Journal of Machine Learning Research 25 (2024) 1-45 Submitted 8/23; Revised 5/24; Published 8/24

Variance estimation in graphs with the fused lasso

Oscar Hernan Madrid Padilla OSCAR.MADRID@STAT.UCLA.EDU
Department of Statistics and Data Science
Univeristy of California, Los Angeles
Los Angeles, CA 90095.

Editor: Pierre Alquier

Abstract

We study the problem of variance estimation in general graph-structured problems. First, we develop a linear time estimator for the homoscedastic case that can consistently estimate the variance in general graphs. We show that our estimator attains minimax rates for the chain and 2D grid graphs when the mean signal has total variation with canonical scaling. Furthermore, we provide general upper bounds on the mean squared error performance of the fused lasso estimator in general graphs under a moment condition and a bound on the tail behavior of the errors. These upper bounds allow us to generalize for broader classes of distributions, such as sub-Exponential, many existing results on the fused lasso that are only known to hold with the assumption that errors are sub-Gaussian random variables. Exploiting our upper bounds, we then study a simple total variation regularization estimator for estimating the signal of variances in the heteroscedastic case. We also provide lower bounds showing that our heteroscedastic variance estimator attains minimax rates for estimating signals of bounded variation in grid graphs, and K-nearest neighbor graphs, and the estimator is consistent for estimating the variances in any connected graph.

Keywords: Total variation, conditional variance estimation, nonparametric regression.

1. Introduction

Consider the problem of estimating signals $\theta^* \in \mathbb{R}^n$ and $v^* \in \mathbb{R}^n_+$, based on data $\{y_i\}_{i=1}^n \subset \mathbb{R}$ generated as

$$y_i \,=\, \theta_i^* + (v_i^*)^{1/2}\epsilon_i, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and $\mathbb{E}(\epsilon_i) = 0$, and $\mathrm{var}(\epsilon_i) = 1$, and where y_i is associated with node i in a connected graph $G = (V, E)$ where $V = \{1, \ldots, n\}$ and $E \subset V \times V$. This class of graph estimation problems has appeared in applications in biology (Tibshirani et al., 2005), image processing (Rudin et al., 1992; Tansey et al., 2017), traffic detection (Wang et al., 2016), among others.

A common method for estimating the signal θ^* is the fused lasso over graphs, also known as (anisotropic) total variation denoising over graphs, independently introduced by Rudin et al. (1992) and (Tibshirani et al., 2005). This consists of solving the optimization problem

$$\hat{\theta} \,:=\, \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\|y-\theta\|^2 \,+\, \lambda \|\nabla_G \theta\|_1 \right\}, \tag{2}$$

where $y = (y_1, \ldots, y_n)^\top$, $\lambda > 0$ is a tuning parameter, and $\nabla_G \in \mathbb{R}^{|E| \times n}$ is the incidence matrix of G. Specifically, each row of ∇_G corresponds to an edge $e = (e^+, e^-) \in E$ and

$$(\nabla_G)_{e,\ell} \begin{cases} 1 & \text{if } \ell = e^+, \\ -1 & \text{if } \ell = e^-, \\ 0 & \text{otherwise.} \end{cases}$$

The motivation behind (2) is to have an estimator that balances between fitting the data well, with the first term in the objective function in (2), and having a small complexity in terms of the quantity $\|\nabla_G \theta\|_1$ which is known as the total variation of the signal θ along the graph G. Intuitively, if the graph G is informative about the signals θ^* and v^*, then we would expect that $\|\nabla_G \theta^*\|_1, \|\nabla_G v^*\|_1 << n$. For instance, suppose that G is constructed as a K-NN graph based on features $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$, and assume that $\theta_i^* = f_0(x_i)$ for all $i = 1, \ldots, n$, and for a smooth function f_0. If K is small, then for $\{i, j\}$ an edge in G, we have that $|\theta_i^* - \theta_j^*| = |f_0(x_i) - f_0(x_j)|$ which would be a small quantity or zero for most edges. Then summing over all the edges, we obtain $\|\nabla_G \theta^*\|_1 << n$. In fact, Madrid Padilla et al. (2020b) showed that $\|\nabla_G \theta^*\|_1 = O_{\mathrm{pr}}(n^{1-1/d})$, ignoring logarithmic factors, provided that f_0 is a piecewise Lipschitz function.

The estimator defined in (2) has attracted a lot of attention in the literature. Specifically, computationally efficient algorithms for chain graphs were developed by Johnson (2013), for grid graphs by Barbero and Sra (2014), and for general graphs by Tansey and Scott (2015); Chambolle and Darbon (2009). Moreover, several authors have studied the statistical properties of (2) in different settings. In particular, Mammen and van de Geer (1997), and Tibshirani (2014) studied slow rates of convergence in chain graphs with signals having bounded variation. Dalalyan et al. (2017); Lin et al. (2017); Guntuboyina et al. (2020), and Ortelli and van de Geer (2021) proved fast rates for piecewise constant signals. Hütter and Rigollet (2016), Sadhanala et al. (2016), Ortelli and van de Geer (2020), and Chatterjee and Goswami (2021b) studied statistical properties of total variation denoising in grid graphs. Padilla et al. (2018), and Ortelli and van de Geer (2018) studied the fused lasso in general graphs. Wang et al. (2016), and Sadhanala et al. (2021) focused on developing higher order versions of total variation denoising.

Despite the tremendous attention from the literature focusing on the fused lasso as defined in (2), most of the statistical work assumes that the errors $\{\epsilon_i\}_{i=1}^n$ are sub-Gaussian when studying the estimator (2). While some works have considered the model in (1) with more arbitrary distributions, such as Madrid-Padilla and Chatterjee (2020) and Ye and Padilla (2021), these efforts have studied the quantile version of (1). Thus, the performance of the estimator defined in (2) is not understood beyond the sub-Gaussian errors assumption.

Additionally, the literature has been silent about estimating the variances in (1). Thus, there is currently no estimator available in the literature for estimating the variances even in the homoscedastic case, where the v_i^* are all equal to some $v_0^* > 0$, when G is a general graph. Estimation of the variance is an important problem because it would allow practitioners the possibility of quantifying the variability of the data in different regions of the graph. For instance, if y_i is the crime rate at location i, then we could have two locations where $\mathbb{E}(y_i) = \mathbb{E}(y_j)$, however, knowing that $\mathrm{var}(y_i) > \mathrm{var}(y_j)$ would be informative about the nature of crime at location i versus location j.

In this paper, we fill the gaps described above regarding mean and variance estimation in general graphs. Our main contributions are listed next.

1.1 Summary of results

We make the following contributions for the model described in (1) with a connected graph G.

1. If the variances satisfy $v_i^* = v_0^*$ for all $i = 1, \ldots, n$, then we show that, under a simple moment condition, there exists an estimator $\hat{v}$ that can be found in linear time, $O(n + |E|)$, and satisfies
$$|\hat{v} - v_0^*| \;=\; O_{\mathrm{pr}}\left(\frac{v_0^*}{n^{1/2}} + \frac{\|\nabla_G \theta^*\|_1}{n}\right). \tag{3}$$
The estimator $\hat{v}$ is based on first running depth-first search (DFS) on the graph G and then using the differences of the y_i's along the ordering. A detailed construction is given in Section 2. Notably, when G is a 1D or 2D grid graph and $\|\nabla_G \theta^*\|_1$ has a canonical scaling, the rate in (3) is minimax optimal. Moreover, our estimator is the first for the problem of estimating the variance in the sequence model where the measurements are collected in a general graph. We also show with experiments in Appendix B.1 that the estimator $\hat{v}$ can be useful for model selection when the goal is to estimate θ^*.

2. For the fused lasso estimator defined in (2), under a moment condition and an assumption stating that
$$\max_{i=1,\ldots,n} \mathrm{pr}(|\epsilon_i| > U_n) \;\to\; 0 \tag{4}$$
fast enough, where $U_n > 0$ is a sequence, we show that:

   (a) For any connected graph, ignoring logarithmic factors, it holds that
$$\frac{\|\hat{\theta} - \theta^*\|^2}{n} \;=\; O_{\mathrm{pr}}\left(\frac{U_n^{4/3}\|\nabla_G \theta^*\|_1^{2/3}}{n^{2/3}} + \frac{U_n^2}{n}\right), \tag{5}$$
   and the same upper bound holds for an estimator that can be found in linear time. Thus, we generalize the conclusions in Theorems 2 and 3 from Padilla et al. (2018) to hold with noise beyond sub-Gaussian noise. For instance, for sub-Exponential noise the term U_n would satisfy $U_n = O(\log n)$.

   (b) For the d-dimensional grid graph with $d > 1$ and n nodes, we show that
$$\frac{\|\hat{\theta} - \theta^*\|^2}{n} \;=\; O_{\mathrm{pr}}\left(\frac{U_n\|\nabla_G \theta^*\|_1}{n} + \frac{U_n^2}{n}\right), \tag{6}$$
   if we disregard logarithmic factors. Thus, under the canonical scaling $\|\nabla_G \theta^*\|_1 = O(n^{1-1/d})$, see e.g Sadhanala et al. (2016), the upper bound is minimax optimal thereby generalizing the results from Hütter and Rigollet (2016) to settings with error distributions that satisfy (4).

(c) For K-nearest neighbor (K-NN) graphs constructed with the assumptions from Madrid Padilla et al. (2020b), we show that the fused lasso estimator satisfies that

$$\frac{\|\hat{\theta}-\theta^*\|^2}{n} \;=\; O_{\mathrm{pr}}\left(\frac{U_n}{n^{1/d}}\right), \tag{7}$$

up to logarithmic factors. Hence, we generalize Theorem 2 from Madrid Padilla et al. (2020b) to models with more general error distributions. Moreover, if $U_n = O\{\mathrm{poly}(\log n)\}$ for a polynomial function $\mathrm{poly}(\cdot)$, then the rate in (7) is minimax optimal for classes of bounded variation.

3. In the heteroscedastic setting, where some of the v_i^* can be different from each other, we are the first in the literature to develop an estimator for the vector of variances $v^* \in \mathbb{R}^n$ in general graph structured models. Specifically, we provide a simple estimator $\hat{v}$ of v^* that can be found with the same computational complexity as that of $\hat{\theta}$. For the proposed estimator we show that there exists U_n' satisfying $U_n' = O(1+U_n^2)$ for which the upper bounds in (5)–(7) hold replacing $\|\hat{\theta}-\theta^*\|^2/n$ with $\|\hat{v}-v^*\|^2/n$ and $\|\nabla_G\theta^*\|_1$ with $\|\nabla_G\theta^*\|_1 + \|\nabla_G v^*\|_1$. Our results hold with the same assumptions that those in 2), but with a stronger moment condition presented in Theorem 9. Moreover, when $U_n = O\{\mathrm{poly}(\log n)\}$ and $\|\nabla_G\theta^*\|_1 \asymp \|\nabla_G v^*\|_1$, our variance estimator attains, up to log factors, the same rates as $\hat{\theta}$ attains in (5)–(7). We also show, save by logarithmic factors, that the upper bounds in the case of grid and K-NN graphs are minimax optimal, see Lemmas 13 and 14.

1.2 Other related work

Besides total variation, other popular methods for mean estimation in graph problems include kernels based methods (Smola and Kondor, 2003; Zhu et al., 2003; Zhou et al., 2005), wavelet constructions (Crovella and Kolaczyk, 2003; Coifman and Maggioni, 2006; Gavish et al., 2010; Hammond et al., 2011; Sharpnack et al., 2013; Shuman et al., 2013), tree based estimators (Donoho, 1997; Blanchard et al., 2007; Chatterjee and Goswami, 2021a; Madrid-Padilla et al., 2021b), and ℓ_0-regularization approaches (Fan and Guan, 2018; Yu et al., 2022).

As for variance estimation, some methods estimate the conditional mean and then compute the residuals before proceeding to estimate the conditional variance. Some of these approaches include Hall and Carroll (1989); Fan and Yao (1998). Other methods, as it is the case of our proposed approach, do not consider the residuals. Some of such works include Wang et al. (2008); Cai et al. (2009), which studied rates of convergence for univariate nonparametric regression with Lipschitz classes. Cai and Wang (2008) considered a wavelet thresholding approach also for univariate data. More recently, Shen et al. (2020) considered univariate Hölder functions classes and some homoscedastic multivariate settings.

Finally, total variation denoising methods have become popular as a tool to tackle different statistics and machine learning problems. Ortelli and van de Geer (2020) and Sadhanala and Tibshirani (2019) studied additive models, Padilla (2018) proposed a method for graphon estimation, Madrid-Padilla et al. (2021a) considered a method for interpretable causal inference, Dallakyan and Pourahmadi (2022) developed a method for covariance matrix estimation. More recently, Tran et al. (2022) proposed an $\ell_1 + \ell_2$ based penalty over graphs called the Generalized Elastic Net aimed for problems where features are associated with the nodes of graph.

1.3 Notation

Throughout, for a vector $v \in \mathbb{R}^n$, we define its ℓ_1, ℓ_2 and ℓ_∞ norms as $\|v\|_1 = \sum_{i=1}^n |v_i|$, $\|v\| = (\sum_{i=1}^n |v_i|)^{1/2}$, $\|v\|_\infty = \max_{i=1,\ldots,n} |v_i|$, respectively. Given a sequence of random variables X_n and a sequence of positive numbers a_n, we write $X_n = O_{\text{pr}}(a_n)$ if for every $t > 0$ there exists $C > 0$ such that $\text{pr}(X_n > Ca_n) < t$ for all n. For two sequences a_n and b_n we write $a_n \asymp b_n$ if there exists positive constants c and C such that $ca_n \leq b_n \leq Ca_n$ for all n. A d-dimensional grid graph of size $n = m^d$ is constructed as the d-dimensional lattice $\{1, \ldots, m\}^d$, where $i, j \in \{1, \ldots, m\}^d$ are connected if and only if $\|i - j\|_1 = 1$. We also write $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $\bar{a} = \frac{1}{n}\sum_{i=1}^n a_i$ for a vector $a \in \mathbb{R}^n$. For a function $f : [0,1]^d \to \mathbb{R}$, we write $\|f\|_2 := \sqrt{\int_{[0,1]^d} f(x)^2 dx}$.

1.4 Outline

The rest of the paper is organized as follows. In Section 2 we introduce the estimator for the homoscedastic case and show an upper bound on its performance. In Section 3 we start by defining our estimator for the heteroscedastic case. In Section 3.1 we provide a general upper bound for the fused lasso estimator. Then we apply our new result in Section 3.2 to obtain general upper bounds for our variances estimator in the heteroscedastic case, and conclude by providing matching lower bounds. Section 4 contains numerical evaluations of the proposed methods in both simulated and real data. All the proofs are deferred to the Appendix.

2. Homoscedastic case

This section considers the homoscedastic case, which means that $v_i^* = v_0^*$ for all i. We now give a motivation on how an estimator of the variance in the homoscedastic setting can be used for model selection of (2). Specifically, if $\hat{v}$ is an estimator of v_0^*, then following Tibshirani and Taylor (2012) and denoting $\hat{\theta}_\lambda$ the solution to (2), we can define

$$\widehat{\text{Risk}}(\lambda) := \|y - \hat{\theta}_\lambda\|^2 + 2\hat{v}\widehat{\text{df}}_\lambda,$$

where $\widehat{\text{df}}_\lambda$ is an estimator of the degrees of freedom corresponding to the model associated with $\hat{\theta}_\lambda$, see Equation (8) in Tibshirani and Taylor (2012). In fact, based on Equation 4 from Tibshirani and Taylor (2012), $\widehat{\text{df}}_\lambda$ can be taken as the number of connected components in G induced by $\hat{\theta}_\lambda$ when removing the edges $(i, j) \in E$ satisfying $(\hat{\theta}_\lambda)_i \neq (\hat{\theta}_\lambda)_j$. Hence, in practice one can choose the value of λ that minimizes $\widehat{\text{Risk}}(\lambda)$ or some variant of it, such as the one we consider in Section 4.1. Therefore, for model selection, it is convenient to estimate v_0^*.

Before providing our estimator of v_0^*, we state the statistical assumption needed to arrive at our main result of this section.

Assumption 1 *We assume that $v_i^* = v_0^*$ for $i = 1, \ldots, n$, and*

$$\max_{i=1,\ldots,n} \mathbb{E}(\epsilon_i^4) = O(1).$$

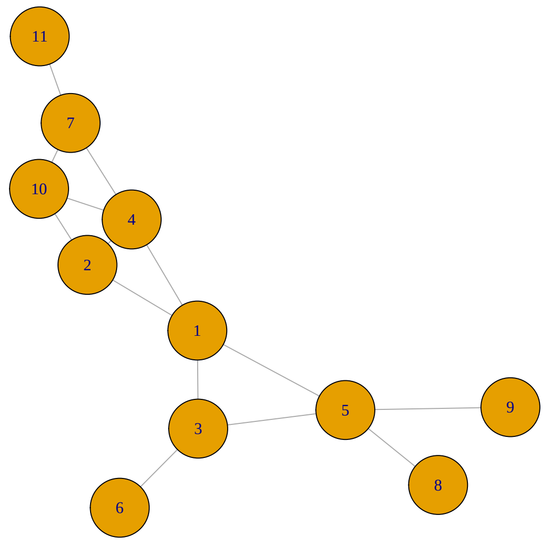

Figure 1: An example of a graph G. Running DFS starting with the node 1 produces the ordering $1, 3, 6, 5, 8, 9, 4, 7, 11, 10, 2$.

Thus, we simply require that the fourth moments of the errors are uniformly bounded. We are now in position to define our estimator. This is given as

$$\hat{v} \;:=\; \frac{1}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor - 1} \{y_{\sigma(2i)} - y_{\sigma(2i-1)}\}^2, \tag{8}$$

where $\sigma(1), \ldots, \sigma(n)$ are the nodes in G visited in order according to the DFS algorithm in the graph G, see Tarjan (1972). The DFS algorithm proceeds as follows:
Procedure DFS(G, v):
Step 1: Label v as discovered.
Step 2: For all w such that $(w, v) \in E$ do
If vertex w is not label then recursively call $DFS(G, w)$.

Figure 1 shows an example of a graph and a potential run of DFS. Clearly, by construction of DFS, the function σ is a bijection from $\{1, \ldots, n\}$ onto itself, and the DFS ordering is not unique. Hence, we propose to select the DFS by randomly choosing the start of the algorithm.

Notice that the total computational complexity for computing $\hat{v}$ is $O(n + |E|)$, which comes from computing the DFS order. Moreover, the estimator $\hat{v}$ does not require any tuning parameters to be specified.

The construction in (8) can be motivated as follows. First, recall that by Lemma 1 in Padilla et al. (2018), it holds

$$\sum_{i=1}^{n-1} |\theta^*_{\sigma(i)} - \theta^*_{\sigma(i+1)}| \;\leq\; 2\|\nabla_G \theta^*\|_1.$$

Hence, if $\|\nabla_G \theta^*\|_1$ is small relative to n, then the signal θ^* is well behaved in the order given by DFS. Our resulting estimator defined in (8) is then obtained by applying the idea of taking differences from Rice (1984), see also Dette et al. (1998) and Tong and Wang (2005).

Theorem 1 *Suppose that Assumption 1 holds and $\|\epsilon\|_\infty = O_{\mathrm{pr}}(U_n)$ for some positive sequence U_n. Then*

$$|v_0^* - \hat{v}| := O_{\mathrm{pr}}\left[\frac{v_0^*}{n^{1/2}} + \frac{\{U_n (v_0^*)^{1/2} + \|\theta^*\|_\infty\}\|\nabla_G \theta^*\|_1}{n}\right]. \tag{9}$$

Remark 2 *Consider the case where G is the chain graph, and suppose that $\theta_i^* = f^*(i/n)$ for $i = 1, \ldots, n$, for a function $f^* : [0,1] \to \mathbb{R}$, bounded and of bounded total variation. Thus, $f^* \in \mathcal{C}$ where*

$$\mathcal{C} := \{f : [0,1] \to \mathbb{R} : \|f\|_\infty \leq C_1,\ \mathrm{TV}(f) \leq C_2\},$$

where C_1 and C_2 are positive constants, and $\mathrm{TV}(f)$ is the total variation defined as

$$\mathrm{TV}(f) := \sup_{0 \leq a_1 < \ldots < a_m \leq 1,\ m \in \mathbb{N}} \sum_{j=1}^{m-1} |f(a_j) - f(a_{j+1})|,$$

see the discussion about functions of bounded total variation in Tibshirani (2014). Then $\max\{\|\theta^*\|_\infty, \|\nabla_G \theta^*\|_1\} = O(1)$. Hence, provided that $v_0^* = O(1)$ and $U_n = O\{\mathrm{poly}(\log n)\}$ for $\mathrm{poly}(\cdot)$ some polynomial function, we obtain that

$$|v_0^* - \hat{v}| := O_{\mathrm{pr}}(n^{-1/2}), \tag{10}$$

if we ignore logarithmic factors. Therefore, from Proposition 3 in Shen et al. (2020), the rate in (10) is minimax optimal in the class $\mathcal{C}$. This follows since $\mathcal{C}$ is a larger class than that considered in Proposition 3 in Shen et al. (2020) for the case corresponding to bounded Lipschitz continuous functions.

Remark 3 *If G is the 2D grid graph, then it is well known that $\|\nabla_G \theta^*\|_1 \asymp n^{1/2}$ is the canonical scaling, see Sadhanala et al. (2016) and our discussion in Appendix A. Hence, if $\max\{v_0^*, \|\theta^*\|_\infty\} = O(1)$, and $U_n = O\{\mathrm{poly}(\log n)\}$, then (10) holds. Therefore, as in Remark 2, by Proposition 3 from Shen et al. (2020), $\hat{v}$ attains minimax rates when θ^* is in the class*

$$\{\theta : \|\theta\|_\infty \leq C_1,\ \|\nabla_G \theta\|_1 \leq C_1 n^{1/2}\}$$

for positive constants C_1 and C_2.

Finally, for a general graph G, if the graph does capture smoothness of the true signal in the sense that $U_n \|\nabla_G \theta^*\|_1 / n \to 0$, then, as long as $\max\{v_0^*, \|\theta^*\|_\infty\} = O(1)$, the upper bound in Theorem 1 shows that $\hat{v}$ is a consistent estimator of v_0^*.

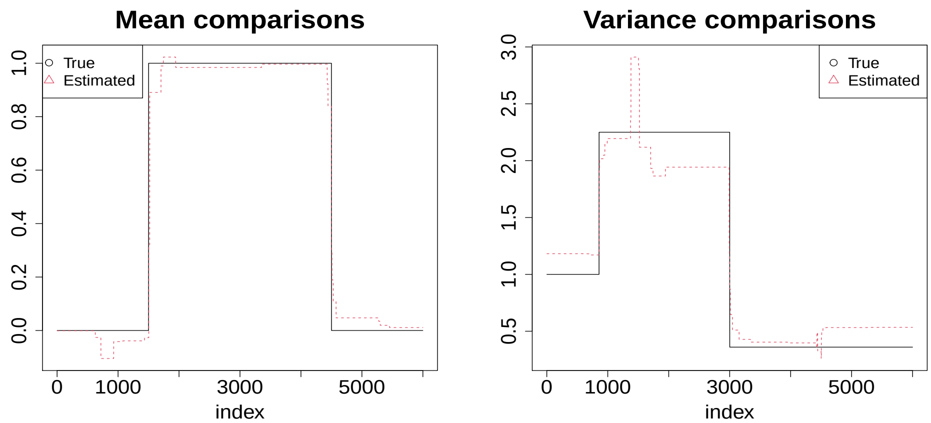

Figure 2: The left panel shows comparisons of the true and estimated means for Example 1 in the text. The right panel shows the corresponding variance comparisons.

3. Heteroscedastic case

We now study the heteroscedastic setting. Hence, we do not longer require that all the variances are equal. To estimate the signal $v^* \in \mathbb{R}^n$, we recall the identity

$$v_i^* \;=\; \mathrm{var}(y_i) \;=\; \mathbb{E}(y_i^2) - \{\mathbb{E}(y_i)\}^2.$$

Therefore, it is natural to estimate v_i^* with

$$\hat{v}_i \;=\; \hat{\gamma}_i - (\hat{\theta}_i)^2, \tag{11}$$

where $\hat{\gamma}_i$ is an estimator of $\gamma_i^* := \mathbb{E}(y_i^2)$, and $\hat{\theta}$ is the fused lasso estimator defined in (2). As an estimator for γ^*, we propose

$$\hat{\gamma} \;:=\; \underset{\gamma \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i^2 - \gamma_i)^2 + \lambda' \sum_{(i,j)\in E} |\gamma_i - \gamma_j| \right\} \tag{12}$$

for a tuning parameter $\lambda' > 0$.

Notice that $\hat{v}$ can be found with the same order of computational cost that it is required for finding $\hat{\theta}$. In practice, this can be done using the algorithm from Chambolle and Darbon (2009). As for parameter tuning, we give details about choosing λ' in practice in Section 4.1.

To illustrate the behavior of the estimator defined in (11)–(12), we now consider a simple numerical example. More comprehensive evaluations are given in Section 4.

Example 1 *We set $n = 6000$ and generate data according to the model given by (1) with $\epsilon_i \overset{ind}{\sim} N(0,1)$ for $i = 1, \ldots, n$, and $\theta^*, v^* \in \mathbb{R}^n$ satisfying*

$$\theta_i^* \;=\; \begin{cases} 1 & \text{if } n/4 < i \leq 3n/4, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$v_i^* = \begin{cases} 1 & \text{if } i \leq \lfloor n/7 \rfloor, \\ 1.5^2 & \text{if } \lfloor n/7 \rfloor < i \leq n/2, \\ 0.6^2 & \text{otherwise.} \end{cases}$$

Given the data $\{y_i\}_{i=1}^n$, we run the estimator defined in (11)–(12) with tuning parameter choices as discussed in Section 4.1. The results are displayed in Figure 2, where we see that the estimated means and variances are reasonably close to the corresponding true parameters.

3.1 A general result for fused lasso estimator

Before presenting our main result for the estimator $\hat{v}$ defined in (11)–(12), we provide a general upper bound for the fused lasso estimator that holds under very weak assumptions and generalizes existing work in Hütter and Rigollet (2016), Padilla et al. (2018) and Madrid Padilla et al. (2020b).

Theorem 4 *Consider data $\{o_i\}_{i=1}^n$ generated as $o_i = \beta_i^* + \varepsilon_i$ for some $\beta^* \in \mathbb{R}^n$ and $\varepsilon_1, \ldots, \varepsilon_n$ independent random variables satisfying satisfying $\mathbb{E}(\varepsilon_i) = 0$ for $i = 1, \ldots, n$, and $\max_{i=1,\ldots,n} \mathbb{E}(\varepsilon_i^4) = O(1)$. Let $\hat{\beta}$ be defined as*

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^n (o_i - \beta_i)^2 + \lambda \sum_{(i,j) \in E} |\beta_i - \beta_j| \right\}. \tag{13}$$

The following results hold:

1. ***General graphs.** For any connected graph G, if for a positive sequence U_n holds that*

$$n^{1/2} U_n^{-1} \{\log(en)\}^{-1/2} \max_{i=1,\ldots,n} \{\mathrm{pr}(|\varepsilon_i| > U_n)\}^{1/4} \to 0, \tag{14}$$

   then

$$\frac{\|\hat{\beta} - \beta^*\|^2}{n} = O_{\mathrm{pr}} \left\{ \frac{U_n^{4/3} (\log n)^{1/3} \|\nabla_G \beta^*\|_1^{2/3}}{n^{2/3}} + \frac{U_n^2 \log n}{n} \right\}, \tag{15}$$

   for a choice of λ satisfying $\lambda \asymp U_n^{4/3} (n \log n)^{1/3} \|\nabla_G \beta^*\|_1^{-1/3}$.

2. ***Grid graphs.** Let G be the d-dimensional grid graph with $d > 1$. Suppose that for a positive sequence U_n we have that*

$$\max_{i=1,\ldots,n} n^{1/2} U_n^{-1} \{\mathrm{pr}(|\varepsilon_i| > U_n)\}^{1/4} \to 0. \tag{16}$$

   Then there exists a choice of λ satisfying

$$\lambda \asymp U_n \phi_n + U_n^2 \|\nabla_G \beta^*\|_1^{-1}$$

   such that

$$\frac{\|\hat{\beta} - \beta^*\|^2}{n} = O_{\mathrm{pr}} \left(\frac{U_n \phi_n \|\nabla_G \beta^*\|_1}{n} + \frac{U_n^2}{n} \right), \tag{17}$$

where $\phi_n = C\log n$ *if* $d=2$ *and* $\phi_n = C(\log n)^{1/2}$ *otherwise, for some constant* $C>0$. *for some constant* $C>0$.

3. ***K-NN graphs.*** *Suppose that in addition to the measurements* $(o_1,\ldots,o_n)^\top$ *we are also given covariates* $\{x_i\}_{i=1}^n \subset \mathcal{X}$, *where* x_i *corresponds to* o_i, *and* $\mathcal{X}$ *is a metric space with metric* $\mathrm{dist}(\cdot)$. *Suppose that* $\{(x_i,o_i)\}_{i=1}^n$ *satisfy the assumptions from Madrid Padilla et al. (2020b), see Appendix E. In particular,* $\mathcal{X}$ *is homeomorphic to* $[0,1]^d$. *In addition, assume that* $K \asymp \log^{1+2r} n$ *for some* $r>0$ *in the construction of the K-NN graph G, and that for a positive sequence* U_n *we have that*

$$n^{1/2}U_n^{-1}K^{-1/2}\max_{i=1,\ldots,n}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4} \to 0. \tag{18}$$

Consider

$$\lambda \asymp \|\nabla_G\beta^*\|_1^{-1}\left[(\mathrm{poly}(\log n)n^{1-1/d}U_n)^{1/2} + K^{1/2}U_n + (U_nK^{1/2}\mathrm{poly}(\log n)\,n^{1-1/d}\phi_n)^{1/2}\right]^2 \tag{19}$$

where $\mathrm{poly}(\cdot)$ *is a polynomial function, and* ϕ_n *is defined as in the case of grid graphs above. Then*

$$\frac{\|\hat\beta-\beta^*\|^2}{n} = O_{\mathrm{pr}}\left\{\frac{U_n\mathrm{poly}_2(\log n)}{n^{1/d}}\right\}, \tag{20}$$

where $\mathrm{poly}_2(\cdot)$ *is another polynomial function.*

Remark 5 *Let us now elaborate on (14), (16) and (18). Suppose, for instance, that* ε_i *is sub-Exponential(a), for some constant* $a>0$. *Then the usual sub-Exponential tail inequality can be written as*

$$\mathrm{pr}(|\varepsilon_i|>t) \le 2\exp(-t/a), \quad \textit{for all } t>0,$$

see Proposition 2.7.1 in Vershynin (2018). Hence, taking $U_n = 4a\log n$ *it follows that (14), (16) and (18) immediately hold. More generally, if*

$$\mathrm{pr}(|\varepsilon_i|>t) \le c_1\exp(-t^\alpha/c_2), \quad \textit{for all } t>0,$$

for positive constants c_1, c_2, *and* α, *then taking* $U_n = 4c_2(\log n)^{1/\alpha}$, *we obtain that (14), (16) and (18) all hold.*

Remark 6 *Remark 5 gives a family of examples where* U_n *can be taken as a power function of* $\log n$. *More generally, if* $U_n = O\{\mathrm{poly}(\log n)\}$, *for a polynomial function* $\mathrm{poly}(\cdot)$, *then up to logarithmic factors, Theorem 4 gives the same rates as in several existing works on the fused lasso, but now we allow for more general error distributions than sub-Gaussian. Specifically:*

1. *For a connected graph G, (15) generalizes the upper bound in Theorem 4 in Padilla et al. (2018). Moreover, the same upper bound in (15) also holds if we replace the fused lasso estimator* $\hat\beta$ *(13) with the DFS fused lasso estimator from Padilla et al. (2018).*

2. *For a d-dimensional grid graph G, the rate in (17) matches that in Corollary 5 from Hütter and Rigollet (2016).*

3. *For nonparameteric regression, (20) gives the same minimax rate as Theorem 2 in Madrid Padilla et al. (2020b) for classes of piecewise functions.*

Remark 7 *As stated before, Theorem 4 is the first result for fused lasso in general graph models where the error terms can be non-Sub-Gaussian, yet the estimator still uses the ℓ_2 loss. The proof of Theorem 4 relies on Theorem 15 in Appendix D. The latter basically allow us to control the quantity* $\mathrm{pr}(\|\hat{\beta}-\beta^*\| > \eta)$*, for* $\eta > 0$*, in terms of the process*

$$\frac{1}{\eta^2}\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_G\beta\|_1\lesssim\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\right] \tag{21}$$

where $\xi_1,\ldots,\xi_n$ *are independent Rademacher random variables independent of* $\{\varepsilon_i\}_{i=1}^n$*. This general result holds for arbitrary sequences* U_n *and it is key given that the random variables* $\xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}$ *for* $i=1,\ldots,n$ *are uniformly bounded. Hence, we do not need to control the standard process*

$$\mathbb{E}\left\{\sup_{\beta\in\Lambda\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_G\beta\|_1\lesssim\|\nabla_G\beta^*\|_1}\varepsilon^\top(\beta-\beta^*)\right\}$$

as it is the case in the analysis in Guntuboyina et al. (2020), which is only able to handle sub-Gaussian random variables ε_i*,* $i=1,\ldots,n$*. With this challenge overcome, the proof of Theorem 15 continues by controlling additional terms, that account for the case* $\xi_i\varepsilon_i 1_{\{|\varepsilon_i|>U_n\}}$ *for* $i=1,\ldots,n$*, of the form*

$$\frac{n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{\eta}. \tag{22}$$

With Theorem 15 in hand, the proof of Theorem 4 continues by deriving upper bounds for the quantities (21) and (22). The analysis for (21) is done customizing for general graphs, grid graphs, and K-NN graphs.

Remark 8 *We now proceed to discuss the choice of the tuning parameters in Theorem 4. As pointed out by one of the reviewers, the corresponding choices of tuning parameters depend on the unknow quantity* $\|\nabla_G\beta^*\|_1$*. Here, we expand on this aspect and point out some relaxations.*

- *First, in order to obtain (15), which holds for general graphs, our choice of* λ *is basically the same as in Theorem 4 from Padilla et al. (2018).*
- *In the context of the d-dimensional grid graph, as explained in Section A, the canonical scaling for* $\|\nabla_G\beta^*\|_1$ *is* $n^{1-1/d}$*. Hence, supposing that* $\|\nabla_G\beta^*\|_1\geq 1$*, our proof can be modified, see Page 34, so that if*

$$\lambda \asymp U_n\phi_n + U_n^2,$$

we have that

$$\frac{\|\hat{\beta}-\beta^*\|^2}{n} = O_{\mathrm{pr}}\left(\frac{\max\{U_n\phi_n, U_n^2\}\|\nabla_G\beta^*\|_1}{n}\right). \tag{23}$$

Thus, ignoring factors that depend on U_n and ϕ_n, we can get the same rate as in Theorem 4 for a choice of tuning parameter that does not involve $\|\nabla_G \beta^*\|_1$. This is in the spirit of the findings in Hütter and Rigollet (2016), where the choice of tuning parameter does not depend on $\|\nabla_G \beta^*\|_1$.

- *Finally, for the setting of K-NN graphs, in the choice of λ that depends on $\|\nabla_G \beta^*\|_1$ in (19), we can replace $\|\nabla_G \beta^*\|_1$ with $n^{1-1/d}$ to attain the rate in (20). This can be done provided that $c_1 n^{1-1/d} \leq \|\nabla_G \beta^*\|_1 \leq c_2 n^{1-1/d}$ for constants $c_1, c_2 > 0$, with probability approaching one.*

3.2 Fused lasso for variance estimation

We are now ready to state our main result regarding the estimator $\hat{v}$ defined in (11)–(12). Notably, our result shows that the estimator $\hat{v}$ enjoys similar properties as the original fused lasso in general graphs, d-dimensional grids, and K-NN graphs. The conclusion of our result follows from an application of Theorem 4 to $\hat{\theta}$ defined in (2) and $\hat{\gamma}$ defined in (12).

Theorem 9 *Consider data $\{y_i\}_{i=1}^n$ generated as in (1) and suppose that $\max_{i=1,\ldots,n} \mathbb{E}(\epsilon_i^8) < \infty$. Then the estimator $\hat{v}$ satisfies the following.*

- ***General graphs.*** *Let G be any connected graph and assume that (14) holds with $\{\epsilon_i\}_{i=1}^n$ instead of $\{\varepsilon_i\}_{i=1}^n$. Then for choices of λ and λ' satisfying*

$$\lambda \asymp U_n^{4/3} (n \log n)^{1/3} \|\nabla_G \theta^*\|_1^{-1/3}$$

and $\lambda' \asymp \{\|v^*\|_1^{1/2} \|\theta^*\|_\infty U_n + \|v^*\|_\infty (1+U_n^2)\}^{4/3} (n \log n)^{1/3} \|\nabla_G \gamma^*\|_1^{-1/3}$, we have that

$$\frac{1}{n}\|\hat{v} - v^*\|^2 \quad = \quad O_{\mathrm{pr}}\left\{ \frac{(\|\theta^*\|_\infty^2 + 1)(U_n')^{4/3} (\log n)^{1/3} (\|\nabla_G v^*\|_1 + \|\theta^*\|_\infty \|\nabla_G \theta^*\|_1)^{2/3}}{n^{2/3}} + \frac{(\|\theta^*\|_\infty^2 + 1)(U_n')^2 \log n}{n} \right\} \tag{24}$$

where

$$U_n' \;:=\; (2\|v^*\|_\infty^{1/2} \|\theta^*\|_\infty + 1) U_n + \|v^*\|_\infty U_n^2 + \|v^*\|_\infty. \tag{25}$$

- ***Grid graphs.*** *Let G be the d-dimensional grid graph with $d > 1$. Suppose that the sequence $\{\epsilon_i\}_{i=1}^n$ satisfies (16). Then there exists tuning parameter choices satisfying*

$$\lambda \asymp U_n \phi_n + U_n^2 \|\nabla_G \theta^*\|_1^{-1}, \quad \text{and} \quad \lambda' \asymp U_n' \phi_n + (U_n')^2 \|\nabla_G \gamma^*\|_1^{-1}$$

for which

$$\frac{\|\hat{v} - v^*\|^2}{n} = O_{\mathrm{pr}}\left\{ \frac{(\|\theta^*\|_\infty^2 + 1) U_n' \phi_n \left(\|\nabla_G v^*\|_1 + \|\theta^*\|_\infty \|\nabla_G \theta^*\|_1\right)}{n} + \frac{(\|\theta^*\|_\infty^2 + 1)(U_n')^2}{n} \right\}, \tag{26}$$

with U_n' as in (25) and ϕ_n as in Theorem 4.

- ***K-NN graphs.*** *Suppose that in addition to the measurements* $(y_1, \ldots, y_n)^\top$ *we are also given covariates* $\{x_i\}_{i=1}^n \subset \mathcal{X}$*, where* x_i *corresponds to* y_i*, and* $\mathcal{X}$ *is a metric space with metric* $\mathrm{dist}(\cdot)$*. Suppose that* $\{(x_i, y_i)\}_{i=1}^n$ *satisfy the assumptions from Madrid Padilla et al. (2020b) stated in Appendix E. In addition, assume that* $K \asymp \log^{1+2r} n$ *for some* $r > 0$ *in the construction of the* K*-NN graph* G*, and (18) holds for* $\{\epsilon_i\}_{i=1}^n$*. Then for choices of* λ *and* λ' *satisfying*

$$\lambda \asymp \|\nabla_G \theta^*\|_1^{-1} \left[(\mathrm{poly}(\log n) n^{1-1/d} U_n)^{1/2} \, + \, K^{1/2} U_n \, + \, (U_n K^{1/2} \mathrm{poly}(\log n)\, n^{1-1/d} \phi_n)^{1/2} \right]^2$$

and

$$\lambda' \asymp \|\nabla_G \gamma^*\|_1^{-1} \left[(\mathrm{poly}(\log n) n^{1-1/d} U_n')^{1/2} \, + \, K^{1/2} U_n' \, + \, (U_n' K^{1/2} \mathrm{poly}(\log n)\, n^{1-1/d} \phi_n)^{1/2} \right]^2$$

for a polynomial function $\mathrm{poly}(\cdot)$*, it holds that*

$$\frac{\|\hat{v} - v^*\|^2}{n} = O_{\mathrm{pr}} \left\{ \frac{(\|\theta^*\|_\infty^2 + 1) U_n' \phi_n \mathrm{poly}(\log n)}{n^{1/d}} \right\}, \tag{27}$$

with U_n' *as in (25),* ϕ_n *as in the previous case of grid graphs, and* $\mathrm{poly}(\cdot)$ *is a polynomial function.*

Remark 10 *Consider the setting in which* $\max\{\|\theta^*\|_\infty, \|v^*\|_\infty\} = O(1)$*, and* $U_n = O\{\mathrm{poly}(\log n)\}$*, for* $\mathrm{poly}(\cdot)$ *a polynomial function. Then, ignoring logarithmic factors, Theorem 9 implies the following:*

1. *For a connected graph* G *the estimator* $\hat{v}$ *satisfies*

$$\frac{\|\hat{v} - v^*\|^2}{n} = O_{\mathrm{pr}} \left\{ \frac{(\|\nabla_G \theta^*\|_1 + \|\nabla_G v^*\|_1)^{2/3}}{n^{2/3}} \right\}.$$

Hence, for the chain graph and the canonical setting in which $\max\{\|\nabla_G \theta^*\|_1, \|\nabla_G v^*\|_1\} = O(1)$*, the estimator* $\hat{v}$ *attains the rate* $n^{-2/3}$*, which is minimax optimal in the class*

$$\{(v, \theta) \, : \, \max\{\|\nabla_G \theta^*\|_1, \|\nabla_G v^*\|_1\} \, \leq \, C_1, \, \max\{\|\theta^*\|_\infty, \|v^*\|_\infty\} \, \leq \, C_1\}$$

for some constants $C_1, C_2 > 0$*, see Theorem 4 in Shen et al. (2020).*

2. *If* $d > 1$*, then for the* d*-dimensional grid graph, we obtain that*

$$\frac{\|\hat{v} - v^*\|^2}{n} = O_{\mathrm{pr}} \left(n^{-1/d} \right), \tag{28}$$

under the canonical scaling, (Sadhanala et al. (2016), see also Appendix A) $\|\nabla_G \theta^*\|_1, \|\nabla_G v^*\|_1 \asymp n^{1-1/d}$*. Hence, from Lemma 13 below, for estimating* v^**,* $\hat{v}$ *attains the minimax rate under the canonical scaling.*

3. *For the* K*-NN graph,* $\hat{v}$ *also attains the rate* $n^{-1/d}$ *for estimating piecewise Lipschitz functions, thereby maintaining the same adaptivity properties of* $\hat{\theta}$ *studied in Madrid Padilla et al. (2020b). Moreover, from Lemma 14, the rate* $n^{-1/d}$ *matches the minimax rate for estimating the signal of variances when this is constructed based on the evaluations of a piecewise Lipschitz function.*

Remark 11 *Given that our proposed estimator defined in (11) is based on estimating* $\gamma_i^* = \mathbb{E}(y_i^*)$ *and* $\theta_i^* = \mathbb{E}(y_i)$ *for* $i = 1, \ldots, n$*, the first step in the proof of Theorem 9 is to establish Lemma 16 which states that the total variation of* γ^* *is bounded by the total variation of the variance signal* v^* *and the total variation of the mean signal* θ^**:*

$$\|\nabla_G \gamma^*\|_1 \lesssim \|\nabla_G v^*\|_1 + \|\nabla_G \theta^*\|_1.$$

Thus, if v^* *and* θ^* *both have small total variation along the graph* G*, then the same can be said about the signal* γ^**, which justifies our construction in (12). Then the proof of Theorem 9 continues by showing that*

$$\frac{1}{n}\|\hat{v} - v^*\|^2 \lesssim \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\gamma}_i - \gamma_i^*\right)^2 + \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i^*\right)^2,$$

and then applying Theorem 4 separately with the choices $\beta^* = \theta^*$ *and* $\beta^* = \gamma^*$*. The latter has an additional small challenge, addressed in Lemma 17, concerning the behavior of the tails of the random variables* y_i^2*.*

Remark 12 *Just as in Remark 8, for the* d*-dimensional grid graph, we can attain the rate in (28) for choices of tuning parameters that do not depend on* $\|\nabla_G \theta^*\|_1$ *and* $\|\nabla_G \gamma^*\|_1$*. Specifically, for choices satisfying*

$$\lambda \asymp U_n \phi_n + U_n^2, \;\; \textit{and} \;\; \lambda' \asymp U_n' \phi_n + (U_n')^2.$$

Next, we justify the second conclusion in Remark 10 concerning the minimax optimality of $\hat{v}$ under canonical scaling. This is presented in the next lemma.

Lemma 13 *Let* G *be the* d*-dimensional grid graph and* $c \in (0, 1)$ *a constant and let*

$$\Theta = \{\theta \in \mathbb{R}^n : \|\nabla_G \theta\|_1 \leq c n^{1-1/d}, \; \|\theta\|_\infty \leq c\}.$$

Consider the collection of estimators given as

$$\mathcal{F} \;\; := \;\; \left\{ v : \mathbb{R}^n \to \mathbb{R}^n \; \textit{measurable} \right\}.$$

Then there exists a constant $C > 0$ *depending on* c *and* d *such that*

$$\inf_{\tilde{v} \in \mathcal{F}} \; \sup_{\theta^*, v^* \in \Theta, \, v_i^* \in (\frac{c^2}{8}, \frac{3c^2}{8})} \mathbb{E}\left(\frac{1}{n}\|\tilde{v}(y) - v^*\|^2\right) \geq \frac{C}{n^{1/d}},$$

for data generated as $y_i = \theta_i^* + \sqrt{v_i^*}\epsilon_i$*, with* $\epsilon_i \overset{ind}{\sim} N(0, 1)$*, for* $i = 1, \ldots, n$*.*

Finally, we conclude our theory section with a lower bound that justifies our assertion that $\hat{v}$ is minimax optimal when using a K-NN graph for estimating a piecewise Lipschitz signal.

Lemma 14 *Consider the class of piecewise Lipschitz functions $\mathcal{F}(L_0)$, defined in Appendix E, for a constant $L_0 \in (0,1)$. Suppose that, for functions $f_0, g_0 \in \mathcal{F}(L_0)$ with $g_0 \geq 0$, the data are generated as*

$$y_i \,=\, f_0(x_i) \,+\, \sqrt{g_0(x_i)}\epsilon_i,$$

where $\epsilon_i \overset{ind}{\sim} N(0,1)$ and $x_i \overset{ind}{\sim} U[0,1]^d$, for $i = 1,\ldots,n$. Then for a constant $C > 0$ depending on L_0, we have that

$$\inf_{\tilde{g}\ estimator} \ \sup_{f_0, g_0 \in \mathcal{F}(L_0)} \mathbb{E}\left(\|\tilde{g} - g_0\|_2^2\right) \,\geq\, \frac{C}{n^{1/d}}.$$

4. Experiments

4.1 Heteroscedastic estimator: Tuning parameters

We now discuss how to choose the tuning parameters for the estimator $\hat{v}$ defined in (11)–(12). Let $\hat{\theta}(\lambda)$ and $\hat{v}(\lambda')$ the estimates based on choices λ and λ'. Notice that $\hat{v}(\lambda')$ depends on λ but we do not make this dependence explicit to avoid overloading the notation.

To choose λ, inspired by Tibshirani and Taylor (2012), we use a Bayesian information criterion given as

$$\widehat{\mathrm{BIC}}(\lambda) \,:=\, \|y - \hat{\theta}(\lambda)\|^2 + \widehat{\mathrm{df}}(\lambda)\log n \tag{29}$$

where $\widehat{\mathrm{df}}(\lambda)$ is the number of connected components induced by $\hat{\theta}(\lambda)$ in the graph G. Then we select the value of λ that minimizes $\widehat{\mathrm{BIC}}(\lambda)$.

Once $\hat{\theta}(\lambda)$ has been computed, we proceed to select λ' for (12). We let $\hat{\gamma}(\lambda')$ be the solution to (12) and $\widetilde{\mathrm{df}}(\lambda')$ be the number of connected components in G induced by $\hat{\gamma}(\lambda')$. Then we define

$$\widetilde{\mathrm{BIC}}(\lambda') \,:=\, \sum_{i=1}^{n}[\min\{q, y_i^2\} - \hat{\gamma}(\lambda')_i]^2 + \widetilde{\mathrm{df}}(\lambda')\log n \tag{30}$$

where q is the 0.95-quantile of the data $\{y_i^2\}_{i=1}^n$. We use $\min\{q, y_i^2\}$ in (30) to avoid the influence of outliers in the model selection step. With the above score in hand, we choose the value of λ' that minimizes $\widetilde{\mathrm{BIC}}(\lambda')$. In all our experiments, we select λ and λ' from the set $\{10^1, 10^2, 10^3, 10^4, 10^5\}$.

4.2 Homoscedastic case simulations

We start by considering settings where the variance, denoted as v_0^*, is constant across the different nodes i. As benchmarks, we consider the estimator defined in (8) which we refer as homoscedastic estimator (Hom.), the heteroscedastic estimator (Het.) defined (11)–(12), and the U-statistic based local polynomial estimator defined in Shen et al. (2020) (U-LP).

For our comparisons, we generate data from the model in (1) with $\epsilon_i \overset{\text{ind}}{\sim} N(0,1)$ and $v_i^* = v_0^*$ for $i = 1,\ldots,n$. We consider 2-dimensional grid graphs G with $n \in \{100^2, 200^2, 300^2, 400^2\}$, and we identify the nodes of G with elements of the set $\{1,\ldots n^{1/2}\} \times \{1,\ldots,n^{1/2}\}$. Then we consider values of v_0^* in $\{0.5, 1, 1.5, 2\}$ and three different scenarios for the signal θ^*. Next, we describe the choices of θ^* that we consider.

		Scenario 1			Scenario 2			Scenario 3		
n	v_0	U-LP	Hom.	Het.	U-LP	Hom.	Het.	U-LP	Hom.	Het.
100^2	0.25	0.33	**0.26**	1.15	0.17	**0.16**	0.17	0.43	**0.28**	2.25
200^2	0.25	0.14	**0.12**	1.121	0.21	**0.10**	0.11	0.39	**0.11**	0.94
300^2	0.25	0.15	**0.08**	0.97	0.19	**0.09**	**0.09**	0.44	**0.08**	0.47
400^2	0.25	0.14	**0.07**	0.70	0.22	0.09	**0.08**	0.49	**0.06**	0.28
100^2	0.5	1.12	**1.10**	1.11	1.13	**1.11**	1.24	5.22	**1.12**	2.65
200^2	0.5	0.52	**0.44**	1.23	1.21	**0.62**	1.34	4.99	**0.44**	0.94
300^2	0.5	0.62	**0.32**	0.97	1.19	**0.34**	1.15	4.84	**0.36**	0.48
400^2	0.5	0.61	**0.20**	0.70	1.38	**0.25**	0.91	4.96	**0.23**	0.29
100^2	0.75	2.79	2.72	**1.24**	1.36	2.40	**1.23**	5.06	3.06	**2.52**
200^2	0.75	1.20	1.19	**1.18**	1.31	**1.27**	1.40	4.89	1.49	**0.91**
300^2	0.75	0.69	**0.68**	1.01	1.43	**0.78**	1.24	4.76	0.78	**0.48**
400^2	0.75	0.68	**0.55**	0.69	1.33	**0.58**	0.92	5.30	0.59	**0.30**
100^2	1.0	3.42	3.94	**1.18**	2.73	2.63	**1.28**	5.46	3.76	**2.60**
200^2	1.0	2.31	2.22	**1.28**	1.58	2.23	**1.38**	4.81	2.03	**0.96**
300^2	1.0	0.76	**0.65**	0.94	1.35	**0.94**	1.22	.4.94	1.06	**0.51**
400^2	1.0	0.57	0.47	**0.74**	1.29	**0.89**	0.95	4.81	1.03	**0.29**

Table 1: Performance evaluations of the competing methods for the different settings described in the text. We report 100 multiplied by the average mean squared error, averaging over 200 Monte Carlo simulations.

Scenario 1. For $k, l \in \{1, \ldots, n^{1/2}\}$, we let

$$\theta^*_{k,l} = \begin{cases} 1 & \text{if } \ |k - n/2| < n/4, \text{ and } |l - n/2| < n/8, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 2. We set

$$\theta^*_{k,l} = \begin{cases} 1 & \text{if } \ (k - n/4)^2 + (l - n/4)^2 < (n/5)^2, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 3. In this scenario we set

$$\theta^*_{k,l} = \begin{cases} 1 & \text{if } \ k < n/2 \text{ and } l < n/2, \\ 0 & \text{otherwise.} \end{cases}$$

For each scenario and value of the model parameters, we generate 200 data sets and, for each data set, compute the different estimators. We compute the Hom. estimator with a random DFS, and the Het. estimator with tuning parameters chosen as in Section 4.1. As for the U-LP estimator, we follow the construction in Section 4.1 from Shen et al. (2020). First, we identify the nodes of the 2-dimensional grid graph with elements of the interval $[0,1]^2$, such that (i,j) in the grid graph corresponds to $X_{(j-1)n^{1/2}+i} := (i/n^{1/2}, j/n^{1/2}) \in [0,1]^2$ for $(i,j) \in \{1\ldots, n^{1/2}\} \times \{1, \ldots, n^{1/2}\}$. We also let $Y_{(j-1)n^{1/2}+i} = y_{(i,j)}$ where $y_{(i,j)}$ is the observation associated with (i,j) in the 2-dimensional grid graph. Then we recall that the estimator in Shen et al. (2020) in this context becomes

$$\hat{v}_{\text{original}} = \frac{\binom{n}{2}^{-1} \sum_{k<\ell} \mathcal{K}_{h_1}(X_{k,1} - X_{\ell,1}) \cdot \mathcal{K}_{h_2}(X_{k,2} - X_{\ell,2})(Y_k - Y_\ell)^2}{\binom{n}{2}^{-1} \sum_{k<\ell} \mathcal{K}_{h_1}(X_{k,1} - X_{\ell,1}) \cdot \mathcal{K}_{h_2}(X_{k,2} - X_{\ell,2})}, \tag{31}$$

where $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is a kernel, $h_1, h_2 > 0$ are bandwidths, and $\mathcal{K}_h(\cdot) := \mathcal{K}(\cdot/h)/h$. Notice that computing $\hat{v}_{\text{original}}$ involves $O(n^2)$ operations which quickly becomes intractable. Hence, we approximate (31) with

$$\hat{v}_{\text{approx}} = \frac{N^{-1}\sum_{s=1}^{N} \mathcal{K}_{h_1}(X_{k_s,1} - X_{\ell_s,1}) \cdot \mathcal{K}_{h_2}(X_{k_s,2} - X_{\ell_s,2})(Y_{k_s} - Y_{\ell_s})^2}{N^{-1}\sum_{s=1}^{N} \mathcal{K}_{h_1}(X_{k_s,1} - X_{\ell_s,1}) \cdot \mathcal{K}_{h_2}(X_{k_s,2} - X_{\ell_s,2})}, \tag{32}$$

where $(k_1, \ell_1), \ldots, (k_N, \ell_N)$ are independent draws from the uniform distribution in $\{1, \ldots, n\} \times \{1, \ldots, n\}$. The resulting estimator is the one that we consider as competitor in representation of the method from Shen et al. (2020). In our simulations, we set $N = 5000$, $\mathcal{K}$ is the Gaussian kernel, and $h_1 = h_2 = h$. We allow $h \in \{2^{-10}, 2^{-9}, \ldots, 2^{-1}\}$ and report results for the choice of h that gives the best performance in terms of estimating the true parameter v_0^*.

We use the mean squared error as a measure of performance for the different estimators. For the methods Hom. and U-LP which only compute a single estimator, denoting the output of the method as $\hat{v} \in \mathbb{R}$, we compute the average of $|\hat{v} - v_0^*|^2$ across the 200 replicates. For the method Het. that produces a vector $\hat{v} \in \mathbb{R}^n$, we compute the average of

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_i - v_0^*)^2 \tag{33}$$

over the 200 Monte Carlo simulations. The results can be seen in Table 1, where observe that our proposed estimators Hom. and Het. outperform the competitor in all of the instances considered. This does not come as a surprise since the true mean in each scenario is piecewise constant, making it challenging for the kernel based method from Shen et al. (2020), while both of our proposed methods are better suited for handling piecewise constant signals for both the mean and variance vectors.

4.3 Heteroscedastic case: 2D Grid graphs

In our next experiment, we consider generative models where the true graph is a 2D grid graph of size $n^{1/2} \times n^{1/2}$. We generate data similarly to Section 4.2 with the difference that the variance is now non-constant. Specifically, the data are generated as

$$y_i = \theta_i^* + \sqrt{v_i^*}\epsilon_i$$

with $\epsilon_i \overset{ind}{\sim} N(0, 1)$. The scenarios we consider are:

Scenario 4. The signal θ^* is taken as in Scenario 3 from Section 4.2, and we let

$$v_{k,l}^* = \begin{cases} 1.75 & \text{if } |k - n/2| < n/3, \text{ and } |l - n/2| < n/3, \\ 1 & \text{otherwise,} \end{cases}$$

for $k, l \in \{1, \ldots, n^{1/2}\}$.

Scenario 5. We set

$$v_{k,l}^* = \begin{cases} 1.5 & \text{if } (k - n/2)^2 + (l - n/2)^2 < (n/4)^2, \\ 0.5 & \text{otherwise.} \end{cases}$$

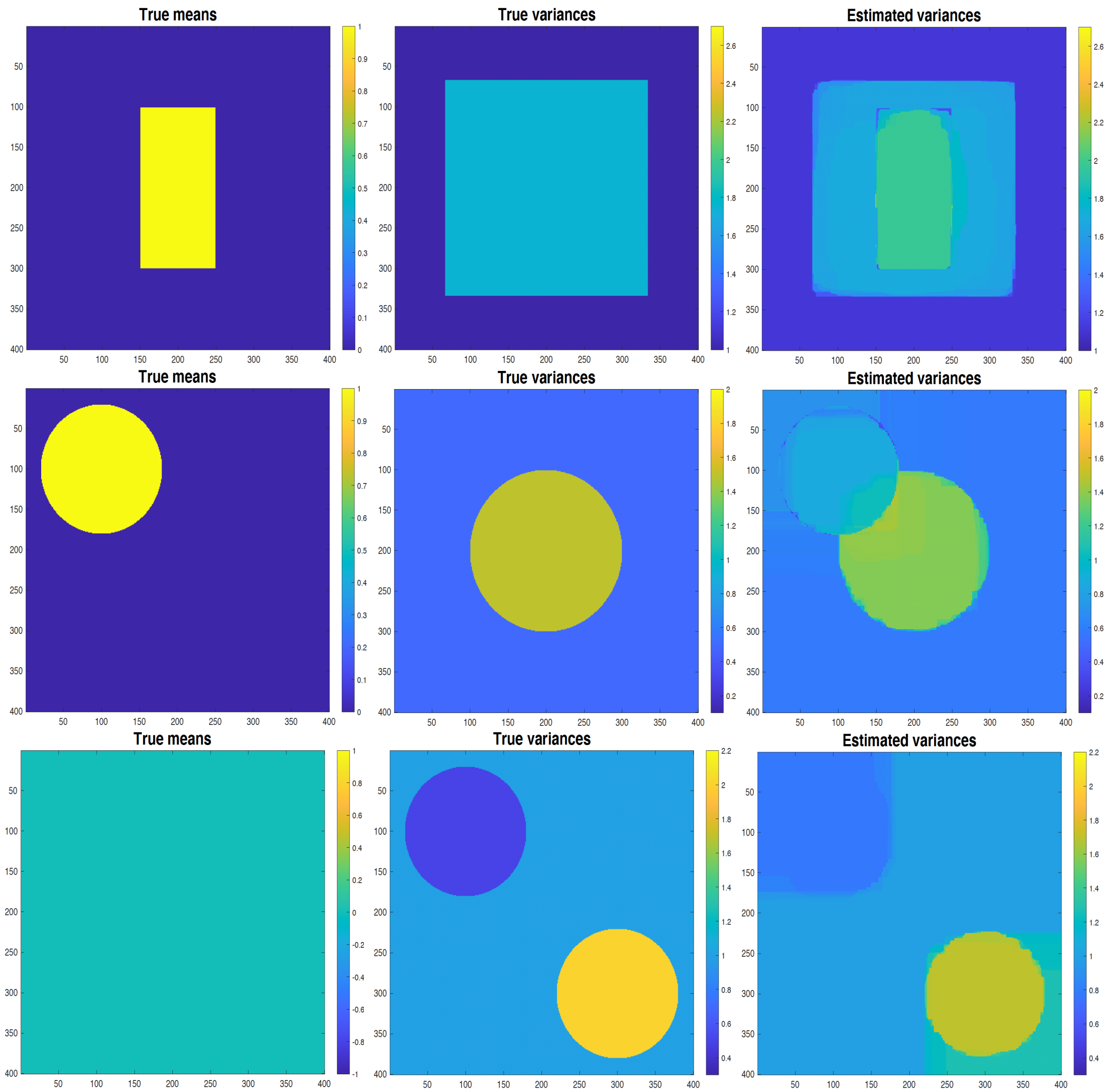

Figure 3: Each row corresponds to one scenario, with the top row corresponding to Scenario 4, the middle to Scenario 5, and the bottom to Scenario 6. The left column depicts the signals θ^*, the middle column the signals v^*, and the right column the estimated $\hat{v}$ with our method in (11)–(12).

	Scenario 4			Scenario 5			Scenario 6		
n	L. Pol.	Laplacian S.	Het.	L. Pol.	Laplacian S.	Het.	L. Pol.	Laplacian S.	Het.
100^2	3.25	11.02	**1.34**	2.18	4.91	**1.57**	2.40	9.49	**1.22**
200^2	3.27	10.71	**0.52**	2.17	4.91	**0.75**	2.44	8.09	**0.72**
300^2	3.21	6.82	**0.29**	2.15	4.43	**0.43**	2.41	7.87	**0.42**
400^2	3.15	6.14	**0.18**	2.14	3.92	**0.28**	2.39	7.41	**0.29**

Table 2: Performance evaluations of the competing methods for the different settings described in the text. We report 10 multiplied by the average mean squared error, averaging over 200 Monte Carlo simulations.

for $k, l \in \{1, \ldots, n^{1/2}\}$ and take θ^* as in Scenario 2 in Section 4.2.

Scenario 6. We let $\theta^* = 0 \in \mathbb{R}^{n^{1/2} \times n^{1/2}}$ and

$$v^*_{k,l} \;=\; \begin{cases} 0.5 & \text{if } \; (k - n/4)^2 + (l - n/4)^2 < (n/5)^2, \\ 2 & \text{if } \; (k - 3n/4)^2 + (l - 3n/4)^2 < (n/5)^2, \\ 1 & \text{otherwise.} \end{cases}$$

for $k, l \in \{1, \ldots, n^{1/2}\}$.

As for benchmarks, we compare our estimator Het. defined in (11)–(12) with the local polynomial regression (L. Pol.) method from Fan and Yao (1998), and the Laplacian smoothing estimator (Laplacian S.), see e.g. Smola and Kondor (2003). For our method Het. the tuning parameters are selected as in Section 4.1. As for the method L. Pol., we use the function *loess* from the *R* package *stats* with the default choices of input. However, for large values of n ($n \geq 10000$) the computational complexity of this function becomes challenging, and hence we average 10 estimates each of which is obtained by fitting the estimator to randomly selected subsets of the data with size 5000.

As for the Laplacian S. estimator, we first define

$$\tilde{\theta} \;:=\; \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2}\|y - \theta\|^2 \;+\; \eta \sum_{(i,j) \in G} |\theta_i - \theta_j|^2 \right\}, \tag{34}$$

where $\eta > 0$ is a tuning parameter, and G is the 2-dimensional grid graph. Thus, the only difference with the estimator $\hat{\theta}$ defined in (2) is in the penalty with (34) using the square of the absolute value of the difference of the signal values, along the edges of the graph. Once $\tilde{\theta}$ has been constructed, we define

$$\tilde{v}_i \;=\; \hat{\gamma}_i - (\tilde{\theta}_i)^2, \tag{35}$$

where

$$\hat{\gamma} \;:=\; \underset{\gamma \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i^2 - \gamma_i)^2 + \eta' \sum_{(i,j) \in E} |\gamma_i - \gamma_j|^2 \right\} \tag{36}$$

for a tuning parameter $\eta' > 0$. The final estimator $\tilde{\gamma}$ is the one that we refer to as Laplacian S., where the tuning parameters are chosen with BIC as in Section 4.1.

For each scenario and value of the tuning parameters, and for each data set, we compute the estimator $\hat{v}$ and choose the tuning parameters as in Section 4.1. We then report

$$\frac{1}{n}\|\hat{v} - v^*\|^2$$

averaging over 200 Monte Carlo simulations. The results in Table 2 show an excellent performance of our estimator, which becomes more evident as n grows. This goes in line with our findings in Theorem 9.

Finally, Figure 3 provides visualizations of Scenarios 4–6 and the corresponding estimates $\hat{v}$ for one instance of $n = 400^2$. There, we can see that $\hat{v}$ is a reasonable estimator of v^*, although $\hat{v}$ is affected by the bias induced by $\hat{\theta}$ which comes from Equation (11).

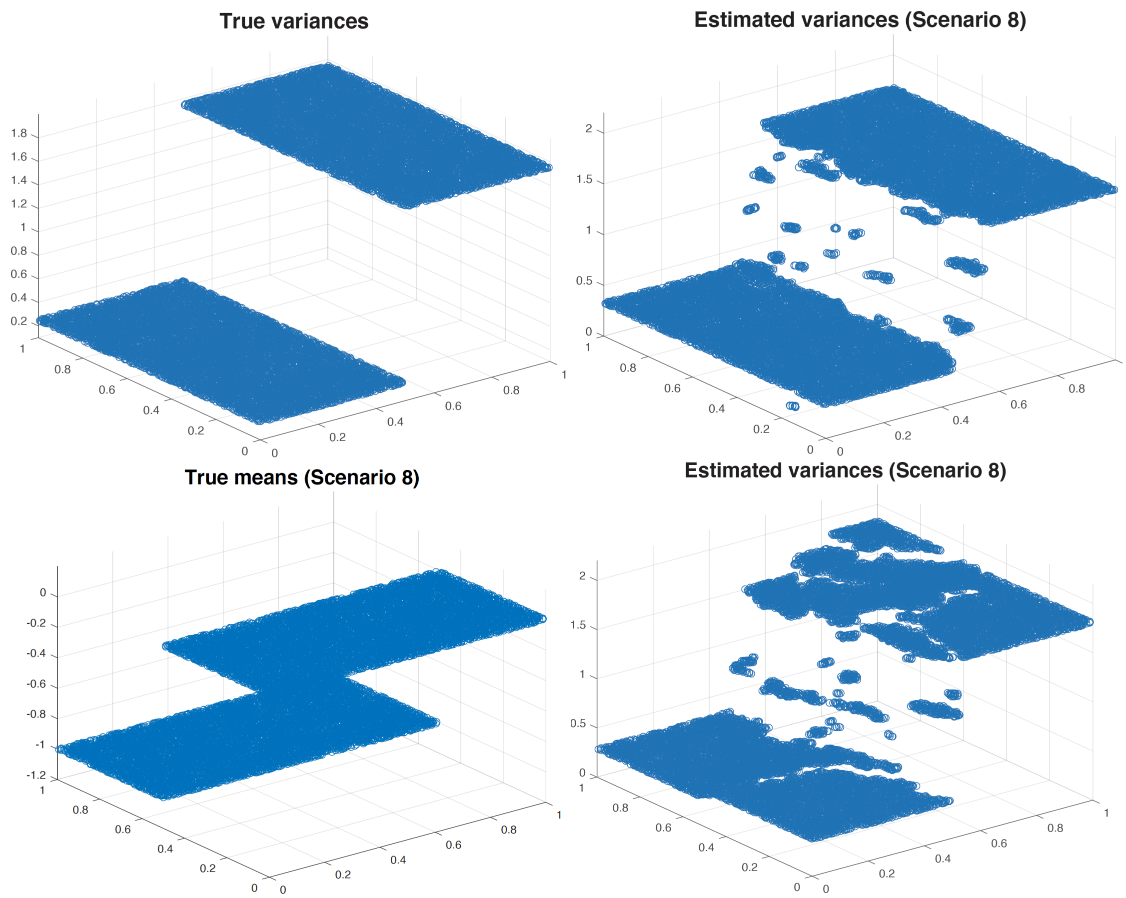

Figure 4: For $n = 20000$ and $d = 2$, the top left panel shows a scatter plot of $\{(x_{i,1}, x_{i,2}, v_i^*)\}_{i=1}^n$ for one instance of Scenarios 7 and 8. The top right panel displays the corresponding scatter plot of $\{(x_{i,1}, x_{i,2}, \hat{v}_i)\}_{i=1}^n$ for Scenario 7. The bottom left panel is the scatter plot of $\{(x_{i,1}, x_{i,2}, f_0(x_i))\}_{i=1}^n$ for Scenario 8, and the bottom right panel shows the scatter plot of $\{(x_{i,1}, x_{i,2}, \hat{v}_i)\}_{i=1}^n$ for Scenario 8. Here, $\hat{v}$ is our Het. estimator defined in (11)–(12) with the K-NN graph.

		Scenario 7			Scenario 8		
n	d	L. Pol.	Laplacian S.	Het.	L. Pol.	Laplacian S.	Het
5000	2	1.06	1.73	**0.59**	1.06	5.27	**0.87**
10000	2	1.04	1.65	**0.40**	1.01	5.11	**0.58**
15000	2	1.01	1.57	**0.34**	1.04	5.26	**0.45**
20000	2	1.25	1.53	**0.27**	1.08	5.05	**0.39**
5000	3	1.48	1.49	**1.45**	**1.86**	4.21	1.88
10000	3	1.38	1.40	**1.05**	1.61	4.11	**1.54**
15000	3	1.34	1.48	**0.92**	1.40	4.11	**1.22**
20000	3	1.42	1.26	**0.89**	1.39	4.37	**1.12**

Table 3: Performance evaluations of the competing methods for the different settings described in the text. We report 10 multiplied by the average mean squared error, averaging over 200 Monte Carlo simulations.

4.4 Heteroscedastic case: K-NN graphs

In this experiment we consider a nonparametric regression setting. Specifically, we generate data from the model

$$\begin{array}{rcl} y_i & = & f_0(x_i) + (v_i^*)^{1/2}\epsilon_i, \\ v_i^* & = & g_0(x_i) \\ \epsilon_i & \overset{\text{ind}}{\sim} & N(0,1), \\ x_i & \overset{\text{ind}}{\sim} & U[0,1]^d, \end{array}$$

where $U[0,1]^d$ is the uniform distribution. In our simulations, we consider $d \in \{2,3\}$, $n \in \{500, 10000, 15000, 20000\}$, and difference choices of f_0 and g_0. The functions f_0 and g_0 are taken from the following scenarios:

Scenario 7. In this scenario, we let $f_0(z) = 0$ for all $z = (z_1, \ldots, z_d)^\top \in \mathbb{R}^d$ and

$$g_0(z) \;=\; \begin{cases} 1.75 & \text{if } \; z_1 > 0.5, \\ 0.25 & \text{otherwise.} \end{cases}$$

Scenario 8. We let g_0 as in Scenario 7, and let

$$f_0(z) \;=\; \begin{cases} 0 & \text{if } \; z_2 > 0.5, \\ -1 & \text{otherwise,} \end{cases}$$

for $z \in \mathbb{R}^d$

Based on the above scenarios, we generate 200 data sets and compute the mean squared error of our estimator in (11)–(12) averaging over all the repetitions. Our estimator is computed using the K-NN graph with $K = 5$. Table 3 seems to corroborate our findings in Theorem 9 as our method's performance appears to improve with sample size but worsens when d increases.

Finally, Figure 4 provides a visualization of the true signals and the estimated variances for one instance with $n = 12000$ and $d = 2$.

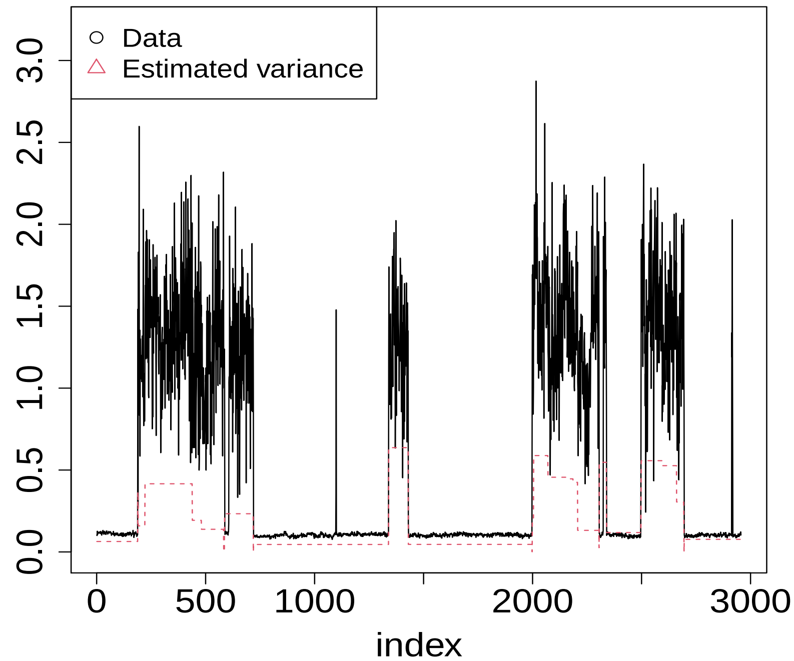

Figure 5: Ion channels data and estimated variances

4.5 Ion channels data

We now validate our method using a real data example. Specifically, we consider the Ion channels data used by Jula Vanegas et al. (2021). The original data was produced by the Steinem Lab (Institute of Organic and Biomolecular Chemistry, University of Gottingen). As explained by Jula Vanegas et al. (2021), Ion channels are a class of proteins expressed by all cells that create pathways for ions to pass through the cell membrane. The data consist of a single ion channel of the bacterial porin PorB, a bacterium related to Neisseria gonorrhoeae.

Although the original data consists of 600000 time instances. We proceed as in Cappello et al. (2021) and construct a signal $y \in \mathbb{R}^{2048}$. The resulting data are plotted in Figure 5. There, we also see the estimated variances using our proposed method for the heteroscedastic case on the 1D chain graph. We see that our method seems to capture the heteroscedastic nature of the data.

5. Conclusion

In this paper, we have studied the problem of estimating the variance in general graph denoising problems. We have proposed and analyzed estimators for both the homoscedastic and heteroscedastic cases. In studying the latter, we also proved generalizations of known bounds for the fused lasso estimator to models beyond sub-Gaussian errors.

Many research directions are left open in this work. One particular problem is to generalize our results to higher order versions of total variation for estimating the vector of variances. Constructing higher order versions of total variation is challenging in the case of estimating the mean in general graph-structured problems, and we expect it to be even more challenging for the variance case. Therefore, we leave this for future work.

Appendix A. Canonical scaling

In previous sections of the paper we made reference to the fact that the canonical scaling of the total variation in a d-dimensional grid graph is $O(n^{1-1/d})$. Thus, for the 1D chain graph we obtain that the canonical scaling is $O(1)$ and $O(n^{1/2})$ for the 2D grid graph. We now justify this by following the discussion from Sadhanala et al. (2017).

To start, consider a d-dimensional grid graph given as $G = (V, E)$, with $V = \{1, \ldots, n\}$. We let $N = n^{1/d}$ and construct the d-dimensional lattice $Z_d = \{(i_1/N, \ldots, i_d/N) \,:\, i_1, \ldots, i_d \in \{1, \ldots, N\}\} \subset [0, 1]^d$. Then we can index the components of a vector $\theta \in \mathbb{R}^n$ by the lattice locations, $\theta(a)$, $a \in Z_d$. Then, the total variation of θ along the graph G is given by

$$\|\nabla_G \theta\|_1 \;=\; \frac{1}{2} \sum_{a \in Z_d} \sum_{b \in Z_d} |\theta(a) - \theta(b)| 1_{\{\|a-b\| = \frac{1}{N}\}}.$$

Notice that the factor $1/2$ appears because we are counting every edge exactly twice. Next, assume that $\theta(a) = f(a)$ for a function $f \,:\, [0, 1]^d \to \mathbb{R}$ such that

$$\|f(a) - f(b)\| \;\leq\; L\|a - b\|,$$

for all $a, b \in [0, 1]^d$ and for a constant $L > 0$. Thus, f is an L-Lipschitz function. It follows that

$$\begin{array}{rcl} \|\nabla_G \theta\|_1 & \leq & \displaystyle\frac{1}{2} \sum_{a \in Z_d} \sum_{b \in Z_d} L\|a - b\| 1_{\{\|a-b\| = \frac{1}{N}\}} \\ & \lesssim & \displaystyle\frac{1}{2} \sum_{a \in Z_d} \frac{dL}{N} \\ & = & \displaystyle\frac{dnL}{2N} \\ & = & O(n^{1-1/d}), \end{array}$$

and so $O(n^{1-1/d})$ makes sense as a canonical scaling for $\|\nabla_G \theta\|_1$ when G is a d-dimensional grid graph.

Appendix B. Additional experiments

B.1 Model selection

Recall that in Section 2 we motivated our homoscedastic estimator as being potentially useful for model selection. In this section, we evaluate a BIC criterion for model selection for the purpose of mean estimation in a 2-dimensional grid graph. For this evaluation, we consider Scenarios 1–3 from Section 4.2. For each scenario, and $\sigma \in \{0.25, 0.5, 0.75, 1\}$, we generate 200 data sets and compare the performance of $\hat{\theta}(\lambda)$, the solution to (2) with two choices of λ. The first choice of λ is taken as optimal, thus, as

$$\lambda \,:=\, \underset{\lambda\in\Lambda}{\arg\min}\|\theta^* - \theta(\lambda)\|^2,$$

where $\Lambda = \{10^1, 10^2, 10^3, 10^4, 10^5\}$. The second choice of λ is set to

$$\lambda \,:=\, \underset{\lambda\in\Lambda}{\arg\min}\{\|y - \hat{\theta}_\lambda\|^2 \,+\, \log(n)\hat{v}\widehat{\mathrm{df}}_\lambda\},$$

where $\widehat{\mathrm{df}}_\lambda$ is the number of connected components induced by $\hat{\theta}(\lambda)$ in the 2-dimensional grid graph, and where $\hat{v}$ is the homoscedastic estimator defined in (8).

We find that in all the instances considered, both choices of λ coincide, suggesting that in practice the estimator $\hat{v}$ can be useful for mean estimation with a BIC criterion that typically produces the optimal choice of λ.

B.2 Mean estimation wih Laplace errors

Notice that in the second step of our heteroscedastic estimator, Equation (12), we actually estimate $\gamma_i^* = \mathbb{E}(y_i^2)$ for $i = 1, \ldots, n$. Thus, we estimate the mean of the random variables $\{y_i^2\}_{i=1}^n$, which are not sub-Gaussian, even if the $\{y_i\}_{i=1}^n$ are sub-Gaussian random variables. In our experiments, the key is that our analysis in Theorem 4 can be used when $\{y_i\}_{i=1}^n$ are sub-Gaussian as in such case the random variables $\{y_i^2\}_{i=1}^n$ are sub-Exponential.

We now evaluate the validity of Thereom 4 in a simulation setting where we estimate the mean of the random variables but with $y_i - \theta_i^*$ following a Laplace distribution. Specifically, we consider the same setting as in Section 4.3, but focusing on estimating $\theta^* \in \mathbb{R}^n$, and with data generated as

$$y_i \,=\, \theta_i^* \,+\, \sqrt{v_i^*}\epsilon_i$$

with $\epsilon_i \overset{ind}{\sim} \mathrm{Laplace}(0, 1)$. The methods that we compare are the gaph fused lasso (GFL) defined in (2) with the Laplacian S. estimator defined in (36), and where we choose the tuning parameters as in Section 4.1.

The results on Table 4 seem to provide additional evidence in favor of Theorem 4. In particular, the GFL outperforms Laplacian S., and the performance of GFL improves as n increases which is what it is expected in light of Theorem 4.

Appendix C. Proof of Theorem 1

Proof First, we observe that

$$\begin{aligned}
|v_0^* - \hat{v}| \;\; \leq \;\; & \left| v_0^* \,-\, \frac{v_0^*}{2(\lfloor n/2\rfloor - 1)} \sum_{i=1}^{\lfloor n/2\rfloor - 1} \{\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}\}^2 \right| + \\
& \left| \frac{v_0^*}{2(\lfloor n/2\rfloor - 1)} \sum_{i=1}^{\lfloor n/2\rfloor - 1} \{\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}\}^2 - \right. \\
& \left. \frac{1}{2(\lfloor n/2\rfloor - 1)} \sum_{i=1}^{\lfloor n/2\rfloor - 1} \{(v_0^*)^{1/2}\epsilon_{\sigma(2i)} - (v_0^*)^{1/2}\epsilon_{\sigma(2i-1)} + \theta^*_{\sigma(2i)} - \theta^*_{\sigma(2i-1)}\}^2 \right|.
\end{aligned}$$

	Scenario 4		Scenario 5		Scenario 6	
n	Laplacian S.	GFL	Laplacian S.	GFL	Laplacian S.	GFL
100^2	1.12	**1.08**	1.12	**0.91**	2.7×10^{-3}	$\mathbf{9.87 \times 10^{-4}}$
200^2	1.13	**0.52**	1.12	**0.54**	2.7×10^{-3}	$\mathbf{2.89 \times 10^{-4}}$
300^2	1.10	**0.23**	1.12	**0.32**	2.7×10^{-3}	$\mathbf{1.44 \times 10^{-4}}$
400^2	1.11	**0.13**	1.11	**0.18**	2.7×10^{-3}	$\mathbf{6.82 \times 10^{-5}}$

Table 4: Performance evaluations of the competing methods for the different settings described in the text. We report 10 multiplied by the average mean squared error, averaging over 200 Monte Carlo simulations.

Next, using the identity $a^2 - (a+b)^2 = -b(2a+b)$ we obtain that

$$\begin{aligned} |v_0^* - \hat{v}| \;\; &\leq \;\; \left| v_0^* - \frac{v_0^*}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} \{\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}\}^2 \right| + \\ &\frac{1}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} |\theta^*_{\sigma(2i)} - \theta^*_{\sigma(2i-1)}| \, |2(v_0^*)^{1/2}(\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}) + (\theta^*_{\sigma(2i)} - \theta^*_{\sigma(2i-1)})| \\ &\leq \;\; \left| v_0^* - \frac{v_0^*}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} \{\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}\}^2 \right| + \\ &\frac{\{4\|\epsilon\|_\infty (v_0^*)^{1/2} + 2\|\theta^*\|_\infty\}}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} |\theta^*_{\sigma(2i)} - \theta^*_{\sigma(2i-1)}| \\ &\leq \;\; v_0^* \left| 1 - \frac{1}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} (\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)})^2 \right| + \\ &\frac{\{4\|\epsilon\|_\infty (v_0^*)^{1/2} + 4\|\theta^*\|_\infty\} \|\nabla_G \theta^*\|_1}{2(\lfloor n/2 \rfloor - 1)} \end{aligned} \tag{37}$$

where the last inequality follows from Lemma 1 in Padilla et al. (2018). Finally, notice that

$$\mathbb{E}\left[\frac{1}{2(\lfloor n/2 \rfloor - 1)} \sum_{i=1}^{\lfloor n/2 \rfloor -1} \{\epsilon_{\sigma(2i)} - \epsilon_{\sigma(2i-1)}\}^2 \right] \; = \; 1 \tag{38}$$

and

$$
\begin{aligned}
\mathrm{var}\left[\frac{1}{2(\lfloor n/2\rfloor-1)}\sum_{i=1}^{\lfloor n/2\rfloor-1}\{\epsilon_{\sigma(2i)}-\epsilon_{\sigma(2i-1)}\}^2\right] &= \frac{1}{4(\lfloor n/2\rfloor-1)^2}\sum_{i=1}^{\lfloor n/2\rfloor-1}\mathrm{var}\left[\{\epsilon_{\sigma(2i)}-\epsilon_{\sigma(2i-1)}\}^2\right]\\
&\le \frac{1}{4(\lfloor n/2\rfloor-1)^2}\sum_{i=1}^{\lfloor n/2\rfloor-1}\mathbb{E}\left[\{\epsilon_{\sigma(2i)}-\epsilon_{\sigma(2i-1)}\}^4\right]\\
&\le \frac{1}{4(\lfloor n/2\rfloor-1)^2}\sum_{i=1}^{\lfloor n/2\rfloor-1}\mathbb{E}\left\{8\epsilon_{\sigma(2i)}^4+8\epsilon_{\sigma(2i-1)}^4\right\}\\
&\le \frac{4}{(\lfloor n/2\rfloor-1)}\sup_{i=1,\ldots,n}\mathbb{E}(\epsilon_i^4)
\end{aligned}
\tag{39}
$$

where the second inequality follows from the inequality $(a+b)^4 \le 8a^4+8b^4$. Combining (37)–(39) with the Chebyshev's inequality we conclude the proof. ■

Appendix D. A general upper bound

Theorem 15 *Consider data $\{o_i\}_{i=1}^n$ generated as $o_i = \beta_i^* + \varepsilon_i$ for some $\beta^* \in \mathbb{R}^n$ and $\varepsilon_1,\ldots,\varepsilon_n$ independent random variables satisfying satisfying $\mathbb{E}(\varepsilon_i)=0$ and*

$$
\max_{i=1,\ldots,n}\mathbb{E}(\varepsilon_i^4) \;=\; O(1).
$$

Let $\hat\beta$ be defined as

$$
\hat\beta \;:=\; \arg\min_{\beta\in\mathbb{R}^n}\left\{\frac{1}{2}\sum_{i=1}^n(o_i-\beta_i)^2+\lambda\sum_{(i,j)\in E}|\beta_i-\beta_j|\right\}.
$$

Let $\eta>0$. Then for any sequence $U_n>0$ and for any $\eta>0$, it holds that

$$
\begin{aligned}
\mathrm{pr}(\|\hat\beta-\beta^*\|>\eta) \;\le\;& \frac{16n^{1/2}\max\limits_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{\eta}+\\
&\frac{16}{\eta^2}\,\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\le\eta,\,\|\nabla_G\beta\|_1\le 5\|\nabla_G\beta^*\|_1}\sum_{i=1}^n\xi_i\varepsilon_i 1_{\{|\varepsilon_i|\le U_n\}}(\beta_i-\beta_i^*)\right],
\end{aligned}
$$

where $\xi_1,\ldots,\xi_n$ are independent Rademacher random variables independent of $\{\varepsilon_i\}_{i=1}^n$, provided that

$$
\lambda \;=\; \frac{\eta^2}{4\|\nabla_G\beta^*\|_1}. \tag{40}
$$

Proof First, notice that by convexity and the basic inequality we have that

$$
\frac{1}{2}\sum_{i=1}^n(o_i-\beta_i)^2+\lambda\|\nabla_G\beta\|_1 \;\le\; \frac{1}{2}\sum_{i=1}^n(o_i-\beta_i^*)^2+\lambda\|\nabla_G\beta^*\|_1 \tag{41}
$$

for any $\beta \in \Lambda := \{s\hat{\beta} + (1-s)\beta^* \,:\, s \in [0,1]\}$. Then

$$\|\nabla_G \beta\|_1 \,\leq\, \|\nabla_G \beta\|_1 + \frac{\|\beta - \beta^*\|^2}{2\lambda} \,\leq\, \frac{\varepsilon^\top(\beta-\beta^*)}{\lambda} + \|\nabla_G \beta^*\|_1 \tag{42}$$

for all $\beta \in \Lambda$. This implies

$$\begin{array}{rcl} \|\nabla_G(\beta - \beta^*)\|_1 & \leq & \|\nabla_G \beta\|_1 + \|\nabla_G \beta^*\|_1 \\ & \leq & \dfrac{\varepsilon^\top(\beta-\beta^*)}{\lambda} + 2\|\nabla_G \beta^*\|_1, \end{array} \tag{43}$$

for all $\beta \in \Lambda$.

Next, let $\beta \in \Lambda$ and suppose that $\|\beta - \beta^*\|^2 \leq \eta^2$, and $\|\nabla_G \beta\|_1 \geq 5\|\nabla_G \beta^*\|_1$. Then

$$\|\nabla_G(\beta - \beta^*)\|_1 \,\geq\, \|\nabla_G \beta\|_1 - \|\nabla_G \beta^*\|_1 \,\geq\, 4\|\nabla_G \beta^*\|_1.$$

Hence, setting

$$s \,:=\, \frac{4\|\nabla_G \beta^*\|_1}{\|\nabla_G(\beta - \beta^*)\|_1},$$

clearly $s \in [0,1]$, and we let

$$\beta' := s\beta + (1-s)\beta^* \in \Lambda.$$

Then

$$\|\beta' - \beta^*\|^2 \,\leq\, \|\beta - \beta^*\|^2 \,\leq\, \eta^2,$$

and

$$\begin{array}{rcl} \|\nabla_G(\beta' - \beta^*)\|_1 & = & s\|\nabla_G(\beta - \beta^*)\|_1 \\ & = & 4\|\nabla_G \beta^*\|_1. \end{array}$$

Therefore, from (43),

$$4\|\nabla_G \beta^*\|_1 \,=\, \|\nabla_G(\beta' - \beta^*)\|_1 \,\leq\, \frac{\varepsilon^\top(\beta'-\beta^*)}{\lambda} + 2\|\nabla_G \beta^*\|_1$$

which implies

$$2\|\nabla_G \beta^*\|_1 \,\leq\, \frac{\varepsilon^\top(\beta'-\beta^*)}{\lambda}.$$

Hence, if we take

$$\lambda \,=\, \frac{\eta^2}{4\|\nabla_G \beta^*\|_1}$$

we obtain

$$\frac{\eta^2}{2} \,\leq\, \varepsilon^\top(\beta' - \beta^*).$$

As a result, the events

$$\Omega_1 \,:=\, \left\{ \sup_{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta} \|\nabla_G \beta\|_1 \,\geq\, 5\|\nabla_G \beta^*\|_1 \right\}$$

and

$$\Omega_2 \,:=\, \left\{ \sup_{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G(\beta-\beta^*)\|_1 \leq 4\|\nabla_G \beta^*\|_1} \varepsilon^\top(\beta - \beta^*) \,\geq\, \frac{\eta^2}{2} \right\}$$

satisfy that $\Omega_1 \subset \Omega_2$. And so,

$$\mathrm{pr}(\Omega_1) \,\leq\, \mathrm{pr}(\Omega_2). \tag{44}$$

Next, suppose that $\|\hat{\beta} - \beta^*\| > \eta$. Then there exists $\beta \in \Lambda$ such that $\|\beta - \beta^*\| = \eta$ and so (42) implies that

$$\frac{\eta^2}{2} \,\leq\, \varepsilon^\top(\beta - \beta^*) \,+\, \lambda\|\nabla_G \beta^*\|_1 \,-\, \lambda\|\nabla_G \beta\|_1.$$

Hence, given our choice of λ, we obtain that

$$\frac{\eta^2}{4} \,\leq\, \varepsilon^\top(\beta - \beta^*),$$

for some $\beta \in \Lambda$, provided that $\|\hat{\beta} - \beta^*\| > \eta$.

The above implies that

$$\begin{array}{lll} \mathrm{pr}(\|\hat{\beta} - \beta^*\| > \eta) & \leq & \mathrm{pr}(\{\|\hat{\beta} - \beta^*\| > \eta\} \cap \Omega_1^c) \,+\, \mathrm{pr}(\Omega_1) \\ & \leq & \mathrm{pr}(\{\|\hat{\beta} - \beta^*\| > \eta\} \cap \Omega_1^c) \,+\, \mathrm{pr}(\Omega_2) \\ & \leq & \mathrm{pr}\left\{ \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \varepsilon^\top(\beta - \beta^*) \,\geq\, \frac{\eta^2}{4} \right\} \,+\, \mathrm{pr}(\Omega_2) \\ & \leq & 2\,\mathrm{pr}\left\{ \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \varepsilon^\top(\beta - \beta^*) \,\geq\, \frac{\eta^2}{4} \right\} \\ & \leq & \dfrac{8}{\eta^2}\,\mathbb{E}\left\{ \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \varepsilon^\top(\beta - \beta^*) \right\} \\ & =: & A_1, \end{array} \tag{45}$$

where the second inequality follows from (44), and third from the discussion above, the fourth from the definition of Ω_2, and the last inequality from Markov's inequality. Next, notice that

$$\begin{array}{lll} A_1 & \leq & \dfrac{8}{\eta^2}\,\mathbb{E}\left[\underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \displaystyle\sum_{i=1}^n \varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}}(\beta_i - \beta_i^*) \right] + \\ & & \dfrac{8}{\eta^2}\,\mathbb{E}\left[\underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \displaystyle\sum_{i=1}^n \varepsilon_i 1_{\{|\varepsilon_i| > U_n\}}(\beta_i - \beta_i^*) \right] \\ & \leq & \dfrac{8}{\eta^2}\,\mathbb{E}\left\{ \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \displaystyle\sum_{i=1}^n (\varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}} - \mathbb{E}[\varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}}])(\beta_i - \beta_i^*) \right\} + \\ & & \dfrac{8}{\eta^2} \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \displaystyle\sum_{i=1}^n \mathbb{E}[\varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}}](\beta_i - \beta_i^*) + \\ & & \dfrac{8}{\eta^2}\,\mathbb{E}\left[\underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \displaystyle\sum_{i=1}^n \varepsilon_i 1_{\{|\varepsilon_i| > U_n\}}(\beta_i - \beta_i^*) \right]. \\ & =: & A_2 + A_3 + A_4, \end{array}$$

Next, we proceed to bound A_2, A_3 and A_4. To bound A_3, notice that since $\mathbb{E}(\varepsilon_i) = 0$ then

$$A_3 \;=\; \frac{8}{\eta^2} \underset{\beta \in \Lambda \,:\, \|\beta-\beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1}{\sup} \sum_{i=1}^n -\mathbb{E}[\varepsilon_i 1_{\{|\varepsilon_i| > U_n\}}](\beta_i - \beta_i^*).$$

Hence,

$$
\begin{aligned}
A_3 &\leq \frac{8n^{1/2}}{\eta} \max_{i=1,\ldots,n} |\mathbb{E}[\varepsilon_i 1_{\{|\varepsilon_i|>U_n\}}]| \\
&\leq \frac{8n^{1/2}}{\eta} \max_{i=1,\ldots,n} \left(\mathbb{E}(\varepsilon_i^2)\, \mathbb{E}[1_{\{|\varepsilon_i|>U_n\}}]\right)^{1/2} \\
&= \frac{8n^{1/2}}{\eta} \left\{ \max_{i=1,\ldots,n} \mathbb{E}(\varepsilon_i^2)\, \mathrm{pr}(|\varepsilon_i| > U_n) \right\}^{1/2},
\end{aligned} \tag{46}
$$

where the first and second inequalities follow from Cauchy–Schwarz inequality.

To bound A_4, we observe that

$$
\begin{aligned}
A_4 &\leq \frac{8}{\eta^2}\, \mathbb{E}\Bigg(\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \left[\sum_{i=1}^n \varepsilon_i^2 1_{\{|\varepsilon_i|>U_n\}}\right]^{1/2} \|\beta-\beta^*\| \Bigg) \\
&\leq \frac{8}{\eta}\, \mathbb{E}\Bigg(\left[\sum_{i=1}^n \varepsilon_i^2 1_{\{|\varepsilon_i|>U_n\}}\right]^{1/2} \Bigg) \\
&\leq \frac{8}{\eta} \left(\mathbb{E}\left[\sum_{i=1}^n \varepsilon_i^2 1_{\{|\varepsilon_i|>U_n\}}\right] \right)^{1/2} \\
&\leq \frac{8n^{1/2}}{\eta} \left(\max_{i=1,\ldots,n} \mathbb{E}\left[\varepsilon_i^2 1_{\{|\varepsilon_i|>U_n\}}\right] \right)^{1/2} \\
&\leq \frac{8n^{1/2} \max_{i=1,\ldots,n} \{\mathbb{E}(\varepsilon_i^4)\}^{1/4} \{\mathrm{pr}(|\varepsilon_i| > U_n)\}^{1/4}}{\eta}.
\end{aligned} \tag{47}
$$

Let us now proceed to bound A_2. Let $\varepsilon_1', \ldots, \varepsilon_n'$ independent copies of $\varepsilon_1, \ldots, \varepsilon_n$. Then for independent Rademacher random variables $\xi_1, \ldots, \xi_n$, it holds that

$$
\begin{aligned}
A_2 &\leq \frac{8}{\eta^2}\, \mathbb{E}\left(\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n [\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}} - \varepsilon_i' 1_{\{|\varepsilon_i'|\leq U_n\}}](\beta_i - \beta_i^*) \right) \\
&= \frac{8}{\eta^2}\, \mathbb{E}\left(\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n \xi_i [\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}} - \varepsilon_i' 1_{\{|\varepsilon_i'|\leq U_n\}}](\beta_i - \beta_i^*) \right) \\
&\leq \frac{8}{\eta^2}\, \mathbb{E}\left(\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n \xi_i [\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}](\beta_i - \beta_i^*) \right) + \\
&\quad\; \frac{8}{\eta^2}\, \mathbb{E}\left(\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n -\xi_i [\varepsilon_i' 1_{\{|\varepsilon_i'|\leq U_n\}}](\beta_i - \beta_i^*) \right) \\
&= \frac{16}{\eta^2}\, \mathbb{E}\left[\sup_{\beta\in\Lambda\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n \xi_i \varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}} (\beta_i - \beta_i^*) \right] \\
&\leq \frac{16}{\eta^2}\, \mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\, \|\beta-\beta^*\|\leq\eta,\, \|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1} \sum_{i=1}^n \xi_i \varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}} (\beta_i - \beta_i^*) \right].
\end{aligned} \tag{48}
$$

The claim then follows.

■

Appendix E. Assumptions for K-NN graph for Theorem 4

We start by explicitly defining the construction of the K-NN graph. Specifically, $(i, j) \in E$ if and only if x_j is among the K-nearest neighbors (with respect to the metric $\mathrm{dist}(\cdot)$ of x_i, or vice versa.

We now state the assumptions from Madrid Padilla et al. (2020b) needed for Theorem 4. Throughout $(\mathcal{X}, \mathrm{dist})$ is a metric space with Borel sets $\mathcal{B}(\mathcal{X})$.

Assumption 2 *The covariates $\{x_i\}_{i=1}^n$ are independent draws from a density p, with respect to the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$, with support $\mathcal{X}$. Furthermore, the density p satisfies $0 < p_{\min} < p(x) < p_{\max}$ for all $x \in \mathcal{X}$, where $p_{\min}$ and $p_{\max}$ are constants.*

Assumption 3 *The base measure satisfies*

$$c_1 r^d \;\leq\; \mu\left[\{q \in \mathcal{X} \,:\, \mathrm{dist}(q, x) \leq r\}\right] \;\leq\; c_2 r^d$$

for all $x \in \mathcal{X}$, and all $0 < r < r_0$, where c_1, c_2 and r_0 are all positive constants, and $d \in \mathbb{N}\backslash\{0\}$ is the intrinsic dimension of $\mathcal{X}$.

Assumption 4 *There exists a homeomorphism (a continuous bijection with a continuous inverse) $h \,:\, \mathcal{X} \rightarrow [0,1]^d$ such that*

$$L_{\min}\mathrm{dist}(x, x') \;\leq\; \|h(x) - h(x')\| \;\leq\; L_{\max}\mathrm{dist}(x, x'), \;\; \forall x, x' \in \mathcal{X},$$

for some positive constants $L_{\min}$ and $L_{\max}$.

For a set $S \subset [0,1]^d$, we let

$$B_t(S) \;:=\; \{q \in [0,1]^d \,:\, \|q - q'\| \leq t \text{ for some } \; q' \in S\}.$$

With this notation, we state our next assumption.

Assumption 5 ***[Piecewise Lipschitz].*** *The parameter β^* satisfies that $\beta^*_i = f_0(x_i)$ for $i = 1, \ldots, n$ for some function f_0, where the following holds for the function f_0.*

1. *f_0 is bounded.*
2. *Let $\partial[0,1]^d$ be the boundary of $[0,1]^d$, and let $\Omega_t = [0,1]^d \backslash B_t(\partial[0,1]^d)$. We assume that there exists a set $\mathcal{S}$ such that:*
   - *(a) The set $\mathcal{S}$ has Lebesgue measure zero.*
   - *(b) For some constants $C_{\mathcal{S}}, t_0 > 0$, we have that*

     $$\mu\left[h^{-1}\left\{B_t(\mathcal{S}) \cup ([0,1]^d \backslash \Omega_t)\right\}\right] \;\leq\; C_{\mathcal{S}} t$$

     for all $0 < t < t_0$.
   - *(c) There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_t \backslash B_t(\mathcal{S})$ then*

     $$|f_0 \circ h^{-1}(z) - f_0 \circ h^{-1}(z')| \;\leq\; L_0 \|z - z'\|.$$

Notation: We denote as $\mathcal{F}(L_0)$ the set of functions $f\,:\,[0,1]^d \rightarrow \mathbb{R}$ that satisfy Assumption 5 with L_0 and such that

$$\sup_{x\in[0,1]^d} |f(x)| \,\leq\, L_0,$$

and $C_{\mathcal{S}} \leq L_0$.

Appendix F. Assumptions for K-NN graph for Theorem 9

We assume that the covariates $\{x_i\}_{i=1}^n$ satisfy Assumptions 2–4. In addition, we assume that Assumption 5 holds replacing β^* with both v^* and θ^*.

Appendix G. Proof of Theorem 4

Proof of (15): First, let G' be a chain graph corresponding to a DFS ordering in G. Based of Theorem 15, we first need to bound

$$B_1 := \frac{16}{\eta^2}\,\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_G\beta\|_1\leq 5\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\right]. \tag{49}$$

To bound this, we recall Lemma 1 in Padilla et al. (2018) which implies that $\|\nabla_{G'}\beta\|_1 \leq 2\|\nabla_G\beta\|_1$ for all $\beta\in\mathbb{R}^n$. Hence,

$$\begin{aligned}
B_1 &\leq \frac{16}{\eta^2}\,\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 10\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\right]\\
&= \frac{16}{\eta^2}\mathbb{E}\left(\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 10\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\,\middle|\,\varepsilon\right]\right)\\
&= \frac{16U_n}{\eta^2}\mathbb{E}\left(\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 10\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\frac{\varepsilon_i}{U_n} 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\,\middle|\,\varepsilon\right]\right).
\end{aligned} \tag{50}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 10\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\frac{\varepsilon_i}{U_n} 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\,\middle|\,\varepsilon\right]\\
&\leq \mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta-\beta^*\|\leq\eta,\,\|\nabla_{G'}(\beta-\beta^*)\|_1\leq 11\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\frac{\varepsilon_i}{U_n} 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i-\beta_i^*)\,\middle|\,\varepsilon\right]\\
&\leq \mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 11\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\frac{\varepsilon_i}{U_n} 1_{\{|\varepsilon_i|\leq U_n\}}\beta_i\,\middle|\,\varepsilon\right]\\
&\leq \mathbb{E}\left[\sup_{\beta\in\mathbb{R}^n\,:\,\|\beta\|\leq\eta,\,\|\nabla_{G'}\beta\|_1\leq 11\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \xi_i\beta_i\,\middle|\,\varepsilon\right]
\end{aligned}$$

where the last inequality follows from Theorem 4.12 in Ledoux and Talagrand (1991). As a result, letting $\tilde{\xi}_i \overset{\text{ind}}{\sim} N(0,1)$ for $i = 1, \dots, n$, we obtain that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sup_{\beta\in\mathbb{R}^n \,:\, \|\beta-\beta^*\|\leq \eta,\, \|\nabla_{G'}\beta\|_1 \leq 10\|\nabla_G \beta^*\|_1} \sum_{i=1}^n \xi_i \frac{\varepsilon_i}{U_n} 1_{\{|\varepsilon_i|\leq U_n\}}(\beta_i - \beta_i^*) \,\Bigg|\, \varepsilon \Bigg] \\
&\leq \mathbb{E}\Bigg(\sup_{\beta\in\mathbb{R}^n \,:\, \|\beta\|\leq \eta,\, \|\nabla_{G'}\beta\|_1 \leq 11\|\nabla_G \beta^*\|_1} \sum_{i=1}^n \xi_i \beta_i \Bigg) \\
&\leq \left(\frac{\pi}{2}\right)^{1/2} \mathbb{E}\Bigg(\sup_{\beta\in\mathbb{R}^n \,:\, \|\beta\|\leq \eta,\, \|\nabla_{G'}\beta\|_1 \leq 11\|\nabla_G \beta^*\|_1} \sum_{i=1}^n \tilde{\xi}_i \beta_i \Bigg) \\
&\leq C_1 \Bigg\{ \eta \left(\frac{11\|\nabla_G \beta^*\|_1 n^{1/2}}{\eta} \right)^{1/2} + \eta\{\log(en)\}^{1/2} \Bigg\}
\end{aligned}$$

where the second inequality follows fromt a well known fact bounding Rademacher Width by Gaussian Width; e.g see Page 132 in Wainwright (2019), and the last by Lemma B.1 from Guntuboyina et al. (2020). This implies that

$$B_1 \leq \frac{16U_n}{\eta^2} C_1 \Bigg[\eta \left(\frac{11\|\nabla_G \beta^*\|_1 n^{1/2}}{\eta} \right)^{1/2} + \eta\{\log(en)\}^{1/2} \Bigg].$$

Therefore, given $a \in (0,1)$, we let

$$\eta := \frac{4}{a} \left[16^{2/3} n^{1/6} (\log n)^{1/6} (11)^{1/3} \|\nabla_G \beta^*\|_1^{1/3} U_n^{2/3} (C_1)^{2/3} + 16 C_1 U_n \{\log(en)\}^{1/2} \right]$$

and λ as in (40). Hence,

$$\begin{aligned}
B_1 &\leq \frac{16C_1 U_n (11\|\nabla_G \beta^*\|_1)^{1/2} n^{1/4}}{\eta^{3/2}} + \frac{16 C_1 U_n \{\log(en)\}^{1/2}}{\eta} \\
&\leq \frac{a^{3/2}}{4^{3/2}} \frac{16C_1 U_n (11\|\nabla_G \beta^*\|_1)^{1/2} n^{1/4}}{\left\{ 16^{2/3} n^{1/6} (\log n)^{1/6} (11)^{1/3} \|\nabla_G \beta^*\|_1^{1/3} U_n^{2/3} (C_1)^{2/3} \right\}^{3/2}} + \frac{a}{4} \\
&\leq \frac{a}{2}.
\end{aligned} \tag{51}$$

Furthermore, by Theorem 15, we must bound

$$B_2 := \frac{16 n^{1/2} \max_{i=1,\dots,n} \{\mathbb{E}(\varepsilon_i^4)\}^{1/4} \{\text{pr}(|\varepsilon_i| > U_n)\}^{1/4}}{\eta}. \tag{52}$$

However, given our definition of η, we obtain that

$$\begin{aligned}
B_2 &\leq \frac{a}{4} \frac{16 n^{1/2} \max_{i=1,\dots,n} \{\mathbb{E}(\varepsilon_i^4)\}^{1/4} \{\text{pr}(|\varepsilon_i| > U_n)\}^{1/4}}{\left[16^{2/3} n^{1/6} (\log n)^{1/6} (11)^{1/3} \|\nabla_G \beta^*\|_1^{1/3} U_n^{2/3} (C_1)^{2/3} + 16 C_1 U_n \{\log(en)\}^{1/2} \right]} \\
&\leq \frac{a}{4}
\end{aligned}$$

where the last inequality follows from (14). The conclusion of the Theorem follows from Theorem 15.

Proof of rate (17): As before, we first bound B_1 as defined in (49). Towards that end, let ∇_G^{+} the pseudo inverse of ∇_G, and Π the orthogonal projection onto the span of $(1,\ldots,1)^\top \in \mathbb{R}^n$. Then notice that

$$
\begin{aligned}
B_1 \;&\leq\; \frac{16U_n}{\eta^2}\,\mathbb{E}\left(\sup_{\delta\in\mathbb{R}^n\,:\,\|\delta\|\leq\eta,\,\|\nabla_G\delta\|_1\leq 6\|\nabla_G\beta^*\|_1}\sum_{i=1}^n \frac{\xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}}{U_n}\delta_i\right)\\
&\leq\; \frac{16U_n}{\eta^2}\,\mathbb{E}\left(\sup_{\delta\in\mathbb{R}^n\,:\,\|\delta\|\leq\eta,\,\|\nabla_G\delta\|_1\leq 6\|\nabla_G\beta^*\|_1}\tilde{\varepsilon}^\top\nabla_G^{+}\nabla_G\delta\right) \;+\\
&\quad\;\; \frac{16U_n}{\eta^2}\,\mathbb{E}\left(\sup_{\delta\in\mathbb{R}^n\,:\,\|\delta\|\leq\eta,\,\|\nabla_G\delta\|_1\leq 6\|\nabla_G\beta^*\|_1}\tilde{\varepsilon}^\top\Pi\delta\right),
\end{aligned}
$$

where $\tilde{\varepsilon}_i = \xi_i\varepsilon_i 1_{\{|\varepsilon_i|\leq U_n\}}/U_n$ for $i=1,\ldots,n$. Next, we observe that by Hölder's inequality and Cauchy–Schwarz inequality, it holds that,

$$
\begin{aligned}
B_1 \;&\leq\; \frac{96U_n\|\nabla_G\beta^*\|_1}{\eta^2}\mathbb{E}\left(\|(\nabla_G^{+})^\top\tilde{\varepsilon}\|_\infty\right) \;+\; \frac{16U_n}{\eta^2}\,\mathbb{E}\left\{\sup_{\delta\in\mathbb{R}^n\,:\,\|\delta\|\leq\eta,\,\|\nabla_G\delta\|_1\leq 6\|\nabla_G\beta^*\|_1}\tilde{\varepsilon}^\top\Pi\delta\right\}\\
&=\; \frac{96U_n\|\nabla_G\beta^*\|_1}{\eta^2}\mathbb{E}\left(\|(\nabla_G^{+})^\top\tilde{\varepsilon}\|_\infty\right) \;+\; \frac{16U_n}{\eta^2}\sup_{\delta\,:\,\|\delta\|\leq\eta}\left(\frac{1}{n^{1/2}}\sum_{i=1}^n\delta_i\right)\mathbb{E}\left(\left|\frac{1}{n^{1/2}}\sum_{i=1}^n\tilde{\varepsilon}_i\right|\right)\\
&\leq\; \frac{96U_n\|\nabla_G\beta^*\|_1}{\eta^2}\mathbb{E}\left(\|(\nabla_G^{+})^\top\tilde{\varepsilon}\|_\infty\right) \;+\; \frac{16U_n}{\eta}\mathbb{E}\left(\left|\frac{1}{n^{1/2}}\sum_{i=1}^n\tilde{\varepsilon}_i\right|\right)\\
&\leq\; \frac{96U_n\|\nabla_G\beta^*\|_1}{\eta^2}\mathbb{E}\left(\|(\nabla_G^{+})^\top\tilde{\varepsilon}\|_\infty\right) \;+\; \frac{16U_n}{\eta}\\
&\leq\; \frac{cU_n\|\nabla_G\beta^*\|_1}{\eta^2}\sqrt{\log n}\cdot\max_j\|(\nabla_G^{+})_{,j}\| \;+\; \frac{16U_n}{\eta}
\end{aligned}
$$

where the third and last inequalities follow from basic properties of Sub-Gaussian random variables, and where $c>0$ is a constant. Next, by Propositions 4 and 6 from Hütter and Rigollet (2016), we obtain that

$$
\max_j\|(\nabla_G^{+})_{,j}\| \;\leq\; \phi_n \;:=\; \begin{cases} C(\log n)^{1/2} & \text{if } d=2,\\ C, & \end{cases} \tag{53}
$$

for some constant $C>0$. Therefore,

$$
B_1 \;\leq\; \frac{cU_n\sqrt{\log n}\|\nabla_G\beta^*\|_1\phi_n}{\eta^2} \;+\; \frac{16U_n}{\eta}. \tag{54}
$$

Hence, for a given $a\in(0,1)$, we let

$$
\eta \;:=\; \frac{2}{a^{1/2}}(cU_n\sqrt{\log n}\|\nabla_G\beta^*\|_1\phi_n)^{1/2} + \frac{4}{a}\cdot 16U_n \tag{55}
$$

and λ as in (40).

Therefore,

$$
B_1 \;\leq\; \frac{a}{4}+\frac{a}{4} \;=\; \frac{a}{2}.
$$

Moreover, from (52), we have that

$$\begin{aligned} B_2 &= \frac{16n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{\eta} \\ &\leq \frac{a}{4}\frac{n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{U_n} \\ &\leq \frac{a}{4}, \end{aligned} \tag{56}$$

where the last inequality follows from (16) and the fact that $\max_{i=1,\ldots,n}\mathbb{E}(\varepsilon_i^4) = O(1)$.

Therefore,

$$\mathrm{pr}(\|\hat{\beta}-\beta^*\| > \eta) \leq \frac{3a}{4}.$$

This proves (17).

Case $\|\nabla_G\beta^*\|_1 \geq 1$. Suppose now that $\|\nabla_G\beta^*\|_1 \geq 1$. Then instead of setting η as in (55), we let

$$\eta := \frac{2}{a^{1/2}}(96U_n\|\nabla_G\beta^*\|_1\phi_n)^{1/2} + \frac{4\|\nabla_G\beta^*\|_1^{1/2}}{a}\cdot 16U_n$$

and λ as in (40), or

$$\lambda := \frac{\eta^2}{4\|\nabla_G\beta^*\|_1} = \frac{1}{4}\left[\frac{2}{a^{1/2}}(96U_n\phi_n)^{1/2} + \frac{4}{a}\cdot 16U_n\right]^2.$$

Hence, from (54)

$$B_1 \leq \frac{a^2}{4} + \frac{a}{4\|\nabla_G\beta^*\|_1^{1/2}} \leq \frac{a^2}{4} + \frac{a}{4} \leq \frac{a}{2}.$$

Furthermore, as in (56),

$$\begin{aligned} B_2 &= \frac{16n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{\eta} \\ &\leq \frac{a}{4\|\nabla_G\beta^*\|_1^{1/2}}\frac{n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{U_n} \\ &\leq \frac{a}{4}\frac{n^{1/2}\max_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i|>U_n)\}^{1/4}}{U_n} \\ &\leq \frac{a}{4}, \end{aligned}$$

where the last inequality follows from (16) and the fact that $\max_{i=1,\ldots,n}\mathbb{E}(\varepsilon_i^4) = O(1)$. As a before, we arrive at

$$\mathrm{pr}(\|\hat{\beta}-\beta^*\| > \eta) \leq \frac{3a}{4}.$$

Proof of (20): First, by Madrid Padilla et al. (2020a), there exists N satisfying $N \asymp (n/K)^{1/d}$ and functions $I : \mathbb{R}^n \to \mathbb{R}^n$, and $\tilde{I} : \mathbb{R}^n \to \mathbb{R}^{N^d}$ satisfying the properties below.

- **[Lemma 8 in Madrid Padilla et al. (2020b)].** Let $\mathcal{E}_1$ be the event such that

$$|e^\top\{\beta - I(\beta)\}| \;\leq\; 2\|e\|_\infty \|\nabla_G \beta\|_1, \quad \forall \beta, e \in \mathbb{R}^n, \tag{57}$$

and there exists a d-dimensional lattice G' with N^d nodes such that

$$\|\nabla_{G'} \tilde{I}(\beta)\|_1 \;\leq\; \|\nabla_G \beta\|_1, \quad \forall \beta \in \mathbb{R}^n. \tag{58}$$

Then $\mathrm{pr}(\mathcal{E}_1) \to 1$.

- **[Lemmas 7, 8 and 10 in Madrid Padilla et al. (2020b)].** Le e be any vector of mean zero independent subg-Gaussian(σ^2), then there exists $\tilde{e}$ a vector of mean zero independent sub-Gaussian(σ^2) random variables and a constant $C > 0$(not depending on e) such that the event $\mathcal{E}_2$ given as

$$\begin{aligned} \mathcal{E}_2 \;\;:=\;\; & \Big\{ e^\top\{I(\beta) - I(\beta^*)\} \;\leq\; CK^{1/2}\Big\{ \|\Pi\tilde{e}\|_2 \|\beta - \beta^*\| \;+\; \|(\nabla_{G'})^+ \tilde{e}\|_\infty (\|\nabla_G \beta^*\|_1 + \|\nabla_G \beta\|_1) \Big\}, \\ & \forall \beta, e \in \mathbb{R}^n \Big\} \end{aligned} \tag{59}$$

satisfies $\mathrm{pr}(\mathcal{E}_2) \to 1$.

- **Theorem 2 in Madrid Padilla et al. (2020b)].** It holds that for some constant $C_2 > 0$ the event

$$\mathcal{E}_3 \;:=\; \Big\{ \|\nabla_G \beta^*\|_1 \;\leq\; C_2 \mathrm{poly}(\log n) n^{1-1/d} \Big\}$$

satisfies $\mathrm{pr}(\mathcal{E}_3) \to 1$, where $\mathrm{poly}(\cdot)$ is a polynomial function.

Let $\mathcal{E}_4 = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ Notice that as in Theorem 15, with the choice

$$\lambda \;=\; \frac{\eta^2}{4\|\nabla_G \beta^*\|_1},$$

we have that

$$\begin{aligned} \mathrm{pr}(\|\hat{\beta} - \beta^*\| > \eta | \mathcal{E}_4) \;\;\leq\;\; & \frac{16 n^{1/2} \max\limits_{i=1,\ldots,n} \{\mathbb{E}(\varepsilon_i^4)\}^{1/4} \{\mathrm{pr}(|\varepsilon_i| > U_n)\}^{1/4}}{\eta} + \\ & \frac{16 U_n}{\eta^2}\, \mathbb{E}\left[\sup_{\beta \in \mathbb{R}^n \,:\, \|\beta - \beta^*\| \leq \eta,\, \|\nabla_G \beta\|_1 \leq 5 \|\nabla_G \beta^*\|_1} \sum_{i=1}^n \frac{\xi_i \varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}}}{U_n} (\beta_i - \beta_i^*) \,\middle|\, \mathcal{E}_4 \right] \\ =: \;\; & T_1 + T_2. \end{aligned} \tag{60}$$

Next we bound T_1 and T_2. To bound T_2, we define

$$e_i \;:=\; \frac{\xi_i \varepsilon_i 1_{\{|\varepsilon_i| \leq U_n\}}}{U_n},$$

and notice that $\mathbb{E}(e_i|\mathcal{E}_4) = \mathbb{E}(e_i) = 0$, and e_i is sub-Gaussian(1) for $i = 1, \ldots, n$. It follows that if $\mathcal{E}_4$ holds, then

$$\begin{array}{lll} e^\top(\beta - \beta^*) & = & e^\top\{\beta - I(\beta)\} + e^\top\{I(\beta) - I(\beta^*)\} - e^\top\{\beta^* - I(\beta^*)\} \\ & \leq & 2\|\nabla_G \beta^*\|_1 + CK^{1/2}\left[\|\Pi\tilde{e}\|_2\|\beta - \beta^*\| + \|(\nabla_{G'})^+\tilde{e}\|_\infty(\|\nabla_G\beta^*\|_1 + \|\nabla_G\beta\|_1)\right] + \\ & & 2\|\nabla_G\beta\|_1. \end{array}$$

Therefore,

$$\begin{array}{lll} T_2 & \leq & \dfrac{16U_n}{\eta^2}\left\{12C_2\mathrm{poly}(\log n)n^{1-1/d} + CK^{1/2}\left[\eta\mathbb{E}(\|\Pi\tilde{e}\|) + 6C_2\mathrm{poly}(\log n)n^{1-1/d}E(\|(\nabla_{G'})^+\tilde{e}\|_\infty)\right]\right\} \\ & \leq & \dfrac{16U_n}{\eta^2}\left(12C_2\mathrm{poly}(\log n)n^{1-1/d} + CK^{1/2}\left[\eta\mathbb{E}\left|\dfrac{1}{n^{1/2}}\displaystyle\sum_{i=1}^n \tilde{e}_i\right| + 6C_2\mathrm{poly}(\log n)n^{1-1/d}\max_j\|(\nabla_G^+)_{,j}\|\right]\right) \\ & \leq & \dfrac{16U_n}{\eta^2}\left[12C_2\mathrm{poly}(\log n)n^{1-1/d} + CK^{1/2}\left[\eta + 6C_2\mathrm{poly}(\log n)n^{1-1/d}\max_j\|(\nabla_G^+)_{,j}\|\right]\right] \\ & \leq & \dfrac{16U_n}{\eta^2}\left[12C_2\mathrm{poly}(\log n)n^{1-1/d} + CK^{1/2}\left[\eta + 6C_2\mathrm{poly}(\log n)n^{1-1/d}\phi_n\right]\right] \end{array}$$

where the second and third inequalities follow from Sub-Gaussian maximal inequality, and the last from (53). Then for a given $a \in (0,1)$, we set

$$\begin{array}{lll} \eta & = & \frac{6^{1/2}}{a^{1/2}}(16\times 12C_2\mathrm{poly}(\log n)n^{1-1/d}U_n)^{1/2} + \frac{6}{a}(16CK^{1/2}U_n) + \\ & & \frac{6}{a}(16\times 6CC_2U_nK^{1/2}\mathrm{poly}(\log n)\, n^{1-1/d}\phi_n)^{1/2} \end{array}$$

and so

$$T_2 \leq \frac{a}{2}. \tag{61}$$

Furthermore,

$$\begin{array}{lll} T_1 & \leq & \dfrac{a}{6}\dfrac{16n^{1/2}\max\limits_{i=1,\ldots,n}\{\mathbb{E}(\varepsilon_i^4)\}^{1/4}\{\mathrm{pr}(|\varepsilon_i| > U_n)\}^{1/4}}{16CK^{1/2}U_n} \\ & \leq & \dfrac{a}{6}. \end{array} \tag{62}$$

The claim then follows. ∎

Appendix H. Auxiliary lemmas for proof of Theorem 9

Lemma 16 *Let $\gamma_i^* = \mathbb{E}(y_i^2)$ for $i = 1, \ldots, n$. Then*

$$\sum_{(i,j)\in E}|\gamma_i^* - \gamma_j^*| \leq \sum_{(i,j)\in E}|v_i^* - v_j^*| + 2\|\theta^*\|_\infty \sum_{(i,j)\in E}|\theta_i^* - \theta_j^*|.$$

Proof Notice that

$$\begin{aligned}\sum_{(i,j)\in E} |\gamma_i^* - \gamma_j^*| &= \sum_{(i,j)\in E} |\{v_i^* + (\theta_i^*)^2\} - \{v_j^* + (\theta_j^*)^2\}| \\ &\leq \sum_{(i,j)\in E} |v_i^* - v_j^*| + \sum_{(i,j)\in E} |\theta_i^* - \theta_j^*|(|\theta_i^*| + |\theta_j^*|) \\ &\leq \sum_{(i,j)\in E} |v_i^* - v_j^*| + 2\|\theta^*\|_\infty \sum_{(i,j)\in E} |\theta_i^* - \theta_j^*|\end{aligned}$$

and the claim follows. ■

Lemma 17 *For any $U_n > 0$ we have that*

$$\mathrm{pr}\left(|y_i^2 - E(y_i^2)| > 2\|v^*\|_\infty^{1/2}\|\theta^*\|_\infty U_n + \|v^*\|_\infty(1+U_n^2)\right) \leq \mathrm{pr}(|\epsilon_i| > U_n)$$

for $i = 1, \ldots, n$.

Proof Simply observe that

$$\begin{aligned}y_i^2 - \mathbb{E}(y_i^2) &= \{\theta_i^* + (v_i^*)^{1/2}\epsilon_i\}^2 - \mathbb{E}[\{\theta_i^* + (v_i^*)^{1/2}\epsilon_i\}^2] \\ &= 2(v_i^*)^{1/2}\theta_i^*\epsilon_i + v_i^*\epsilon_i^2 - v_i^*\end{aligned}$$

and hence

$$|y_i^2 - \mathbb{E}(y_i^2)| \leq 2\|v^*\|_\infty^{1/2}\|\theta^*\|_\infty|\epsilon_i| + \|v^*\|_\infty|\epsilon_i|^2 + \|v^*\|_\infty$$

and so the claim follows.

■

Appendix I. Proof of Theorem 9

Proof First notice that

$$\begin{aligned}\frac{1}{n}\|\hat{v} - v^*\|^2 &= \frac{1}{n}\sum_{i=1}^n \left[\{\hat{\gamma}_i - (\hat{\theta}_i)^2\} - \{\gamma_i^* - (\theta_i^*)^2\}\right]^2 \\ &\leq \frac{2}{n}\sum_{i=1}^n \left(\hat{\gamma}_i - \gamma_i^*\right)^2 + \frac{2}{n}\sum_{i=1}^n \left\{(\hat{\theta}_i)^2 - (\theta_i^*)^2\right\}^2 \\ &\leq \frac{2}{n}\sum_{i=1}^n \left(\hat{\gamma}_i - \gamma_i^*\right)^2 + \frac{8\|\theta^*\|_\infty^2}{n}\sum_{i=1}^n \left(\hat{\theta}_i - \theta_i^*\right)^2\end{aligned}$$

and so each conclusion of the theorem follows applying Theorem 4, Lemma 16, and Lemma 17. Specifically, it is clear that the generative model and θ^* satisfy the conditions of Theorem 4. As for the estimation of γ^*, letting $r_i = y_i^2 - \mathbb{E}(y_i^2)$, for $i = 1, \ldots, n$, we need to verify the tail conditions , e.g. (14), for $\{r_i\}_{i=1}^n$.

Proof of (24). Notice that by Lemma 17,

$$\frac{n^{1/4}\max\limits_{i=1,\ldots,n}\{\mathrm{pr}(|r_i|>U_n')\}^{1/4}}{U_n'\{\log(en)\}^{1/2}} \;\le\; \frac{n^{1/4}\max\limits_{i=1,\ldots,n}\{\mathrm{pr}(|r_i|>2\|v^*\|_\infty^{1/2}\|\theta^*\|_\infty U_n+\|v^*\|_\infty U_n^2+\|v^*\|_\infty)\}^{1/4}}{U_n\{\log(en)\}^{1/2}} \;\le\; \frac{n^{1/4}\max\limits_{i=1,\ldots,n}\{\mathrm{pr}(|\epsilon_i|>U_n)\}^{1/4}}{U_n\{\log(en)\}^{1/2}} \to 0.$$

Therefore, by Lemma 16 and Theorem 4,

$$\frac{1}{n}\sum_{i=1}^n\left(\hat{\theta}_i-\theta_i^*\right)^2 \,=\, O_{\mathrm{pr}}\left\{\frac{U_n^{4/3}(\log n)^{1/3}\|\nabla_G\theta^*\|_1^{2/3}}{n^{2/3}}+\frac{U_n^2\log n}{n}\right\},$$

and

$$\frac{1}{n}\sum_{i=1}^n\left(\hat{\gamma}_i-\gamma_i^*\right)^2 \,=\, O_{\mathrm{pr}}\left\{\frac{(U_n')^{4/3}(\log n)^{1/3}\left(\|\nabla_G v^*\|_1+\|\theta^*\|_\infty\|\nabla_G\theta^*\|_1\right)^{2/3}}{n^{2/3}}+\frac{(U_n')^2\log n}{n}\right\},$$

and so the claim (24) follows.

The proof of (26) and (27) follow similarly. ■

Appendix J. Lower bounds

J.1 Proof of Lemma 13

Proof

We notice that

$$\begin{aligned}
&\inf_{\tilde v\in\mathcal{F}}\;\sup_{\theta^*,v^*\in\Theta,\;v_i^*\in(\frac{c^2}{8},\frac{3c^2}{8}),\;y_i=\theta_i^*+\sqrt{v_i^*}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\|\tilde v(y)-v^*\|^2\right)\\
&\ge \inf_{\tilde v\in\mathcal{F}}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)-(c^2/2-(\theta_i^*)^2))^2\right)\\
&= \inf_{\tilde v\in\mathcal{F}}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)-(\theta_i^*)^2)^2\right)\\
&\ge \inf_{\tilde v\in\mathcal{F},\,\tilde v_i(\cdot)\in[\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}]}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)^2-(\theta_i^*)^2)^2\right)\\
&\ge \inf_{\tilde v\in\mathcal{F},\,\tilde v_i(\cdot)\in[\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}]}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\Bigg(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)-\theta_i^*)^2\cdot\\
&\qquad\qquad\qquad\qquad\min_{j=1,\ldots n}(\tilde v_j(y)+\theta_j^*)^2\Bigg)\\
&\ge \frac{c^2}{32}\inf_{\tilde v\in\mathcal{F},\,\tilde v_i(\cdot)\in[\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}]}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)-\theta_i^*))^2\right)\\
&\ge \frac{c^2}{32}\inf_{\tilde v\in\mathcal{F}}\;\sup_{\theta^*\in\Theta,\;\theta_i^*\in(\frac{c}{\sqrt8},\frac{\sqrt3 c}{\sqrt8}),\;y_i=\theta_i^*+\sqrt{\frac{c^2}{2}-(\theta_i^*)^2}\epsilon_i,\;\epsilon_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)}\mathbb{E}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\tilde v_i(y)-\theta_i^*))^2\right)
\end{aligned}\tag{63}$$

Next, let $d_{\max}$ be the maximum degree of any node in G and consider distinct $a_1, \ldots, a_m \in [n^{1/d}] \times \ldots \times [n^{1/d}]$, for $m \in \mathbb{N}$ with $m \asymp n^{1-1/d}$ and $d_{\max} \cdot m \leq n^{1-1/d}$, such that for all $j, j' \in \{1, \ldots, m\}$ it holds that a_j and $a_{j'}$ are not connected by an edge in the d-dimensional grid graph associated with $[n^{1/d}] \times \ldots \times [n^{1/d}]$. Then for $\eta \in \{-1,1\}^m$ let $\theta_\eta \in \mathbb{R}^n$ be given as

$$(\theta_\eta)_i \;=\; \begin{cases} \frac{c}{\sqrt{8}} \left[\frac{\eta_j(\sqrt{3}-1)}{4} + \frac{(1+\sqrt{3})}{2} \right] & \text{if} \quad i = a_{_j}, \quad j \in \{1, \ldots, m\}, \\ \frac{c}{\sqrt{8}} \cdot \frac{(1+\sqrt{3})}{2} & \text{otherwise.} \end{cases}$$

Notice that by construction $(\theta_\eta)_i \in (\frac{c}{\sqrt{8}}, \frac{\sqrt{3}c}{\sqrt{8}})$ for all i and $\eta \in \{-1,1\}^m$. Moreover,

$$\|\nabla_G \theta_\eta\|_1 \;\leq\; \frac{d_{\max} \cdot c}{\sqrt{8}} \cdot \frac{m(\sqrt{3}-1)}{4} \;\leq\; c n^{1-1/d}.$$

In addition, if $\eta, \eta' \in \{-1,1\}^m$ such that $\|\eta - \eta'\|_1 = 2$, then

$$\|\theta_\eta - \theta_{\eta'}\| \;=\; \frac{c}{\sqrt{8}} \cdot \frac{(\sqrt{3}-1)}{2}.$$

Also, denoting by P_η and $P_{\eta'}$ the distributions $N(\theta_\eta, \mathrm{diag}(\frac{c^2}{2} - (\theta_\eta)_i^2))$ and $N(\theta_{\eta'}, \mathrm{diag}(\frac{c^2}{2} - (\theta_{\eta'})_i^2))$, respectively, we obtain that

$$\begin{aligned} \mathrm{TV}(P_\eta, P_{\eta'}) \;\leq\; \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P_\eta, P_{\eta'})} \;\; &\leq\;\; \tfrac{1}{2} \Bigg[\big(\tfrac{4-[3/4+\sqrt{3}/4]^2}{4-[1/4+3\sqrt{3}/4]^2} - 1 \big) + \tfrac{8}{c^2} \cdot \|\theta_\eta - \theta_{\eta'}\|^2 + \\ &\qquad\quad \log\big(\tfrac{4-[3/4+\sqrt{3}/4]^2}{4-[1/4+3\sqrt{3}/4]^2} \big) \Bigg]^{1/2} \\ &<\;\; \tfrac{1}{2} \sqrt{ \Big[0.63 + \big(\tfrac{(\sqrt{3}-1)}{2} \big)^2 \; + \; \log(1.63) \Big] } \\ &<\;\; 0.6 \end{aligned}$$

where the first inequality follows from Pinsker's inequality, and the second by the usual formula for KL distance between multivariate normal distributions. Therefore, by Assouad's lemma, Lemma 2 in Yu (1997), we obtain that

$$\begin{aligned} &\inf_{\tilde{v} \in \mathcal{F}} \sup_{\theta^* \in \Theta,\; \theta_i^* \in (\frac{c}{\sqrt{8}}, \frac{\sqrt{3}c}{\sqrt{8}}),\;\; y_i = \theta_i^* + \epsilon_i,\; \epsilon_i \overset{\text{ind}}{\sim} \mathrm{N}(0, \frac{c^2}{2} - (\theta_i^*)^2)} \mathbb{E}\left(\textstyle\sum_{i=1}^n (\tilde{v}_i(y) - \theta_i^*)^2 \right) \\ &\gtrsim \tfrac{m}{2} \cdot (1 - 0.6) \\ &\gtrsim n^{1-1/d}. \end{aligned} \tag{64}$$

Hence, from (63) and (64), we arrive at

$$\inf_{\tilde{v} \in \mathcal{F}} \sup_{\theta^*, v^* \in \Theta,\; v_i^* \in (\frac{c^2}{8}, \frac{3c^2}{8}),\;\; y_i = \theta_i^* + \sqrt{v_i^*}\epsilon_i,\; \epsilon_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)} \mathbb{E}\left(\frac{1}{n} \|\tilde{v}(y) - v^*\|^2 \right) \;\gtrsim\; \frac{1}{n^{1/d}}.$$

$\blacksquare$

J.2 Proof of Lemma 14

Proof Let $\mathcal{I}$ be a rectangular partition of $[0,1]^d$ such that $|\mathcal{I}| = n$ and each rectangle $I \in \mathcal{I}$ has sides of lenght in $[c_1 n^{-1/d}, c_2 n^{-1/d}]$ for some positive constants $c_1, c_2 > 0$. Notice that

$$\frac{c_1^d}{n} \leq \mathrm{vol}(I) \leq \frac{c_2^d}{n} \tag{65}$$

for all $I \in \mathcal{I}$. Also, for each rectangle $I \in \mathcal{I}$ let I° and ∂I_l be its interior and boundary, respectively. Let also be $\{I_l\}_{l=1}^m \subset \mathcal{I}$ such that

$$I_l^{\circ} \cap I_{l'}^{\circ} = \emptyset$$

for $l \neq l'$ and such that $m \asymp n^{1-1/d}$. Moreover, the rectangles $\{I_l\}_{l=1}^m$ can be chosen such that their union is a rectangle.

Then for fixed $a > b > 0$, define for every $\eta \in \{-1,1\}^m$ the function $g_\eta : [0,1]^d \to \mathbb{R}$,

$$g_\eta(x) = \begin{cases} a + \eta_j b & \text{if } x \in I_j^{\circ} \\ a & \text{otherwise.} \end{cases}$$

We claim that $g_\eta \in \mathcal{F}(L_0)$. Towards that end notice that g is piecewise constant with boundary $S = \cup_{l=1}^m \partial I_l$. Moreover, for $t \in (0,1)$, since $\cup_{l=1}^m I_l$ is rectangle, we obtain that

$$\mathrm{vol}(B_t(S)) \leq \sum_{l=1}^m c_3 [d \left(\frac{1}{n^{1/d}}\right)^{d-1} t] + c_3 d t^d \leq c_4 t$$

for some positive constants $c_3, c_4 > 0$ that can depend on d. Hence, for a and b small enough, it holds that $g_\eta \in \mathcal{F}(L_0)$ for all $\eta \in \{-1,1\}^m$. Moreover, for any η, η' such that $\|\eta - \eta'\|_1 = 2$, we have that

$$\|g_\eta - g_\eta'\|_2^2 = \int_{I_j} 4b^2 dx = 4b^2 \mathrm{vol}(I_j) \geq \frac{c_1^d 4b^2}{n}$$

for $j \in \{1,\ldots,m\}$ such that $\eta_j \neq \eta_j'$. Moreover, let P_η be the distribution associated with $(x,y) = (x_1,\ldots,x_n,y_1,\ldots,y_n)$ and p_η its corresponding density when $f_0 = 0$ and $g_0 = g_\eta$. The objects $P_{\eta'}$ and $p_{\eta'}$ are defined accordingly. Then, since $x_i \overset{ind}{\sim} \mathrm{Uniform}[0,1]^d$,

$$\begin{aligned}
\mathrm{TV}(P_\eta, P_{\eta'}) &= \frac{1}{2} \int_{[0,1]^{nd} \times \mathbb{R}^n} |p_\eta(x,y) - p_{\eta'}(x,y)| d(x \times y) \\
&= \frac{1}{2} \int_{[0,1]^d} \cdots \int_{[0,1]^d} \left[\int_{\mathbb{R}^n} |p_\eta(y|x) - p_{\eta'}(y|x)| dy \right] dx_1 \ldots dx_n \\
&= \int_{[0,1]^d} \cdots \int_{[0,1]^d} \left[\mathrm{TV}(P_\eta(\cdot|x), P_{\eta'}(\cdot|x)) \right] dx_1 \ldots dx_n \\
&\leq \frac{3}{2} \int_{[0,1]^d} \cdots \int_{[0,1]^d} \left[\frac{\left(\sum_{i=1}^n (g_\eta(x_i) - g_{\eta'}(x_i))^2\right)^{1/2}}{\inf_{q \in [0,1]^d} g_\eta(q)} \right] dx_1 \ldots dx_n \\
&= \frac{3}{2(a-b)} \int_{[0,1]^d} \cdots \int_{[0,1]^d} \left[\left(\sum_{i=1}^n (g_\eta(x_i) - g_{\eta'}(x_i))^2 \right)^{1/2} \right] dx_1 \ldots dx_n
\end{aligned}$$

where the inequality follows from Theorem 1.1 in Devroye et al. (2018). Hence,

$$\begin{aligned}
\mathrm{TV}(P_\eta, P_{\eta'}) &\leq \frac{6b}{2(a-b)} \int_{[0,1]^d} \cdots \int_{[0,1]^d} |\{i \,:\, x_i \in I_j\}| \, dx_1 \ldots dx_n \\
&= \frac{6b}{2(a-b)} \mathbb{E}\left(|\{i \,:\, x_i \in I_j\}|\right) \\
&= \frac{6b}{2(a-b)} n \mathrm{vol}(I_j) \\
&\leq \frac{6bc_2^d}{2(a-b)} \\
&\leq \tfrac{1}{2}
\end{aligned}$$

where the first inequality follows from (65), and where the last inequality follows by choosing $b = a/2$ and c_2 small enough in the construction of $\mathcal{I}$.

Therefore, by Assouad's lemma, see Lemma 2 in Yu (1997), we obtain that

$$\begin{aligned}
\inf_{\tilde{g}} \sup_{f_0, g_0 \in \mathcal{F}(L_0),\; y_i = f_0(x_i) + \sqrt{g_0(x_i)}\epsilon_i,\; \epsilon_i \overset{\mathrm{ind}}{\sim} \mathrm{N}(0,1)} \mathbb{E}\left(\|\tilde{g} - g_0\|_2^2\right) &\gtrsim m\left(\tfrac{4b^2}{n}\right) \cdot \left(1 - \tfrac{1}{2}\right) \\
&\gtrsim n^{-1/d}.
\end{aligned} \tag{66}$$

■

References

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

Gilles Blanchard, Christin Schäfer, Yves Rozenholc, and K-R Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007.

T Tony Cai and Lie Wang. Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics*, 36(5):2025–2054, 2008.

T Tony Cai, Michael Levine, and Lie Wang. Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, 100(1): 126–136, 2009.

Lorenzo Cappello, Oscar Hernan Madrid Padilla, and Julia A Palacios. Scalable bayesian change point detection with spike and slab priors. *arXiv preprint arXiv:2106.10383*, 2021.

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.

Sabyasachi Chatterjee and Subhajit Goswami. Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. *The Annals of Statistics*, 49(5):2531–2551, 2021a.

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2d total variation denoising. *IEEE Transactions on Information Theory*, 67(6):4060–4091, 2021b.

Ronald R Coifman and Mauro Maggioni. Diffusion wavelets. *Applied and computational harmonic analysis*, 21(1):53–94, 2006.

Mark Crovella and Eric Kolaczyk. Graph wavelets for spatial traffic analysis. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 3, pages 1848–1857. IEEE, 2003.

Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. 2017.

Aramayis Dallakyan and Mohsen Pourahmadi. Fused-lasso regularized cholesky factors of large nonstationary covariance matrices of replicated time series. *Journal of Computational and Graphical Statistics*, (just-accepted):1–27, 2022.

Holger Dette, Axel Munk, and Thorsten Wagner. Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):751–764, 1998.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

David L Donoho. Cart and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5): 1870–1911, 1997.

Jianqing Fan and Qiwei Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 1998.

Zhou Fan and Leying Guan. Approximate l0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217–3245, 2018.

Matan Gavish, Boaz Nadler, and Ronald R Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *ICML*, 2010.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. 2020.

Peter Hall and Raymond J Carroll. Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):3–14, 1989.

David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146. PMLR, 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.

Laura Jula Vanegas, Merle Behr, and Axel Munk. Multiscale quantile segmentation. *Journal of the American Statistical Association*, pages 1–14, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

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 6884–6893, 2017.

Oscar Hernan Madrid-Padilla and Sabyasachi Chatterjee. Risk bounds for quantile trend filtering. *Biometrika*, 2020.

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

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

Oscar Hernan Madrid-Padilla, Yanzhen Chen, and Gabriel Ruiz. A causal fused lasso for interpretable heterogeneous treatment effects estimation. *arXiv preprint arXiv:2110.00901*, 2021a.

Oscar Hernan Madrid-Padilla, Yi Yu, and Alessandro Rinaldo. Lattice partition recovery with dyadic cart. *Advances in Neural Information Processing Systems*, 34:26143–26155, 2021b.

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

Francesco Ortelli and Sara van de Geer. Adaptive rates for total variation image denoising. *Journal of Machine Learning Research*, 21:247, 2020.

Francesco Ortelli and Sara van de Geer. Prediction bounds for higher order total variation regularized least squares. *The Annals of Statistics*, 49(5):2755–2773, 2021.

Oscar Hernan Madrid Padilla. Graphon estimation via nearest neighbor algorithm and 2d fused lasso denoising. *arXiv preprint arXiv:1805.07042*, 2018.

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

John Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230, 1984.

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. *The Annals of Statistics*, 47(6):3032–3068, 2019.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. *Advances in Neural Information Processing Systems*, 29, 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. *Advances in Neural Information Processing Systems*, 30, 2017.

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

James Sharpnack, Aarti Singh, and Akshay Krishnamurthy. Detecting activations over graphs using spanning tree wavelet bases. In *Artificial intelligence and statistics*, pages 536–544. PMLR, 2013.

Yandi Shen, Chao Gao, Daniela Witten, and Fang Han. Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618, 2020.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 144–158. Springer, 2003.

Wesley Tansey and James Scott. A fast and flexible algorithm for the graph-fused lasso. *arXiv preprint arXiv:1505.06475*, 2015.

Wesley Tansey, Oluwasanmi Koyejo, Russell A Poldrack, and James G Scott. False discovery rate smoothing. *To appear in Journal of the American Statistical Association*, 2017.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

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. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Tiejun Tong and Yuedong Wang. Estimating residual variance in nonparametric regression using least squares. *Biometrika*, 92(4):821–830, 2005.

Huy Tran, Sansen Wei, and Claire Donnat. The generalized elastic net for least squares regression with network-aligned signal and correlated design. *arXiv preprint arXiv:2211.00292*, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Lie Wang, Lawrence D Brown, T Tony Cai, and Michael Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008.

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.

Steven Siwei Ye and Oscar Hernan Madrid Padilla. Non-parametric quantile regression via the k-nn fused lasso. *Journal of Machine Learning Research*, 22:111–1, 2021.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

Yi Yu, Oscar Madrid, and Alessandro Rinaldo. Optimal partition recovery in general graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 4339–4358. PMLR, 2022.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22nd international conference on Machine learning*, pages 1036–1043, 2005.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.