The Annals of Statistics
2018, Vol. 46, No. 6B, 3217–3245
https://doi.org/10.1214/17-AOS1656

APPROXIMATE ℓ_0-PENALIZED ESTIMATION OF PIECEWISE-CONSTANT SIGNALS ON GRAPHS

BY ZHOU FAN[1] AND LEYING GUAN

Stanford University

We study recovery of piecewise-constant signals on graphs by the estimator minimizing an l_0-edge-penalized objective. Although exact minimization of this objective may be computationally intractable, we show that the same statistical risk guarantees are achieved by the α-expansion algorithm which computes an approximate minimizer in polynomial time. We establish that for graphs with small average vertex degree, these guarantees are minimax rate-optimal over classes of edge-sparse signals. For spatially inhomogeneous graphs, we propose minimization of an edge-weighted objective where each edge is weighted by its effective resistance or another measure of its contribution to the graph's connectivity. We establish minimax optimality of the resulting estimators over corresponding edge-weighted sparsity classes. We show theoretically that these risk guarantees are not always achieved by the estimator minimizing the l_1/total-variation relaxation, and empirically that the l_0-based estimates are more accurate in high signal-to-noise settings.

1. Introduction. Let $G = (V, E)$ be a known (undirected) graph, with vertices $V := \{1, \dots, n\}$ and edge set E. At each vertex $i \in \{1, \dots, n\}$, an unknown signal value $\mu_{0,i}$ is observed with noise:

$$Y_i = \mu_{0,i} + \varepsilon_i.$$

For simplicity, we assume $\varepsilon_1, \dots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2)$ and G is fully connected. This paper studies the problem of estimating the true signal vector $\mu_0 := (\mu_{0,1}, \dots, \mu_{0,n})$ from observed data $Y := (Y_1, \dots, Y_n)$, when μ_0 is (or is well approximated by) a piecewise-constant signal over G. Informally, this will mean that the set of edges $\{i, j\} \in E$ where $\mu_{0,i} \neq \mu_{0,j}$ is a small subset of all edges.

Examples of this problem occur in a number of applications:

• *Multiple changepoint detection.* The graph G is a linear chain with n vertices and $n - 1$ edges, which identifies a sequential order to the observations. The signal μ_0 is piecewise constant in the sense $\mu_{0,i} \neq \mu_{0,i+1}$ for a small number of changepoints i.

Received April 2017; revised September 2017.

[1]Supported by a Hertz Foundation Fellowship and an NDSEG Fellowship (DoD, Air Force Office of Scientific Research, 32 CFR 168a).

MSC2010 subject classifications. 62G05.

Key words and phrases. Approximation algorithm, graph cut, effective resistance, total-variation denoising.

- *Image segmentation.* The graph G is a 2-D (or 3-D) lattice graph, and μ_0 corresponds to the pixels (or voxels) of a digital image. The assumption of piecewise-constancy implies that μ_0 has regions of approximately constant pixel value.
- *Anomaly identification in networks.* The graph G represents a network. The signal μ_0 indicates locations of anomalous clusters of vertices, for example, representing individuals infected by a disease or a computer virus. Piecewise-constancy of μ_0 reflects the assumption that the anomaly spreads along the network connections.

Early and pioneering works include [8, 18, 69] on multiple changepoint detection and [9, 26] on image segmentation. For general graphs and networks, [1–3, 5, 56, 57] studied related hypothesis testing problems, and [34, 47, 66] also recently considered estimation. We discuss some connections of our work to this existing literature in Section 1.1.

The focus of our paper is the method of "l_0-edge-denoising," which seeks to estimate μ_0 by the values $\mu \in \mathbb{R}^n$ minimizing the residual squared error $\frac{1}{2}\|Y-\mu\|^2$ plus a penalty for each edge $\{i, j\} \in E$ where $\mu_i \neq \mu_j$. (Here and throughout, $\|\cdot\|$ without a subscript denotes the standard Euclidean norm.) More formally, this estimate minimizes the objective function

$$\text{(L0)} \qquad F_0(\mu) := \frac{1}{2}\|Y-\mu\|^2 + \lambda\|D\mu\|_0, \qquad \|D\mu\|_0 := \sum_{\{i,j\}\in E} \mathbb{1}\{\mu_i \neq \mu_j\}.$$

Here, $D : \mathbb{R}^n \to \mathbb{R}^E$ denotes a vertex-edge incidence matrix with entries in $\{-1, 0, 1\}$ that maps $\mu \in \mathbb{R}^n$ to the vector of edge differences $(\mu_i - \mu_j)_{\{i,j\}\in E}$ (with an arbitrary sign for each edge). The penalty term $\|D\mu\|_0$ denotes the usual "l_0-norm" of $D\mu$, and λ is a user-specified tuning parameter that controls the magnitude of this penalty.

For reasons to be discussed, we will also consider procedures that seek to minimize a more general weighted version of the above objective function,

$$\text{(W)} \quad F_{\mathrm{w}}(\mu) := \frac{1}{2}\|Y-\mu\|^2 + \lambda\|D\mu\|_{\mathrm{w}}, \qquad \|D\mu\|_{\mathrm{w}} := \sum_{\{i,j\}\in E} \mathrm{w}(i,j)\mathbb{1}\{\mu_i \neq \mu_j\},$$

where $\mathrm{w} : E \to \mathbb{R}_+$ assigns a nonnegative weight to each edge. This allows possibly different penalty values to be applied to different edges of the graph.

The combinatorial nature of (L0) and (W) render exact minimization of these objectives computationally intractable for general graphs. A primary purpose of this paper is to show, however, that approximate minimization is sufficient to obtain statistically rate-optimal guarantees. We study one such approximation algorithm by Boykov, Veksler and Zabih [13], suggest its use in minimizing (W) for applications involving inhomogeneous networks, and provide a unified analysis of minimax squared-error risk for this estimation problem over edge-sparse signal classes on general graphs.

We summarize our results as follows:

1. A polynomial-time algorithm using the α-expansion procedure of [13] yields approximate minimizers $\hat{\mu}$ of (L0) and (W) that achieve the same statistical risk guarantees as the exact minimizers, up to constant factors. Computation of $\hat{\mu}$ is reasonably efficient in practice and yields good empirical signal recovery in our tested examples. In this sense, inference based on minimizing (L0) or (W) is computationally tractable, even for large graphs.

2. For any graph G, the estimate $\hat{\mu}$ (exactly or approximately) minimizing (L0) with $\lambda \asymp \sigma^2 \log|E|$ satisfies an "edge-sparsity" oracle inequality

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \lesssim \inf_{\mu \in \mathbb{R}^n} \|\mu - \mu_0\|^2 + \sigma^2 \max\big(\|D\mu\|_0, 1\big) \log|E|. \tag{1}$$

This bounds the squared-error risk of $\hat{\mu}$ in terms of the approximability of μ_0 by any piecewise-constant signal μ. If it is known that $\|D\mu_0\|_0 \le s$, then setting instead $\lambda \asymp \sigma^2(1 + \log \frac{|E|}{s})$ yields

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \lesssim \sigma^2 s\left(1 + \log \frac{|E|}{s}\right). \tag{2}$$

The risk bound (2) is rate-optimal in a minimax sense over the edge-sparse signal class $\{\mu_0 : \|D\mu_0\|_0 \le s\}$ up to a multiplicative factor depending on the mean vertex degree of G.

3. An alternative to minimizing (L0) is to minimize its l_1/total-variation relaxation,

$$F_1(\mu) := \frac{1}{2}\|Y - \mu\|^2 + \lambda \|D\mu\|_1, \qquad \|D\mu\|_1 := \sum_{\{i,j\} \in E} |\mu_i - \mu_j|. \tag{TV}$$

One advantage of this approach is that (TV) is convex and can be exactly minimized in polynomial time. However, whether the risk guarantees (1) and (2) hold for $\hat{\mu}$ minimizing (TV) depends on properties of the graph. In particular, they do not hold for the linear chain graph, where instead

$$\inf_{\lambda \ge 0} \sup_{\mu_0 : \|D\mu_0\|_0 \le s} \mathbb{E}\big[\|\hat{\mu}^\lambda - \mu_0\|^2\big] \gtrsim \sigma^2 (\log n)^{-5} \sqrt{sn}, \tag{3}$$

$\hat{\mu}^\lambda$ denoting the minimizer of (TV) for each λ. This result is connected to the "slow rate" of convergence in prediction risk for the Lasso [17, 62] in certain linear regression settings with correlated predictors.

4. When G has regions of differing connectivity, $\|D\mu\|_0$ is not a spatially homogeneous measure of complexity, and it may be more appropriate to minimize the edge-weighted objective (W) where $\mathrm{w}(i, j)$ measures the contribution of edge $\{i, j\}$ to the connectivity of the graph. One such weighting, inspired by the analyses in [56, 57], weighs each edge by its effective resistance when G is viewed as an electrical resistor network. In simulations on real networks, this weighting can yield a substantial reduction in error over minimizers of the unweighted objective (L0). For general weightings $\mathrm{w} : E \to \mathbb{R}_+$ belonging to the spanning tree polytope

of G, the guarantee (2) holds over the larger class $\{\mu_0 : \|D\mu_0\|_w \le s\}$ for $\hat{\mu}$ minimizing (W), and this guarantee is minimax rate-optimal up to a graph-independent constant factor, for all graphs.

We provide a more detailed discussion of these results in Sections 2 to 5. Simulations comparing minimization of (L0), (W) and (TV) over several graphs are presented in Section 6. Proofs are deferred to the Appendices in the Supplementary Material [25].

1.1. *Related work.* For changepoint problems where G is the linear chain, (L0) may be exactly minimized by dynamic programming in quadratic time [6, 35, 67]. Pruning ideas may reduce runtime to be near-linear in practice [38]. Correct changepoint recovery and distributional properties of $\hat{\mu}$ minimizing (L0) were studied asymptotically in [70, 71] when the number of true changepoints is fixed. Nonasymptotic risk bounds similar to (1) and (2) were established for estimators minimizing similar objectives in [11, 43]; we discuss this further below. Extension to the recovery of piecewise-constant functions over a continuous interval was considered in [14].

In image applications where G is the 2-D lattice, (L0) is closely related to the Mumford–Shah functional [50] and Ising/Potts-model energies for discrete Markov random fields [26]. In the latter discrete setting, where each μ_i is allowed to take value in a finite set of "labels," a variety of algorithms seek to minimize (L0) using minimum s-t cuts on augmented graphs; see [39] and the contained references for a review. For an Ising model with only two distinct labels, [29] showed that the exact minimizer may be computed via a single minimum s-t cut. For more than two distinct labels, exact minimization of (L0) is NP-hard [13]. We analyze a graph-cut algorithm from [13] that applies to more than two labels, where the exact minimization property is replaced by an approximation guarantee. We show that the deterministic guarantee of this algorithm implies rate-optimal statistical risk bounds, for the 2-D lattice as well as for general graphs.

For an arbitrary graph G, the estimators $\hat{\mu}$ exactly minimizing (L0) and (W) are examples of general model-complexity penalized estimators studied in [7, 11]. The penalties we impose may be smaller than those needed for the analyses of [7, 11] by logarithmic factors, and we instead control the supremum of a certain Gaussian process using an argument specialized to our graph-based problem. A theoretical focus of [7, 11] was on adaptive attainment of minimax rates over families of models. For example, for the linear chain graph, [11, 43] considered penalties increasing but concave in the number of changepoints, and the resulting estimates achieve the guarantee (2) simultaneously for all s. Instead of using such a penalty, which poses additional computational challenges, we will apply a data-driven procedure to choose λ, although we will not study the adaptivity properties of the procedure in this paper.

The method of l_0-edge-denoising and the characterization of signal complexity by $\|D\mu_0\|_0$ are "nonparametric" in the sense of [2, 4]. This is in contrast to

methods that employ additional prior knowledge about μ_0, for instance, that its constant regions belong to parametric classes of shapes [4], are thick and blob-like in nature [2] or have sufficiently smooth boundaries when G is embedded in a Euclidean space [22, 40]. In this regard, our study is more closely related to the hypothesis testing work of [5, 56, 57] in similar nonparametric contexts. An advantage of this perspective is that the inference algorithm is broadly applicable to general graphs and networks, where appropriate notions of boundary smoothness or support constraints for μ_0 are less naturally defined. A disadvantage is that such an approach may not be statistically optimal in more specialized settings when such prior assumptions hold true.

A connection between this problem, effective edge resistances, and graph spanning trees emerged in the analyses of [56, 57]. In [56], a procedure was proposed to construct an orthonormal wavelet basis over a spanning tree of G and to perform inference by thresholding in this basis. Our proposal to minimize (W) for $\mathrm{w} : E \to \mathbb{R}_+$ in the spanning tree polytope of G may be viewed as a derandomization of this idea when the spanning tree is chosen at random; we discuss this connection in Remark 5.7. Sampling edges by effective resistances is also a popular method of graph sparsification [58, 59], and effective-resistance edge weighting may be viewed as a derandomization of procedures such as in [53] that operate on a randomly sparsified graph.

There is a large body of literature on the l_1-relaxation (TV). This method and generalizations were suggested in different contexts and guises for the linear chain graph in [17, 21, 42, 48, 63] and also studied theoretically in [19, 30, 31, 44, 51]. For 2-D lattice graphs in image denoising, variants of (TV) were proposed and studied in [16, 28, 52]. For more general graphs, this method and generalizations have been studied in [33, 34, 41, 54, 55, 60, 64, 66], among others. In particular, [15, 20, 68] developed algorithms for minimizing (TV) and related objectives also using iterated graph cuts, although these algorithms yield exact solutions and are different from the algorithm we study. A body of theoretical work establishes that $\hat{\mu}$ minimizing (TV) is (or is nearly) minimax rate-optimal over signal classes of bounded variation, $\{\mu_0 : \|D\mu_0\|_1 \le s\}$, for the linear chain graph and higher-dimensional lattices [34, 48, 54, 66]. Several risk bounds over the exact-sparsity classes $\{\mu_0 : \|D\mu_0\|_0 \le s\}$ that we consider were also established for the linear chain graph in [19, 30, 44] and for general graphs in [34, 47]; we discuss some of these results in Section 4. We believe that benefits of using effective resistance weighted edge penalties may also apply to the l_1/TV setting, and we leave further exploration of this to future work.

1.2. *Notation and conventions.* We assume throughout that G is fully connected with $n \ge 3$ vertices. Theoretical results are nonasymptotic, in the sense that they are valid for all finite n and s with universal constants $C, c > 0$ independent of n, s, and the graph G. For positive a and b, we write informally $a \lesssim b$ if $a \le Cb$ and $a \asymp b$ if $ca \le b \le Ca$ for universal constants $C, c > 0$ and all $n \ge 3$.

For a vector v, $\|v\| := (\sum_i v_i^2)^{1/2}$ is the Euclidean norm, $\|v\|_0 := |\{i : v_i \neq 0\}|$ the "l_0-norm," $\|v\|_1 := \sum_i |v_i|$ the l_1-norm and $\|v\|_\infty := \max_i |v_i|$ the l_∞-norm. For vectors v and w, $\langle v, w\rangle = \sum_i v_i w_i$ is the Euclidean inner-product.

$\mathbb{R}_+$ denotes the nonnegative reals. For an edge weighting $\mathrm{w} : E \to \mathbb{R}_+$, $\mathrm{w}(i, j)$ is shorthand for $\mathrm{w}(\{i, j\})$, and we denote $\mathrm{w}(E') := \sum_{\{i,j\}\in E'} \mathrm{w}(i, j)$ for any edge subset $E' \subseteq E$. For two edge weightings $\mathrm{w}, \mathrm{r} : E \to \mathbb{R}_+$, we write $\mathrm{w} \geq \mathrm{r}$ if $\mathrm{w}(i, j) \geq \mathrm{r}(i, j)$ for all edges $\{i, j\} \in E$. For $v \in \mathbb{R}^E$, $\|v\|_\mathrm{w} := \sum_{\{i,j\}\in E} \mathrm{w}(i, j)v_{\{i,j\}}$ denotes the l_0-norm weighted by w.

$\mathbb{1}\{\cdot\}$ denotes the indicator function, that is, $\mathbb{1}\{\mathcal{E}\} = 1$ if condition $\mathcal{E}$ is true and 0 otherwise.

2. Approximation algorithm. As discussed in Section 1.1, whether (L0) and (W) may be minimized exactly in polynomial time depends on the graph G. However, good approximation of the solution is tractable for any graph. We review in this section one approach that achieves such an approximation, based on discretizing the range of values of the entries of μ and applying the α-expansion local move of [13] for discrete Markov random fields. We describe the algorithm for (W), as (L0) is a special case.

The fundamental property of this algorithm will be that its output is a $(\tau, \delta\mathbb{Z})$-local-minimizer for the objective function (W), defined as follows.

DEFINITION 2.1. For $\delta > 0$, denote by

$$\delta\mathbb{Z} := \{\ldots, -3\delta, -2\delta, -\delta, 0, \delta, 2\delta, 3\delta, \ldots\}$$

the set of all integer multiples of δ. For any $\mu \in \mathbb{R}^n$, a *$\delta\mathbb{Z}$-expansion* of μ is any other vector $\tilde{\mu} \in \mathbb{R}^n$ such that there exists a single value $c \in \delta\mathbb{Z}$ for which, for every $i \in \{1, \ldots, n\}$, either $\tilde{\mu}_i = \mu_i$ or $\tilde{\mu}_i = c$. For $\delta > 0$ and $\tau \geq 0$, a $(\tau, \delta\mathbb{Z})$-*local-minimizer* of (W) is any $\mu \in \mathbb{R}^n$ such that for every $\delta\mathbb{Z}$-expansion $\tilde{\mu}$ of μ,

$$F_\mathrm{w}(\mu) - \tau \leq F_\mathrm{w}(\tilde{\mu}).$$

More informally, a $\delta\mathbb{Z}$-expansion of μ can replace any subset of vertex values by a single new value $c \in \delta\mathbb{Z}$, and a $(\tau, \delta\mathbb{Z})$-local-minimizer is such that no further $\delta\mathbb{Z}$-expansion reduces the objective value by more than τ. This definition does not require $(\tau, \delta\mathbb{Z})$-local-minimizers to have all entries belonging to $\delta\mathbb{Z}$; hence, in particular, a global minimizer of (W) is also a $(\tau, \delta\mathbb{Z})$-local-minimizer for any $\delta > 0$ and $\tau \geq 0$. We define analogously $(\tau, \delta\mathbb{Z})$-local-minimizers for (L0).

The α-expansion procedure of [13] may be used to compute a $(\tau, \delta\mathbb{Z})$-local-minimizer efficiently with graph cuts. We review this procedure and how we apply it to our problem in Algorithm 1. We will use a small discretization δ so as to yield a good solution to the original continuous problem.

The following propositions verify that this algorithm returns a $(\tau, \delta\mathbb{Z})$-local-minimizer in polynomial time; proofs are contained in Appendix S1.

Algorithm 1 Algorithm to compute a $(\tau, \delta\mathbb{Z})$-local-minimizer of (W)

Let $\bar{Y}$, $Y_{\min}$, and $Y_{\max}$ be the mean, minimum, and maximum of Y, rounded to $\delta\mathbb{Z}$.
Initialize $\hat{\mu} \in \mathbb{R}^n$ by setting $\hat{\mu}_i = \bar{Y}$ for all $i \in \{1, \ldots, n\}$.
loop
  for each $c \in \delta\mathbb{Z} \cap [Y_{\min}, Y_{\max}]$ **do**
    Compute the best $\delta\mathbb{Z}$-expansion $\tilde{\mu}$ of $\hat{\mu}$ with new value c using Algorithm 2.
    If $F_{\mathrm{w}}(\tilde{\mu}) \le F_{\mathrm{w}}(\hat{\mu}) - \tau$, then set $\hat{\mu} = \tilde{\mu}$.
  end for
  If $\hat{\mu}$ was unchanged, then return $\hat{\mu}$.
end loop

PROPOSITION 2.2. *Algorithm 1, using Edmonds–Karp or Dinic's algorithm for solving minimum s-t cut, has worst-case runtime* $O(n|E|^3 \times (Y_{\max} - Y_{\min})^3/(\delta\tau))$.

PROPOSITION 2.3. *Among all $\delta\mathbb{Z}$-expansions of $\hat{\mu}$ with new value c, the vector $\tilde{\mu}$ returned by Algorithm 2 has lowest objective value in* (W). *The estimate $\hat{\mu}$ returned by Algorithm 1 is a $(\tau, \delta\mathbb{Z})$-local-minimizer of* (W).

In particular, if $Y_{\max} - Y_{\min}$, $1/\delta$, and $1/\tau$ are polynomial in n, then Algorithm 1 is polynomial-time in n. We will use the Boykov–Kolmogorov algorithm [12] instead of Edmonds–Karp or Dinic to solve minimum s-t cut. This has slower worst-case runtime but is much faster in practice on our tested examples. We have found Algorithm 1 to be fast in practice, even with $\tau = 0$, and we discuss empirical runtime in Section 6.2.

The l_2 vertex-cost $(Y_i - \mu_i)^2$ and l_0 edge cost $\mathbb{1}\{\mu_i \neq \mu_j\}$ of (L0) and (W) are not intrinsic to this algorithm, and the same method may be applied to approximately minimize

$$F(\mu) = \sum_{i=1}^{n} c_i(Y_i, \mu_i) + \sum_{\{i,j\} \in E} c_{i,j}(\mu_i, \mu_j)$$

for any vertex cost functions c_i and edge cost functions $c_{i,j}$ such that each $c_{i,j}$ satisfies a triangle inequality. Thus the algorithm is easily applicable to other likelihood models and forms of edge penalties.

3. Theoretical guarantees for l_0 denoising. In this section, we describe squared-error risk guarantees for $\hat{\mu}$ (exactly or approximately) minimizing (L0). Although these results are corollaries of those in Section 5 for the weighted objective (W), we state them here separately as they are simpler to understand and also

Algorithm 2 α-expansion subroutine [13]

Construct the following edge-weighted augmentation $G_{c,\hat{\mu}}$ of G:
  Introduce a source vertex s and a sink vertex t.
  Connect s to each $i \in \{1, \dots, n\}$ with weight $\frac{1}{2}(Y_i - c)^2$.
  Connect t to each $i \in \{1, \dots, n\}$ with weight $\frac{1}{2}(Y_i - \hat{\mu}_i)^2$ if $\hat{\mu}_i \neq c$, or weight ∞ if $\hat{\mu}_i = c$.
  for each edge $\{i, j\} \in E$ **do**
    if $\hat{\mu}_i = \hat{\mu}_j$ **then**
      Assign weight $\lambda \mathrm{w}(i, j) \mathbb{1}\{\hat{\mu}_i \neq c\}$ to $\{i, j\}$.
    else
      Introduce a new vertex $\mathsf{a}_{i,j}$.
      Replace edge $\{i, j\}$ by the three edges $\{i, \mathsf{a}_{i,j}\}$, $\{j, \mathsf{a}_{i,j}\}$, and $\{\mathsf{t}, \mathsf{a}_{i,j}\}$, with weights $\lambda \mathrm{w}(i, j) \mathbb{1}\{\hat{\mu}_i \neq c\}$, $\lambda \mathrm{w}(i, j) \mathbb{1}\{\hat{\mu}_j \neq c\}$, and $\lambda \mathrm{w}(i, j)$, respectively.
    end if
  end for
Find the minimum s-t cut (S, T) of $G_{c,\hat{\mu}}$ such that $\mathsf{s} \in S, \mathsf{t} \in T$.
For each vertex $i \in \{1, \dots, n\}$, set $\tilde{\mu}_i = c$ if i belongs to T and $\tilde{\mu}_i = \hat{\mu}_i$ otherwise.
Return $\tilde{\mu}$.

provide a basis for comparison with total-variation denoising discussed in the next section. We defer discussion of the proofs to Section 5.

Recall Definition 2.1 of $(\tau, \delta\mathbb{Z})$-local-minimizers, which include both the exact minimizer and the estimator computed by Algorithm 1. A sparsity-oracle inequality for any such minimizer holds when the penalty λ in (L0) is set to a "universal" level $C\sigma^2 \log |E|$.

THEOREM 3.1. *Let $\delta \leq \sigma/\sqrt{n}$ and $\tau \leq \sigma^2$. For any $\eta > 0$, there exist constants $C_\eta, C'_\eta > 0$ depending only on η such that if $\lambda \geq C_\eta \sigma^2 \log |E|$ and $\hat{\mu}$ is any $(\tau, \delta\mathbb{Z})$-local-minimizer of* (L0), *then*

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \leq \inf_{\mu \in \mathbb{R}^n} (1 + \eta)\|\mu - \mu_0\|^2 + C'_\eta \lambda \max\big(\|D\mu\|_0, 1\big). \tag{4}$$

The upper bound in (4) trades off the edge-sparsity of μ and its approximation of the true signal μ_0. Setting $\lambda = C_\eta \sigma^2 \log |E|$ yields the guarantee (1) described in the Introduction. If μ_0 is exactly edge-sparse with $\|D\mu_0\|_0 = s$, then evaluating (4) at $\mu = \mu_0$ yields a risk bound of order $\sigma^2 s \log |E|$. When s is known, we may obtain the tighter guarantee (2) by using a smaller penalty.

THEOREM 3.2. *Let $\delta \leq \sigma/\sqrt{n}$ and $\tau \leq \sigma^2$. There exist universal constants $C, C' > 0$ such that for any $s \in [1, |E|]$, if $\lambda \geq C\sigma^2(1 + \log \frac{|E|}{s})$ and $\hat{\mu}$ is any*

$(\tau, \delta\mathbb{Z})$-*local-minimizer of* (L0), *then*

$$\sup_{\mu_0:\|D\mu_0\|_0\le s} \mathbb{E}\big[\|\hat{\mu}-\mu_0\|^2\big] \le C'\lambda s. \tag{5}$$

Theorems 3.1 and 3.2 are analogous to estimation guarantees in the sparse normal-means problem: For estimating a signal $\mu_0 \in \mathbb{R}^n$ with at most k nonzero entries, asymptotically if $n \to \infty$ and $k/n \to 0$, then

$$\inf_{\hat{\mu}} \sup_{\mu_0:\|\mu_0\|\le k} \mathbb{E}\big[\|\hat{\mu}-Y\|^2\big] \sim 2\sigma^2 k \log\frac{n}{k}. \tag{6}$$

This risk is achieved by $\hat{\mu} = \operatorname{argmin}_\mu \frac{1}{2}\|Y-\mu\|^2 + \lambda\|\mu\|_0$ for $\lambda = \sigma^2 \log\frac{n}{k}$, corresponding to entrywise hard-thresholding at level $\sqrt{2\lambda}$ [36], Theorem 8.20. Setting $\lambda = \sigma^2 \log n$ hard-thresholds instead at the universal level $\sqrt{2\sigma^2 \log n}$, and Lemma 1 of [23] implies an oracle bound

$$\mathbb{E}\big[\|\hat{\mu}-\mu_0\|^2\big] \le \inf_{\mu\in\mathbb{R}^n} 1.2\|\mu-\mu_0\|^2 + \sigma^2(2\log n+1)\big(\|\mu\|_0+1\big)$$

for any true signal $\mu_0 \in \mathbb{R}^n$.

When there is an underlying graph G, the sparsity condition $\|\mu_0\|_0 \le k$ is a notion of vertex sparsity, in contrast to our notion of edge-sparsity. The edge-sparsity of a "typical" piecewise-constant signal may be graph-dependent. For example, if G is a K-dimensional lattice graph with side length $n^{1/K}$ and μ_0 consists of two constant pieces separated by a smooth boundary, then $s \asymp n^{1-1/K}$. For such choices of μ_0 and for $K \ge 2$, the risk in (5) grows polynomially in n and does not represent a parametric rate. On the other hand, vertex-sparse signals are also edge-sparse for low-degree graphs. This containment may be used to show, when G has bounded average degree, that the above nonparametric rate is optimal in a minimax sense over $\{\mu_0 : \|D\mu_0\|_0 \le s\}$.

THEOREM 3.3. *Suppose G has average vertex degree d. There exists a universal constant $c > 0$ such that for any $s \in [4d, |E|]$,*

$$\inf_{\hat{\mu}} \sup_{\mu_0:\|D\mu_0\|_0\le s} \mathbb{E}\big[\|\hat{\mu}-\mu_0\|^2\big] \ge c\sigma^2 \frac{s}{d}\left(1+\log\frac{|E|}{s}\right), \tag{7}$$

where the infimum is taken over all possible estimators $\hat{\mu} := \hat{\mu}(Y)$.

When the average degree d is not small, there is a gap between (5) and (7) of order d, which we will discuss in Section 5.

REMARK 3.4. We assume $s \ge 4d$ for (7) so that the result does not depend on the exact structure of near-minimum cuts in G. For example, if vertices $\{1,\dots,n-1\}$ are connected in a single cycle and vertex n is connected to vertex 1 by a single edge, then for $s = 1$, any μ_0 with $\|D\mu_0\|_0 \le 1$ must be constant

over vertices $\{1, \dots, n-1\}$ and take a possibly different value on vertex n. The minimax risk over this class is then $2\sigma^2$, rather than order $\sigma^2 \log n$. Considering the graph tensor product of this example with the complete graph on d vertices, a similar argument shows that a general lower bound must restrict to $s \geq cd$ for some small constant $c > 0$.

While our main focus is estimation, let us state a result relevant to testing:

THEOREM 3.5. *Let $\delta \leq \sigma/\sqrt{n}$ and $\tau \leq \sigma^2$, and suppose μ_0 is constant over G. There exist universal constants $C, C' > 0$ such that if $\lambda \geq C\sigma^2 \log|E|$ and $\hat{\mu}$ is any $(\tau, \delta\mathbb{Z})$-local-minimizer of (L0), then*

$$\mathbb{P}[\hat{\mu} \text{ is constant over } G] \geq 1 - C'n^{-3}.$$

This implies that we may test the null hypothesis

$$H_0 : \mu_0 \text{ is constant} \tag{8}$$

by setting $\lambda \asymp \sigma^2 \log|E|$ and rejecting H_0 if $\hat{\mu}$ is not constant. Denoting by $P^{\perp}$ the orthogonal projection onto the space orthogonal to the all-1's vector, since

$$\min_{\mu:\mu \text{ is constant}} \|\mu - \mu_0\|^2 = \|P^{\perp}\mu_0\|^2,$$

the risk bound (5) [or more precisely, Lemma S2.3(b) which establishes an analogous bound in probability] implies that this test can distinguish a non-constant alternative μ_0 with probability approaching 1 as long as $\|P^{\perp}\mu_0\|^2 \geq C\sigma^2\|D\mu_0\|_0 \log|E|$, for a universal constant $C > 0$.

4. Comparison with l_1/total-variation denoising. We compare the guarantees of the preceding section with those attainable by $\hat{\mu}$ minimizing (TV). Theoretical risk bounds for the TV-penalized estimator have been established for both piecewise-constant classes $\{\mu_0 : \|D\mu_0\|_0 \leq s\}$ and bounded-variation classes $\{\mu_0 : \|D\mu_0\|_1 \leq s\}$, and we focus our comparison on the former. We will empirically explore in Section 6 some trade-offs between the l_0 and TV approaches for signals that are both piecewise-constant and of small total-variation norm.

One general risk bound for $\hat{\mu}$ minimizing (TV) was established in [34]. For an arbitrary graph G, let $D : \mathbb{R}^n \to \mathbb{R}^E$ be its vertex-edge incidence matrix, $S = D^{\dagger}$ the Moore–Penrose pseudo-inverse of D, and ρ the maximum Euclidean norm of any column of S. Theorem 2 of [34] implies, for the estimator $\hat{\mu}$ minimizing (TV) with the choice $\lambda = \sigma\rho\sqrt{2(1 + \log(|E|/\delta))}$, and for any $\mu \in \mathbb{R}^n$, with probability at least $1 - 2\delta$,

$$\|\hat{\mu} - \mu_0\|^2 \leq \|\mu - \mu_0\|^2 + 8\sigma^2\left(\frac{\rho^2\|D\mu\|_0}{\kappa^2}\left(1 + \log\frac{|E|}{\delta}\right) + \log\frac{e}{\delta}\right), \tag{9}$$

where κ is a compatibility constant bounded as $\kappa^{-2} \le 4\min(d, \|D\mu\|_0)$ and d is the mean vertex degree of G. (The result of [34] is more general, involving both $\|D\mu\|_1$ and $\|D\mu\|_0$, and we have specialized to the "pure-l_0" setting.)

An important difference between this result and Theorem 3.1 is the appearance of ρ^2, which is graph-dependent. Assuming G has small average degree d, the above guarantee is similar to Theorem 3.1 if ρ^2 is small. It is shown in [34] that $\rho^2 \lesssim 1$ for 3-D (and higher-dimensional) lattice graphs and $\rho^2 \lesssim \log n$ for 2-D lattice graphs, indicating that $\hat{\mu}$ is nearly rate-optimal over $\{\mu_0 : \|D\mu_0\|_0 \le s\}$ for these graphs. However, for example, when G is the linear chain, $\rho^2 \asymp n$ and the bound (9) is larger than those of the preceding section by a factor of n.

More specialized analyses were performed for the linear chain in [19, 44], where sharper results were obtained that depend on the minimum spacing $\Delta(\mu_0)$ between two changepoints of μ_0. More precisely, denoting by $1 \le i_1 < \cdots < i_s < n$ the values i for which $\mu_{0,i} \ne \mu_{0,i+1}$ and letting $i_0 := 0$ and $i_{s+1} := n$, define $\Delta(\mu_0) := \min_{0 \le r \le s} i_{r+1} - i_r$. Then Theorem 4 of [44] shows, if $\|D\mu_0\|_0 = s$ and $\lambda = \sigma(n\Delta(\mu_0))^{1/4}$, then

$$\mathbb{E}[\|\hat{\mu} - \mu_0\|^2] \lesssim \sigma^2 s\big((\log s + \log\log n)\log n + \sqrt{n/\Delta(\mu_0)}\big).$$

If $\Delta(\mu_0) \gtrsim n/(s+1)$ so that changepoints are nearly equally-spaced, then this bound is of order $s^{3/2}$ times logarithmic factors, and furthermore this has been improved to the optimal bound $\mathbb{E}[\|\hat{\mu} - \mu_0\|^2] \lesssim \sigma^2 s \log(1 + n/s)$ in [30]. However, if $\Delta(\mu_0) \lesssim n^\alpha$ for any $\alpha < 1$, then the above bound differs from the guarantee of Theorem 3.2 by a factor of roughly $n^{(1-\alpha)/2}$, and in the worst case this suboptimality is of order $\sqrt{n}$.

It has been conjectured, for example in Remark 3 of [44] and Remark 2.3 of [30], that this suboptimality is not an artifact of the theoretical analysis, but rather that the TV-penalized estimate $\hat{\mu}$ exhibits a slower rate of convergence when the equal spacing condition $\Delta(\mu_0) \gtrsim n/(s+1)$ is not met. We provide in this section a theoretical validation of this conjecture; proofs are given in Appendix S3.

First, suppose the true signal μ_0 is constant and equal to zero.

THEOREM 4.1. *Let G be the linear chain graph with n vertices, and suppose $\mu_0 = 0$. There exists a constant $c > 0$ such that for any fixed $\lambda \in [0, \sigma\sqrt{cn/\log n}]$, if $\hat{\mu}$ is the minimizer of* (TV), *then the following hold*:

(a) *For some constants $C, c' > 0$, letting $\hat{k} := \|D\hat{\mu}\|_0 + 1$ be the number of constant intervals of $\hat{\mu}$,*

$$\mathbb{P}\left[\hat{k} > \frac{cn}{\max(\lambda^2/\sigma^2, 1)\log n}\right] \ge 1 - Ce^{-c'n/\max(\lambda^2/\sigma^2, 1)}.$$

(b) *For some constant $c' > 0$, the squared-error risk of $\hat{\mu}$ satisfies*

$$\mathbb{E}[\|\hat{\mu}\|^2] \ge \frac{c'\sigma^2 n}{\max(\lambda^2/\sigma^2, 1)(\log n)^4}.$$

Hence, if $\mu_0 = 0$ and $\lambda \asymp \sigma n^{\alpha}$ for any $\alpha < 1/2$, then the number of changepoints and the squared-error risk of the TV-penalized estimator $\hat{\mu}$ are (up to logarithmic factors) at least of order $n^{1-2\alpha}$ and $\sigma^2 n^{1-2\alpha}$, respectively. As a consequence, we obtain the following lower bound in a minimax sense.

THEOREM 4.2. *Let G be the linear chain graph with n vertices, and let $\Delta(\mu_0)$ denote the minimum distance between two changepoints in μ_0. For each fixed $\lambda \geq 0$, let $\hat{\mu}^{\lambda}$ denote the minimizer of (TV) for this λ. Then there exists a constant $c > 0$ such that for any $s \in [2, n-1]$ and $\Delta \leq n/(s+1)$,*

$$\inf_{\lambda \geq 0} \sup_{\mu_0 : \|D\mu_0\|_0 \leq s, \Delta(\mu_0) \geq \Delta} \mathbb{E}\big[\|\hat{\mu}^{\lambda} - \mu_0\|^2\big] \geq \frac{c\sigma^2}{(\log n)^5} \sqrt{\frac{sn}{\Delta}}.$$

In particular, setting $\Delta = 1$ removes restrictions on the minimum spacing between changepoints and yields (3) stated in the Introduction.

Theorem 4.2 may be reinterpreted in the context of the Lasso estimate for sparse linear regression: Setting $\beta_0 := D\mu_0$ and

$$X := D^{\dagger} = \begin{pmatrix} -\frac{n-1}{n} & -\frac{n-2}{n} & -\frac{n-3}{n} & \cdots & -\frac{1}{n} \\ \frac{1}{n} & -\frac{n-2}{n} & -\frac{n-3}{n} & \cdots & -\frac{1}{n} \\ \frac{1}{n} & \frac{2}{n} & -\frac{n-3}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{2}{n} & \frac{3}{n} & \cdots & \frac{n-1}{n} \end{pmatrix} \in \mathbb{R}^{n \times (n-1)},$$

minimizing (TV) is equivalent to minimizing the Lasso objective

$$\frac{1}{2}\|\tilde{Y} - X\beta\|^2 + \lambda\|\beta\|_1$$

over $\beta \in \mathbb{R}^{n-1}$, where $\tilde{Y} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$ denotes Y centered by its mean. The maximum column norm of X is $\sqrt{n/4}$, the error $\|\hat{\mu} - \mu_0\|^2$ corresponds to n times the "prediction loss" $n^{-1}\|X\hat{\beta} - X\beta_0\|^2$, and in this context Theorem 4.2 (with $\Delta = 1$) implies

$$\inf_{\lambda \geq 0} \sup_{\beta_0 : \|\beta_0\|_0 \leq s} \mathbb{E}\left[\frac{1}{n}\big\|X\hat{\beta}^{\lambda} - X\beta_0\big\|^2\right] \geq \frac{c\sigma^2}{(\log n)^5}\sqrt{\frac{s}{n}}.$$

Hence the minimax prediction risk for the Lasso estimate over $\{\beta_0 : \|\beta_0\|_0 \leq s\}$ decays essentially no faster than order $n^{-1/2}$. This is in contrast to the faster rate of n^{-1} that is achievable when X has well-behaved restricted eigenvalue constants (see, e.g., [10, 65] and the references contained therein).

More generally, for any connected graph G, noting that $D^\dagger \in \mathbb{R}^{n\times E}$ is of rank $n-1$ with range orthogonal to the all-1's vector, minimizing (TV) is equivalent to minimizing

$$\frac{1}{2}\|\tilde{Y} - D^\dagger\beta\|^2 + \lambda\|\beta\|_1 \qquad \text{subject to } \beta \in \text{range}(D),$$

where range(D) denotes the column span of D in $\mathbb{R}^E$. The results of the two preceding sections imply that whenever G has small average degree, the "fast" optimal rate for prediction risk over the class $\{\beta_0 \in \text{range}(D) : \|\beta_0\|_0 \le s\}$ for the above problem is attainable in polynomial time, even if it is not achieved by the l_1 estimator. This may be contrasted with the negative results of [72, 73], which show that there exist adversarial design matrices X for sparse regression where such fast rates are not achieved by a broad class of M-estimators or by any polynomial-time algorithm returning an s-sparse output.

5. Edge-weighting for inhomogeneous graphs. In this section, we generalize the results of Section 3 by considering (exact or approximate) minimizers $\hat{\mu}$ of the edge-weighted objective (W). Proofs are contained in Appendix S2, with a brief summary of proof ideas at the end of this section.

We motivate our discussion by the following example, which examines the factor-d gap between the upper and lower bounds of (5) and (7).

EXAMPLE 5.1. Let G be the complete graph on n vertices. Then the average vertex degree of G is $d = n-1$, and (7) implies

$$\inf_{\hat{\mu}} \sup_{\mu_0:\|D\mu_0\|_0\le s} \mathbb{E}[\|\hat{\mu} - \mu_0\|^2] \gtrsim \sigma^2 \frac{s}{n}\left(1 + \log\frac{|E|}{s}\right). \tag{10}$$

This lower bound is in fact tight, and the upper bound of (5) is loose by a factor of n: Theorem 5.5 below will imply that setting $\lambda \asymp \frac{\sigma^2}{n}(1+\log\frac{|E|}{s})$ in (L0) achieves the above level of risk, when G is the complete graph.

On the other hand, let G be a "tadpole" graph consisting of a linear chain of $n/2$ vertices with one endpoint connected by an edge to a clique of $n/2$ remaining vertices. The average vertex degree of G in this case is $d = (n+1)/2$, so a direct application of (7) still yields (10). However, by restricting to the subclass of signals that take a constant value on the $n/2$-clique, it is clear that the minimax risk over $\{\mu_0 : \|D\mu_0\|_0 \le s\}$ is at least that of estimating the signal over only the linear chain portion of G with $n/2$ vertices. The lower bound (7) applied to only this subgraph implies that in this case, the upper bound (5) is tight up to a constant factor, and the lower bound (10) is loose by a factor of $1/n$.

This example highlights the problem that the complexity measure $\|D\mu_0\|_0$ is not necessarily spatially homogeneous over G. For example, when G is the tadpole

graph, a signal μ_0 that is constant over all but one vertex belonging to the $n/2$-clique has $\|D\mu_0\|_0 = n/2 - 1$, but a signal taking a different value at each of the vertices of the linear chain also has $\|D\mu_0\|_0 = n/2 - 1$. A theoretical consequence is that the minimax risk over $\{\mu_0 : \|D\mu_0\|_0 \le s\}$ is controlled by the least well-connected portion of the graph. A practical consequence is that any choice of λ will either oversmooth the signal over the $n/2$-clique or undersmooth the signal over the linear chain, and no single choice of λ leads to good signal recovery in both of these regions.

While the tadpole graph is an extreme example, the same phenomenon arises in any graph with regions of varying connectivity. In such applications, we propose to consider the weighted objective function (W) where each edge is weighted by a measure of its contribution to the connectivity of G. We believe both that minimizing this weighted objective is usually a more reasonable procedure in practice and that the value $\|D\mu_0\|_{\mathrm{w}}$ provides a better indication of the complexity of the piecewise-constant signal μ_0.

One specific weighting $\mathrm{w} : E \to \mathbb{R}_+$ that implements this idea is to weigh each edge by its effective resistance.

DEFINITION 5.2. Let G be a connected graph and $\{i, j\}$ an edge in G. The *effective resistance* $\mathrm{r}(i, j)$ of this edge has the following four equivalent definitions:

1. $\mathrm{r}(i, j)$ is the effective electrical resistance measured across vertices i and j when G represents an electrical network where each edge is a resistor with resistance 1.
2. Let L be the (unweighted) Laplacian matrix of G, $L^\dagger$ the pseudo-inverse of L, and e_i the basis vector with ith entry 1 and remaining entries 0. Then $\mathrm{r}(i, j) = (e_i - e_j)L^\dagger(e_i - e_j)$.
3. Consider a simple random walk on G starting at vertex i, and let t be the number of steps taken to reach vertex j and then return to vertex i for the first time. Then $\mathrm{r}(i, j) = \frac{1}{2|E|}\mathbb{E}[t]$.
4. Let T be (the edges of) a random spanning tree of G chosen uniformly from the set of all spanning trees of G. Then $\mathrm{r}(i, j) = \mathbb{P}[\{i, j\} \in T]$.

For verification of the equivalence of these definitions, see [27, 46]. In practice, $\mathrm{r}(i, j)$ may be computed via the second characterization using fast Laplacian solvers [45, 59].

The fourth characterization above describes one sense in which $\mathrm{r}(i, j)$ measures the "contribution" of edge $\{i, j\}$ to the connectivity of G: For example, if removing $\{i, j\}$ breaks G into two disconnected components, then every spanning tree T of G must contain $\{i, j\}$, so $\mathrm{r}(i, j) = 1$. Conversely, if there are many short alternative paths from i to j not using edge $\{i, j\}$, then $\mathrm{r}(i, j)$ is much smaller than 1.

More generally, the contribution of each edge to the graph connectivity may be measured by any weighting belonging to the spanning tree polytope of G.

DEFINITION 5.3. A weighting $\mathrm{w} : E \to \mathbb{R}_+$ is a *spanning tree weighting* if there exists a spanning tree T of G such that $\mathrm{w}(i, j) = 1$ if $\{i, j\} \in T$ and $\mathrm{w}(i, j) = 0$ otherwise. The *spanning tree polytope* $\mathcal{ST}(G)$ is the convex hull of all spanning tree weightings.

[With a slight abuse of notation, we will henceforth denote general weightings by w and any weighting in $\mathcal{ST}(G)$ by r.] Thus, if $\mathrm{r} \in \mathcal{ST}(G)$, then there exist spanning trees $T_1, \dots, T_M$ of G and $\lambda_1, \dots, \lambda_M > 0$ with $\sum_{m=1}^M \lambda_m = 1$ such that

$$\mathrm{r}(i, j) = \sum_{m=1}^{M} \lambda_m \mathbb{1}\{\{i, j\} \in T_m\}$$

for every edge $\{i, j\} \in E$. The weighting r is thus identified with the probability distribution of a random spanning tree T, where $T = T_m$ with probability λ_m. This distribution satisfies the property, for all $\{i, j\} \in E$,

$$\mathrm{r}(i, j) = \mathbb{P}[\{i, j\} \in T].$$

For any subset of edges $E' \subseteq E$ and weighting $\mathrm{w} : E \to \mathbb{R}_+$, let us denote $\mathrm{w}(E') = \sum_{\{i,j\} \in E'} \mathrm{w}(i, j)$ as the total weight of these edges. Then the above implies, for the random spanning tree T associated to $\mathrm{r} \in \mathcal{ST}(G)$,

$$\mathrm{r}(E') = \mathbb{E}[|E' \cap T|]. \tag{11}$$

The effective resistance weighting of Definition 5.2 corresponds to the uniform distribution for T.

The results below describe the squared-error risk of the estimator $\hat{\mu}$ that (exactly or approximately) minimizes (W) for any edge-weighting $\mathrm{w} : E \to \mathbb{R}_+$. We derive, for all graphs G, minimax upper and lower bounds on this risk over the class $\{\mu_0 : \|D\mu_0\|_{\mathrm{w}} \le s\}$. The tightness of these bounds will depend on how close w is to the spanning tree polytope $\mathcal{ST}(G)$; for effective resistance weighting, or more generally for any $\mathrm{r} \in \mathcal{ST}(G)$, these bounds are tight up to a universal constant factor independent of the graph.

Let us make a remark about scaling, which is important for the interpretation of the below results: As rescaling w by $c > 0$ and λ by $1/c$ leads to the same penalty in (W), we will state all of our results, for simplicity and without loss of generality, under a scaling such that $\mathrm{w} \ge \mathrm{r}$ for some $\mathrm{r} \in \mathcal{ST}(G)$, meaning $\mathrm{w}(i, j) \ge \mathrm{r}(i, j)$ for every edge. For any w (where G remains connected by edges with positive weight), there is a smallest constant $c > 0$ for which this property holds for $\tilde{\mathrm{w}} = c\mathrm{w}$; the below results yield the tightest risk bounds when applied to $\tilde{\mathrm{w}}$ scaled in this way. Whenever $\mathrm{w} \ge \mathrm{r}$ for some $\mathrm{r} \in \mathcal{ST}(G)$, (11) implies that the total weight of all edges satisfies

$$\mathrm{w}(E) \ge \mathrm{r}(E) = n - 1, \tag{12}$$

since every spanning tree has $n - 1$ edges. The ratio $\mathrm{w}(E)/(n - 1)$ provides a measure of the distance of w to $\mathcal{ST}(G)$. Furthermore, if E' is any subset of edges

whose removal disconnects G into $k+1$ connected components, then every spanning tree contains at least k edges of E', so $\mathrm{w}(E') \geq \mathrm{r}(E') \geq k$. In particular,

$$\mu \in \mathbb{R}^n \text{ has } k+1 \text{ distinct values} \quad \Rightarrow \quad \|D\mu\|_{\mathrm{w}} \geq k. \tag{13}$$

The following results generalize Theorems 3.1, 3.2, and 3.3.

THEOREM 5.4. *Let* $\mathrm{w} : E \to \mathbb{R}_+$ *be such that* $\mathrm{w} \geq \mathrm{r}$ *for some* $\mathrm{r} \in \mathcal{ST}(G)$, *and let* $\delta \leq \sigma/\sqrt{n}$ *and* $\tau \leq \sigma^2$. *For any* $\eta > 0$, *there exist constants* $C_\eta, C'_\eta > 0$ *depending only on* η *such that if* $\lambda \geq C_\eta \sigma^2 \log \mathrm{w}(E)$ *and* $\hat{\mu}$ *is any* $(\tau, \delta\mathbb{Z})$*-local-minimizer of* (W), *then*

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \leq \inf_{\mu} (1+\eta)\|\mu - \mu_0\|^2 + C'_\eta \lambda \max\big(\|D\mu\|_{\mathrm{w}}, 1\big). \tag{14}$$

THEOREM 5.5. *Let* $\mathrm{w} : E \to \mathbb{R}_+$ *be such that* $\mathrm{w} \geq \mathrm{r}$ *for some* $\mathrm{r} \in \mathcal{ST}(G)$, *and let* $\delta \leq \sigma/\sqrt{n}$ *and* $\tau \leq \sigma^2$. *There exist universal constants* $C, C' > 0$ *such that for any* $s \in [1, \mathrm{w}(E)]$, *if* $\lambda \geq C\sigma^2(1 + \log \frac{\mathrm{w}(E)}{s})$ *and* $\hat{\mu}$ *is any* $(\tau, \delta\mathbb{Z})$*-local-minimizer of* (W), *then*

$$\sup_{\mu_0 : \|D\mu_0\|_{\mathrm{w}} \leq s} \mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \leq C' \lambda s. \tag{15}$$

Conversely, there exists a universal constant $c > 0$ *such that for any* $s \in [\frac{4\mathrm{w}(E)}{n}, \mathrm{w}(E)]$,

$$\inf_{\hat{\mu}} \sup_{\mu_0 : \|D\mu_0\|_{\mathrm{w}} \leq s} \mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \geq c\sigma^2 s \frac{n}{\mathrm{w}(E)} \left(1 + \log \frac{\mathrm{w}(E)}{s}\right),$$

where the infimum is taken over all possible estimators $\hat{\mu} := \hat{\mu}(Y)$.

The restriction to $s \gtrsim \mathrm{w}(E)/n$ in the lower bound is necessary for generality of the result to all graphs G, for the same reason as in Remark 3.4.

The minimax upper and lower bounds above differ by the factor $n/\mathrm{w}(E)$. Recall from (12) that $\mathrm{w}(E) \geq n-1$, with $\mathrm{w}(E) = n-1$ precisely when $\mathrm{w} \in \mathcal{ST}(G)$. Hence the above immediately implies the following corollary.

COROLLARY 5.6. *If* $\mathrm{w} = \mathrm{r}$ *where* $\mathrm{r}(i,j)$ *is the effective resistance of each edge* $\{i, j\}$, *or more generally where* $\mathrm{r} \in \mathcal{ST}(G)$, *then for any* $s \in [4, n-1]$,

$$\inf_{\hat{\mu}} \sup_{\mu_0 : \|D\mu_0\|_{\mathrm{w}} \leq s} \mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \asymp \sigma^2 s \left(1 + \log \frac{n}{s}\right). \tag{16}$$

REMARK 5.7. One may compare (15) with a guarantee achieved by the wavelet spanning tree method of [56]: In this method, for a fixed spanning tree T of G, an orthonormal basis of Haar-like wavelet functions is constructed over T such that a signal μ_0 cutting s edges of T has a representation of sparsity

$s(\log d_{\max}(T))(\log n)$ in this basis, where $d_{\max}(T)$ is the maximal vertex degree of T. The corresponding wavelet thresholding estimator then satisfies

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \lesssim \sigma^2 s\big(\log d_{\max}(T)\big)(\log n)^2.$$

If T is chosen at random from the spanning tree distribution corresponding to any weighting $\mathrm{w} \in \mathcal{ST}(G)$, then bounding $d_{\max}(T) \le d_{\max}(G)$ and averaging over the random choice of T yields

$$\mathbb{E}\big[\|\hat{\mu} - \mu_0\|^2\big] \lesssim \sigma^2\big(\log d_{\max}(G)\big)(\log n)^2 \|D\mu_0\|_{\mathrm{w}},$$

which agrees with (15) up to extra logarithmic factors. Whereas this defines a randomized algorithm and the above risk is averaged also over the algorithm execution, minimizing (W) for $\mathrm{w} \in \mathcal{ST}(G)$ directly penalizes the number of edges cut by $\hat{\mu}$ in the average spanning tree, and thus may be interpreted loosely as a derandomization of this wavelet approach.

Finally, we state a result of relevance to testing the null hypothesis (8).

THEOREM 5.8. *Let* $\mathrm{w} : E \to \mathbb{R}_+$ *be such that* $\mathrm{w} \ge \mathrm{r}$ *for some* $\mathrm{r} \in \mathcal{ST}(G)$, *and let* $\delta \le \sigma/\sqrt{n}$ *and* $\tau \le \sigma^2$. *There exist universal constants* $C, C' > 0$ *such that if* μ_0 *is constant over* G, $\lambda \ge C\sigma^2 \log \mathrm{w}(E)$ *and* $\hat{\mu}$ *is any* $(\tau, \delta\mathbb{Z})$*-local-minimizer of* (W), *then*

$$\mathbb{P}[\hat{\mu} \textit{ is constant over } G] \ge 1 - C'n^{-3}.$$

Thus we may test H_0 in (8) by setting $\lambda \asymp \sigma^2 \log \mathrm{w}(E)$ and rejecting H_0 if $\hat{\mu}$ minimizing (W) is not constant. Denoting by $P^{\perp}$ the projection orthogonal to the all-1's vector, the risk bound (15) [or more precisely, the probability guarantee of Lemma S2.3(b)] implies that this test can distinguish a nonconstant alternative μ_0 with probability approaching 1 as long as $\|P^{\perp}\mu_0\|^2 \ge C\sigma^2\|D\mu_0\|_{\mathrm{w}} \log \mathrm{w}(E)$, for a universal constant $C > 0$. When $\mathrm{w} : E \to \mathbb{R}_+$ is the effective resistance weighting, this recovers a similar detection threshold as established for the tests in [56, 57].

In the case of uniform edge weights $\mathrm{w} \equiv 1$, it is clear that $\mathrm{w}(E) = |E|$ and $\mathrm{w} \ge \mathrm{r}$ for all $\mathrm{r} \in \mathcal{ST}(G)$. Then Theorems 3.1, 3.2, 3.3 and 3.5 follow directly by specializing these results. If there exists $\mathrm{r} \in \mathcal{ST}(G)$ such that $\mathrm{r}(i, j) < 1$ for every edge, then the results of Section 3 are trivially strengthened by rescaling λ by $\max_{\{i,j\}\in E} \mathrm{r}(i, j)$. For example, if G is the complete graph, then every edge has effective resistance $\mathrm{r}(i, j) = 2/n$, and Theorems 5.4 and 5.5 imply that λ may in fact be set to $Cn^{-1}\log|E|$ and $Cn^{-1}\log|E|/s$ in Theorems 3.1 and 3.2, respectively, as claimed in Example 5.1.

We prove Theorems 5.4, 5.5 and 5.8 in Appendix S2. The upper bound in Theorem 5.5 uses the idea of [37], Theorem 3.3, for bounding the number of small graph cuts by controlling the number of cut edges in a given spanning tree. We

apply this idea in Lemma S2.2, controlling a supremum over all small cuts by selecting a random spanning tree according to the weighting r and taking a union bound over cuts of this tree. In conjunction with a Chernoff bound and a standard Cauchy–Schwarz argument, this establishes (14) and (15) for the exact minimizer of (W) with high probability. We obtain bounds in expectation using Holder's inequality to control the risk on the complementary low-probability event. The extension to approximate minimizers uses the factor-2 approximation guarantee for the alpha-expansion algorithm established in [13]. However, whereas the optimal objective value for (W) is usually dominated by the squared-error term, we verify in Lemma S2.1 that the approximation factor applies not to this term but only to the l_0 penalty, and it holds not only with respect to the global minimizer of (W) but also with respect to any candidate vector μ. This yields (14) and (15) for local minimizers. Theorem 5.8 uses the preceding risk bounds together with the observation that the optimal constant estimate is within one alpha-expansion from any vector μ. Finally, the lower bound in Theorem 5.5 follows from an embedding of vertex-sparse vectors into $\{\mu_0 : \|D\mu_0\|_w \leq s\}$ and a standard lower bound for sparse normal-means; similar arguments were used in [54, 56].

6. Simulations. We study empirically the squared-errors of the approximate minimizers of (L0) and (W) as returned by Algorithm 1, as well as the exact minimizer of (TV) (computed using the pygfl library [60]). We denote these estimates by $\hat{\mu}^{L0}$, $\hat{\mu}^{W}$ and $\hat{\mu}^{TV}$. We consider piecewise-constant signals over various graphs, corrupted by Gaussian noise for various noise levels σ. We report in each setting the standardized mean-squared-error

$$\text{(st.MSE)} \qquad \frac{1}{n\sigma^2}\|\hat{\mu} - \mu_0\|^2.$$

Due to this normalization by σ^2, one may equivalently interpret these results as for a fixed noise level σ under various rescalings of the true signal μ_0.

6.1. *Parameter tuning.* For Algorithm 1, we fix throughout $\delta = 0.01$ and $\tau = 0$. This value of δ may be larger than that prescribed by the theory of the preceding section, but represents a compromise to yield faster runtime.

We select λ by minimizing an empirical estimate of $\mathbb{E}[\|\hat{\mu} - \mu_0\|^2]$. Typically, cross-validation is used to obtain such an estimate. However, we observe that naive cross-validation does not necessarily work well for all graphs and signals. (Consider, e.g., a case where the primary contribution to error comes from vertices i near the boundaries of the constant pieces of μ_0, and estimation of these values $\mu_{0,i}$ is more difficult when Y_i is removed.) We instead use the following procedure based on [32, 61]:

1. Compute an estimate $\hat{\sigma}$ for σ. Set $\alpha = 0.04$.
2. For repetitions $b = 1, \ldots, B$:

(a) Generate $z = (z_1, \ldots, z_n) \sim \text{Normal}(0, \alpha\hat{\sigma}^2\text{Id})$, and set $Y^* = Y + z$ and $Y^{**} = Y - z/\alpha$.

(b) For each λ, compute $\hat{\mu}$ based on data Y^*, and compute $\text{err}^{(b)}(\lambda) = \|\hat{\mu} - Y^{**}\|^2$.

3. Choose λ that minimizes the average error $\overline{\text{err}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \text{err}^{(b)}(\lambda)$.

This is motivated by the insight that if $\hat{\sigma} = \sigma$, then Y^* and Y^{**} are independent, so $\mathbb{E}[\text{err}^{(b)}(\lambda)] = n\sigma^2(1 + \alpha^{-1}) + \mathbb{E}[\|\hat{\mu} - \mu_0\|^2]$. Hence $\overline{\text{err}}(\lambda)$ estimates a constant plus the risk of $\hat{\mu}$ applied to data at the slightly elevated noise level $\sigma(1 + \sqrt{\alpha}) = 1.2\sigma$. Due to this elevation in noise level, this procedure has a slight tendency to oversmooth.

For each edge $\{i, j\}$ where $\mu_{0,i} = \mu_{0,j}$, we have $Y_i - Y_j \sim \text{Normal}(0, 2\sigma^2)$. Hence σ may be estimated from the edge differences $(Y_i - Y_j)_{\{i,j\}\in E}$ by identifying a normal mixture component corresponding to this subset of values; we used the method of [24] as implemented in the locfdr R package. Increasing B reduces the variability of the selection procedure. For the smaller graphs (linear chain, Oldenburg, Gnutella P2P) we set $B = 20$, and for the larger graphs (2-D cow, San Francisco, Enron email) we set $B = 5$.

We will report both the st.MSE achieved using this method, as well as the best-attained st.MSE corresponding to retrospective optimal tuning of λ. For (TV), λ may alternatively be selected by minimizing Stein's unbiased risk estimate (SURE) using the simple degrees-of-freedom formula derived in [64]. We found results of the SURE approach to be very close to those obtained using the above procedure.

6.2. *Empirical runtime.* For Algorithm 1, we computed minimum s-t cuts using the method of [12]. The outer loop required no more than 15 iterations, and typically fewer than 10 iterations, in all tested examples. Table 1 displays the average runtimes of this algorithm on our personal computer for computing $\hat{\mu}^{\text{L0}}$ with a single value of λ. The runtimes of the algorithm for computing $\hat{\mu}^{\text{W}}$ were comparable, although computing effective resistance weights required an additional a priori cost of 10 seconds, 3 hours, 45 seconds and 30 minutes for the four networks in the order listed, using the approxCholLap method of the Laplacians-0.2.0 Julia package with error tolerance 0.01. (The effective resistance computation is a one-time cost per network, reusable across different λ values and data vectors Y.)

In our experiments, the runtimes in Table 1 were roughly comparable to those of the TV denoising algorithm of [60], but slower than more optimized TV denoising algorithms on lattice graphs. Parameter tuning using the approach of Section 6.1 is much slower as it requires running the method multiple times over a range of λ values, although this computation is easily parallelized.

6.3. *Linear chain graph.* Two signals on a linear chain graph with $n = 1000$ vertices are depicted in Figure 1. The first signal has 19 equally-spaced break

TABLE 1
Average computational time of Algorithm 1 for one value of λ

	Graph					
	1-D	**Cow**	**Oldenburg**	**San Fran.**	**Gnutella**	**Enron**
Runtime (seconds)	0.13	45	0.7	40	4	240

points, while the second has 20 break points at unequal spacing. We studied recovery for noise levels $\sigma = 0.1$ to $\sigma = 1$. Figure 1 displays one instance of simulated noise and the resulting estimates $\hat{\mu}^{L0}$ and $\hat{\mu}^{TV}$. In both examples, for data-tuned λ, $\hat{\mu}^{L0}$ tends to over-smooth (missing two and four changepoints, resp.) and $\hat{\mu}^{TV}$ tends to undersmooth.

Figure 2 plots st.MSE comparisons for $\hat{\mu}^{L0}$ and $\hat{\mu}^{TV}$. The $\hat{\mu}^{L0}$ estimate achieves significantly smaller risk than $\hat{\mu}^{TV}$ at higher signal-to-noise regimes, for example those displayed in Figure 1, while $\hat{\mu}^{TV}$ becomes competitive or better in lower signal-to-noise regimes, corresponding to lower values of normalized total-variation $\|D\mu_0\|_1/\sigma$ for the true signal. Figure 3 presents a different example to further explore this trade-off, in which the normalized total-variation of the signal is fixed at $\|D\mu_0\|_1/\sigma = 20$, and we increase the number of changepoints while simultaneously decreasing the jump sizes. (Changepoints are equally spaced, and distinct signal values are normally distributed.) The estimate $\hat{\mu}^{L0}$ is better under strong edge-sparsity, while $\hat{\mu}^{TV}$ becomes better as we transition to weaker edge-sparsity.

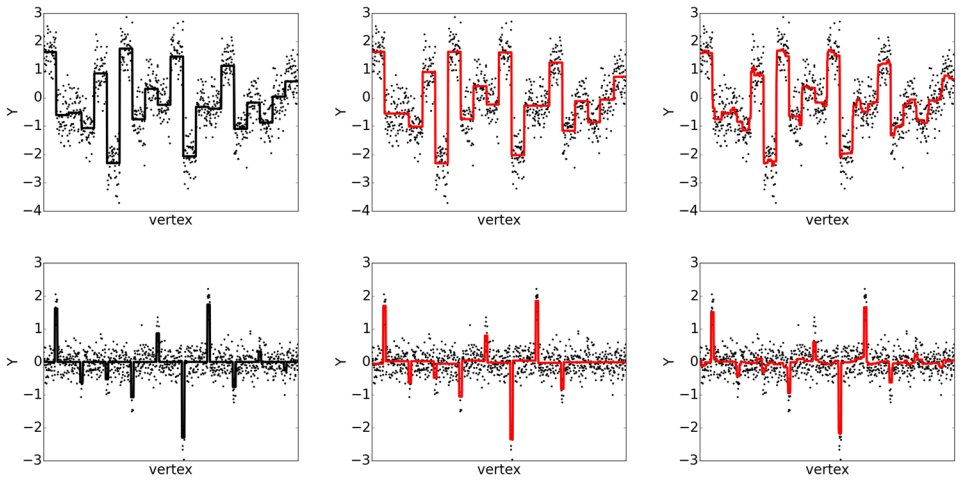

FIG. 1. *Signals on a linear chain graph. Top*: *Equally-spaced breaks*, $\sigma = 0.5$. *Bottom*: *Unequally-spaced breaks*, $\sigma = 0.3$. *The true signal* μ_0 *is displayed on the left*, $\hat{\mu}^{L0}$ *in the middle, and* $\hat{\mu}^{TV}$ *on the right* (*both with data-tuned* λ).

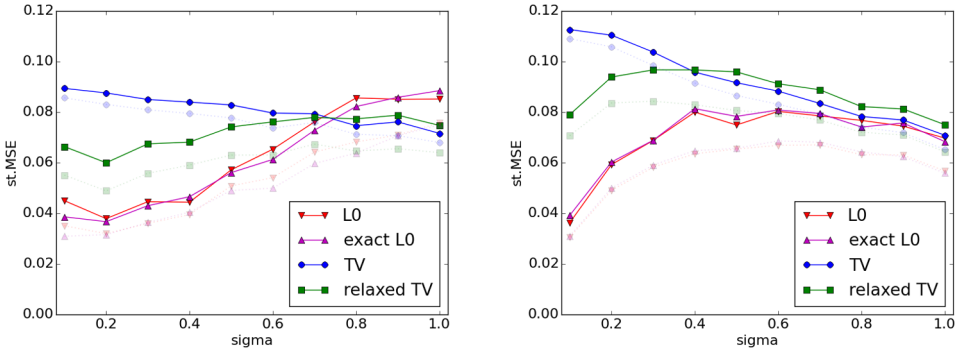

FIG. 2. *Comparisons of st.MSE in* (*left*) *the equally-spaced and* (*right*) *unequally-spaced examples of Figure* 1, *for* $\hat{\mu}^{\mathrm{L0}}$, $\hat{\mu}^{\mathrm{TV}}$, *the exact minimizer of* (L0), *and* $\hat{\mu}^{\mathrm{TV,relaxed}}$. *Solid lines correspond to data-tuned* λ, *and dashed transparent lines to best-achieved error. All errors are averaged over* 100 *replicates of the simulated noise.*

For the linear chain, we may compare $\hat{\mu}^{\mathrm{L0}}$ with the exact minimizer of (L0) (computed using PELT in the changepoint R package [38]). Algorithm 1 achieves risk comparable to the exact minimizer in all tested settings. One may ask, at the higher signal-to-noise regimes, how much of the sub-optimality of $\hat{\mu}^{\mathrm{TV}}$ is due to estimator bias incurred by shrinkage. To address this, we computed also the "relaxed" TV estimate

$$\hat{\mu}^{\mathrm{TV,relaxed}} = \alpha\hat{\mu}^{\mathrm{TV}} + (1-\alpha)\hat{\mu}^{\mathrm{TV,debiased}},$$

where $\alpha \in \{0, 0.1, 0.2, \ldots, 1\}$ is an additional tuning parameter, and where $\hat{\mu}^{\mathrm{TV,debiased}}$ replaces each constant interval of $\hat{\mu}^{\mathrm{TV}}$ with the mean of Y over this interval. The error of $\hat{\mu}^{\mathrm{TV}}$ at high signal-to-noise is partially reduced, but not to the same levels as $\hat{\mu}^{\mathrm{L0}}$.

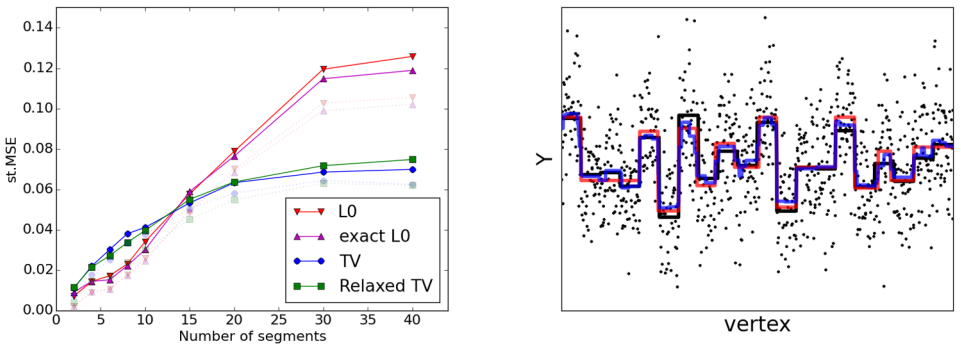

FIG. 3. *Left*: *Comparisons of st.MSE for signals of fixed total-variation* $\|D\mu_0\|_1/\sigma = 20$ *and increasing numbers of segments/decreasing jump sizes. Right*: *Raw data and true signal in black,* $\hat{\mu}^{\mathrm{L0}}$ *in red, and* $\hat{\mu}^{\mathrm{TV}}$ *in blue, for the signal with* 20 *segments.*

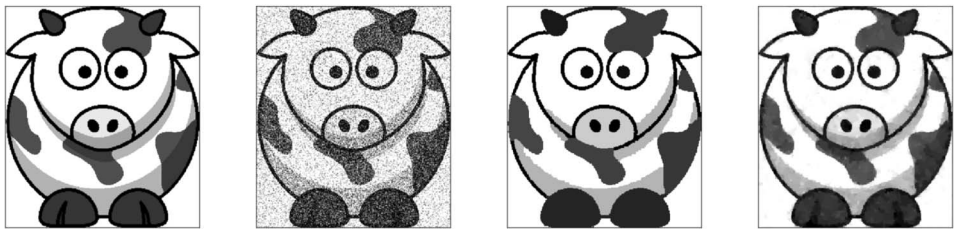

FIG. 4. *Far left*: *Original image, with pixel values normalized to* $[0, 1]$. (*middle left*) *Noisy image,* $\sigma = 0.3$. *Middle right*: $\hat{\mu}^{\mathrm{L0}}$, *and far right*: $\hat{\mu}^{\mathrm{TV}}$, *both with data-tuned* λ.

6.4. *2-D lattice graph*. Figure 4 displays a cartoon gray-scale image of a cow, represented by its pixel values on a 2-D lattice graph of size 320×283. Pure white corresponds to $\mu_0 = 1$, and pure black to $\mu_0 = 0$. The figure also displays $\hat{\mu}^{\mathrm{L0}}$ and $\hat{\mu}^{\mathrm{TV}}$ when the image is contaminated by noise at level $\sigma = 0.3$. We again observe that $\hat{\mu}^{\mathrm{L0}}$ oversmooths, missing details in the cow's feet, right horn and the shadows of the image. In contrast, $\hat{\mu}^{\mathrm{TV}}$ undersmooths and returns a blotchy cow.

Table 2 reports st.MSE comparisons for $\sigma = 0.1$ to $\sigma = 0.5$. At the level $\sigma = 0.3$ displayed in Figure 4, the st.MSE of $\hat{\mu}^{\mathrm{L0}}$ is slightly greater than that of $\hat{\mu}^{\mathrm{TV}}$. At higher signal-to-noise levels, $\hat{\mu}^{\mathrm{L0}}$ is better, while it is worse at lower signal-to-noise.

6.5. *Road and digital networks*. We tested signal recovery over four real networks: the Oldenburg and San Francisco road networks from www.cs.utah.edu/~lifeifei/SpatialDataset.htm, and the Gnutella08 peer-to-peer network and Enron email network from snap.stanford.edu/data. Duplicate edges were removed, and only the largest connected component of each network was retained.

For each network, we simulated an epidemic according to a simple graph-based discrete-time SI model [49], randomly selecting a source vertex to infect at time $t = 0$ and, for each of T timesteps, allowing each infected vertex to independently

TABLE 2
Comparison of st.MSE for the cow image of Figure 4. Nonparenthesized values correspond to data-tuned λ, *and parenthesized values to best-attained error*

	sigma				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
$\hat{\mu}^{\mathrm{L0}}$ st.MSE	0.041 (0.039)	0.067 (0.067)	0.076 (0.071)	0.081 (0.081)	0.082 (0.082)
$\hat{\mu}^{\mathrm{TV}}$ st.MSE	0.083 (0.083)	0.075 (0.075)	0.065 (0.065)	0.061 (0.059)	0.054 (0.053)

TABLE 3

For each network: Number of total vertices, total edges, variability of effective edge resistances (measured by standard deviation/mean) and numbers of infected vertices and cut edges corresponding to the signal at three observation times

Network	Verts.	Edges	Res. var.	Inf. T_1	Cut T_1	Inf. T_2	Cut T_2	Inf. T_3	Cut T_3
Oldenburg	6105	7029	0.118	57	16	515	75	2108	164
San Fran.	174,956	221,802	0.203	8574	306	27,724	562	70,925	774
Gnutella	6299	20,776	0.826	19	91	67	477	271	3894
Enron	33,696	180,811	1.297	179	7319	2564	74,117	16,868	29,253

infect each noninfected neighbor with probability 0.5. We associated the values $\mu_0 = 1.005$ and $\mu_0 = 0.005$ to infected and noninfected vertices. For each network, we considered three signals corresponding to observations of the epidemic at three different times T. Various properties of these networks and signals are summarized in Table 3.

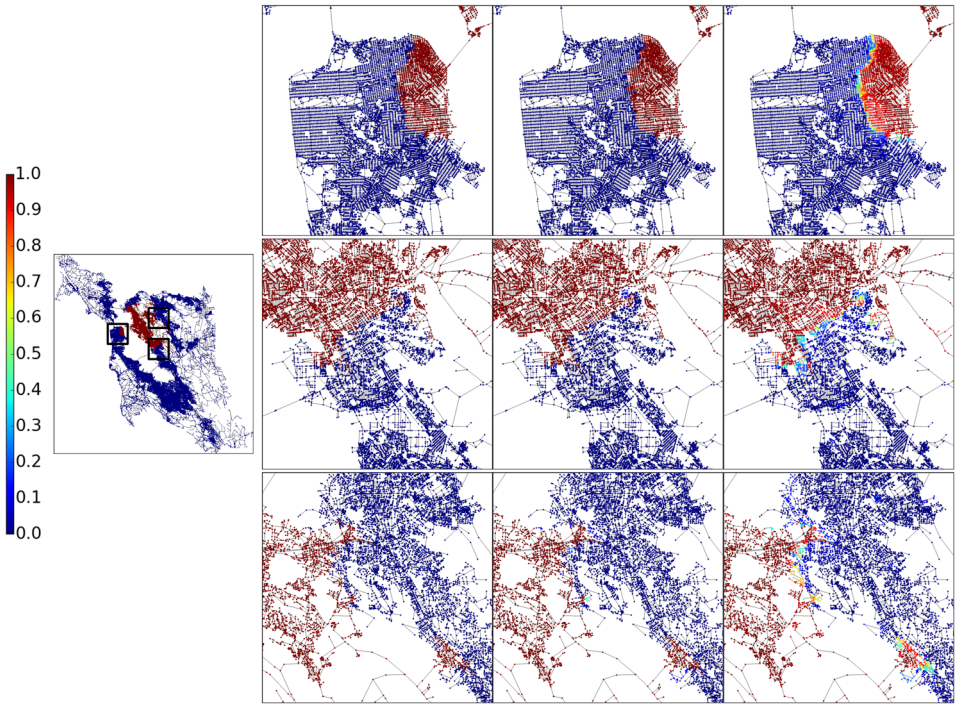

FIG. 5. *A simulated epidemic signal over the San Francisco road network is displayed on the far left. Three columns display the true signal (left), $\hat{\mu}^{\mathrm{W}}$ (middle) and $\hat{\mu}^{\mathrm{TV}}$ (right) for the three boxed areas of the map. Both $\hat{\mu}^{\mathrm{W}}$ and $\hat{\mu}^{\mathrm{TV}}$ are estimated on noisy data with $\sigma = 0.3$, using data-tuned λ, and $\hat{\mu}^{\mathrm{W}}$ uses effective-resistance edge weighting.*

For noise level $\sigma = 0.3$, the simulated signal, $\hat{\mu}^{\mathrm{TV}}$, and $\hat{\mu}^{\mathrm{W}}$ computed with effective-resistance edge weights are depicted in Figure 5 for the San Francisco road network at observation time T_2. The most difficult regions to estimate are the signal boundaries; we zoom in on three regions of the map where $\hat{\mu}^{\mathrm{TV}}$ is inaccurate at these boundaries, but $\hat{\mu}^{\mathrm{W}}$ is mostly correct. At this noise level, the st.MSE of $\hat{\mu}^{\mathrm{TV}}$ exceeds $\hat{\mu}^{\mathrm{W}}$ by a factor of about 2.

Figure 6 displays st.MSE comparisons for $\hat{\mu}^{\mathrm{L0}}$, $\hat{\mu}^{\mathrm{TV}}$ and $\hat{\mu}^{\mathrm{W}}$ with effective-resistance weighting at noise levels $\sigma = 0.1$ to $\sigma = 0.5$ in each example. We

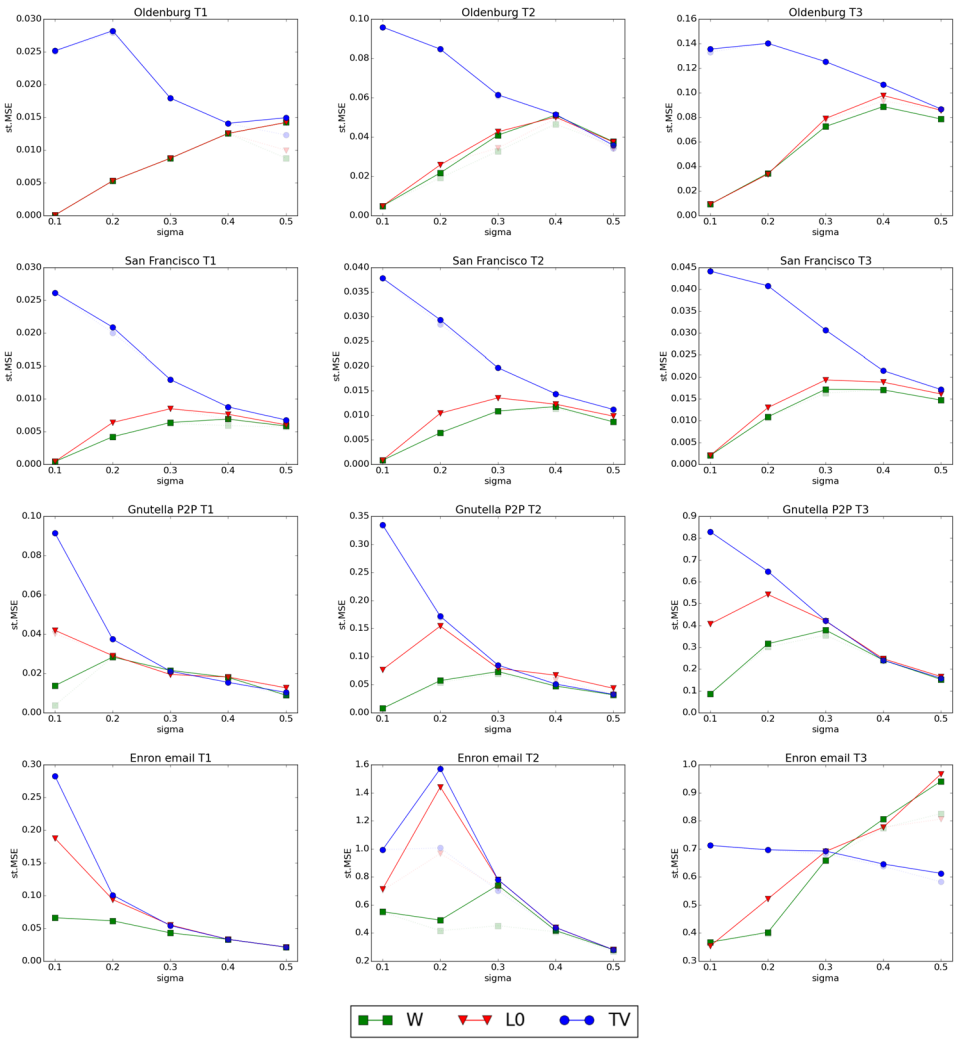

FIG. 6. *Comparisons of st.MSE for recovery of epidemic signals on four networks, for* $\hat{\mu}^{\mathrm{L0}}$, $\hat{\mu}^{\mathrm{TV}}$ *and* $\hat{\mu}^{\mathrm{W}}$ *with effective-resistance edge weights. Solid lines correspond to data-tuned* λ*, and dashed transparent lines to best-achieved error.*

observe that $\hat{\mu}^{\mathrm{W}}$ is not substantially worse than $\hat{\mu}^{\mathrm{L0}}$ in any tested setting, and that in the Gnutella and Enron digital networks where there is large variation in effective edge resistances, $\hat{\mu}^{\mathrm{W}}$ is sometimes substantially better. At the tested noise levels, these methods are (with the exception of Enron T_3) not substantially worse than $\hat{\mu}^{\mathrm{TV}}$, and can be substantially better in the lower noise settings.

7. Conclusion. We have studied estimation of piecewise-constant signals over arbitrary graphs using an l_0 edge penalty, establishing minimax rate-optimal statistical guarantees for the local minimizer computed by an approximation algorithm for minimizing this objective. We have shown theoretically that the same guarantees are not necessarily achieved by total-variation denoising, and empirically that l_0-penalization may be more effective in high signal-to-noise settings. For application to networks with regions of varying connectivity, we have proposed minimization of an edge-weighted objective, which achieves better empirical performance in tested examples and leads to theoretical guarantees that are spatially uniform over all graphs.

We note that while Algorithm 1 is provably polynomial-time, discretization of the continuous parameter domain yields poor worst-case runtime and may be computationally costly to extend to likelihood models with multi-dimensional parameters. The development of faster nondiscretized algorithms is an interesting direction for future work. Finally, our problem may be reformulated as sparse regression with particular graph-based designs, and we believe it is an interesting question whether similar computational ideas may be used to achieve better prediction error in sparse regression for more general families of designs.

Acknowledgments. We would like to thank Professors Amin Saberi, James Sharpnack and Emmanuel Candes for helpful and inspiring discussions on this problem and related literature.

SUPPLEMENTARY MATERIAL

Supplementary Appendices (DOI: 10.1214/17-AOS1656SUPP; .pdf). The supplementary appendices contain proofs of theoretical results.

REFERENCES

[1] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847

[3] ARIAS-CASTRO, E., CANDÈS, E. J., HELGASON, H. and ZEITOUNI, O. (2008). Searching for a trail of evidence in a maze. *Ann. Statist.* **36** 1726–1757. MR2435454

[4] ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425.
[5] ARIAS-CASTRO, E. and GRIMMETT, G. R. (2013). Cluster detection in networks using percolation. *Bernoulli* **19** 676–719. MR3037169
[6] AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. MR0978902
[7] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.
[8] BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319.
[9] BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840
[10] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469
[11] BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73.
[12] BOYKOV, Y. and KOLMOGOROV, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1124–1137.
[13] BOYKOV, Y., VEKSLER, O. and ZABIH, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23** 1222–1239.
[14] BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348
[15] CHAMBOLLE, A. (2005). Total variation minimization and a class of binary MRF models. In *EMMCVPR* 2005 136–152. Springer, Berlin.
[16] CHAMBOLLE, A. and LIONS, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numer. Math.* **76** 167–188. MR1440119
[17] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159.
[18] CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Stat.* **35** 999–1018.
[19] DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784
[20] DARBON, J. and SIGELLE, M. (2005). A fast and exact algorithm for total variation minimization. In *Iberian Conference on Pattern Recognition and Image Analysis* 351–359. Springer, Berlin.
[21] DAVIES, P. L. and KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. MR1833958
[22] DONOHO, D. L. (1999). Wedgelets: Nearly minimax estimation of edges. *Ann. Statist.* **27** 859–897. MR1724034
[23] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.
[24] EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104.
[25] FAN, Z. and GUAN, L. (2018). Supplement to "Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs." DOI:10.1214/17-AOS1656SUPP.
[26] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
[27] GHOSH, A., BOYD, S. and SABERI, A. (2008). Minimizing effective resistance of a graph. *SIAM Rev.* **50** 37–66. MR2403057

[28] GOLDSTEIN, T. and OSHER, S. (2009). The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.* **2** 323–343.

[29] GREIG, D. M., PORTEOUS, B. T. and SEHEULT, A. H. (1989). Exact maximum a posteriori estimation for binary images. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **51** 271–279.

[30] GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2017). Spatial adaptation in trend filtering. Available at arXiv:1702.05113.

[31] HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

[32] HARRIS, X. T. (2016). Prediction error after model search. Available at arXiv:1610.06107.

[33] HOEFLING, H. (2010). A path algorithm for the fused lasso signal approximator. *J. Comput. Graph. Statist.* **19** 984–1006. Supplementary materials available online. MR2791265

[34] HÜTTER, J.-C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In *Conf. Learning Theory* 1115–1146.

[35] JACKSON, B., SCARGLE, J. D. et al. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.

[36] JOHNSTONE, I. (2015). *Gaussian Estimation*: *Sequence and Wavelet Models*. Available at statweb.stanford.edu/~imj/GE09-08-15.pdf.

[37] KARGER, D. R. and STEIN, C. (1996). A new approach to the minimum cut problem. *J. ACM* **43** 601–640.

[38] KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

[39] KOLMOGOROV, V. and ZABIN, R. (2004). What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 147–159.

[40] KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statistics* **82**. Springer, New York.

[41] KOVAC, A. and SMITH, A. D. (2011). Nonparametric regression on a graph. *J. Comput. Graph. Statist.* **20** 432–447.

[42] LAND, S. R. and FRIEDMAN, J. H. (1997). Variable fusion: A new adaptive signal regression method. Technical Report 656, Dept. Statistics, Carnegie Mellon Univ., Pittsburgh, PA.

[43] LEBARBIER, É. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Process.* **85** 717–736.

[44] LIN, K., SHARPNACK, J., RINALDO, A. and TIBSHIRANI, R. J. (2016). Approximate recovery in changepoint problems, from ℓ_2 estimation error rates. Available at arXiv:1606.06746.

[45] LIVNE, O. E. and BRANDT, A. (2012). Lean algebraic multigrid (LAMG): Fast graph Laplacian linear solver. *SIAM J. Sci. Comput.* **34** B499–B522. MR2970416

[46] LOVÁSZ, L. (1996). Random walks on graphs: A survey. In *Combinatorics*: *Paul Erdős Is Eighty*, *Vol.* 2 (*Keszthely*, 1993). *Bolyai Soc. Math. Stud.* **2** 353–397. János Bolyai Math. Soc., Budapest. MR1395866

[47] MADRID PADILLA, O. H., SCOTT, J. G., SHARPNACK, J. and TIBSHIRANI, R. J. (2016). The DFS fused lasso: Nearly optimal linear-time denoising over graphs and trees. Available at arXiv:1608.03384.

[48] MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931

[49] MOORE, C. and NEWMAN, M. E. (2000). Epidemics and percolation in small-world networks. *Phys. Rev. E* **61** 5678–5682.

[50] MUMFORD, D. and SHAH, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* **42** 577–685.

[51] RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37** 2922–2952. MR2541451

[52] RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268. Experimental mathematics: Computational issues in nonlinear science (Los Alamos, NM, 1991). MR3363401

[53] SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. (2016). Graph sparsification approaches for Laplacian smoothing. In *Int. Conf. Artific. Intell. Statist.* 1250–1259.

[54] SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Adv. Neural Inform. Process. Syst.* 3513–3521.

[55] SHARPNACK, J., RINALDO, A. and SINGH, A. (2012). Sparsistency of the edge lasso over graphs. In *Int. Conf. Artific. Intell. Statist.* 1028–1036.

[56] SHARPNACK, J., SINGH, A. and RINALDO, A. (2013). Detecting activations over graphs using spanning tree wavelet bases. In *Int. Conf. Artific. Intell. Statist.* 545–553.

[57] SHARPNACK, J. L., KRISHNAMURTHY, A. and SINGH, A. (2013). Near-optimal anomaly detection in graphs using Lovasz extended scan statistic. In *Adv. Neural Inform. Process. Syst.* 1959–1967.

[58] SPIELMAN, D. A. and SRIVASTAVA, N. (2011). Graph sparsification by effective resistances. *SIAM J. Comput.* **40** 1913–1926.

[59] SPIELMAN, D. A. and TENG, S.-H. (2004). Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *ACM Symp. Theory Comput.* 81–90. ACM, New York.

[60] TANSEY, W. and SCOTT, J. G. (2015). A fast and flexible algorithm for the graph-fused lasso. Available at arXiv:1505.06475.

[61] TIAN, X. and TAYLOR, J. E. (2015). Selective inference with a randomized response. Available at arXiv:1507.06739.

[62] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288.

[63] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

[64] TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

[65] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[66] WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17** Paper No. 105. MR3543511

[67] WINKLER, G. and LIEBSCHER, V. (2002). Smoothers for discontinuous signals. *J. Nonparametr. Stat.* **14** 203–222.

[68] XIN, B., KAWAHARA, Y., WANG, Y. and GAO, W. (2014). Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease. In *Proc. Assoc. Adv. Artific. Intell. Conf.* 2163–2169.

[69] YAO, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Ann. Statist.* **12** 1434–1447. MR0760698

[70] YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373

[71] YAO, Y.-C. and AU, S.-T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613

[72] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conf. Learning Theory* **35** 1–28.

[73] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electron. J. Stat.* **11** 752–799.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: zhoufan@stanford.edu
lguan@stanford.edu