \, \exp\left(\frac{-1}{1-\|x\|_2^2}\right) & \text{if } \|x\|_2 < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

is called a test function if C_1 is chosen such that $\int \psi(z)\, dz = 1$. We also denote, for $s \in \mathbb{N}$, by $\mathbf{1}_s$ the vector $\mathbf{1}_s := (1, \ldots, 1)^T \in \mathbb{R}^s$. Moreover, for vectors $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^t$, we write $w = (u^T, v^T)^T \in \mathbb{R}^{s+t}$ for the concatenation of u and v.

Furthermore, we denote by $L_1(\Omega)$ the set of Lebesgue measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\int_\Omega |f(x)| \mu(dx) < \infty,$$

where μ is the Lebesgue measure in Ω. We also denote by $L_\infty(\Omega)$ the set of measurable functions $f : \Omega \to \mathbb{R}$, such that $|f|$ is bounded except perhaps in set of μ-measure zero. The supremum norm in $L_\infty(\Omega)$ is denoted by $\|\cdot\|_{L_\infty(\Omega)}$.

B Definition of Bounded Variation

We now review some notation regarding general spaces of functions of bounded variation. Let $\Omega \subset \mathbb{R}^d$ be an open set. Recall that the divergence of a function $g : \Omega \to \mathbb{R}$ is defined as

$$\mathrm{div}(g)(x) \;=\; \sum_{j=1}^{d} \frac{\partial g(x)}{\partial x_j}, \quad \forall x \in \Omega,$$

provided that the partial derivatives involved exist.

Definition 2. A function $f : \Omega \to \mathbb{R}$ has bounded variation if

$$|f|_{\mathrm{BV}(\Omega)} := \sup \left\{ \int_{\Omega} f(x) \operatorname{div}(g)(x)\, dx ;\ g \in C_c^1(\Omega, \mathbb{R}^d),\ \|g\|_\infty \leq 1 \right\},$$

is finite, where

$$\|g\|_\infty := \left\| \left(\sum_{j=1}^{d} g_j^2 \right)^{1/2} \right\|_{L_\infty(\Omega)}.$$

The space of functions of bounded variation on Ω is denoted by $\mathrm{BV}(\Omega)$. We refer the reader to Ziemer (2012) for a full description of the class of functions of bounded variation.

C Discussion of Assumptions 4–5

C.1 Piecewise Lipschitz Condition and Assumption 5

To better understand the definition of $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ (see (8) in the main paper), notice that we can view $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ as a discretization over the set $\Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S})$ of the integral of the function

$$m(z_A, \epsilon) = \frac{1}{\epsilon^{d+1}} \int_{B_\epsilon(z_A)} |g(z_A) - g(z)| dz. \tag{19}$$

Recall that z_A is a Lebesgue point of g if $\lim_{\epsilon \to 0}\ \epsilon\, m(z_A, \epsilon) = 0$ (see Definition 2.18 in Giaquinta and Modica, 2010). Furthermore, if $g \in L_1(\mathbb{R}^d)$ then almost all points are Lebesgue points. This follows from the Lebesgue differentiation theorem; see Theorem 2.16 in Giaquinta and Modica (2010). Thus, if $g \in L_1(\mathbb{R}^d)$, then loosely speaking each term in the right hand side of (8) in the main paper is $O(\epsilon^{d-1})$, for any configuration of points z_A. In Assumption 5, we further require $S_1(f_0 \circ h^{-1}, \mathcal{P}_{\epsilon,\mathcal{S}})$ to be bounded.

Next, we show that piecewise Lipschitz continuity implies Assumption 5. First, note that if g is piecewise Lipschitz, then $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$, defined in (8) of the paper, is bounded. To see this, assume that g is piecewise Lipschitz with the set $\mathcal{S}$. Then

$$S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}}) \leq \sum_{A \in \mathcal{P}_{\epsilon,\mathcal{S}}} \sup_{z_A \in A} \int_{B_\epsilon(z_A)} \frac{|g(z_A) - g(z)|}{\|z_A - z\|_2} dz \leq \frac{L_0\, \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)} < \infty, \tag{20}$$

where the second-to-last inequality follows from the fact that the volume of a d-dimensional ball of radius ϵ is $\pi^{d/2}\epsilon^d/\Gamma(d/2+1)$, as well as the fact that there are at most on the order of ϵ^{-d} elements of $\mathcal{P}_{\epsilon,S}$.

Furthermore, with a similar argument to that in (20), we can show that if g is piecewise Lipschitz, then $S_2(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ is bounded. Therefore, if $f_0 \circ h^{-1}$ is piecewise Lipschitz, then $f_0 \circ h^{-1}$ satisfies Assumption 5.

C.2 Generic Functions that Do Not Satisfy Definition 1

We now present a general condition on f_0 that implies that f_0 is not piecewise Lipschitz.

Let $f_0 : [0,1]^d \to \mathbb{R}$ such that f_0 is differentiable in $(0,1)^d$. Suppose that the gradient of f_0 is continuous and unbounded. Then, we claim that f_0 is not piecewise Lipschitz. To see this, notice that there exists $t^{(0)} \in [0,1]^d$ such that

$$\lim_{t \to t^{(0)}} \|\nabla f_0(t)\|_\infty = \infty.$$

Hence, without lost of generality

$$\lim_{t \to t^{(0)}} \left| \frac{\partial f_0(t)}{\partial t_1} \right| = \infty.$$

Now, let $S \subset [0,1]^d$ satisfy all but the third condition of Definition 1. Then there exists a positive decreasing sequence $\{\epsilon_m\}_{m=1}^{\infty}$, and $t^{(m,1)}, t^{(m,2)} \in \{(0,1)^d \backslash B_{\epsilon_m}(S)\} \cap B_{C\epsilon_m^{1/d}}(t^{(0)})$ such that $\|t^{(m,1)} - t^{(m,2)}\|_2 < C\epsilon_m^{1/d} < 1$ for some constant C, $t_j^{(m,1)} = t_j^{(m,2)}$ for $j \in \{2, \dots, d\}$, and such that the segment connecting $t^{(m,1)}$ to $t^{(m,2)}$ does not contain elements in $B_{\epsilon_m}(S)$. Hence, by the mean value theorem,

$$\frac{|f_0(t^{(m,1)}) - f_0(t^{(m,2)})|}{\|t^{(m,1)} - t^{(m,2)}\|_2} = \left| \frac{\partial f_0(t^{(m,1,2)})}{\partial t_1} \right|,$$

where $t^{(m,1,2)}$ is a point in the segment connecting $t^{(m,1)}$ and $t^{(m,2)}$. By the continuity of ∇f_0, we can make the right-hand side of the previous inequality as large as desired. Therefore, the function f_0 is not piecewise Lipschitz.

C.3 Assumptions 4–5 and Definition 1 Induce Different Function Classes

We start by providing a stronger condition that Assumption 5. Specifically, assume that for small enough ϵ it holds that

$$\sum_{A \in \mathcal{P}_{\epsilon,\emptyset}} \text{Vol}(A) \sup_{z \in B_{2\epsilon}(A)} \|\nabla f_0(z)\|_1 < C, \tag{21}$$

for a constant C. Then we claim that f_0 satisfies Assumption 5. To verify this, we note that by the mean value theorem,

$$S_1(f_0, \mathcal{P}_{\epsilon,\emptyset}) \leq \sum_{A \in \mathcal{P}_{\epsilon,\emptyset}} C_2 \text{Vol}(A) \sup_{z \in B_{\epsilon}(A)} \|\nabla f_0(z)\|_1 < C_2 C, \tag{22}$$

for some constant $C_2 > 0$. Moreover,

$$T(f_0, z_A) \leq \sup_{z \in B_{2\epsilon}(z_A)} \text{Vol}(B_\epsilon(0)) \frac{\|\nabla f_0(z)\|_1}{\epsilon^d} d\|\nabla \phi\|_\infty \leq \text{Vol}(A) \sup_{z \in B_{2\epsilon}(A)} \frac{\|\nabla f_0(z)\|_1}{\epsilon^d} d\|\nabla \phi\|_\infty,$$

which implies that for some constant $C_3 > 0$

$$S_2(f_0, \mathcal{P}_{\epsilon,\emptyset}) \leq C_3 \sum_{A \in \mathcal{P}_{\epsilon,\emptyset}} \text{Vol}(A) \sup_{z \in B_{2\epsilon}(A)} \|\nabla f_0(z)\|_1 < C_3 C. \tag{23}$$

Therefore, combining (22) and (23), we obtain that the condition defined in (21) implies Assumption 5.

Next, we exploit (21) to show that Assumptions 4–5 and Definition 1 induce different classes of functions.

Example 1. Consider the function $f_0(x) = \sum_{j=1}^d \sqrt{x_j}$, $x \in [0,1]^d$. Then f_0 satisfies Assumptions 4–5. However, f_0 is not piecewise Lipschitz.

To verify Example 1, we check that (21) holds for $f_0 : [0,1]^d \to \mathbb{R}$ defined as $f_0(x) := \sum_{j=1}^d \sqrt{x_j}$.

To that end, notice that for some constant $C_4 > 0$, we have that

$$\begin{aligned}
\sum_{A \in \mathcal{P}_{\epsilon,\emptyset}} \mathrm{Vol}(A) \sup_{z \in B_{2\epsilon}(A)} \|\nabla f_0(z)\|_1 &\leq C_4 \left[\sum_{j_1=1}^{1/\epsilon} \ldots \sum_{j_d=1}^{1/\epsilon} \epsilon^d \left(\frac{1}{\sqrt{\epsilon j_1}} + \ldots + \frac{1}{\sqrt{\epsilon j_d)}} \right) \right] \\
&\leq C_4 d \sqrt{\epsilon} \sum_{j_1=1}^{1/\epsilon} \frac{1}{\sqrt{j_1}} \\
&\leq C_4 d \sqrt{\epsilon} \sum_{j_1=1}^{1/\epsilon} \frac{2}{\sqrt{j_1} + \sqrt{j_1 - 1}} \\
&\leq C_4 d \sqrt{\epsilon} \sum_{j_1=1}^{1/\epsilon} 2 \left[\sqrt{j_1} - \sqrt{j_1 - 1} \right] \\
&\leq C_4 d \sqrt{\epsilon} \sqrt{\tfrac{1}{\epsilon}} \\
&\leq C_4 d.
\end{aligned}$$

Furthermore, it is straightforward to check that f_0 has bounded variation, and hence f_0 satisfies Assumption 4. However,

$$\lim_{t \to 0} \left| \frac{\partial f_0(t)}{\partial t_1} \right| = \infty,$$

which by the argument in Section C.2 implies that f_0 is not piecewise Lipschitz.

D Outline of the Proof of Theorem 1

We start by discussing how to embed a mesh on a K-NN graph generated under Assumptions 1–3. This construction will then be used in the next sections in order to derive an upper bound for the MSE of the K-NN-FL estimator.

The embedding idea that we will present appeared in the flow-based proof of Theorem 4 from Von Luxburg et al. (2014) regarding commute distance on K-NN-graphs. There, the authors introduced the notion of a valid grid (Definition 17). For a fixed set of design points, a grid graph G is valid if, among other things, G satisfies the following: (i) The grid width is not too small, in the sense that each cell of the grid contains at least one of the design points. (ii) The grid width is not too large: points in the same or neighboring cells of the grid are always connected in the K-NN graph.

The notion of a valid grid was introduced for fixed design points, but through a minor modification, we construct a grid graph that with high probability satisfies the conditions of a valid grid from Von Luxburg et al. (2014). We now proceed to construct a grid embedding that for any signal will lead to a lower bound on the total variation along the K-NN graph.

Given $N \in \mathbb{N}$, in $[0,1]^d$ we construct a d-dimensional grid graph $G_{\text{lat}} = (V_{\text{lat}}, E_{\text{lat}})$, i.e., a lattice graph, with equal side lengths, and total number of nodes $|V_{\text{lat}}| = N^d$. Without loss of generality, we assume that the nodes of the grid correspond to the points

$$P_{\text{lat}}(N) = \left\{ \left(\frac{i_1}{N} - \frac{1}{2N}, \ldots, \frac{i_d}{N} - \frac{1}{2N} \right) : i_1, \ldots, i_d \in \{1, \ldots, N\} \right\}. \tag{24}$$

Notice that $P_{\text{lat}}(N)$ is the set of centers of the elements of the parition $\mathcal{P}_{N^{-1}}$, with the notation of Section 3.1 in the main paper. Moreover, $z, z' \in P_{\text{lat}}(N)$ share an edge in the graph $G_{\text{lat}}(N)$ if only if $\|z - z'\|_2 = N^{-1}$. If the nodes corresponding to z, z' share an edge, then we will write $(z, z') \in E_{\text{lat}}(N)$. Note that the lattice $G_{\text{lat}}(N)$ is constructed in $[0,1]^d$ and not in the set $\mathcal{X}$. However, this lattice can be transformed into

a mesh in the covariate space through the homeomorphism from Assumption 3, by $I(N) = h^{-1}\{P_{\text{lat}}(N)\}$. We can use $I(N)$ to perform a quantization in the domain $\mathcal{X}$ by using the cells associated with $I(N)$. See Alamgir et al. (2014) for more general aspects of quantizations under Assumptions 1–3.

Using this mesh $I(N)$, which depend on the homeomorphism h, for any signal $\theta \in \mathbb{R}^n$, we can construct two vectors denoted by $\theta_I \in \mathbb{R}^n$ and $\theta^I \in \mathbb{R}^{N^d}$. The former ($\theta_I$) is a signal vector that is constant within mesh cells. The latter (θ^I) has coordinates corresponding to the different nodes of the mesh (centers of cells). The precise definitions of θ_I and θ^I are given in Section E. Since θ and θ_I have the same dimension, it is natural to ask how these two relate to each other, at least for the purpose of understanding the empirical process associated with the K-NN-FL estimator. Moreover, given that $\theta^I \in \mathbb{R}^{N^d}$, one can try to relate the total variation of θ^I along a d-dimensional grid with N^d nodes, with the total variation of the original signal θ along the K-NN graph. We proceed to establish these connections next.

Lemma 4. *Assume that K is chosen such that $K/\log n \to \infty$. Then with high probability the following holds. Under Assumptions 1–3, there exists an $N \asymp (n/K)^{1/d}$ such that for the corresponding mesh $I(N)$ we have that*

$$|e^T(\theta - \theta_I)| \;\le\; 2\,\|e\|_\infty\,\|\nabla_{G_K}\theta\|_1, \qquad \forall \theta \in \mathbb{R}^n, \tag{25}$$

for all $e \in \mathbb{R}^n$. Moreover,

$$\|D\,\theta^I\|_1 \;\le\; \|\nabla_{G_K}\,\theta\|_1, \qquad \forall \theta \in \mathbb{R}^n, \tag{26}$$

where D is the incidence matrix of a d-dimensional grid graph $G_{grid} = (V_{grid}, E_{grid})$ with $V_{grid} = [N^d]$.

Lemma 4 immediately provides a path to control the empirical process associated with the K-NN-FL estimator. In particular, by the basic inequality argument (see, for instance, Wang et al. 2016), it is of interest to bound the quantity $\varepsilon^T(\hat{\theta} - \theta^*)$. In our context, this can be done by noticing that

$$\tfrac{1}{n}\varepsilon^T(\hat{\theta} - \theta^*) \;=\; \tfrac{1}{n}\varepsilon^T(\hat{\theta} - \hat{\theta}_I) + \tfrac{1}{n}\varepsilon^T(\hat{\theta}_I - \theta^*_I) + \tfrac{1}{n}\varepsilon^T(\theta^*_I - \theta^*). \tag{27}$$

Hence in the proof of our main theorem stated in Section G, we proceed to bound each term in the right hand side of (27).

Furthermore, Lemma 4 provides lower bounds involving θ_I and θ^I, both of which depend on the homeomorphism h. However, we only need to specify K, not h. In other words, we can avoid constructing the mesh $I(N)$, which would require knowledge of the unknown function h.

E Explicit Construction of Mesh for K-NN Embeddings

We now describe in detail the embedding idea from Section D.

Importantly, $I(N)$ (defined in Section D) can be thought of as a quantization in the domain $\mathcal{X}$; see Alamgir et al. (2014) for more general aspects of quantizations under Assumptions 1–3. For our purposes, it will be crucial to understand the behavior of $\{x_i\}_{i=1}^n$ and their relationship to $I(N)$. This is because we will use a grid embedding in order to analyze the behavior of the K-NN-FL estimator $\hat{\theta}$. With that goal in mind, we define a collection of cells, $\{C(x)\}_{x \in I(N)}$, in $\mathcal{X}$ as

$$C(x) \;=\; h^{-1}\left(\left\{z \in [0,1]^d \;:\; h(x) \;=\; \underset{x' \in P_{\text{lat}}(N)}{\arg\min}\, \|z - x'\|_\infty\right\}\right). \tag{28}$$

Recall that the goal in this paper is to estimate $\theta^* \in \mathbb{R}^n$. However, the mesh construction $I(N)$ has N^d elements, which we denote by $u_1, \ldots, u_{N^d}$. Hence, it is not immediately clear how to evaluate the total variation of θ^* along the graph corresponding to the mesh.

We first define θ_I, a vector constructed from θ that incorporates information about the samples $\{x_i\}_{i=1}^n$ and the cells $\{C(x)\}_{x\in I(N)}$. For $x \in \mathcal{X}$ we write $P_I(x)$ as the point in $I(N)$ such that $x \in C(P_I(x))$. If there is more than one such point, then we arbitrarily pick one. Then we collapse measurements corresponding to different observations x_i that fall in the same cell in $\{C(x)\}_{x\in I(N)}$ into a single value associated with the observation closest to the center of the cell after mapping with the homeomorphism. Thus, for a given $\theta \in \mathbb{R}^n$, we define $\theta_I \in \mathbb{R}^n$ as

$$(\theta_I)_i \;=\; \theta_j \quad \text{where} \quad j \;=\; \operatorname*{arg\,min}_{l\in[n]} \|h(P_I(x_i)) \,-\, h(x_l)\|_\infty. \tag{29}$$

Next we construct θ^I, a mapping from $\mathbb{R}^n$ to $\mathbb{R}^{N^d}$. For any $\theta \in \mathbb{R}^n$, we can induce a signal in $\mathbb{R}^{N^d}$ corresponding to the elements in $I(N)$. We write

$$I_j \;=\; \{i \in [n] \,:\, P_I(x_i) \;=\; u_j\}, \quad j \in [N^d],$$

where as before $u_1, \ldots, u_{N^d}$ are the elements of $I(N)$. If $I_j \neq \emptyset$, then there exists $i_j \in I_j$ such that $(\theta_I)_i \;=\; \theta_{i_j}$ for all $i \in I_j$. Here θ_I is the vector defined in (29). Using this notation, for a vector $\theta \in \mathbb{R}^n$, we write $\theta^I = (\theta_{i_1}, \ldots, \theta_{i_{N^d}})$. We use the convention that $\theta_{i_j} \;=\; 0$ if $I_j \;=\; \emptyset$.

Hence, for a given $\theta \in \mathbb{R}^n$, we have constructed two very intuitive signals $\theta_I \in \mathbb{R}^n$ and $\theta^I \in \mathbb{R}^{N^d}$. The former forces covariates x_i in the same cell to take the same signal value. The latter has coordinates corresponding to the different nodes of the mesh (centers of cells).

F Auxiliary Lemmas for Proof of Theorem 1

In several of the lemmas we will present next, we will implicitly condition on one of the observations, for example x_1. When doing so, we will exploit the fact that the other observations are independent as by our model assumption $x_i \overset{\text{ind}}{\sim} p(x)$, $i = 1, \ldots, n$.

The following lemma is a well-known concentration inequality for binomial random variables. It can be found as Proposition 27 in Von Luxburg et al. (2014).

Lemma 5. *Let m be a Binomial(M, q) random variable. Then, for all $\delta \in (0, 1]$,*

$$\begin{aligned} \mathbb{P}\left(m \;\leq\; (1-\delta)Mq\right) \;&\leq\; \exp\left(-\tfrac{1}{3}\delta^2 Mq\right), \\ \mathbb{P}\left(m \;\geq\; (1+\delta)Mq\right) \;&\leq\; \exp\left(-\tfrac{1}{3}\delta^2 Mq\right). \end{aligned}$$

We now use Lemma 5 to bound the distance between any design point and its K-nearest neighbors.

Lemma 6. ***(See Proposition 30 in Von Luxburg et al. (2014))*** *Denote by $R_K(x)$ the distance from $x \in \mathcal{X}$ to its Kth nearest neighbor in the set $\{x_1, \ldots, x_n\}$. Setting*

$$\begin{aligned} R_{K,\max} \;&=\; \max_{1\leq i\leq n} R_K(x_i), \\ R_{K,\min} \;&=\; \min_{1\leq i\leq n} R_K(x_i), \end{aligned}$$

we have that

$$\text{pr}\left\{ a\left(\frac{K}{n}\right)^{1/d} \;\leq\; R_{K,\min} \;\leq\; R_{K,\max} \;\leq\; \tilde{a}\left(\frac{K}{n}\right)^{1/d} \right\} \;\geq\; 1 \;-\; n\exp(-K/3) \;-\; n\exp(-K/12),$$

under Assumptions 1–3, where $a \;=\; 1/(2\, c_{2,d}\, p_{\max})^{1/d}$, and $\tilde{a} \;=\; 2^{1/d}/(p_{\min}\, c_{1,d})^{1/d}$.

Proof. This proof closely follows that of Proposition 30 in Von Luxburg et al. (2014).

First note that for any $x \in \mathcal{X}$, by Assumptions 1–2 we have

$$\text{pr}\left\{x_1 \in B_r(x)\right\} \;=\; \textstyle\int_{B_r(x)} p(t)\mu(dt) \;\leq\; p_{\max}\,\mu\{B_r(x)\} \;\leq\; c_{2,d}\,r^d\,p_{\max} \;:=\; \mu_{\max} \;<\; 1,$$

for small enough r. We also have that $R_K(x) \leq r$ if and only if there are at least K observations $\{x_i\}$ in $B_r(x)$. Let $V \sim \text{Binomial}(n,\, \mu_{\max})$. Then,

$$\mathbb{P}\left(R_K(x) \leq r\right) \;\leq\; \mathbb{P}\left(V \geq K\right) \;=\; \mathbb{P}\left(V \geq 2\,\mathbb{E}(V)\right),$$

where the last equality follows by choosing $a = 1/(2\,c_{2,d}\,p_{\max})^{1/d}$, and $r = a(K/n)^{1/d}$. Therefore, from Lemma 5 we obtain

$$\begin{aligned}
\text{pr}\left\{R_{K,\min} \leq a(K/n)^{1/d}\right\} \;&\leq\; \text{pr}\left\{\exists i \,:\, R_K(x_i) \leq a(K/n)^{1/d}\right\} \\
&\leq\; n \max_{1\leq i\leq n} \text{pr}\left\{R_K(x_i) \leq r\right\} \\
&\leq\; n\,\exp(-K/3).
\end{aligned}$$

Furthermore,

$$\text{pr}\left\{x_1 \in B_r(x)\right\} \;=\; \int_{B_r(x)} p(t)\mu(dt) \;\geq\; p_{\min}\mu\{B_r(x)\} \;\geq\; c_{1,d}\,r^d\,p_{\min} \;:=\; \mu_{\min} \;>\; 0,$$

and we arrive with a similar argument to

$$\text{pr}\left\{R_{K,\max} \;>\; \tilde{a}(K/n)^{1/d}\right\} \;\leq\; n\,\exp(-K/12), \tag{30}$$

where $\tilde{a} \;=\; 2^{1/d}/(p_{\min}\,c_{1,d})^{1/d}$. □

The upper bound in Lemma 6 allows us to control the maximum distance between x_i and x_j whenever they are connected in the K-NN graph. This maximum distance scales as $(K/n)^{1/d}$. The lower bound, on the other hand, prevents x_i from being arbitrarily close to its Kth nearest neighbor. These properties are particularly important as they will be used to characterize the penalty $\|\nabla_G \theta^*\|_1$.

As explained in Wang et al. (2016), there are different strategies for proving convergence rates in generalized lasso problems such as (2) from the main paper. Our approach here is similar in spirit to Padilla et al. (2018), and it is based on considering a lower bound for the penalty function induced by the K-NN graph. This lower bound will arise by constructing a signal over the grid graph induced by the cells $\{C(x)\}_{x\in I(N)}$ defined in (28). Towards that end, we provide the following lemma characterizing the minimum and maximum number of observations $\{x_i\}$ that fall within each cell $C(x)$. With an abuse of notation, we write $|C(x)| = |\{i \in [n] \,:\, x_i \in C(x)\}|$.

We now present a result related to Proposition 28 in Von Luxburg et al. (2014).

Lemma 7. *Assume that N in the construction of $P_{lat}(N)$ defined in (24) is chosen as*

$$N \;=\; \left\lceil \frac{3\,\sqrt{d}\;(2\,c_{2,d}\,p_{\max})^{1/d}\;n^{1/d}}{L_{\min}\,K^{1/d}} \right\rceil.$$

Then there exist positive constants $\tilde{b}_1$ and $\tilde{b}_2$ depending on $L_{\min}$, $L_{\max}$, d, $p_{\min}$, $p_{\max}$, $c_{1,d}$, and $c_{2,d}$ defined in Assumptions 1–3, such that

$$\begin{aligned}
\text{pr}\left\{\max_{x\in I(N)} |C(x)| \;\geq\; (1+\delta)\,c_{2,d}\,\tilde{b}_1\,K\right\} \;&\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\,\tilde{b}_2\,c_{1,d}\,K\right), \\
\text{pr}\left\{\min_{x\in I(N)} |C(x)| \;\leq\; (1-\delta)\,c_{1,d}\,\tilde{b}_2\,K\right\} \;&\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\,\tilde{b}_2\,c_{1,d}\,K\right),
\end{aligned} \tag{31}$$

for all $\delta \in (0,1)$, with $a = 1/(2\,c_{2,d}\,p_{\max})^{1/d}$, and $\tilde{a} = 2^{1/d}/(p_{\min}\,c_{1,d})^{1/d}$. Moreover, the symmetric K-NN graph has maximum degree $d_{\max}$ satisfying

$$\text{pr}\left(d_{\max} \geq \frac{3}{2}\,p_{\max}\,c_{2,d}\,\tilde{a}^d\,K\right) \leq n\left\{\exp\left(-\frac{K}{12}\right) + \exp\left(-\frac{p_{\min}c_{1,d}\tilde{a}^dK}{24}\right)\right\}.$$

Define the event Ω as: "If $x_i \in C(x_i')$ and $x_j \in C(x_j')$ for $x_i', x_j' \in I(N)$ with $\|h(x_i') - h(x_j')\|_2 \leq N^{-1}$, then x_i and x_j are connected in the K-NN graph". Then,

$$\text{pr}\,(\Omega) \geq 1 - n\,\exp(-K/3).$$

Proof. Let x_i and x_j such that $x_i \in C(x_i')$ and $x_j \in C(x_j')$ for $x_i', x_j' \in I(N)$ with $\|h(x_i') - h(x_j')\|_2 \leq N^{-1}$. Then the following holds using Assumption 3:

$$\begin{aligned}
d_{\mathcal{X}}(x_i,x_j) &\leq L_{\min}^{-1}\|h(x_i) - h(x_j)\|_2 \\
&\leq L_{\min}^{-1}\sqrt{d}\,\|h(x_i) - h(x_j)\|_\infty \\
&\leq L_{\min}^{-1}\sqrt{d}\left\{\|h(x_i) - h(x_i')\|_\infty + \|h(x_i') - h(x_j')\|_\infty + \|h(x_j') - h(x_j)\|_\infty\right\} \\
&< 3\,L_{\min}^{-1}\sqrt{d}\,N^{-1} \\
&\leq a\left(\tfrac{K}{n}\right)^{1/d}
\end{aligned}$$

with $a = 1/(2\,c_{2,d}\,p_{\max})^{1/d}$, where the first inequality follows from Assumption 3. Therefore, as in the proof of Lemma 6,

$$\text{pr}\,(\Omega) \geq \text{pr}\left\{a\left(\frac{K}{n}\right)^{1/d} \leq R_{K,\min}\right\} \geq 1 - n\,\exp(-K/3).$$

Next we proceed to derive an upper bound on the counts $\{C(x)\}$. Assume that $x \in h^{-1}\{P_{\text{lat}}(N)\}$, and $x' \in C(x)$. Then,

$$\begin{aligned}
d_{\mathcal{X}}(x,x') &\leq \tfrac{1}{L_{\min}}\|h(x) - h(x')\|_2 \\
&\leq \tfrac{\sqrt{d}}{L_{\min}}\|h(x) - h(x')\|_\infty \\
&\leq \tfrac{\sqrt{d}}{2\,L_{\min}\,N} \\
&\leq \tfrac{a}{6}\left(\tfrac{K}{n}\right)^{1/d} \\
&= \tfrac{(\tilde{b}_1)^{1/d}}{p_{\max}^{1/d}}\left(\tfrac{K}{n}\right)^{1/d},
\end{aligned}$$

where the first inequality follows from Assumption 3, the second from the definition of $P_{\text{lat}}(N)$, the third one from the choice of N, and $\tilde{b}_1$ is an appropriate constant. Therefore,

$$C(x) \subset B_{\frac{(\tilde{b}_1)^{1/d}}{p_{\max}^{1/d}}\left(\frac{K}{n}\right)^{1/d}}(x). \tag{32}$$

On the other hand, if $\tilde{b}_2$ is such that

$$\frac{(\tilde{b}_2)^{1/d}}{p_{\min}^{1/d}} \leq \frac{a}{2\,L_{\max}}\frac{L_{\min}}{3\sqrt{d}},$$

then if

$$d_{\mathcal{X}}(x,x') \;\leq\; \frac{(\tilde{b}_2)^{1/d}}{p_{\min}^{1/d}} \left(\frac{K}{n}\right)^{1/d},$$

we have that by Assumption 3, for large enough n

$$\|h(x) \,-\, h(x')\|_\infty \;\leq\; \tfrac{1}{2N}.$$

And so,

$$B_{\frac{(\tilde{b}_2)^{1/d}}{p_{\min}^{1/d}}\left(\frac{K}{n}\right)^{1/d}}(x) \,\subset\, C(x). \tag{33}$$

In consequence,

$$\begin{aligned}
\text{pr}\left\{\max_{x\in I(N)} |C(x)| \;\geq\; (1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\right\} \;&\leq\; \sum_{x\in I(N)} \text{pr}\left\{|C(x)| \;\geq\; (1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\right\} \\
&\leq\; \sum_{x\in I(N)} \text{pr}\left[|C(x)| \;\geq\; (1+\delta)\, n\, \text{pr}\{x_1 \in C(x)\}\right] \\
&\leq\; \sum_{x\in I(N)} \exp\left(-\tfrac{1}{3}\delta^2\, \tilde{b}_2\, c_{1,d}\, K\right) \\
&=\; N^d \exp\left(-\tfrac{1}{3}\delta^2\, \tilde{b}_2^d\, c_{1,d}\, K\right),
\end{aligned}$$

where the first inequality follows from a union bound, the second from (32), and the third from (33) combined with Lemma 5.

On the other hand, with a similar argument, we have that

$$\begin{aligned}
\text{pr}\left\{\min_{x\in I(N)} |C(x)| \;\leq\; (1-\delta)\, c_{1,d}\, \tilde{b}_2\, K\right\} \;&\leq\; \sum_{x\in I(N)} \text{pr}\left\{|C(x)| \;\leq\; (1-\delta)\, c_{1,d}\, \tilde{b}_2\, K\right\} \\
&\leq\; \sum_{x\in I(N)} \text{pr}\left[|C(x)| \;\leq\; (1-\delta)\, n\, \text{pr}\{x_1 \in C(x)\}\right] \\
&\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\, \tilde{b}_2\, c_{1,d}\, K\right).
\end{aligned}$$

Next we proceed to find an upper bound on the maximum degree of the K-NN graph. We start by defining the sets

$$B_i(x) \;=\; \left\{j \in [n]\backslash\{i\} \;:\; x_j \in B_{\tilde{a}(K/n)^{1/d}}(x)\right\}, \quad B_i \;=\; \left\{j \in [n]\backslash\{i\} \;:\; x_j \in B_{\tilde{a}(K/n)^{1/d}}(x_i)\right\},$$

for $i \in [n]$, and $x \in \mathcal{X}$, and where $\tilde{a}$ is given as in Lemma 6. Then,

$$|B_i(x)| \;\sim\; \text{Binomial}\left[n-1, \text{pr}\{x_1 \in B_{\tilde{a}(K/n)^{1/d}}(x)\}\right],$$

where

$$p_{\min}\, c_{1,d}\, \tilde{a}^d \frac{K}{n} \;\leq\; \text{pr}\{x_1 \in B_{\tilde{a}(K/n)^{1/d}}(x)\} \;\leq\; p_{\max}\, c_{2,d}\, \tilde{a}^d \frac{K}{n},$$

which implies by Lemma 5 that

$$\text{pr}\left\{|B_i(x)| \;\geq\; \frac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right\} \;\leq\; \text{pr}\left[|B_i(x)| \;\geq\; \frac{3}{2}\text{pr}\{x_1 \in B_{\tilde{a}(K/n)^{1/d}}(x)\}(n-1)\right] \;\leq\; \exp\left(-\frac{p_{\min}\, c_{1,d}\, \tilde{a}^d}{24} K\right).$$

Hence,

$$\mathrm{pr}\left\{|B_i| \;\geq\; \frac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right\} \;=\; \int_{\mathcal{X}} \mathrm{pr}\left\{|B_i(x)| \;\geq\; \frac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right\} p(x)\mu(dx) \;\leq\; \exp\left(-\frac{p_{\min}\, c_{1,d}\, \tilde{a}^d}{24} K\right).$$

Therefore, if d_i is the degree associated with x_i then

$$\begin{aligned}
\mathrm{pr}\left(d_i \;\leq\; \tfrac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right) \;&\geq\; \mathrm{pr}\left[\left|\{j \in [n]\backslash\{i\} \;:\; d_{\mathcal{X}}(x_i, x_j) \leq R_{K,\max}\}\right| \;\leq\; \tfrac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right] \\
&\geq\; \mathrm{pr}\Big[\left|\{j \in [n]\backslash\{i\} \;:\; d_{\mathcal{X}}(x_i, x_j) \leq R_{K,\max}\}\right| \;\leq\; \tfrac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K, \\
&\qquad R_{K,\max} \leq \tilde{a}\left(\tfrac{K}{n}\right)^{1/d}\Big] \\
&\geq\; \mathrm{pr}\Big[\left|\{j \in [n]\backslash\{i\} \;:\; d_{\mathcal{X}}(x_i, x_j) \leq \tilde{a}(K/n)^{1/d}\}\right| \;\leq\; \tfrac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K, \\
&\qquad R_{K,\max} \leq \tilde{a}\left\{\tfrac{K}{n}\right\}^{1/d}\Big] \\
&\geq\; 1 \;-\; \mathrm{pr}\left\{R_{K,\max} > \tilde{a}\left(\tfrac{K}{n}\right)^{1/d}\right\} \;-\; \mathrm{pr}\left\{|B_i| \;>\; \tfrac{3}{2} p_{\max}\, c_{2,d}\, \tilde{a}^d\, K\right\},
\end{aligned}$$

and the claim follows from the previous inequality and Equation (30). □

As stated in Lemma 7, the number of observations x_i that fall within each cell $C(x)$ scales like K, with high probability. We will exploit this fact next in order to obtain an upper bound on the MSE.

Lemma 8. *Assume that the event Ω from Lemma 7 intersected with*

$$\min_{x \in I(N)} |C(x)| \;\geq\; \frac{1}{2}\, c_{1,d}\, \tilde{b}_2\, K \tag{34}$$

holds. Define N as in that lemma and let $I(N)$ be the corresponding mesh. Then for all $e \in \mathbb{R}^n$, it holds that

$$|e^T(\theta \;-\; \theta_I)| \;\leq\; 2\, \|e\|_\infty\, \|\nabla_{G_K}\theta\|_1, \qquad \forall \theta \in \mathbb{R}^n. \tag{35}$$

Moreover,

$$\|D\,\theta^I\|_1 \;\leq\; \|\nabla_{G_K}\,\theta\|_1, \qquad \forall \theta \in \mathbb{R}^n, \tag{36}$$

where D is the incidence matrix of a d-dimensional grid graph $G_{grid} = (V_{grid}, E_{grid})$ with $V_{grid} = [N^d]$, where $(l, l') \in E_{grid}$ if and only if

$$\|h\{P_I(x_{i_l})\} \;-\; h\{P_I(x_{i_{l'}})\}\|_2 \;=\; \frac{1}{N}.$$

Here we use the notation from Section E.

Proof. We start by introducing the notation $x_i' \;=\; P_I(x_i)$. To prove (35) we proceed in cases.

Case 1. If $(\theta_I)_1 \;=\; \theta_1$, then clearly $|e_1|\,|\theta_1 \;-\; (\theta_I)_1| \;=\; 0$.
Case 2. If $(\theta_I)_1 \;=\; \theta_i$, for $i \neq 1$, then

$$\|h(x_1') \;-\; h(x_i)\|_\infty \;\leq\; \|h(x_1') \;-\; h(x_1)\|_\infty \;\leq\; \frac{1}{2\,N}.$$

Thus, $x_1' \;=\; x_i'$, and so $(1, i) \in E_K$ by the assumption that the event Ω holds.

Therefore for every $i \in \{1, \ldots, n\}$, there exists $j_i \in [n]$ such that $(\theta_I)_i \;=\; \theta_{j_i}$ and either $i = j_i$ or $(i, j_i) \in E_K$. Hence,

$$\begin{aligned}
|e^T(\theta \;-\; \theta_I)| \;&\leq\; \textstyle\sum_{i=1}^n |e_i|\,|\theta_i \;-\; \theta_{j_i}| \\
&\leq\; 2\, \|e\|_\infty\, \|\nabla_{G_K}\hat{\theta}\|_1.
\end{aligned}$$

To verify (36), we observe that

$$\|D\,\theta^I\|_1 \;=\; \sum_{(l,l')\in E_{grid}} \left|\theta_{i_l} \,-\, \theta_{i_{l'}}\right|. \tag{37}$$

Now, if $(l,l') \in E_{grid}$, then x_{i_l} and $x_{i_{l'}}$ are in neighboring cells in I. This implies that $(i_l, i_{l'})$ is an edge in the K-NN graph. Thus, every edge in the grid graph G_{grid} corresponds to an edge in the K-NN graph and the mapping is injective. Note that here we have used the fact that (34) ensures that every cell has at least one point, provided that K is large enough. The claim follows. □

Lemma 9. *With the notation from Lemma 8, we have that*

$$\varepsilon^T(\hat{\theta} \,-\, \theta^*) \le 2\,\|\varepsilon\|_\infty \left(\|\nabla_{G_K}\theta^*\|_1 \,+\, \|\nabla_{G_K}\hat{\theta}\|_1\right) \,+\, \varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I),$$

on the event Ω.

Proof. Let us assume that event Ω happens. Then we observe that

$$\varepsilon^T(\hat{\theta} \,-\, \theta^*) \;=\; \varepsilon^T(\hat{\theta} \,-\, \hat{\theta}_I) \,+\, \varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \,+\, \varepsilon^T(\theta^*_I \,-\, \theta^*), \tag{38}$$

and the claim follows by Lemma 8.

□

Lemma 10. *With the notation from Lemma 9, on the event* Ω*, we have that*

$$\varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \;\le\; \max_{u\in I}\sqrt{|C(u)|}\,\left(\|\Pi\tilde{\varepsilon}\|_2\;\|\hat{\theta} \,-\, \theta^*\|_2 \,+\, \|(D^+)^T\,\tilde{\varepsilon}\|_\infty\left[\|\nabla_{G_K}\hat{\theta}\|_1 \,+\, \|\nabla_{G_K}\,\theta^*\|_1\right]\right),$$

where $\tilde{\varepsilon} \in \mathbb{R}^{N^d}$ *is a mean zero vector whose coordinates are independent and sub-Gaussian with the same constants as in (7). Here,* Π *is the orthogonal projection onto the span of* $\mathbf{1} \in \mathbb{R}^{N^d}$*, and* D^+ *is the pseudo-inverse of the incidence matrix* D *from Lemma 8.*

Proof. Here we use the notation from the proof of Lemma 8. Then,

$$\varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \;=\; \sum_{j=1}^{N^d}\sum_{l\in I_j}\varepsilon_l\left(\hat{\theta}_{i_j} \,-\, \theta^*_{i_j}\right) \;=\; \max_{u\in I}|C(u)|^{\frac{1}{2}}\,\tilde{\varepsilon}^T(\hat{\theta}^I \,-\, \theta^{*,I})$$

where

$$\tilde{\varepsilon}_j \;=\; \left\{\max_{u\in I}|C(u)|\right\}^{-1/2}\sum_{l\,\in I_j}\varepsilon_l.$$

Clearly, the $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_{N^d}$ are independent given Ω, and are also sub-Gaussian with the same constants as the original errors $\varepsilon_1, \ldots, \varepsilon_n$.

Moreover, let Π be the orthogonal projection onto the span of $\mathbf{1} \in \mathbb{R}^{N^d}$. Then, by Hölder's inequality, and by the triangle inequality,

$$\begin{aligned}
\varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \;&\le\; \left\{\max_{u\in I}|C(u)|\right\}^{1/2}\left\{\|\Pi\tilde{\varepsilon}\|_2\;\|\hat{\theta}^I \,-\, \theta^{*,I}\|_2 \,+\, \|(D^+)^T\,\tilde{\varepsilon}\|_\infty\,\|D(\hat{\theta}^I \,-\, \theta^{*,I})\|_1\right\} \\
&\le\; \left\{\max_{u\in I}|C(u)|\right\}^{1/2}\left\{\|\Pi\tilde{\varepsilon}\|_2\;\|\hat{\theta}^I \,-\, \theta^{*,I}\|_2 \,+\, \|(D^+)^T\,\tilde{\varepsilon}\|_\infty\left(\|D\hat{\theta}^I\|_1 \,+\, \|D\,\theta^{*,I}\|_1\right)\right\}.
\end{aligned} \tag{39}$$

Next we observe that

$$\|\hat{\theta}^I \,-\, \theta^{*,I}\|_2 \;=\; \left\{\sum_{j=1}^{N^d}\left(\hat{\theta}_{i_j} \,-\, \theta^*_{i_j}\right)^2\right\}^{1/2} \;\leq\; \left\{\sum_{i=1}^{n}\left(\hat{\theta}_i \,-\, \theta^*_i\right)^2\right\}^{1/2}. \tag{40}$$

Therefore, combining (39), (40) and Lemma 8, we arrive at

$$\varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \;\leq\; \max_{u\in I}|C(u)|^{\frac{1}{2}}\,\left\{\|\Pi\tilde{\varepsilon}\|_2\;\|\hat{\theta} \,-\, \theta^*\|_2 \;+\; \|(D^+)^T\,\tilde{\varepsilon}\|_\infty\left(\|\nabla_{G_K}\hat{\theta}\|_1 \,+\, \|\nabla_{G_K}\,\theta^*\|_1\right)\right\}.$$

□

The following lemma results from a similar argument as the proof of Theorem 2 in Hutter and Rigollet (2016).

Lemma 11. *Let $d > 1$ (recall Assumptions 1–3) and let $\delta > 0$. With the notation from the previous lemma, given the event Ω, we have that*

$$\begin{aligned}\varepsilon^T(\hat{\theta}_I \,-\, \theta^*_I) \;\leq\; \{(1+\delta)\,c_{2,d}\,\tilde{b}_1^d\,K\}^{\frac{1}{2}}\Bigg(&2\,\sigma\,\left\{2\,\log\left(\tfrac{e}{\delta}\right)\right\}^{\frac{1}{2}}\,\|\hat{\theta} \,-\, \theta^*\|_2 \;+\; \sigma\,C_1(d)\,\left[\tfrac{2\,\log n}{d}\,\log\left\{\tfrac{C_2(p_{\max},L_{\min},d)\,n}{K\,\delta}\right\}\right]^{\frac{1}{2}}\\ &\cdot\left(\|\nabla_{G_K}\hat{\theta}\|_1 \,+\, \|\nabla_{G_K}\,\theta^*\|_1\right)\Bigg)\end{aligned} \tag{41}$$

with probability at least

$$1 \;-\; 2\delta \;-\; \frac{n}{K\,\tilde{a}^d\,L_{\max}^d}\exp\left(-\frac{1}{3}\delta^2\,\tilde{b}_2\,c_{1,d}\,K\right) \;-\; n\,\exp(-K/3),$$

where $C_1(d) > 0$ is a constant depending on d, and $C_2(p_{\max}, L_{\min}, d) > 0$ is another constant depending on $p_{\max}$, $L_{\min}$, and d.

Proof. First, as in the proof of Theorem 2 from Hutter and Rigollet (2016), and our choice of N, we obtain that given Ω,

$$\begin{aligned}\|(D^+)^T\tilde{\varepsilon}\|_\infty \;&\leq\; \sigma\,C_1(d)\,\left[\tfrac{2\log n}{d}\,\log\left\{\tfrac{C_2(p_{\max},L_{\min},d)\,n}{K\,\delta}\right\}\right]^{\frac{1}{2}},\\ \|\Pi\tilde{\varepsilon}\|_2 \;&\leq\; 2\,\sigma\,\left\{2\,\log\left(\tfrac{e}{\delta}\right)\right\}^{\frac{1}{2}}\end{aligned} \tag{42}$$

with probability at least $1 - 2\,\delta$, where $C(d)$ is constant depending on d, and $C_2(p_{\max}, L_{\min}, d)$ is a constant depending on $p_{\max}$, $L_{\min}$ and d.

On the other hand, by Lemma 7, given the event Ω we have

$$\max_{x\in h^{-1}(P_{\text{lat}})}|C(x)|^{\frac{1}{2}} \;\leq\; \{(1+\delta)\,c_{2,d}\,\tilde{b}_1\,K\}^{\frac{1}{2}}$$

with probability at least

$$1 \;-\; N^d\exp\left(-\frac{1}{3}\delta^2\,\tilde{b}_2\,c_{1,d}\,K\right) \;-\; n\,\exp(-K/3). \tag{43}$$

Therefore, combining (42) and (43), the result follows. □

G Proof of Theorem 1

Instead of proving Theorem 1, we will prove a more general result that holds for a general choice of K. This is given next. The corresponding result for ϵ-NN-FL can be obtained with a similar argument.

Theorem 12. *There exist constants $C_1(d)$ and $C_2(p_{\max}, L_{\min}, d)$, depending on d, $p_{\max}$, and $L_{\min}$, such that with the notation from Lemmas 6 and 7, if λ is chosen as*

$$\lambda \;=\; \sigma\, C_1(d)\, \left[(1+\delta)\, c_{2,d}\, \tilde{b}_1\; \frac{2\,\log n}{d}\; \log\left\{\frac{C_2(p_{\max}, L_{\min}, d)\, n}{K\,\delta}\right\} K\right] \;+\; 8\sigma\,(\log n)^{\frac{1}{2}}\,,$$

then the K-NN-FL estimator $\hat{\theta}$ (3) satisfies

$$\begin{aligned}\|\hat{\theta} \,-\, \theta^*\|_n^2 \;\;\leq\;\; & \|\nabla_{G_K}\,\theta^*\|_1 \left(\frac{4\,\sigma\, C_1(d)}{n}\left[(1+\delta)\, c_{2,d}\, \tilde{b}_1\; \frac{2\,\log n}{d}\; \log\left\{\frac{C_2(p_{\max}, L_{\min}, d)\, n}{K\,\delta}\right\} K\right]^{1/2}\right. \\ & \left. +\; \frac{32\,(\log n)}{n}\right) \;+\; \frac{16\,\sigma^2\,(1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\, \log\left(\frac{e}{\delta}\right)}{n},\end{aligned}$$

with probability at least $\eta_n\,\{1 \,-\, n\,\exp(-K/3)\}$. Here,

$$\eta_n \;=\; 1 \,-\, 2\delta \;\;-\;\; N^d \exp\left(-\frac{1}{3}\delta^2\, \tilde{b}_2\, c_{1,d}\, K\right) \;\;-\;\; n\,\exp(-K/3) \;\;-\;\; \frac{C}{n^7},$$

where C is the constant in (7), and N is given as in Lemma 7. Consequently, taking $K \asymp \log^{1+2r} n$ we obtain the result in Theorem 1.

Proof. We notice by the basic inequality argument, see for instance Wang et al. (2016), that

$$\tfrac{1}{2}\|\hat{\theta} \,-\, \theta^*\|_n^2 \;\;\leq\;\; \tfrac{1}{n}\left\{\varepsilon^T\,(\hat{\theta} \,-\, \theta^*) \;+\; \lambda\left(-\|\nabla_{G_K}\hat{\theta}\|_1 \;+\; \|\nabla_{G_K}\,\theta^*\|_1\right)\right\}. \tag{44}$$

On the other hand, by (7) in the paper, and a union bound,

$$\text{pr}\left\{\max_{1\leq i\leq n}|\varepsilon_i| \;>\; 4\sigma\,(\log n)^{\frac{1}{2}} \;\Big|\; \Omega\right\} \;=\; \text{pr}\left\{\max_{1\leq i\leq n}|\varepsilon_i| \;>\; 4\sigma\,(\log n)^{\frac{1}{2}}\right\} \;\leq\; \frac{C}{n^7}. \tag{45}$$

Therefore, combining (44), (45), Lemma 9, Lemma 10, and Lemma 11 we obtain that conditioning on Ω,

$$\begin{aligned}\tfrac{1}{2}\|\hat{\theta} \,-\, \theta^*\|_n^2 \;\;\leq\;\; & \frac{\left\{(1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\right\}^{\frac{1}{2}}}{n}\left(2\,\sigma\,\left\{2\,\log\left(\tfrac{e}{\delta}\right)\right\}^{\frac{1}{2}}\,\|\hat{\theta} \,-\, \theta^*\|_2 \;+\; \sigma\, C_1(d)\,\left[\frac{2\,\log n}{d}\,\log\left\{\frac{C_2(p_{\max}, L_{\min}, d)\, n}{K\,\delta}\right\}\right]^{\frac{1}{2}}\right. \\ & \left.\cdot\left(\|\nabla_{G_K}\hat{\theta}\|_1 \;+\; \|\nabla_{G_K}\,\theta^*\|_1\right)\right) \;+ \\ & \frac{8\sigma\,(\log n)^{\frac{1}{2}}}{n}\left(\|\nabla_{G_K}\hat{\theta}\|_1 \;+\; \|\nabla_{G_K}\,\theta^*\|_1\right) \;+\; \frac{\lambda}{n}\left(-\|\nabla_{G_K}\hat{\theta}\|_1 \;+\; \|\nabla_{G_K}\,\theta^*\|_1\right) \\ \leq\;\; & \|\nabla_{G_K}\,\theta^*\|_1\left(\frac{\sigma\, C_1(d)}{n}\left[(1+\delta)\, c_{2,d}\, \tilde{b}_1\; \frac{2\,\log n}{d}\; \log\left\{\frac{C_2(p_{\max}, L_{\min}, d)\, n}{K\,\delta}\right\} K\right]^{\frac{1}{2}}\right. \\ & \left. +\; \frac{8\sigma\,\log^{\frac{1}{2}} n}{n}\right) \;+\; 2\,\sigma\,\frac{\left\{2\,(1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\, \log\left(\frac{e}{\delta}\right)\right\}^{\frac{1}{2}}}{n}\|\hat{\theta} \,-\, \theta^*\|_2,\end{aligned}$$

with probability at least η_n. Hence by the inequality $a\,b - 4^{-1}\,b^2 \;\leq\; a^2$, we obtain that conditioning on Ω,

$$\begin{aligned}\tfrac{1}{4}\|\hat{\theta} \,-\, \theta^*\|_n^2 \;\;\leq\;\; & \|\nabla_{G_K}\,\theta^*\|_1\left(\frac{\sigma\, C_1(d)}{n}\left[(1+\delta)\, c_{2,d}\, \tilde{b}_1\; \frac{2\,\log n}{d}\; \log\left\{\frac{C_2(p_{\max}, L_{\min}, d)\, n}{K\,\delta}\right\} K\right]^{\frac{1}{2}}\right. \\ & \left. +\; \frac{8\sigma\,(\log n)^{\frac{1}{2}}}{n}\right) \;+\; \frac{4\,\sigma^2\,(1+\delta)\, c_{2,d}\, \tilde{b}_1\, K\, \log\left(\frac{e}{\delta}\right)}{n},\end{aligned}$$

with high probability. The claim follows. □

H Auxiliary lemmas for Proof of Theorem 2

Throughout we use the notation from Section E. The lemma below resembles Lemma 7 with the difference that we now prove that an edge in the K-NN graph induces an edge in a certain mesh.

Lemma 13. *Assume that N in the construction of G_{lat} is chosen as*

$$N \;=\; \left\lfloor \frac{(\,c_{1,d}\,p_{\min})^{1/d}\;n^{1/d}}{2^{1/d}\,L_{\max}\,K^{1/d}} \right\rfloor.$$

Then there exist positive constants b_1' and b_2' depending on $L_{\min}$, $L_{\max}$, d, $p_{\min}$, $c_{1,d}$, and $c_{2,d}$, such that

$$\begin{aligned} \mathrm{pr}\Big\{ \max_{x\in h^{-1}(P_{lat})} |C(x)| \;\geq\; (1+\delta)\,c_{2,d}\,b_1'\,K \Big\} \;&\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\,b_2'\,c_{1,d}\,K\right), \\ \mathrm{pr}\Big\{ \min_{x\in h^{-1}(P_{lat})} |C(x)| \;\leq\; (1-\delta)\,c_{1,d}\,b_2'\,K \Big\} \;&\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\,b_2'\,c_{1,d}\,K\right), \end{aligned} \tag{46}$$

for all $\delta \in (0,1]$. Moreover, let $\tilde{\Omega}$ denote the event: "For all $i,j \in [n]$, if x_i and x_j are connected in the K-NN graph, then $\|h(x_i') - h(x_j')\|_2 < 2\,N^{-1}$ where $x_i \in C(x_i')$ and $x_j \in C(x_j')$ with $x_i', x_j' \in I(N)$". Then

$$\mathrm{pr}\left(\tilde{\Omega}\right) \;\geq\; 1 \;-\; n\,\exp(-K/3).$$

Proof. Let $i,j \in [n]$ such that x_i and x_j are connected in the K-NN graph where $x_i \in C(x_i')$ and $x_j \in C(x_j')$ with $x_i', x_j' \in I(N)$. Then,

$$\begin{aligned} \|h(x_i') - h(x_j')\|_2 \;&\leq\; \|h(x_i') - h(x_i)\|_2 \;+\; \|h(x_i) - h(x_j)\|_2 \;+\; \|h(x_j) - h(x_j')\|_2 \\ &\leq\; \tfrac{1}{N} \;+\; \|h(x_i) - h(x_j)\|_2 \\ &\leq\; \tfrac{1}{N} \;+\; L_{\max}\,d_{\mathcal{X}}(x_i,x_j) \\ &\leq\; \tfrac{1}{N} \;+\; L_{\max}\,R_{K,\max} \end{aligned}$$

Therefore,

$$\mathrm{pr}\left(\tilde{\Omega}\right) \;\geq\; \mathrm{pr}\left\{R_{K,\max} \leq \tilde{a}(K/n)^{1/d}\right\} \;\geq\; 1 \;-\; n\,\exp(-K/12),$$

where $\tilde{a}$ is given as in Lemma 6.

Next, we derive an upper bound on the counts of the mesh. Assume that $x \in h^{-1}\{P_{\mathrm{lat}}(N)\}$, and $x' \in C(x)$. Then,

$$\begin{aligned} d_{\mathcal{X}}(x,x') \;&\leq\; \tfrac{1}{L_{\min}}\|h(x) - h(x')\|_2 \\ &\leq\; \frac{1}{2\,L_{\min}\,N} \\ &\leq\; \frac{1}{2\,L_{\min}}\,\frac{2^{1+1/d}\,K^{1/d}\,L_{\max}}{n^{1/d}} \\ &=:\; (b_1')^{1/d}\,\left(\tfrac{K}{n}\right)^{1/d}, \end{aligned}$$

where the first inequality follows from Assumption 3, the second from the definition of $P_{\mathrm{lat}}(N)$, and the third one from the choice of N. Therefore,

$$C(x) \;\subset\; B_{(b_1')^{1/d}\left(\frac{K}{n}\right)^{1/d}}(x). \tag{47}$$

On the other hand, we can find b_2' with a similar argument the proof of Lemma 7, and the proof follows the proof of that lemma.

□

Lemma 14. *Suppose that Assumptions 1–3 hold, and choose $K \asymp \log^{1+2r} n$ for some $r > 0$. With $\hat{f}(x)$ the prediction function for the K-NN-FL estimator as defined in (4), and for an appropriate choice of λ, it follows that*

$$\mathbb{E}_{X\sim p}\left|f_0(X) - \hat{f}(X)\right|^2 = O_{pr}\left(\frac{\log^{1+2r} n}{n} + \frac{\log^{1.5+r} n}{n}\|\nabla_{G_K}\theta^*\|_1 + \mathit{AErr}\right), \tag{48}$$

where AErr is the approximation error, defined as

$$\mathit{AErr} = \int \left\{f_0(x) - \frac{1}{K}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i)\right\}^2 p(x)\,\mu(dx).$$

Proof. Throughout we use the notation from Appendix E. We start by noticing that

$$\begin{aligned}
\mathbb{E}_{x\sim p}\left\{f_0(x) - \hat{f}(x)\right\}^2 &= \int \left\{f_0(x) - \hat{f}(x)\right\}^2 p(x)\,\mu(dx) \\
&= \int \Big\{f_0(x) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) \\
&\quad + \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} \hat{f}(x_i)\Big\}^2 p(x)\,\mu(dx) \\
&\le 2\int \left\{f_0(x) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i)\right\}^2 p(x)\,\mu(dx) + \\
&\quad 2\int \left\{\tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} \hat{f}(x_i)\right\}^2 p(x)\,\mu(dx).
\end{aligned}$$

Therefore we proceed to bound the second term in the last inequality. We observe that

$$\begin{aligned}
&\int \left\{\tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} \hat{f}(x_i)\right\}^2 p(x)\,\mu(dx) \\
&= \sum_{x'\in I(N)} \int_{C(x')} \left\{\tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) - \tfrac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} \hat{f}(x_i)\right\}^2 p(x)\,\mu(dx).
\end{aligned} \tag{49}$$

Let $x' \in \mathcal{X}$. Then there exists $u(x') \in I(N)$ such that $x \in C(u(x'))$ and

$$\|h(x') - h(u(x'))\|_2 \le \frac{1}{2\,N}.$$

Moreover, by Lemma 7, with high probability, there exists $i(x') \in [n]$ such that $x_{i(x')} \in C(u(x'))$. Hence,

$$d_{\mathcal{X}}(x', x_{i(x')}) \le \frac{1}{L_{\min}}\,\|h(x') - h(x_{i(x')})\|_2 \le \frac{1}{L_{\min}\,N}.$$

This implies that there exist $x_{i_1}, \ldots, x_{i_K}, i_1, \ldots, i_K \in [n]$ such that

$$d_{\mathcal{X}}(x', x_{i_l}) \le \frac{1}{L_{\min}\,N} + R_{K,\max}, \quad l = 1, \ldots, K.$$

Thus, with high probability, for any $x' \in \mathcal{X}$,

$$\begin{aligned}
\mathcal{N}_K(x') &\subset \left\{ i \,:\, d_{\mathcal{X}}(x', x_i) \,\leq\, \tfrac{1}{L_{\min} N} \,+\, R_{K,\max} \right\} \\
&\subset \left\{ i \,:\, \|h(x') \,-\, h(x_i)\|_2 \,\leq\, \tfrac{L_{\max}}{L_{\min} N} \,+\, L_{\max} R_{K,\max} \right\} \\
&\subset \left\{ i \,:\, \|h(u(x')) \,-\, h(x_i)\|_2 \,\leq\, \left(\tfrac{1}{2} \,+\, \tfrac{L_{\max}}{L_{\min}}\right) \tfrac{1}{N} \,+\, L_{\max} R_{K,\max} \right\}.
\end{aligned} \tag{50}$$

Hence, we set

$$\tilde{\mathcal{N}}(u(x')) \,=\, \left\{ i \,:\, \|h(u(x')) \,-\, h(x_i)\|_2 \,\leq\, \left(\frac{1}{2} \,+\, \frac{L_{\max}}{L_{\min}}\right) \frac{1}{N} \,+\, L_{\max} R_{K,\max} \right\}.$$

As a result, denoting by $u_1, \ldots, u_{N^d}$ the elements of $I(N)$, we have that for $j \in [N^d]$,

$$\begin{aligned}
&\int\limits_{C(u_j)} \left\{ \tfrac{1}{|\mathcal{N}_K(x)|} \sum_{i \in \mathcal{N}_K(x)} f_0(x_i) \,-\, \tfrac{1}{|\mathcal{N}_K(x)|} \sum_{i \in \mathcal{N}_K(x)} \hat{f}(x_i) \right\}^2 p(x)\, \mu(dx) \\
&\leq \int\limits_{C(u_j)} \tfrac{1}{|\mathcal{N}_K(x)|} \sum_{i \in \mathcal{N}_K(x)} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 p(x)\, \mu(dx) \\
&\leq \tfrac{1}{K} \int\limits_{C(u_j)} \sum_{i \in \tilde{\mathcal{N}}(u_j)} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 p(x)\, \mu(dx) \\
&= \tfrac{1}{K} \left[\sum_{i \in \tilde{\mathcal{N}}(u_j)} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 \right] \int\limits_{C(u_j)} p(x)\, \mu(dx).
\end{aligned}$$

The above combined with (49) leads to

$$\begin{aligned}
&\int \left\{ \frac{1}{|\mathcal{N}_K(x)|} \sum_{i \in \mathcal{N}_K(x)} f_0(x_i) \,-\, \frac{1}{|\mathcal{N}_K(x)|} \sum_{i \in \mathcal{N}_K(x)} \hat{f}(x_i) \right\}^2 p(x)\, \mu(dx) \\
&\leq \sum_{j=1}^{N^d} \left(\frac{1}{K} \left[\sum_{i \in \tilde{\mathcal{N}}(u_j)} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 \right] \int\limits_{C(u_j)} p(x)\, \mu(dx) \right) \\
&\leq \left[\sum_{j=1}^{N^d} \frac{1}{K} \sum_{i \in \tilde{\mathcal{N}}(u_j)} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 \right] \max_{1 \leq j \leq N^d} \int\limits_{C(u_j)} p(x)\, \mu(dx) \\
&\leq \left[\sum_{i=1}^{n} \frac{1}{K} \sum_{j \in [N^d] \,:\, \|h(x_i) - h(u_j)\|_2 \,\leq\, \left(\frac{1}{2} + \frac{L_{\max}}{L_{\min}}\right) \frac{1}{N} + L_{\max} R_{K,\max}} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 \right] \max_{1 \leq j \leq N^d} \int\limits_{C(u_j)} p(x)\, \mu(dx) \\
&\leq \left[\sum_{i=1}^{n} \left\{ f_0(x_i) \,-\, \hat{f}(x_i) \right\}^2 \right] \left[\max_{i \in [n]} \left| \left\{ j \,:\, \|h(x_i) \,-\, h(u_j)\|_2 \,\leq\, \left(\frac{1}{2} \,+\, \frac{L_{\max}}{L_{\min}}\right) \frac{1}{N} \,+\, L_{\max} R_{K,\max} \right\} \right| \right] \cdot \\
&\quad \tfrac{1}{K} \max_{1 \leq j \leq N^d} \int\limits_{C(u_j)} p(x)\, \mu(dx).
\end{aligned} \tag{51}$$

Next, for a set $A \subset \mathbb{R}^d$ and positive constant r, we define the packing and external covering numbers as

$$\begin{aligned} \mathrm{N}_r^{\text{pack}}(A) &:= \max\left\{l \in \mathbb{N} : \exists q_1, \ldots, q_l \in A, \text{ such that } \|q_j - q_{j'}\|_2 > r \ \ \forall j \neq j'\right\}, \\ \mathrm{N}_r^{\text{ext}} &:= \min\left\{l \in \mathbb{N} : \exists q_1, \ldots, q_l \in \mathbb{R}^d, \text{such that } \forall x \in A \text{ there exists } l_x \text{ with } \|x - q_{l_x}\|_2 < r\right\}. \end{aligned}$$

Furthermore, by Lemma 7, there exists a constant $\tilde{c}$ such that $R_{K,\max} \leq \tilde{c}/N$ with high probability. This implies that with high probability, for a positive constant $\tilde{C}$, we have that

$$\begin{aligned} &\max_{i\in[n]} \left|\left\{j \in [N^d] : \|h(x_i) - h(u_j)\|_2 \leq \left(\tfrac{1}{2} + \tfrac{L_{\max}}{L_{\min}}\right)\tfrac{1}{N} + L_{\max} R_{K,\max}\right\}\right| \\ &\leq \max_{i\in[n]} \left|\left\{j \in [N^d] : \|h(x_i) - h(u_j)\|_2 \leq \tfrac{\tilde{C}}{N}\right\}\right| \end{aligned} \tag{52}$$

$$\begin{aligned} &\leq \max_{i\in[n]} \mathrm{N}^{\text{pack}}_{\frac{1}{N}}\left(B_{\frac{\tilde{C}}{N}}(h(x_i))\right) \\ &\leq \mathrm{N}^{\text{ext}}_{\frac{1}{N}}(B_{\frac{\tilde{C}}{N}}(0)) \\ &= \mathrm{N}^{\text{ext}}_{1}(B_{\tilde{C}}(0)) \\ &< C', \end{aligned} \tag{53}$$

for some positive constant C', where the first inequality follows from $R_{K,\max} \leq \tilde{c}/N$, the second from the definition of packing number, and the remaining inequalities from well-known properties of packing and external covering numbers.

Therefore, there exists a constant C_1 such that

$$\begin{aligned} &\int \left\{\frac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} f_0(x_i) - \frac{1}{|\mathcal{N}_K(x)|}\sum_{i\in\mathcal{N}_K(x)} \hat{f}(x_i)\right\}^2 p(x)\,\mu(dx) \\ &\leq \left[\sum_{i=1}^n \left\{f_0(x_i) - \hat{f}(x_i)\right\}^2\right] \frac{C_1}{K} \max_{1\leq j\leq N^d} \int_{C(u_j)} p(x)\,\mu(dx) \\ &\leq \left[\sum_{i=1}^n \left\{f_0(x_i) - \hat{f}(x_i)\right\}^2\right] \frac{C_1\, p_{\max}}{K}\, \mu\left(B_{\frac{\tilde{b}_1}{p_{\max}^{1/d}}\left(\frac{K}{n}\right)^{1/d}}(x)\right) \end{aligned}$$

where the last equation follows as in (32). The claim then follows.

□

Lemma 15. *Assume that $g_0 := f_0 \circ h^{-1}$ satisfies Definition 1, i.e. g_0 is piecewise Lipschitz. Then, under Assumptions 1–3,*

$$AErr \;=\; O_{pr}\left(\frac{K^{1/d}}{n^{1/d}}\right),$$

provided that $K/\log n \to \infty$, where AErr was defined in Lemma 14. Consequently, with $K \asymp \log^{1+2r} n$ for some $r > 0$, and for an appropriate choice of λ, we have that

$$\mathbb{E}_{X\sim p}\left|f_0(X) - \hat{f}(X)\right|^2 \;=\; O_{pr}\left\{\frac{\log^{2.5+3r+(1+2r)/d} n}{n^{1/d}}\right\}. \tag{54}$$

Proof. Throughout we use the notation from Lemma 7 and Section E. We denote by $u_1, \dots, u_{N^d}$ the elements of $h^{-1}(P_{\text{lat}}(N))$, and so

$$\begin{aligned}
\text{AErr} &= \int \Big\{ f_0(x) - \tfrac{1}{K} \sum_{i \in \mathcal{N}_K(x)} f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx) \\
&= \sum_{j \in [N^d]} \int_{C(u_j)} \Big\{ f_0(x) - \tfrac{1}{K} \sum_{i \in \mathcal{N}_K(x)} f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx) \\
&\le \sum_{j \in [N^d]} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{1}{K} \Big\{ f_0(x) - f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx) \\
&= \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} = \emptyset} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{1}{K} \Big\{ f_0(x) - f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx) + \\
&\quad \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{1}{K} \Big\{ f_0(x) - f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx),
\end{aligned}$$

where the inequality holds by convexity. Therefore,

$$\begin{aligned}
\text{AErr} &\le 4 \left| \{ j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset \} \right| \, \|f_0\|_\infty^2 \max_{1 \le j \le N^d} \int_{C(u_j)} p(x) \, \mu(dx) + \\
&\quad \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} = \emptyset} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{1}{K} \Big\{ f_0(x) - f_0(x_i) \Big\}^2 \, p(x) \, \mu(dx) \\
&\le \mu \left(B_{\frac{\tilde{b}_1^{1/d}}{p_{\max}^{1/d}} (\frac{K}{n})^{1/d}}(x) \right) \left(4 \, p_{\max} \, \|f_0\|_\infty^2 \right) \left| \{ j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset \} \right| \\
&\quad + \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} = \emptyset} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{1}{K} \Big\{ g_0(h(x)) - g_0(h(x_i)) \Big\}^2 \, p(x) \, \mu(dx) \\
&\le \left(c_{2,d} \, 4 \, \tilde{b}_1 \, \|f_0\|_\infty^2 \right) \tfrac{K}{n} \left| \{ j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset \} \right| \\
&\quad + \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} = \emptyset} \int_{C(u_j)} \sum_{i \in \mathcal{N}_K(x)} \tfrac{L_0}{K} \|h(x) - h(x_i)\|_2^2 \, p(x) \, \mu(dx) \\
&\le \left(c_{2,d} \, 4 \, \tilde{b}^d \, \|f_0\|_\infty^2 \right) \tfrac{K}{n} \left| \{ j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset \} \right| \\
&\quad + L_0 \left\{ \sum_{j \in [N^d] \colon C(u_j) \cap \mathcal{S} = \emptyset} \int_{C(u_j)} p(x) \, \mu(dx) \right\} \left(\tfrac{L_{\max}}{L_{\min}} \tfrac{1}{N} + L_{\max} \, R_{K,\max} \right)^2 \\
&\le \left(c_{2,d} \, 4 \, \tilde{b}^d \, \|f_0\|_\infty^2 \right) \tfrac{K}{n} \left| \{ j \in [N^d] \colon C(u_j) \cap \mathcal{S} \neq \emptyset \} \right| \\
&\quad + L_0 \left(\tfrac{L_{\max}}{L_{\min}} \tfrac{1}{N} + L_{\max} \, R_{K,\max} \right)^2,
\end{aligned}$$

where the first inequality holds by elementary properties of integrals, the second inequality by (32), the third inequality by Assumptions 2–3, the fourth inequality by the same argument as in (50), and the fifth inequality

from properties of integration. The conclusion of the lemma follows from the inequality above combined with Lemma 6 and the proof of Proposition 23 from Hutter and Rigollet (2016) which uses Lemma 8.3 from Arias-Castro et al. (2012).

□

I Proof of Theorem 2

Proof. Combining Lemmas 14 and 15 we obtain that

$$E_{X\sim p}\left\{\left|f_0(X) - \hat{f}(X)\right|^2\right\} = O_{\text{pr}}\left\{\frac{\log^\alpha n}{n^{1/d}}\right\},$$

provide that Assumptions 1–3 hold and f_0 satisfies Definition 1.

Suppose now that Assumptions 1–5 hold. Throughout, we extend the domain of the function g_0 to be $\mathbb{R}^d$ by simply making it take the value zero in $\mathbb{R}^d \backslash [0,1]^d$. We will proceed to construct smooth approximations to g_0 that will allow us to obtain the desired result. To that end, for any $\epsilon > 0$ we construct the regularizer (or mollifier) $g_\epsilon : \mathbb{R}^d \to \mathbb{R}$ defined as

$$g_\epsilon(z) = \psi_\epsilon * g_0(z) = \int \psi_\epsilon(z')\, g_0(z - z')\, dz',$$

where $\psi_\epsilon(z') = \epsilon^{-d}\, \psi(z'/\epsilon)$. Then given Assumption 4, by the proof of Theorem 5.3.5 from Zhu et al. (2003), it follows that there exists a constant C_2 such that

$$\limsup_{\epsilon \to +0} \int_{(0,1)^d} \|\nabla g_\epsilon(z)\|_1\, dz < C_2$$

which implies that there exists $\epsilon_1 > 0$ such that

$$\sup_{0<\epsilon<\epsilon_1} \int_{(0,1)^d} \|\nabla g_\epsilon(z)\|_1\, dz < C_2. \tag{55}$$

Next, for N as in Lemma 13, we set $\epsilon = N^{-1}$ and consider the event

$$\Lambda_\epsilon = \left\{h(x_1) \in B_{4\epsilon}(\mathcal{S}) \cup \left[(0,1)^d \backslash \Omega_{4\epsilon}\right]\right\}, \tag{56}$$

and note that

$$\begin{aligned} \text{pr}(\Lambda_\epsilon) &= \int_{h^{-1}\left(B_{4\epsilon}(\mathcal{S}) \cup (0,1)^d \backslash \Omega_{4\epsilon}\right)} p(z)\, \mu(dz) \\ &\leq p_{\max}\, \mu\left[h^{-1}\left\{B_{4\epsilon}(\mathcal{S}) \cup (0,1)^d \backslash \Omega_{4\epsilon}\right\}\right] \\ &\leq p_{\max}\, C_{\mathcal{S}}\, 4\,\epsilon, \end{aligned} \tag{57}$$

where the last inequality follows from Assumption 5. Defining

$$J = \{\, i \in [n] \;:\; h(x_i) \in \Omega_{4\epsilon} \backslash B_{4\epsilon}(\mathcal{S})\}, \tag{58}$$

by the triangle inequality we have

$$\begin{aligned} &\left| \sum_{(i,j)\in E_K,\, i,j\in J} |\theta_i^* - \theta_j^*| - \sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\}| \right| \\ &= \left| \sum_{(i,j)\in E_K\, i,j\in J} |g_0\{h(x_i)\} - g_0\{h(x_j)\}| - \sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\}| \right| \end{aligned} \tag{59}$$

$$
\begin{aligned}
&\leq \sum_{(i,j)\in E_k\, i,j\in J} \Big[|g_0\{h(x_i)\} - g_\epsilon\{h(x_i)\}| + |g_0\{h(x_j)\} - g_\epsilon\{h(x_j)\}| \Big] \\
&\leq d_{\max} \sum_{i\in J} |g_0\{h(x_i)\} - g_\epsilon\{h(x_i)\}| \\
&\leq d_{\max} \sum_{i\in J} \int \psi_\epsilon\{h(x_i) - z\}\, |g_0\{h(x_i)\} - g_0(z)|\, dz \\
&\leq K\, \tau_d\, C_1\, \epsilon^{-d} \sum_{i\in J} \int_{\|h(x_i)-z\|_2\leq\epsilon} |g_0\{h(x_i)\} - g_0(z)|\, dz,
\end{aligned}
\tag{60}
$$

where τ_d is a positive constant and the second inequality happens with high probability as shown in Lemma 7.

We then bound the last term in (60) using Assumption 5. Thus,

$$
\begin{aligned}
\epsilon^{-d} \sum_{i\in J} \int_{\|h(x_i)-z\|_2\leq\epsilon} |g_0\{h(x_i)\} - g_0(z)|\, dz &= \epsilon^{-d} \sum_{A\in\mathcal{P}_\epsilon} \sum_{i\in J, h(x_i)\in A} \int_{\|h(x_i)-z\|_2\leq\epsilon} |g_0\{h(x_i)\} - g_0(z)|\, dz \\
&\leq \Big[\max_{z\in P_{\text{lat}}(N)} |C\{h^{-1}(z)\}| \Big]\; S_1(g_0, \mathcal{P}_{N^{-1},\mathcal{S}})\, N^d\, \epsilon \\
&\leq \{(1+\delta)\, c_{2,d}\, b_1'\, K\}\; S_1(g_0, \mathcal{P}_{N^{-1},\mathcal{S}})\, N^d\, \epsilon,
\end{aligned}
\tag{61}
$$

with probability at least

$$
1 - \frac{n}{K\, \tilde{a}^d\, L_{\max}^d} \exp\left(-\frac{1}{3}\delta^2\, b_2'\, c_{1,d}\, K \right),
$$

which follows from Lemma 13.

If $h(x_i) \notin \Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, then

$$
|g_0\{h(x_i)\} - g_0\{h(x_j)\}| \leq 2\, \|g\|_{L_\infty(0,1)^d}. \tag{62}
$$

We now proceed to put the different pieces together. Setting $\tilde{n} = |[n]\backslash J|$, we observe that

$$
\tilde{n} \sim \text{Binomial}\left\{ n\,,\, \text{pr}(\Lambda_\epsilon) \right\}.
$$

If

$$
n' \sim \text{Binomial}\left\{ n\,,\, p_{\max}\, C_{\mathcal{S}}\; 4\epsilon \right\},
$$

then by (57), we have

$$
\begin{aligned}
\text{pr}\left(\tilde{n} \geq \tfrac{3}{2}\, n\, p_{\max}\, C_{\mathcal{S}}\, \epsilon\, 4 \right) &\leq \text{pr}\left(n' \geq \tfrac{3}{2}\, n\, p_{\max}\, C_{\mathcal{S}}\, \epsilon\, 4 \right) \\
&\leq \exp\left\{ -\tfrac{1}{12}\, n\, \left(p_{\max}\, C_{\mathcal{S}}\, \epsilon \right) 4 \right\} \\
&= \exp\left[-\tfrac{p_{\max}\, 4\, C_{\mathcal{S}}}{12}\, n\, \left\{ \frac{2^{1/d}\, L_{\max}\, K^{1/d}}{\left(c_{1,d}\, p_{\min} \right)^{1/d}\, n^{1/d}} \right\} \right] \\
&= \exp\left(-\tilde{C}\, n^{1-1/d}\, K^{1/d} \right),
\end{aligned}
\tag{63}
$$

where the first inequality follows from (57), the second from Lemma 5, and $\tilde{C}$ is a positive constant that depends on $p_{\min}$, $p_{\max}$, $L_{\min}$, $L_{\max}$, d, and $C_{\mathcal{S}}$. Consequently, combining the above inequality with (59), (61) and (62) we arrive at

$$\begin{aligned}
\sum_{(i,j)\in E_K} |\theta_i^* - \theta_j^*| &= \sum_{(i,j)\in E_K,\, i,j\in J} |\theta_i^* - \theta_j^*| + \sum_{(i,j)\in E_K,\, i\notin J \text{ or } j\notin J} |\theta_i^* - \theta_j^*| \\
&\leq \sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\}| + \\
&\quad K\,\tau_d\,C_1\,\{(1+\delta)\,c_{2,d}\,b_1'\,K\}\;S_1(g_0,\mathcal{P}_{N^{-1}})\,N^d\,\epsilon \\
&\quad + 2\,\|g_0\|_{L_\infty(0,1)^d} K\,\tau_d\,\tilde{n} \\
&< \sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\}| + C_6\,n^{1-1/d}\,K^{1+1/d} \\
&\quad + 2\,\|g_0\|_{L_\infty(0,1)^d} K\,\tau_d\,\tilde{n},
\end{aligned}$$

for some positive constant $C_6 > 0$, which happens with high probability, see (63). Hence, from (63),

$$\begin{aligned}
\sum_{(i,j)\in E_K} |\theta_i^* - \theta_j^*| \leq \sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\}| + C_6\,n^{1-1/d}\,K^{1+1/d} \\
+ C_7\,n^{1-1/d}\,K^{1+1/d},
\end{aligned} \tag{64}$$

where $C_7 > 0$ is a constant, and the last inequality holds with probability approaching one provided that $K/\log n \to \infty$.

Therefore, it remains to bound the first term in the right hand side of inequality (64). To that end, we notice that if $(i,j) \in E_K$, then from the proof of Lemma 13 we observe that

$$\|h(x_i) - h(x_j)\|_2 \leq L_{\max}\,R_{K,\max} \leq \epsilon,$$

where the last inequality happens with high probability. Hence, with high probability, if z is in the segment connecting $h(x_i)$ and $h(x_j)$, then $z \notin B_{2\epsilon}(\mathcal{S}) \cup ((0,1)^d \backslash \Omega_{2\epsilon})$ provided that $i,j \in J$. As a result, by the mean value theorem, for $i,j \in J$ there exists a $z_{i,j} \in \Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S})$ such that

$$g_\epsilon\{h(x_i)\} - g_\epsilon\{h(x_j)\} = \nabla g_\epsilon(z_{i,j})^T\{h(x_i) - h(x_j)\},$$

and this holds uniformly with probability approaching one.

Then,

$$
\begin{aligned}
&\sum_{(i,j)\in E_K\, i,j\in J} |g_\epsilon\{h(x_i)\} \,-\, g_\epsilon\{h(x_j)\}| \\
&= \sum_{(i,j)\in E_K\, i,j\in J} |\nabla g_\epsilon(z_{i,j})^T\{h(x_i) - h(x_j)\}| \\
&\leq \left\{ \sum_{(i,j)\in E_K\, i,j\in J} \|\nabla g_\epsilon(z_{i,j})\|_1 \right\} \tfrac{2}{N} \\
&= 2 \sum_{A\in\mathcal{P}_\epsilon,\, A\cap\{\Omega_{2\epsilon}\setminus B_{2\epsilon}(\mathcal{S})\}\neq\emptyset} \left[\sum_{(i,j)\in E_K \text{ s.t } i,j\in J, \text{ and } z_{i,j}\in A} \|\nabla g_\epsilon(z_{i,j})\|_1 \,\mathrm{Vol}\{B_\epsilon(z_{i,j})\} \right] \tfrac{C_8\, N^d}{N} \\
&\leq 2 \sum_{A\in\mathcal{P}_\epsilon,\, A\cap\{\Omega_{2\epsilon}\setminus B_{2\epsilon}(\mathcal{S})\}\neq\emptyset} \left\{ \sum_{(i,j)\in E_K \text{ s.t } i,j\in J, \text{ and } z_{i,j}\in A} \int\limits_{B_\epsilon(z_{i,j})} \|\nabla g_\epsilon(z)\|_1\, dz \right\} \tfrac{C_8\, N^d}{N} \,+ \\
&\quad 2 \sum_{A\in\mathcal{P}_\epsilon,\, A\cap\{\Omega_{2\epsilon}\setminus B_{2\epsilon}(\mathcal{S})\}\neq\emptyset} \left\{ \sum_{(i,j)\in E_K \text{ s.t } i,j\in J, \text{ and } z_{i,j}\in A} \int\limits_{B_\epsilon(z_{i,j})} |\|\nabla g_\epsilon(z_{i,j})\|_1 \,-\, \|\nabla g_\epsilon(z)\|_1|\, dz \right\} \tfrac{C_8\, N^d}{N} \\
&=: T_1 \,+\, T_2.
\end{aligned}
\tag{65}
$$

Therefore we proceed to bound T_1 and T_2. Let us assume that $(i,j)\in E_K$ with $i,j\in J,$ and $z_{i,j}\in A$ with $A\in\mathcal{P}_\epsilon$. Then by Lemma 13 we have two cases. Either $h(x_i)$ and $h(x_j)$ are in the same cell (element of $\mathcal{P}_\epsilon$), or $h(x_i)$ and $h(x_j)$ are in adjacent cells. Denoting by $c(A')$ the center of a cell $A'\in\mathcal{P}_\epsilon$, then if $z'\in B_\epsilon(z_{i,j})\cap A'$ it implies that

$$
\|c(A') \,-\, c(A)\|_\infty \;\leq\; \|c(A') \,-\, z'\|_\infty + \|z' \,-\, z_{i,j}\|_\infty + \|z_{i,j} \,-\, c(A)\|_\infty \;\leq\; 2\epsilon.
$$

And if in addition $h(x_i)\in A_i$ and $h(x_j)\in A_j$, then

$$
\|c(A_i) \,-\, c(A)\|_\infty \;\leq\; \|c(A_i) \,-\, h(x_i)\|_\infty + \|h(x_i) \,-\, z_{i,j}\|_\infty + \|z_{i,j} \,-\, c(A)\|_\infty \;\leq\; 2\epsilon, \tag{66}
$$

and the same is true for $c(A_j)$. Hence,

$$
\int\limits_{B_\epsilon(z_{i,j})} \|\nabla g_\epsilon(z)\|_1\, dz \;\leq\; \sum_{A'\in\mathcal{P}_\epsilon\,:\,\|c(A)-c(A')\|_\infty\leq 2\epsilon} \int\limits_{A'} \|\nabla g_\epsilon(z)\|_1\, dz.
$$

Since the previous discussion was for an arbitrary $z_{i,j}$ with $i,j\in J$, we obtain that

$$
\begin{aligned}
T_1 \;&\leq \sum_{A\in\mathcal{P}_\epsilon,\, A\cap\{\Omega_{2\epsilon}\setminus B_{2\epsilon}(\mathcal{S})\}\neq\emptyset} \left\{ \sum_{(i,j)\in E_K \text{ s.t } i,j\in J, \text{ and } z_{i,j}\in A} \;\; \sum_{A'\in\mathcal{P}_\epsilon\,:\,\|c(A)-c(A')\|_\infty\leq 2\epsilon} \int\limits_{A'} \|\nabla g_\epsilon(z)\|_1\, dz \right\} \tfrac{2\,C_8\, N^d}{N} \\
&\leq \tfrac{2\,C_8\, N^d}{N} \sum_{A\in\mathcal{P}_\epsilon} \left[|\{A'\in\mathcal{P}_\epsilon \,:\, \|c(A) \,-\, c(A')\|_\infty \leq 2\epsilon\}| \right]^3 \max_{x\in h^{-1}\{P_{\text{lat}}(N)\}} |C(x)|^2 \int\limits_A \|\nabla g_\epsilon(z)\|_1\, dz \\
&\leq C_9\, N^{d-1} \max_{x\in h^{-1}\{P_{\text{lat}}(N)\}} |C(x)|^2 \sum_{A\in\mathcal{P}_\epsilon} \int\limits_A \|\nabla g_\epsilon(z)\|_1\, dz,
\end{aligned}
$$

where C_9 is positive constant depending on d which can be obtained with an entropy argument similar to (52). As a result, from (55) we obtain

$$T_1 \;\leq\; C_2 C_9\, N^{d-1} \max_{x \in h^{-1}\{P_{\text{lat}}(N)\}} |C(x)|^2. \tag{67}$$

It remains to bound T_2 in (65). Towards that goal, we observe that

$$\begin{aligned}
\int\limits_{B_\epsilon(z_{i,j})} |\|\nabla g_\epsilon(z_{i,j})\|_1 \,-\, \|\nabla g_\epsilon(z)\|_1|\, dz \;&\leq\; \int\limits_{B_\epsilon(z_{i,j})} \sum_{l=1}^{d} \left| \frac{\partial g_\epsilon}{\partial z_l}(z_{i,j}) \,-\, \frac{\partial g_\epsilon}{\partial z_l}(z) \right| dz \\
&=\; \int\limits_{B_\epsilon(z_{i,j})} \sum_{l=1}^{d} \left| \frac{1}{\epsilon^{d+1}} \int\limits_{B_\epsilon(0)} \frac{\partial \psi}{\partial z_l}(z'/\epsilon) \left\{ g_0(z_{i,j} - z') \,-\, g_0(z - z') \right\} dz' \right| dz \\
&\leq\; \int\limits_{B_\epsilon(z_{i,j})} \sum_{l=1}^{d} \left| \frac{1}{\epsilon^{d} \|z_{i,j} \,-\, z\|_2} \int\limits_{B_\epsilon(0)} \frac{\partial \psi(z'/\epsilon)}{\partial z_l} \{ g_0(z_{i,j} - z') - \right. \\
&\qquad\qquad \left. g_0(z - z') \right\} dz' \Bigg| dz,
\end{aligned}$$

which implies that for some $C_{10} > 0$

$$\int\limits_{B_\epsilon(z_{i,j})} |\|\nabla g_\epsilon(z_{i,j})\|_1 \,-\, \|\nabla g_\epsilon(z)\|_1|\, dz \;\leq\; C_{10}\, \epsilon^d \sup_{z \in B_\epsilon(z_{i,j})} \sum_{l=1}^{d} \left| \int\limits_{B_\epsilon(0)} \frac{\partial \psi(z'/\epsilon)}{\partial z_l} \frac{\{ g_0(z_{i,j} - z') \,-\, g_0(z - z') \}}{\epsilon^d \|z_{i,j} \,-\, z\|_2}\, dz' \right|$$

and so

$$\begin{aligned}
&T_2 \\
&\leq \sum_{A \in \mathcal{P}_\epsilon,\; A \cap (\Omega_{2\epsilon} \setminus B_{2\epsilon}(\mathcal{S})) \neq \emptyset} \left[\sum_{(i,j) \in E_K \; i,j \in J,\; z_{i,j} \in A} \sup_{z \in B_\epsilon(z_{i,j})} \sum_{l=1}^{d} \left| \int\limits_{\|z'\|_2 \leq \epsilon} \frac{\partial \psi(z'/\epsilon)}{\partial z_l} \left(\frac{g_0(z_{i,j} - z') \,-\, g_0(z - z')}{\|z - z_{i,j}\|_2\, \epsilon^d} \right) dz' \right| \right] \\
&\quad \cdot 2\, C_{10}\, C_8\, \epsilon^d\, N^{d-1} \\
&= \sum_{A \in \mathcal{P}_\epsilon,\; A \cap (\Omega_{2\epsilon} \setminus B_{2\epsilon}(\mathcal{S})) \neq \emptyset} \left[\sum_{(i,j) \in E_K \; i,j \in J,\; z_{i,j} \in A} T(g_0, z_{i,j}) \epsilon^d \right] \cdot 2\, C_{10}\, C_8\; N^{d-1},
\end{aligned} \tag{68}$$

with $T(g_0, z)$ as in Equation (10) in the main paper. Now, if $z_{i,j} \in A$, $h(x_i) \in A_i$ and $h(x_j) \in A_j$, then just as in (66) we obtain that

$$\max\{\|c(A_i) \,-\, c(A)\|_\infty, \|c(A_j) \,-\, c(A)\|_\infty\} \;\leq\; 2\epsilon.$$

Hence, since by construction we also have $z_{i,j} \in \Omega_{2\epsilon} \setminus B_{2\,\epsilon}(\mathcal{S})$ then

$$\begin{aligned}
\sum_{(i,j) \in E_K \; i,j \in J,\; z_{i,j} \in A} T(g_0, z_{i,j}) \;&\leq\; |\{\{i,j\} \,:\, \max\{\|c(A_i) \,-\, c(A)\|_\infty, \|c(A_j) \,-\, c(A)\|_\infty\} \;\leq\; 2\epsilon\}| \\
&\qquad \sup_{z_A \in A \cap (\Omega_{2\epsilon} \setminus B_{2\,\epsilon}(\mathcal{S}))} T(g_0, z_A),
\end{aligned}$$

which combined with (68) implies that

$$\begin{aligned} T_2 \;\leq\; & \sum_{A\in\mathcal{P}_\epsilon,\; A\cap(\Omega_{2\epsilon}\setminus B_{2\epsilon}(\mathcal{S}))\neq\emptyset} \left[\sup_{z_A\in A\cap(\Omega_{2\epsilon}\setminus B_{2\,\epsilon}(\mathcal{S}))} T(g_0,z_{i,j})\epsilon^d\right]\cdot 2\,C_{10}\,C_8\;N^{d-1}\,\cdot \\ & |\{\{i,j\}\;:\;\max\{\|c(A_i)\;-\;c(A)\|_\infty,\|c(A_j)\;-\;c(A)\|_\infty\}\;\leq\;2\epsilon\}| \\ \leq\; & C_{11}\;N^{d-1}\max_{x\in h^{-1}(P_{\text{lat}}(N))}|C(x)|^2, \end{aligned} \tag{69}$$

for some positive constant C_{11}, where the last inequality follows from Assumption 5 and that fact that for every A the set of pairs of cells with centers within distance 2ϵ is constant. Combining (64), (67), (65), (69) and Lemma 13, we obtain that

$$\sum_{(i,j)\in E_K}|\theta_i^*\;-\;\theta_j^*|\;=\;O_{\text{pr}}(\text{poly}(\log n)\,n^{1-1/d}), \tag{70}$$

when Assumptions 1–5 hold.

To conclude the proof, we proceed to verify (70) when Assumptions 1–3 hold and f_0 satisfies Definition 1. Using the notation from before, we observe that (57) still holds and for J in (58) we have that

$$\begin{aligned} \sum_{(i,j)\in E_K}|\theta_i^*\;-\;\theta_j^*| \;&=\; \sum_{(i,j)\in E_K,\,i,j\in J}|\theta_i^*\;-\;\theta_j^*|\;+\;\sum_{(i,j)\in E_K,\,i\notin J\text{ or }j\notin J}|\theta_i^*\;-\;\theta_j^*| \\ &\leq\; \sum_{(i,j)\in E_K\;i,j\in J}|f_0(x_i)\;-\;f_0(x_j)|\;+\;2\,\|g_0\|_{L_\infty(0,1)^d}K\,\tau_d\,\tilde{n}. \end{aligned} \tag{71}$$

So the claim follows from combining Lemma 6, (63), and the piecewise Lipschitz condition.

□

J Proof of Theorem 3

Proof. Recall

$$\begin{aligned} y_i \;&=\; \theta_i^*\;+\;\varepsilon_i,\quad i=1,\ldots,n, \\ \theta_i^* \;&=\; f_{0,z_i}(x_i), \\ x_i \;&\sim\; p_{z_i}(x), \\ \text{pr}(z_i=l) \;&\sim\; \pi_l^*,\;\text{ for } l=1,\ldots,\ell. \end{aligned} \tag{72}$$

Next we let $A_l\;=\;\{i\in[n]\;:\;z_i=l\}$, and $n_l\;=\;|A_l|$ for $l\;=\;1,\ldots,\ell$. Then by Proposition 27 in Von Luxburg et al. (2014) we have that the event

$$\frac{n\pi_l^*}{2}\;\leq\;n_l\;\leq\;\frac{3n\pi_l^*}{2},\quad\text{for}\;\;l=1,\ldots,\ell, \tag{73}$$

happens with probability at least $1\;-\;2l\exp(-n\min\{\pi_l^*\;:\;l\in[\ell]\}/12)$. Therefore, we will assume that the event defined in (73) holds.

We proceed by conditioning on $\{z_i\}_{i=1}^n$, and for simplicity we omit to make this conditioning explicit. For $i\in[n]$ we write $l(i):=j\in[\ell]$ if $i\in A_j$. We also introduce the following notation:

$$\begin{aligned} N_K(x_i) \;&=\; \left\{\text{ set of nearest neighbors of } x_i \text{ in the } K\text{-NN graph constructed from points } \{x_{i'}\}_{i'=1,\ldots,n}\right\}, \\ \tilde{N}_K(x_i) \;&=\; \left\{\text{ set of nearest neighbors of } x_i \text{ in the } K\text{-NN graph constructed from points } \{x_{i'}\}_{i'\in A_{l(i)}}\right\}, \\ \tilde{R}^l_{K,\max} \;&=\; \max_{i\in A_l}\max_{x\in\tilde{N}_K(x_i)} d_{\mathcal{X}}(x,x_i), \end{aligned}$$

and we denote by $\nabla_{G_K^l}$ the incidence matrix of the K-NN graph corresponding to the points $\{x_i\}_{i\in A_l}$.

Next we observe that just as in the proof of Lemma 7,

$$\text{pr}\left\{\tilde{R}^l_{K,\max} \,>\, \tilde{a}_l(K/n_l)^{1/d_l}\right\} \,\leq\, n_l\,\exp(-K/12), \tag{74}$$

for some positive constant $\tilde{a}_l$. We write $\epsilon_l \,=\, \tilde{a}_l(K/n_l)^{1/d_l}$, and consider the sets

$$\Lambda_l \,=\, \left\{i\in A_l \,:\, \text{such that } N_K(x_i) \,=\, \tilde{N}_K(x_i)\right\}.$$

Our goal is to use these sets in order to split the basic inequality for the K-NN-FL into different processes corresponding to the different sets $\mathcal{X}_l$. To that end let us pick $l\in[\ell]$. We notice that if $\partial\mathcal{X}_l \,=\, \emptyset$, then by Assumption 6 we have that for all $x\in\mathcal{X}_l$ it holds that $B_\epsilon(x)\subset\mathcal{X}_l$ for small enough ϵ. Hence, with high probability, $N_K(x_i) \,=\, \tilde{N}_K(x_i)$ for all $i\in A_l$.

Let us now assume that $\partial\mathcal{X}_l \,\neq\, \emptyset$. Let $i\in A_l$. Then

$$\text{pr}\left\{x_i\in B_{\epsilon_l}(\partial\mathcal{X}_l)\right\} \,\geq\, p_{l,\min}\,\mu_l\left\{B_{\epsilon_l}\left(\partial\mathcal{X}_l\right)\bigcap\mathcal{X}_l\right\} \,\geq\, c_l'\,\epsilon_l^{d_l}, \tag{75}$$

where $p_{l,\min} \,=\, \min_{x\in\mathcal{X}_l} p_l(x)$, and c_l' is a positive constant that exists because $\mathcal{X}_l$ satisfies Assumption 2. On the other hand, (14) implies that for $i\in A_l$ we have

$$\text{pr}\left\{x_i\in B_{\epsilon_l}(\partial\mathcal{X}_l)\right\} \,\leq\, p_{l,\max}\,\mu_l\left\{B_{\epsilon_l}\left(\partial\mathcal{X}_l\right)\bigcap\mathcal{X}_l\right\} \,\leq\, p_{l,\max}\,\tilde{c}_l\,\epsilon_l, \tag{76}$$

where $p_{l,\max} \,=\, \max_{x\in\mathcal{X}_l} p_l(x)$, and $\tilde{c}_l$ is a positive constant.

Therefore, combining (15), (74), (75) , and (76) with Lemma 5, we obtain that

$$\begin{aligned}\text{pr}\left(n_l - |\Lambda_l| \,\leq\, \tfrac{3}{2}p_{l,\max}\,\tilde{c}_l\,n_l\,\epsilon_l\right) \,&\geq\, 1 \,-\, p_{\max}\exp\left(-\tfrac{1}{12}c_l'\,\tilde{a}_l^{d_l}\,K\right) \,-\, n_l\,\exp(-K/12)\\ &\geq\, 1 \,-\, p_{\max}\exp\left(-\tfrac{1}{12}c_l'\,\tilde{a}_l^{d_l}\,K\right) \,-\, \exp(-K/12+\log n).\end{aligned} \tag{77}$$

Next we see how the previous inequality can be used to put an upper bound on the penalty term of the K-NN-FL. For any $\theta\in\mathbb{R}^n$, we have

$$2\|\nabla_{G_K}\theta\|_1 \,=\, \sum_{l=1}^{\ell}\sum_{i=1}^{n_l}\sum_{j\in N_K(x_i)}|\theta_i \,-\, \theta_j| \,=\, \sum_{l=1}^{\ell}\sum_{i=1}^{n_l}\sum_{j\in\tilde{N}_K(x_i)}|\theta_i \,-\, \theta_j| \,+\, R(\theta),$$

where

$$\begin{aligned}|R(\theta)| \,&=\, \left|\sum_{l=1}^{\ell}\sum_{i=1}^{n_l}\sum_{j\in N_K(x_i)}|\theta_i \,-\, \theta_j| \,-\, \sum_{l=1}^{\ell}\sum_{i=1}^{n_l}\sum_{j\in\tilde{N}_K(x_i)}|\theta_i \,-\, \theta_j|\right|\\ &=\, \left|\sum_{l=1}^{\ell}\sum_{i\in[n_l]\backslash\Lambda_l}\sum_{j\in N_K(x_i)}|\theta_i \,-\, \theta_j| \,-\, \sum_{l=1}^{\ell}\sum_{i\in[n_l]\backslash\Lambda_l}\sum_{j\in\tilde{N}_K(x_i)}|\theta_i \,-\, \theta_j|\right|\\ &\leq\, 4\,\tau_d\,\|\theta\|_\infty\,K\sum_{l=1}^{\ell}|[n_l]\backslash\Lambda_l|,\end{aligned} \tag{78}$$

where τ_d is a positive constant depending only on d, and the second inequality follows from the well-known bound on the maximum degree of K-NN graphs, see Corollary 3.23 in Miller et al. (1997). Combining with (77), we obtain that

$$R(\theta) \,-\, R(\hat{\theta}) \,=\, O_{\text{pr}}\left[\left\{(\log n)^{\frac{1}{2}} \,+\, K\,\lambda\right\}\,K^{1+1/d_l}\sum_{l=1}^{\ell}(n\pi_l^*)^{1-1/d_l}\right]. \tag{79}$$

Note that if $\partial\mathcal{X}_l = \emptyset$, then (79) will still hold as $R(\theta) = 0$ for all $\theta \in \mathbb{R}^n$ with high probability. To conclude the proof, we notice by the basic inequality that

$$\begin{aligned}\|\theta^* - \hat{\theta}\|_n^2 &\leq \tfrac{1}{n}\varepsilon^T\left(\hat{\theta} - \theta^*\right) + \tfrac{\lambda}{n}\left(\|\nabla_{G_K}\theta^*\|_1 - \|\nabla_{G_K}\hat{\theta}\|_1\right) \\ &= \sum_{l=1}^{\ell}\varepsilon_{A_l}^T\left(\hat{\theta}_{A_l} - \theta^*_{A_l}\right) + \sum_{l=1}^{\ell}\frac{\lambda}{2n}\left(\|\nabla_{G_K^l}\theta^*_{A_l}\|_1 - \|\nabla_{G_K^l}\hat{\theta}_{A_l}\|_1\right) + \frac{\lambda}{2n}\left\{R(\theta^*) - R(\hat{\theta})\right\},\end{aligned}$$

where the notation x_A denotes the vector x with the coordinates with indices not in A removed. The proof follows from (79) and from bounding each term

$$\frac{1}{n}\varepsilon_{A_l}^T\left(\hat{\theta}_{A_l} - \theta^*_{A_l}\right) + \frac{\lambda}{n}\left(\|\nabla_{G_K^l}\theta^*_{A_l}\|_1 - \|\nabla_{G_K^l}\hat{\theta}_{A_l}\|_1\right),$$

as in the proof of Theorem 12.

□

K Manifold Adaptivity Example

The following example suggests that the ϵ-NN-FL estimator may not be manifold adaptive.

Example 2. For $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \subset \mathbb{R}^3$, suppose that $\mathcal{X}_1 = [0,1]^2 \times \{c\}$ for some $c < 0$, and $\mathcal{X}_2 = [0,1]^3$. Note that the sets $\mathcal{X}_1$ and $\mathcal{X}_2$ satisfy Assumption 6. Let us assume that all of the conditions of Theorem 3 are met with $n_1 := n\pi_1^* \asymp n$, and $n_2 := n\pi_2^* \asymp n^{3/4}$. Motivated by the scaling of ϵ in Theorem 2, we let $\epsilon \asymp \text{poly}(\log n)/n^{1/t}$. We now consider two possibilities for t: $t \in (0,2]$, and $t > 2$.

- For any $t \in (0,2]$, and for any positive constant $a \in (0, 3/4)$, there exists a positive constant $c(a)$ such that $E(\|\hat{\theta}_\epsilon - \theta^*\|_n^2) \geq c(a)\, n^{-1/4-a}$ for large enough n, where $\hat{\theta}_\epsilon$ is the ϵ-NN-FL estimator, see proof below. In other words, if $t < 2$, then ϵ-NN-FL does not achieve a desirable rate. By contrast, with an appropriate choice of λ and K, the K-NN-FL estimator attains the rate $n^{-1/2} \asymp n_1^{1-1/2}/n + n_2^{1-1/3}/n$ (ignoring logarithmic terms) by Theorem 3.
- If $t > 2$, then the minimum degree in the ϵ-graph of the observations in $\mathcal{X}_1$ will be at least $c(t)n^{1-2/t}$, with high probability, for some positive constant $c(t)$. This has two consequences:
  1. Recall that the algorithm from Chambolle and Darbon (2009) has worst-case complexity $O(mn^2)$, where m is the number of edges in the graph (although empirically the algorithm is typically much faster, Wang et al., 2016). Therefore, if t is too large, then the computations involved in applying the fused lasso to the ϵ-NN graph may be too demanding.
  2. In a different context, El Alaoui et al. (2016) argues that it is desirable for geometric graphs (which generalize ϵ-NN graphs) to have degree $\text{poly}(\log n)$. This suggests that using $t > 2$ leads to an ϵ-NN graph that is too dense.

Next, we proceed to justify the lower bound on the MSE of $\hat{\theta}_\epsilon$ when $t \in (0,2]$ in Example 2. We begin by noticing that if $i, j \in A_2$ with $i \neq j$, then for the choice of $\epsilon > 0$ in this example, we have that

$$\text{pr}\{d_{\mathcal{X}}(x_i, x_j) \leq \epsilon\} \leq p_{2,\max}\int_{\mathcal{X}}\text{pr}\{d_{\mathcal{X}}(x, x_j) \leq \epsilon\}\mu_2(dx) \leq \tilde{k}_l\,\frac{(\text{poly}(\log n))^3}{n^{3/t}}.$$

Let $m \asymp n^{3/4-a}$, and $j_1 < j_2 < \ldots < j_m$ elements of A_2. Then the event

$$\Lambda = \cap_{s=1}^m\{d_{\mathcal{X}}(x_{j_s}, x_l) > \epsilon,\ \forall l \in A_2\backslash\{j_s\}\},$$

satisfies

$$\text{pr}\,(\Lambda) \;\geq\; 1 \;-\; c_2\, n^{3/2-3/t-a}\,(\text{poly}(\log n))^3,$$

for some positive constant c_2. Therefore,

$$\mathbb{E}\left\{\sum_{i=1}^{n}(\theta_i^* \;-\; \hat{\theta}_{\epsilon,i})^2\right\} \;\geq\; \mathbb{E}\left\{\sum_{i=1}^{n}(\theta_i^* \;-\; \hat{\theta}_{\epsilon,i})^2 \mid \Lambda\right\} \text{pr}(\Lambda) \;\geq\; m\,\sigma^2[1 - c_s\, n^{3/2-3/t-a}\,\{\text{poly}(\log n)\}^3] \;\geq\; C_1\, n^{3/4-a}$$

for some positive constant C_1 if n is large enough.

L Proving Theorems 1–2 for ϵ-NN graphs

We start by giving an overview of how the conclusion of Theorem 1 in the main paper can be obtained for ϵ-NN graphs. The general idea is described next. The first step is to obtain a lemma that controls the maximum and minimum degrees of the ϵ-NN graph.

Lemma 16. ***(See Proposition 29 in Von Luxburg et al. (2014)).*** *Suppose that Assumptions 1–3 hold. Let $d_{\min}$ and $d_{\max}$ denote the minimum and maximum degrees of an ϵ-NN graph, respectively. Then for all $\delta \in (0,1)$ we have that*

$$\begin{aligned} pr\left\{d_{\max} \geq (1+\delta)n\epsilon^d c_{\max}\right\} &\leq \exp\left(-\tfrac{\delta^2 n\epsilon^d c_{\max}}{3}\right), \\ pr\left\{d_{\max} \leq (1-\delta)n\epsilon^d c_{\min}\right\} &\leq \exp\left(-\tfrac{\delta^2 n\epsilon^d c_{\min}}{3}\right), \end{aligned}$$

for positive constants $c_{\min}$ and $c_{\max}$.

Next we present a lemma controlling the maximum and minimum counts of the mesh. This is similar to Lemma 7.

Lemma 17. *Let $\epsilon \asymp \log^{(1+2r)/d} n/n^{1/d}$, then there exists N such that if*

$$N \;\asymp\; \left\lceil \frac{1}{\epsilon} \right\rceil$$

in the construction of $P_{lat}(N)$ defined in (24), then the following properties hold:

- *There exist positive constants b_2' and b_2' such that*

$$\begin{aligned} \text{pr}\left\{\max_{x\in I(N)}|C(x)| \;\geq\; (1+\delta)\, b_2' \;\log^{1+2r} n\right\} &\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\, b_2' \;\log^{1+2r} n\right), \\ \text{pr}\left\{\min_{x\in I(N)}|C(x)| \;\leq\; (1-\delta)\, b_1' \;\log^{1+2r} n\right\} &\leq\; N^d \exp\left(-\tfrac{1}{3}\delta^2\, b_1' \;\log^{1+2r} n\right), \end{aligned}$$

  for all $\delta \in (0,1)$.

- *Define the event Ω as: "If $x_i \in C(x_i')$ and $x_j \in C(x_j')$ for $x_i', x_j' \in I(N)$ with $\|h(x_i') \;-\; h(x_j')\|_2 \leq N^{-1}$, then x_i and x_j are connected in the ϵ-NN graph". Then for some constant $c > 0$,*

$$\text{pr}\,(\Omega) \;\geq\; 1 \;-\; n\,\exp(-c\log^{1+2r} n/3).$$

The proof of Theorem 1 results from setting ϵ as in Lemma 17, and then following the steps in the proof of Theorem 1 in the main document. Specifically, we can follow the proofs of Lemmas 8–11, and Theorem 12, by replacing ∇_{G_K} with ∇_{G_ϵ}, and K with $c \log^{1+2r} n$ for a constant c.

To prove the conclusion of Theorem 2 for the ϵ-NN graph, we need lemmas similar to those in Section H. The first of these lemmas is given next.

Lemma 18. *Let $\epsilon \asymp \log^{(1+2r)/d} n/n^{1/d}$. Then there exists N in the construction of G_{lat} with*

$$N \asymp \lceil \epsilon^{-1} \rceil,$$

and positive constants b_1' and b_2' depending on $L_{\min}$, $L_{\max}$, d, $p_{\min}$, $c_{1,d}$, and $c_{2,d}$, such that

$$\begin{aligned} \text{pr}\left\{ \max_{x \in h^{-1}(P_{lat})} |C(x)| \;\geq\; (1+\delta)\, c_{2,d}\, b_1'\, \log^{(1+2r)} n \right\} &\leq N^d \exp\left(-\tfrac{1}{3}\delta^2\, b_2'\, c_{1,d}\, \log^{(1+2r)} n \right), \\ \text{pr}\left\{ \min_{x \in h^{-1}(P_{lat})} |C(x)| \;\leq\; (1-\delta)\, c_{1,d}\, b_2'\, \log^{(1+2r)} n \right\} &\leq N^d \exp\left(-\tfrac{1}{3}\delta^2\, b_2'\, c_{1,d}\, \log^{(1+2r)} n \right), \end{aligned} \tag{80}$$

for all $\delta \in (0,1]$. Moreover, let $\tilde{\Omega}$ denote the event: "For all $i, j \in [n]$, if x_i and x_j are connected in the ϵ-NN graph, then $\|h(x_i') - h(x_j')\|_2 < 2\, N^{-1}$ where $x_i \in C(x_i')$ and $x_j \in C(x_j')$ with $x_i', x_j' \in I(N)$". Then

$$\text{pr}\left(\tilde{\Omega}\right) \;\geq\; 1 \;-\; n\, \exp(-\log^{(1+2r)} n/3).$$

Using Lemma 18, we can obtain the conclusions of Lemmas 14–15 and Theorem 2 for the ϵ-NN graph. This can be done with minor modifications to the proofs in Section H.

Acknowledgements

JS is partially supported by NSF Grant DMS-1712996. DW is partially supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and a Simons Investigator Award in Mathematical Modeling of Living Systems.

References

Morteza Alamgir, Gábor Lugosi, and Ulrike Luxburg. Density-preserving quantization with application to graph downsampling. In *Conference on Learning Theory*, pages 543–559, 2014.

Ery Arias-Castro, Joseph Salmon, and Rebecca Willett. Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods. *SIAM Journal on Imaging Sciences*, 5(3):944–992, 2012.

Álvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *arXiv preprint arXiv:1411.0589*, 2014.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. Institute of Mathematical Statistics, 2007.

Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9): 1124–1137, 2004.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

Rui M Castro, Rebecca Willett, and Robert Nowak. Faster rates in regression via active learning. In *Tech. Rep., University of Wisconsin, Madison, June 2005, ECE-05-3 Technical Report (available at http://homepages.cae.wisc.edu/ rcastro/ECE-05-3.pdf)*., 2005.

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

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

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Ming-Yen Cheng and Hau-tieng Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

Sanjoy Dasgupta. Consistency of nearest neighbor classification under selective sampling. In *Conference on Learning Theory*, pages 18–1, 2012.

Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems*, pages 2555–2563, 2014.

Sanjoy Dasgupta and Kaushik Sinha. Randomized partition trees for exact nearest neighbor search. In *Conference on Learning Theory*, pages 317–337, 2013.

P. Laurie Davies and Arne Kovac. Local extremes, runs, strings and multiresolution. *Annals of Statistics*, 29(1):1–65, 2001.

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

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

Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of\ell_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

Abderrahim Elmoataz, Olivier Lezoray, and Sébastien Bougleux. Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. *IEEE transactions on Image Processing*, 17(7):1047–1060, 2008.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

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

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 (TOMS)*, 3(3):209–226, 1977.

Mariano Giaquinta and Giuseppe Modica. *Mathematical Analysis: An Introduction to Functions of Several Variables*. Springer Science & Business Media, 2010.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Spatial adaptation in trend filtering. *arXiv preprint arXiv:1702.05113*, 2017.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Martin T Hagan, Howard B Demuth, Mark H Beale, and Orlando De Jesús. *Neural Network Design*, volume 20. Pws Pub. Boston, 1996.

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

Tomislav Hengl, Gerard BM Heuvelink, and David G Rossiter. About regression-kriging: from equations to case studies. *Computers & geosciences*, 33(10):1301–1315, 2007.

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

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. *Annual Conference on Learning Theory*, 29:1115–1146, 2016.

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

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

Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. Active nearest-neighbor learning in metric spaces. In *Advances in Neural Information Processing Systems*, pages 856–864, 2016.

Samory Kpotufe. Escaping the curse of dimensionality with a tree-based regressor. *arXiv preprint arXiv:0902.3453*, 2009.

Samory Kpotufe. K-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

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

Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6887–6896, 2017.

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

Gary L Miller, Shang-Hua Teng, William Thurston, and Stephen A Vavasis. Separators for sphere-packings and nearest neighbor graphs. *Journal of the ACM (JACM)*, 44(1):1–29, 1997.

Francesco Ortelli and Sara van de Geer. On the total variation regularized estimator over a class of tree graphs. *Electronic Journal of Statistics*, 12(2):4517–4570, 2018.

Oscar Hernan Madrid Padilla, James G Scott, James Sharpnack, and Ryan J Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18(176):1–36, 2018.

Ashley Petersen, Noah Simon, and Daniela Witten. Convex regression with interpretable sharp partitions. *The Journal of Machine Learning Research*, 17(1):3240–3270, 2016a.

Ashley Petersen, Daniela Witten, and Noah Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025, 2016b.

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

Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *arXiv preprint arXiv:1702.05037*, 2017.

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*, pages 3513–3521, 2016.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James L Sharpnack, and Ryan J Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, pages 5796–5806, 2017.

Shashank Singh and Barnabás Póczos. Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv preprint arXiv:1603.08578*, 2016.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

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

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*,

42(1):285–323, 2014.

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

Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large graphs are often misleading. *Journal of Machine Learning Research*, 15:1751–1798, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

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.

Yun Yang and David B Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2):876–905, 2016.

Yun Yang, Surya T Tokdar, et al. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.

Chi Zhang, Feifei Li, and Jeffrey Jestes. Efficient parallel knn joins for large data in mapreduce. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 38–49. ACM, 2012.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *International Conference on Machine Learning*, 20:912–919, 2003.

William P Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120. Springer Science & Business Media, 2012.