Approximate Recovery in Changepoint Problems, from ℓ_2 Estimation Error Rates

Kevin Lin[1] James Sharpnack[2] Alessandro Rinaldo[1]
Ryan J. Tibshirani[1]

[1]Carnegie Mellon University, [2]University of California at Davis

Abstract

In the 1-dimensional multiple changepoint detection problem, we prove that any procedure with a fast enough ℓ_2 error rate, in terms of its estimation of the underlying piecewise constant mean vector, automatically has an (approximate) changepoint screening property—specifically, each true jump in the underlying mean vector has an estimated jump nearby. We also show, again assuming only knowledge of the ℓ_2 error rate, that a simple post-processing step can be used to eliminate spurious estimated changepoints, and thus delivers an (approximate) changepoint recovery property—specifically, in addition to the screening property described above, we are assured that each estimated jump has a true jump nearby. As a special case, we focus on the application of these results to the 1-dimensional fused lasso, i.e., 1-dimensional total variation denoising, and compare the implications with existing results from the literature. We also study extensions to related problems, such as changepoint detection over graphs.

Keywords: *changepoint detection*, *fused lasso*, *total variation denoising*, *approximate recovery*

1 Introduction

Consider the 1-dimensional multiple changepoint model

$$y_i = \theta_{0,i} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ϵ_i, $i = 1, \ldots, n$ are i.i.d. errors, and $\theta_{0,i}$, $i = 1, \ldots, n$ is a piecewise constant mean sequence, having a set of changepoints

$$S_0 = \big\{ i \in \{1, \ldots, n-1\} : \theta_{0,i} \neq \theta_{0,i+1} \big\}.$$

This is a well-studied problem, and there is a large body of literature on estimation of the piecewise constant mean vector $\theta_0 \in \mathbb{R}^n$ in this model, as well as detection of its changepoints θ_0. Though estimation (of θ_0) and detection (of its changepoints) are clearly related pursuits, they are different enough that most works on the changepoint problem are focused on one or the other. For example, 1-dimensional total variation denoising (Rudin et al. 1992), i.e., the 1-dimensional fused lasso (Tibshirani et al. 2005), has been primarily studied from the perspective of its estimation properties. Meanwhile, segmentation methods like binary segmentation (Vostrikova 1981, Venkatraman 1992) and wild binary segmentation (Fryzlewicz 2014) have been mostly studied for their detection properties. In this paper, we assert that the estimation and detection problems are very closely linked, in the following sense: any estimator with ℓ_2 estimation error guarantees automatically has certain approximate changepoint detection guarantees, and not surprisingly, a faster ℓ_2 estimation error rate here translates into a stronger statement about approximate detection. We use this general link to establish new approximate changepoint recovery results for the 1d fused lasso, an estimator that is given central focus in our work.

1.1 Background and related work

Given a data vector $y \in \mathbb{R}^n$ from a model as in (1), the *1-dimensional fused lasso* (1d fused lasso, or simply fused lasso) estimate is defined by

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \; \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} |\theta_i - \theta_{i+1}|, \tag{2}$$

where $\lambda \geq 0$ serves as a tuning parameter. This was proposed by Tibshirani et al. (2005)[1], though the same idea had been proposed in signal processing much earlier, under the name *total variation (TV) denoising*, by Rudin et al. (1992). There has been plenty of statistical theory developed for the fused lasso, e.g., Mammen & van de Geer (1997), Davies & Kovac (2001), Rinaldo (2009), Harchaoui & Levy-Leduc (2010), Qian & Jia (2012), Rojas & Wahlberg (2014), Dalalyan et al. (2014). In particular, Mammen & van de Geer (1997) and Dalalyan et al. (2014) derived ℓ_2 error rates for the fused lasso, under different settings (different assumptions on θ_0). We review these in Section 2.1, and compare the latter to a new related result that we establish in Section 3. Harchaoui & Levy-Leduc (2010) and Rojas & Wahlberg (2014) derived approximate changepoint recovery properties for the fused lasso. We review these in Section 2.2, and compare them to our own results on approximate recovery in Section 4.

The literature on general approaches in multiple changepoint detection is enormous, and we do not give an extensive overview, but we do summarize some relevant work of Donoho & Johnstone (1994, 1998), Fryzlewicz (2007), Boysen et al. (2009), Fryzlewicz (2014), Frick et al. (2014), Fryzlewicz (2016) in Sections 2.1 and 2.2, and revisit these results in more detail in Section 4 when we compare them to our new approximate changepoint recovery results. Extensions of the fused lasso such as trend filtering and the graph-based fused lasso have been analyzed by, e.g, Sharpnack et al. (2012), Tibshirani (2014), Wang et al. (2016), Hutter & Rigollet (2016), which are discussed later in Section 6.

1.2 Notation

For a vector $\theta \in \mathbb{R}^n$, we write $S(\theta)$ for the set of its changepoint indices, i.e.,

$$S(\theta) = \big\{ i \in \{1, \dots, n-1\} : \theta_i \neq \theta_{i+1} \big\}.$$

We abbreviate $S_0 = S(\theta_0)$ and $\widehat{S} = S(\widehat{\theta})$ for the changepoints of the mean θ_0 in (1), and the fused lasso estimate $\widehat{\theta}$ in (2), respectively. Throughout, we will use the words "changepoint" and "jump" interchangeably. We will also make use of the following quantities defined in terms of θ_0. The size of S_0 is denoted $s_0 = |S_0|$. For convenience, we write $S_0 = \{t_1, \dots, t_{s_0}\}$, where $1 \leq t_1 < \dots < t_{s_0} < n$, and by convention, $t_0 = 0$, $t_{s_0+1} = n$. The smallest distance between jumps in θ_0 is denoted by

$$W_n = \min_{i=0,1\dots,s_0} (t_{i+1} - t_i), \tag{3}$$

and the smallest distance between consecutive levels of θ_0 by

$$H_n = \min_{i \in S_0} |\theta_{0,i+1} - \theta_{0,i}|. \tag{4}$$

Our notation here makes the dependence of W_n, H_n on n explicit (of course stemming from the fact that the mean vector θ_0 itself changes with n, though for simplicity we suppress this notationally.)

For a matrix $D \in \mathbb{R}^{m \times n}$, we write D_S to extract rows of D indexed by a subset $S \subseteq \{1, \dots, m\}$, and D_{-S} as shorthand for D_{-S}, where $-S = \{1, \dots, m\} \setminus S$. Unless otherwise specified, the notation

[1] In Tibshirani et al. (2005), the authors actually used an additional ℓ_1 penalty on θ itself, to induce componentwise sparsity; here we do not consider this extension, and simply refer to the estimator in (2) as the fused lasso.

$D \in \mathbb{R}^{(n-1)\times n}$ will be used to denote the difference operator

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{5}$$

For a vector $x \in \mathbb{R}^n$, we define its scaled ℓ_2 norm $\|x\|_n = \|x\|_2/\sqrt{n}$, and its discrete total variation

$$\mathrm{TV}(x) = \sum_{i=1}^{n-1} |x_i - x_{i+1}| = \|Dx\|_1. \tag{6}$$

For two discrete sets A, B, we define the metrics

$$d(A|B) = \max_{b\in B} \min_{a\in A} |a-b| \quad \text{and} \quad d_H(A,B) = \max\big\{d(A|B), d(B|A)\big\}. \tag{7}$$

The former metric can seen as a one-sided screening distance from B to A, measuring the furthest distance of an element in B to its closest element in A. The latter metric is traditionally known as the Hausdorff distance between A and B. Note that if A is empty, then we have $d(A|B) = 0$, and if B is empty, then $d(A|B) = \infty$; this makes $d_H(A,B) = \infty$ if either A or B are empty.

For deterministic sequences a_n, b_n we write $a_n = O(b_n)$ to denote that a_n/b_n is bounded for large enough n, $a_n = \Omega(b_n)$ to denote that b_n/a_n is bounded for large enough n, and $a_n = \Theta(b_n)$ to denote that both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We also write $a_n = o(b_n)$ to denote that $a_n/b_n \to 0$, and $a_n = \omega(b_n)$ to denote that $b_n/a_n \to 0$. Finally, we write $A_n = O_{\mathbb{P}}(B_n)$ for random sequences A_n, B_n to denote that A_n/B_n is bounded in probability, and $A_n = o_{\mathbb{P}}(B_n)$ to denote that $A_n/B_n \to 0$ in probability.

1.3 Summary of results

A summary of our contributions is as follows.

- **New ℓ_2 error analysis for the fused lasso, under strong sparsity.** In Section 3, we give a new ℓ_2 estimation error analysis for the fused lasso, in the case $s_0 = O(1)$, which we refer to as the "strong sparsity" case. Theorem 4 provides the bound

  $$\|\widehat{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}\bigg(\frac{\log n \log\log n}{n}\bigg),$$

  for the fused lasso estimate $\widehat{\theta}$ in (2). This is sharper than the previously established error rate of $\log^2 n/n$, from Dalalyan et al. (2014), for the fused lasso under strong sparsity, and quite close to the "oracle" rate of $\log n/n$ under strong sparsity, as we discuss in Remark 10. Our theorem also applies beyond the case of a constant sparsity level s_0, and gives an explicit error bound in terms of s_0. We believe that the proof of Theorem 4 is interesting in its own right, as it leverages a new quantity that we call a *lower interpolant* to approximate the fused lasso estimate in a certain sense using $2s_0 + 2$ piecewise monotonic segments, which allows for finer control of the sub-Gaussian complexity.

- **Bound on the screening distance, based on ℓ_2 error.** In Section 4.1, we derive a bound on the sreening distance from S_0 to the detected changepoints $S(\widetilde{\theta})$ of any estimator $\widetilde{\theta}$, given a bound on its ℓ_2 error rate $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. Specifically, in Theorem 8, we show that

  $$d\big(S(\widetilde{\theta}) \,|\, S_0\big) = O_{\mathbb{P}}\bigg(\frac{nR_n}{H_n^2}\bigg).$$

To emphasize, this bound on the screening distance is agnostic about the details of the estimator $\widetilde{\theta}$, provided that its ℓ_2 error rate R_n is known. As two principal applications, we plug in the known error rate R_n for the fused lasso under two different settings—weak and strong sparsity—to derive new screening results on the fused lasso in Corollaries 9 and 10. Perhaps surprisingly (since these screening bounds are not based on fine-grained analysis of the fused lasso, but on achieved ℓ_2 rates alone), these results provide interesting conclusions in each of their own settings, as we discuss in Remarks 14 and 15.

- **Bound on the Hausdorff distance, based on ℓ_2 error and a post-processing step.** In Section 4.2, we give a bound on the Hausdorff distance between S_0 and the detected changepoints $S(\widetilde{\theta})$ of any estimator $\widetilde{\theta}$, given a bound $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$ and a simple filtering-based technique to remove spurious changepoints in $S(\widetilde{\theta})$ that occur far away from elements of S_0. In particular, Theorem 11 states that the filtered set $S_F(\widetilde{\theta})$ of changepoints satisfies

$$\mathbb{P}\left(d_H\big(S_F(\widetilde{\theta}), S_0\big) \leq \frac{nR_n\nu_n}{H_n^2} \right) \to 1 \quad \text{as } n \to \infty,$$

  where ν_n is any diverging sequence (i.e., diverging as slowly as desirable). As two applications, we consider post-processing the changepoints from the fused lasso estimator, in the weak and strong sparsity settings, in Corollaries 14 and 15. We compare these to existing approximate changepoint recovery results in the literature in Remarks 19 and 20; the summary is that under strong sparsity, our result on the post-processed fused lasso is comparable with the best known recovery results, but under weak sparsity, our result is worse than the guarantees given in Frick et al. (2014), Fryzlewicz (2014, 2016) for other changepoint estimators. It should be reiterated that, unlike other results in the literature which are based on detailed analyses of specific changepoint estimators, our results are generic and based only on ℓ_2 error properties, making them widely applicable. Therefore, a lack in sharpness in some cases, such as the weak sparsity case, is perhaps not unexpected.

- **Practical guidelines for post-processing.** In Section 4.3, we present a modification of the aforementioned post-processing rule, which guarantees that the filtered set has at most $3\widetilde{s} + 2$ elements, where $\widetilde{s} = |S(\widetilde{\theta})|$. In Section 5, we describe a data-driven procedure to determine an appropriate threshold level for the filter, and we also conduct detailed empirical investigations of our proposals.

- **Extension to piecewise linear segmentation, and graph changepoint detection.** In Section 6, we give extensions of our screening results to two related settings: piecewise linear segmentation and changepoint detection over graphs. For piecewise linear segmentation, the main screening result is in Theorem 19, and its specialization to the trend filtering estimator is in Corollary 21; for graph changepoint detection, the main result is in Theorem 23, and its specialization to the 2d fused lasso estimator is in Corollary 24.

2 Preliminary review of existing theory

We review existing statistical theory for the fused lasso, first on ℓ_2 estimation error, and then on (approximate) changepoint recovery.

2.1 Review: ℓ_2 estimation error

We begin by describing two major results on the quantity

$$\|\widehat{\theta} - \theta_0\|_n^2,$$

the squared ℓ_2 estimation error between the fused lasso estimate $\widehat{\theta}$ in (2) and the mean θ_0 in (1). In somewhat of an abuse of notation, we will simply refer to the above quantity as the ℓ_2 estimation error, or estimation error for short.

The first result, from Mammen & van de Geer (1997), studies what may be called the "weak sparsity" case, in which the total variation of θ_0 is controlled. Before stating this, we recall that a random variable Z is said to have a mean zero sub-Gaussian distribution provided that

$$\mathbb{E}(Z) = 0 \quad \text{and} \quad \mathbb{P}(|Z| > t) \leq M \exp\big(-t^2/(2\sigma^2)\big) \quad \text{for } t \geq 0, \tag{8}$$

for some constants $M, \sigma > 0$.

Theorem 1 (Fused lasso error rate, weak sparsity setting, Theorem 10 of Mammen & van de Geer 1997). *Assume the data model in (1), with errors ϵ_i, $i = 1, \ldots, n$ i.i.d. from a sub-Gaussian distribution as in (8). Also assume that $\mathrm{TV}(\theta_0) \leq C_n$, for a nondecreasing sequence C_n. Then for a choice of tuning parameter $\lambda = \Theta(n^{1/3} C_n^{-1/3})$, the fused lasso estimate $\widehat{\theta}$ in (2) satisfies*

$$\|\widehat{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(n^{-2/3} C_n^{2/3}).$$

Remark 1 (Consistency, optimality). This shows that the fused lasso estimator is consistent when $C_n = o(n)$. When $C_n = O(1)$, its estimation error rate is $n^{-2/3}$, which is in fact the minimax optimal rate as θ_0 varies over the class of signals with bounded total variation, i.e., $\theta_0 \in \{\theta \in \mathbb{R}^n : \mathrm{TV}(\theta) \leq C\}$ for a constant $C > 0$ (Donoho & Johnstone 1998). For explanations of the above theorem and this minimax result, in notation that is more consistent with that of the current paper, see Tibshirani (2014).

The second result, from Dalalyan et al. (2014), studies what may be called the "strong sparsity" case, in which the number of changepoints s_0 in θ_0 is controlled.

Theorem 2 (Fused lasso error rate, strong sparsity setting, Proposition 4 of Dalalyan et al. 2014). *Assume the data model in (1), with errors ϵ_i, $i = 1, \ldots, n$ drawn i.i.d. from $N(0, \sigma^2)$. Then for a choice of tuning parameter $\lambda = \sqrt{2n \log(n/\delta)}$, the fused lasso estimate $\widehat{\theta}$ in (2) satisfies*

$$\|\widehat{\theta} - \theta_0\|_n^2 \leq c \frac{s_0 \log(n/\delta)}{n} \left(\log n + \frac{n}{W_n} \right),$$

with probability at least $1 - 2\delta$, for all $\delta > 0$ and all $n \geq N$, where $c, N > 0$ are constants, and recall W_n is the minimum distance between jumps in θ_0, as in (3).

Remark 2 (The roles of s_0, W_n). When the number of jumps s_0 in θ_0 grows quickly enough with n, the error rate in Theorem 2 will become worse than that in Theorem 1. Given s_0 jumps, in the best case, the minimum gap W_n between jumps scales as $W_n = \Theta(n/s_0)$, which delivers a rate of $s_0 \log^2 n/n + s_0^2 \log n/n$ in Theorem 2. When s_0 scales faster than $n^{1/6} (\log n)^{-1/2}$, we can see that this is slower than the $n^{-2/3}$ rate delivered by Theorem 1 (assuming $C_n = O(1)$).

Of course, Theorem 2 is most useful when $s_0 = O(1)$. When this is true, and additionally $W_n = \Theta(n)$, we see that the theorem implies that the fused lasso has ℓ_2 error $\|\widehat{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(\log^2 n/n)$. This is a very fast rate, nearly equal to the "parametric rate" of $1/n$ associated with estimating a finite-dimensional parameter.

Remark 3 (Alternative fused lasso error rate, strong sparsity setting). When $s_0 = O(1)$, Proposition 2 in Harchaoui & Levy-Leduc (2010) proves that the fused lasso has estimation error $\|\widehat{\theta} - \theta_0\|_n^2 = O(\log n/n)$ with probability approaching 1, under a choice $\lambda = \Theta(\sqrt{\log n/n^3})$. But the authors must also assume that the number of changepoints in the fused lasso estimate $\widehat{\theta}$, which we might denote as $\widehat{s} = |\widehat{S}|$, is bounded with probability tending to 1. This seems to be an unrealistic assumption, given the required scaling for λ. Theoretically, we remark that such a small choice of λ, on the order of $\sqrt{\log n/n^3}$, does not match the much larger choices dictated by Theorems 2 and 4, both on the order of approximately $\sqrt{n}$. Empirically, when λ scales as $\sqrt{\log n/n^3}$, we find that the number of estimated changepoints in $\widehat{\theta}$ often grows very large, even when θ_0 has few jumps and the signal-to-noise ratio is quite high.

Remark 4 (**Comparable error rates of other estimators**)**.** Various other estimators obtain comparable estimation error rates to those descibed above for the fused lasso. The Potts estimator, defined by replacing the ℓ_1 penalty $\sum_{i=1}^{n-1} |\theta_i - \theta_{i+1}|$ in (2) with the ℓ_0 penalty $\sum_{i=1}^{n-1} 1\{\theta_i \neq \theta_{i+1}\}$, and denoted say by $\widehat{\theta}^{\text{Potts}}$, has been shown to satisfy $\|\widehat{\theta}^{\text{Potts}} - \theta_0\|_n^2 = O((\log n/n)^{2/3})$ a.s. when $\text{TV}(\theta_0) = O(1)$, and $\|\widehat{\theta}^{\text{Potts}} - \theta_0\|_n^2 = O(\log n/n)$ a.s. when $s_0 = O(1)$, by Boysen et al. (2009). Wavelet denoising (under weak conditions on the wavelet basis), denoted by $\widehat{\theta}^{\text{wav}}$, has been shown to satisfy $\mathbb{E}\|\widehat{\theta}^{\text{wav}} - \theta_0\|_n^2 = O(n^{-2/3})$ when $\text{TV}(\theta_0) = O(1)$, by Donoho & Johnstone (1998), and $\mathbb{E}\|\widehat{\theta}^{\text{wav}} - \theta_0\|_n^2 = O(\log^2 n/n)$ when $s_0 = O(1)$, by Donoho & Johnstone (1994). Combining unbalanced Haar (UH) wavelets with a basis selection method, Fryzlewicz (2007) gave an estimator $\widehat{\theta}^{\text{UH}}$ with $\mathbb{E}\|\widehat{\theta}^{\text{UH}} - \theta_0\|_n^2 = O(\log^2 n/n^{2/3})$ when $\text{TV}(\theta_0) = O(1)$, and $\mathbb{E}\|\widehat{\theta}^{\text{UH}} - \theta_0\|_n^2 = O(\log^2 n/n)$ when $s_0 = O(1)$. Though they are not written in this form, the results in Fryzlewicz (2016) imply that his "tail-greedy" unbalanced Haar (TGUH) estimator, $\widehat{\theta}^{\text{TGUH}}$, satisfies $\|\widehat{\theta}^{\text{TGUH}} - \theta_0\|_n^2 = O(\log^2 n/n)$ with probability tending to 1, when $s_0 = O(1)$.

2.2 Review: changepoint recovery

Next, we review the relevant results on the quantities

$$d(\widehat{S}\,|\,S_0) \quad \text{or} \quad d_H(\widehat{S}, S_0).$$

The former is the screening distance from S_0 to the set of changepoints $\widehat{S} = S(\widehat{\theta})$ in the fused lasso estimate $\widehat{\theta}$ in (2); the latter is the Hausdorff distance between S_0 and $\widehat{S}$; recall, both metrics were defined in (7). We use the term "approximate screening" to mean that $d(\widehat{S}\,|\,S_0)$ is controlled, and "approximate recovery" to mean that $d_H(\widehat{S}, S_0)$ is controlled, though often times we will drop the word "approximate" from either term, for brevity. Below we summarize two results from Harchaoui & Levy-Leduc (2010).

Theorem 3 (**Fused lasso approximate screening and recovery results, strong sparsity setting, Propositions 3 and 4 of Harchaoui & Levy-Leduc 2010**)**.** *Assume the data model in (1), where the errors ϵ_i, $i = 1, \dots, n$ are i.i.d. from a sub-Gaussian distribution as in (8). Assume also that (i) $s_0 = O(1)$, (ii) $\mathbb{P}(\widehat{s} \geq s_0) \to 1$, and that r_n is a sequence satisfying (iii) $r_n \leq W_n$, (iv) $r_n = \omega(\max\{\log n/H_n^2, \lambda/H_n\})$, where, recall, W_n is the minimum distance between changepoints in θ_0, as defined in (3), and H_n is the minimum gap between levels of θ_0, as defined in (4). Then the fused lasso estimate $\widehat{\theta}$ in (2) with tuning parameter λ satisfies*

$$\mathbb{P}\Big(d(\widehat{S}\,|\,S_0) \leq r_n\Big) \to 1 \quad \textit{as} \quad n \to \infty.$$

Under assumptions (i), (ii') $\mathbb{P}(\widehat{s} = s_0) \to 1$, (iii), and (iv') $r_n = \omega(\max\{\log n/H_n^2, \log(n^5/\lambda^2)/H_n^2\})$, we instead have

$$\mathbb{P}\Big(d_H(\widehat{S}, S_0) \leq r_n\Big) \to 1 \quad \textit{as} \quad n \to \infty.$$

Remark 5 (**Stringency of conditions**)**.** Assumption (ii') in the above result, needed for the bound on the Hausdorff distance, states that the number of estimated changepoints $\widehat{s}$ in $\widehat{\theta}$ equals the number of changepoints s_0 in θ_0 with probability tending to 1, which is of course a very strong assumption. Assumption (ii), needed for the bound on the one-sided screening distance, states that $\widehat{s} \geq s_0$, which is itself fairly strong, though believable if λ is chosen to be small enough.

The tuning parameter λ in fact plays an important role in the achieved rates r_n in Theorem 3. In order to satisfy condition (iv) on r_n, a choice of $\lambda = \log n/H_n$ gives the tightest possible scaling for r_n. Then we require r_n to grow faster than $\log n$, e.g., at the rate $\log^2 n$. This choice is basically the same as that discussed in Harchaoui & Levy-Leduc (2010); with this choice, their results show that the screening distance achieved by the fused lasso estimator is at most $\log^2 n$, with probability tending to 1. However, the choice of λ here is worrisome—it is considerably smaller than the choices known to achieve reasonable error rates, specifically in the strong sparsity setting, with $s_0 = O(1)$, where we expect λ to scale at something like a $\sqrt{n}$ rate (see Theorems 2 and 4). With such a small choice of $\lambda = \log n / H_n$, there would likely be a very large number of estimated changepoints in $\widehat{\theta}$, rendering a quantity like the screening distance uninteresting.[2] The same critique could made be about condition (iv'), needed for the bound on the Hausdorff distance. In particular, condition (ii') seems unrealistic unless λ is chosen to be much larger.

For the reasons just described, we will consider a larger scaling of $\lambda = \Theta(\sqrt{n})$, when comparing Theorem 3 to our new results on changepoint screening and recovery in Sections 4.1 and 4.2.

Remark 6 (**Other fused lasso recovery results**)**.** Several other results have appeared in the literature regarding changepoint recovery for the fused lasso. Rinaldo (2009) studied exact recovery of changepoints (in which the achieved Hausdorff distance would be zero). There is an error in the proof of his Theorem 2.3, which invalidates the result.[3] Qian & Jia (2012) studied a modification of the fused lasso defined by transforming the fused lasso problem (2) into a lasso problem with particular design matrix X, and then applying a step that "preconditions" y and X. The authors concluded that exact recovery is possible with probability at tending to 1, as long as the minimum signal gap and tuning parameter satisfy $H_n \geq \lambda = \omega(\sqrt{\log n})$. This is a very strong requirement on the scaling of the signal gap H_n; Sharpnack et al. (2012) showed that, when $H_n \geq \omega(\sqrt{\log n})$, even simple pairwise thresholding (i.e., thresholding based on the observed absolute differences $|y_i - y_{i+1}|$, $i = 1, \ldots, n-1$) achieves exact recovery. Most recently, Rojas & Wahlberg (2014) established an impossibility result for the fused lasso estimator when θ_0 exhibits a "staircase" pattern, which means that $D_{S_0}\theta_0$ has two consecutive positive or negative values; specifically, these authors proved that for such a staircase pattern, the quantity $d_H(\widehat{S}, S_0)/n$ remains bounded away from zero with nonzero asymptotic probability. For non-staircase patterns in θ_0, the authors also showed, under certain assumptions, that $d_H(\widehat{S}, S_0)/n$ converges to zero in probability.

Remark 7 (**Comparable recovery properties of other estimators**)**.** It is worth describing relevant changepoint recovery properties of various methods in the literature. Boysen et al. (2009) showed that the Potts estimator, denoted by $\widehat{\theta}^{\mathrm{Potts}}$, satisfies $d_H(S(\widehat{\theta}^{\mathrm{Potts}}), S_0) = O(\log n)$ a.s., when $s_0 = O(1)$. Frick et al. (2014) proposed a simultaneous multiscale changepoint estimator (SMUCE), using an ℓ_0-penalized optimization problem (like the Potts estimator), and under weak assumptions on W_n, H_n, proved that their estimator $\widehat{\theta}^{\mathrm{SMUCE}}$ satisfies $d_H(S(\widehat{\theta}^{\mathrm{SMUCE}}), S_0) = O(\log n / H_n^2)$ with probability tending to 1. There is quite a large body of literature on binary segmentation (BS). To the best of our knowledge, the sharpest analysis for BS is in Fryzlewicz (2014), who also proposed and analyzed a "wild" (i.e., randomized) variant of the method (WBS). Denoting these two estimators by $\widehat{\theta}^{\mathrm{BS}}$ and $\widehat{\theta}^{\mathrm{WBS}}$, Fryzlewicz (2014) established that $d_H(S(\widehat{\theta}^{\mathrm{BS}}), S_0) = O(n^2 \log n / (W_n^2 H_n^2))$ and $d_H(S(\widehat{\theta}^{\mathrm{BS}}), S_0) = O(\log n / H_n^2)$, both with probability tending to 1, and both under certain restrictions on W_n, H_n, these restrictions being stronger for BS than for WBS. Very recently, Fryzlewicz (2016) proved that his tail-greedy unbiased Haar estimator, denoted by $\widehat{\theta}^{\mathrm{TGUH}}$, satisfies $d_H(S(\widehat{\theta}^{\mathrm{TGUH}}), S_0) = O(\log^2 n)$ with probability tending to 1, under weak conditions on W_n, H_n. All of these results will be revisited in greater detail in Remarks 19 and 20.

Lastly, it should be noted that many of the methods described here also come with a guarantee (under possibly additional conditions) that they correctly identify the number of changepoints s_0 in θ_0, with probability tending to 1. We refer the reader to the references above, for details.

[2]Of course, with $\lambda = 0$, the screening distance achieved by the fused lasso is trivially zero. Hence, when studying screening distance, it is implicitly understood that some other aspect of $\widehat{\theta}$ must be kept in balance. In our work, we study screening distances while maintaining that $\widehat{\theta}$ must exhibit good estimation performance, as measured by its ℓ_2 error rate.

[3]See the correction note posted at `http://www.stat.cmu.edu/~arinaldo/Fused_Correction.pdf`.

3 Error analysis under strong sparsity

In this section, we derive a new ℓ_2 estimation error bound for the fused lasso in the strong sparsity case, improving on the result of Dalalyan et al. (2014) stated in Theorem 2. Our proof is based on the concept of a *lower interpolant*, which as far as we can tell is a new idea that may be of interest in its own right. We first state our error bound.

Theorem 4 (**Fused lasso error rate, strong sparsity setting**). *Assume the data model in (1), with errors ϵ_i, $i = 1, \ldots, n$ i.i.d. from a sub-Gaussian distribution as in (8). Then for a choice of tuning parameter $\lambda = (nW_n)^{1/4}$, the the fused lasso estimate $\widehat{\theta}$ in (2) satisfies*

$$\|\widehat{\theta} - \theta_0\|_n^2 \le \gamma^2 c \frac{s_0}{n} \left((\log s_0 + \log\log n) \log n + \sqrt{\frac{n}{W_n}} \right),$$

with probability at least $1 - \exp(-C\gamma)$, for all $\gamma > 1$ and $n \ge N$, where $c, C, N > 0$ are constants.

Remark 8 (**The roles of s_0, W_n**). When s_0 grows quickly enough with n, the error rate provided by the above theorem will become worse than the weak sparsity rate in Theorem 1. Given s_0 evenly spaced jumps, so that $W_n = \Theta(n/s_0)$, the rate in Theorem 4 is $(\log s_0 + \log\log n) s_0 \log n/n + s_0^{3/2}/n$; when s_0 grows faster than $n^{2/9}$, this is slower than the $n^{-2/3}$ rate in Theorem 1 (assuming $C_n = O(1)$). Theorem 4 gives the fastest rate when $s_0 = O(1)$, $W_n = \Theta(n)$, this being $(\log n \log\log n)/n)$, an improvement over the rate in Theorem 2. This comparison, and the comparison to other results in the literature, will be drawn out in more detail in the last remark of this section.

Remark 9 (**Expectation bound**). An expectation bound follows more of less directly from the high probability bound in Theorem 4. Define the random vairable

$$M = \frac{\|\widehat{\theta} - \theta_0\|_2^2}{c^2 s_0 ((\log s_0 + \log\log n) \log n + \sqrt{n/W_n})},$$

which we know has the tail bound $\mathbb{P}(U > u) \le \exp(-C\sqrt{u})$ for $u > 1$, and observe that

$$\mathbb{E}(U) = \int_0^\infty \mathbb{P}(U > u)\, du \le 1 + \int_1^\infty \exp(-C\sqrt{u})\, du.$$

The right-hand side is a finite constant, and this gives the result

$$\mathbb{E}\|\widehat{\theta} - \theta_0\|_n^2 \le c \frac{s_0}{n} \left((\log s_0 + \log\log n) \log n + \sqrt{\frac{n}{W_n}} \right),$$

where the constant $c > 0$ is adjusted to be larger, as needed.

Here is an overview of the proof of Theorem 4. The details are deferred until Appendix A, and the proofs of the lemmas stated below are given in Appendix B. We consider a decomposition

$$\widehat{\theta} - \theta_0 = P_0(\widehat{\theta} - \theta_0) + P_1 \widehat{\theta},$$

where P_0 is the projection matrix onto the piecewise constant structure inherent to the mean θ_0, and $P_1 = I - P_1$. To give more detail, recall that we write $S_0 = \{t_1, \ldots, t_{s_0}\}$ for the changepoints in θ_0, ordered as in $t_1 < \ldots < t_{s_0}$, and we write $t_0 = 0$ and $t_{s_0+1} = n$ for convenience. Furthermore, define $B_j = \{t_j + 1, \ldots, t_{j+1}\}$, and write $\mathbb{1}_{B_j} \in \mathbb{R}^n$ for the indicator of block B_j, for $j = 0, \ldots, s_0$. With this notational setup, we may now define P_0 as the projection onto the $(s_0 + 1)$-dimensional linear subspace $\mathcal{R} = \mathrm{span}\{\mathbb{1}_{B_0}, \ldots, \mathbb{1}_{B_{s_0}}\}$. It is common practice (e.g., see van de Geer (2000)) to bound the estimation error by bounding the empirical process term $\epsilon^\top(\widehat{\theta}-\theta_0)$, where $\epsilon \in \mathbb{R}^n$ is the vector of errors in the data model (1). Using the decomposition above, this becomes

$$\epsilon^\top(\widehat{\theta}-\theta_0) = \epsilon^\top\widehat{\delta} + \epsilon^\top\widehat{x},$$

where we define $\widehat{\delta} = P_0(\widehat{\theta}-\theta_0)$ and $\widehat{x} = P_1\widehat{\theta}$. The parameter $\widehat{\delta}$ lies in an (s_0+1)-dimensional space, which makes bounding $\epsilon^\top\widehat{\delta}$ relatively easy. Bounding the term $\epsilon^\top\widehat{x}$ requires a much more intricate argument, which is spelled out in the following lemmas. Lemma 5 is a deterministic result ensuring the existence of what we call the *lower interpolant* $\widehat{z}$ to the vector $\widehat{x}$. This interpolant approximates $\widehat{x}$ using roughly $2s_0+2$ monotonic pieces, and its empirical process term $\epsilon^\top\widehat{z}$ can be finely controlled, as shown in Lemma 6. The residual from the interpolant approximation, denoted $\widehat{w} = \widehat{x}-\widehat{z}$, has an empirical process term $\epsilon^\top\widehat{w}$ that is more crudely controlled, in Lemma 7. Put together, as in $\epsilon^\top\widehat{x} = \epsilon^\top\widehat{z} + \epsilon^\top\widehat{w}$, gives the final control on $\epsilon^\top\widehat{x}$.

Before stating Lemma 5, we define the class of vectors containing the lower interpolant. Given any collection of changepoints $t_1 < \ldots < t_{s_0}$ (and $t_0 = 0$, $t_{s_0+1} = n$), let $\mathcal{M}$ be the set of "piecewise monotonic" vectors $z \in \mathbb{R}^n$, with the following properties, for each $i = 0,\ldots,s_0$:

(i) there exists a point t_i' such that $t_i + 1 \le t_i' \le t_{i+1}$, and the absolute value $|z_j|$ is nonincreasing over the segment $j \in \{t_i+1,\ldots,t_i'\}$, and nondecreasing over the segment $j \in \{t_i',\ldots,t_{i+1}\}$;

(ii) the signs remain constant on the monotone pieces,

$$\begin{aligned} &\mathrm{sign}(z_{t_i}) \cdot \mathrm{sign}(z_j) \ge 0, \quad j = t_i+1,\ldots,t_i', \\ &\mathrm{sign}(z_{t_{i+1}}) \cdot \mathrm{sign}(z_j) \ge 0, \quad j = t_i'+1,\ldots,t_{i+1}. \end{aligned}$$

Now we state our lemma that characterizes the lower interpolant.

Lemma 5. *Given changepoints $t_0 < \ldots < t_{s_0+1}$, and any $x \in \mathbb{R}^n$, there exists a vector $z \in \mathcal{M}$ (not necessarily unique), such that the following statements hold:*

$$\|D_{-S_0}x\|_1 = \|D_{-S_0}z\|_1 + \|D_{-S_0}(x-z)\|_1, \tag{9}$$

$$\|D_{S_0}x\|_1 = \|D_{S_0}z\|_1 \le \|D_{-S_0}z\|_1 + \frac{4\sqrt{s_0}}{\sqrt{W_n}}\|z\|_2, \tag{10}$$

$$\|z\|_2 \le \|x\|_2 \quad \text{and} \quad \|x-z\|_2 \le \|x\|_2, \tag{11}$$

where $D \in \mathbb{R}^{(n-1)\times n}$ is the difference matrix in (5). We call a vector z with these properties a lower interpolant *to x.*

Loosely speaking, the lower interpolant $\widehat{z}$ can be visualized by taking a string that lies initially on top of $\widehat{x}$, is nailed down at the changepoints $t_0,\ldots t_{s_0+1}$, and then pulled taut while maintaining that it is not greater (elementwise) than $\widehat{x}$, in magnitude. Here "pulling taut" means that $\|D\widehat{z}\|_1$ is made small. Figure 1 provides illustrations of the interpolant $\widehat{z}$ to $\widehat{x}$ for a few examples.

Note that $\widehat{z}$ consists of $2s_0+2$ monotonic pieces. This special structure leads to a sharp concentration inequality. The next lemma is the primary contributor to the fast rate given in Theorem 4.

Lemma 6. *Given changepoints $t_1 < \ldots < t_{s_0}$, there exists constants $c_I, C_I, N_I > 0$ such that when $\epsilon \in \mathbb{R}^n$ has i.i.d. sub-Gaussian components satisfying (8),*

$$\mathbb{P}\left(\sup_{z\in\mathcal{M}} \frac{|\epsilon^\top z|}{\|z\|_2} > \gamma c_I \sqrt{(\log s_0 + \log\log n)s_0\log n}\right) \le 2\exp\big(-C_I\gamma^2 c_I^2(\log s_0 + \log\log n)\big),$$

for any $\gamma > 1$, and $n \ge N_I$.

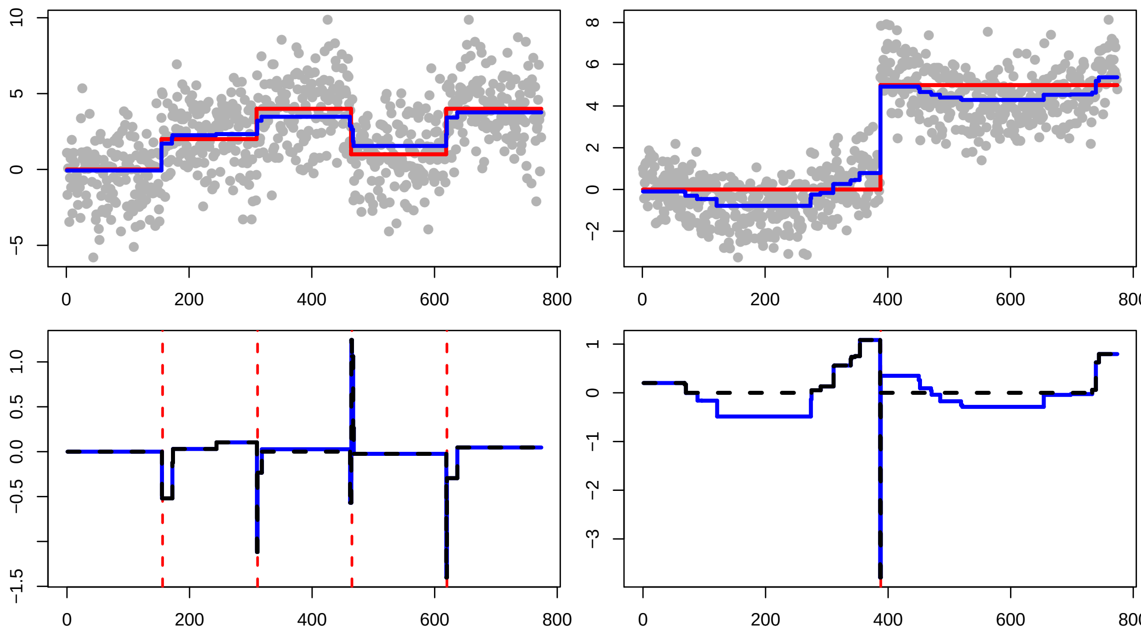

Figure 1: *The lower interpolants for two examples (in the left and right columns), each with $n = 800$ points. In the top row, the data y (in gray) and underlying signal θ_0 (red) are plotted across the locations $1, \ldots, n$. Also shown is the fused lasso estimate $\widehat{\theta}$ (blue). In the bottom row, the error vector $\widehat{x} = P_1\widehat{\theta}$ is plotted (blue) as well as the interpolant (black), and the dotted vertical lines (red) denote the changepoints $t_1, \ldots t_{s_0}$ of θ_0.*

Finally, the following lemma controls the residuals, $\widehat{w} = \widehat{x} - \widehat{z}$.

Lemma 7. *Given changepoints $t_1 < \ldots < t_{s_0}$, there exists constants $c_R, C_R > 0$ such that when $\epsilon \in \mathbb{R}^n$ has i.i.d. sub-Gaussian components satisfying (8),*

$$\mathbb{P}\left(\sup_{w \in \mathcal{R}^\perp} \frac{|\epsilon^\top w|}{\sqrt{\|D_{-S_0} w\|_1 \|w\|_2}} > \gamma c_R (n s_0)^{1/4} \right) \leq 2 \exp(-C_R \gamma^2 c_R^2 \sqrt{s_0}),$$

for any $\gamma > 1$, where $\mathcal{R}^\perp$ is the orthogonal complement of $\mathcal{R} = \mathrm{span}\{\mathbb{1}_{B_0}, \ldots, \mathbb{1}_{B_{s_0}}\}$.

We conclude this section with a remark comparing Theorem 4 to related results in the literature.

Remark 10 (Comparison to Theorem 2). We compare Theorem 4 to the strong sparsity result in Dalalyan et al. (2014), as stated in Theorem 2. For any s_0, W_n, the former rate is sharper than the latter, since $\log s_0 \leq \log n$ and $\sqrt{n/W_n} \leq n/W_n$. Moreover, when $s_0 = O(1)$ and $W_n = \Theta(n)$, the rates are $(\log n \log\log n)/n$ versus $\log^2 n/n$, in Theorems 4 and 2 respectively. Furthermore, in this setting, we note that the scaling of the tuning parameter investigated by both theorems is $\lambda = \Theta(\sqrt{n})$.

As discussed in Remarks 3 and 4, essentially all rates from comparable estimators in the strong sparsity case scale as $\log^2 n/n$, with the exception being the Potts estimator, which has a rate of $\log n/n$. Therefore the improvement from $\log^2 n/n$ to $(\log n \log\log n)/n$ offered by Theorem 4 could certainly be viewed as nontrivial. An error rate faster than $\log n/n$ in the strong sparsity case seems likely unattainable by any method, as Donoho & Johnstone (1994) showed that an oracle wavelet estimator (that is allowed the optimal choice of wavelet threshold for each problem instance) still has an expected estimation error on the order of $\log n/n$.

4 Approximate changepoint screening and recovery

We develop results on approximate screening of changepoints by the fused lasso, and approximate recovery of changepoints after a post-processing step has been applied to the fused lasso estimate. A distinctive feature of our results is that their proofs based on only the ℓ_2 estimation error rates achieved by the fused lasso. In fact, in their most general form, our results imply certain changepoint screening and recovery properties for any estimation method that has a known ℓ_2 error rate, which clearly has implications well beyond the fused lasso.

4.1 Results on approximate changepoint screening

We present a theorem that takes a general estimator $\widetilde{\theta}$ of θ_0, with a known ℓ_2 estimation error rate, and infers a bound on the screening distance between changepoints of θ_0 and those of $\widetilde{\theta}$.

Theorem 8 (Generic screening result). *Let $\theta_0 \in \mathbb{R}^n$ be a piecewise constant vector, and $\widetilde{\theta} \in \mathbb{R}^n$ be an estimator that satisfies the error bound $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. Assume that $nR_n/H_n^2 = o(W_n)$, where, recall, H_n is the minimum gap between adjacent levels of θ_0, defined in (4), and W_n is the minimum distance between adjacent changepoints of θ_0, defined in (3). Then*

$$d\big(S(\widetilde{\theta}) \,|\, S_0\big) = O_{\mathbb{P}}\bigg(\frac{nR_n}{H_n^2}\bigg),$$

where $S(\widetilde{\theta})$ is the set of changepoints in $\widetilde{\theta}$, $S_0 = S(\theta_0)$ is the set of changepoints in θ_0, and $d(\,\cdot\,|\,\cdot\,)$ is the one-sided screening distance, as defined in (7).

Proof. The proof is derived from the ℓ_2 rate. Fix any $\epsilon > 0$, $C > 0$. By assumption, we know that there is an integer $N_1 > 0$ such that

$$\mathbb{P}\bigg(\|\widetilde{\theta} - \theta_0\|_n^2 > \frac{C}{4} R_n\bigg) \le \epsilon,$$

for all $n \ge N_1$. We also know that there is an integer $N_2 > 0$ such that $CnR_n/H_n^2 \le W_n$ for all $n \ge N_2$. Let $N = \max\{N_1, N_2\}$, take $n \ge N$, and let $r_n = \lfloor CnR_n/H_n^2 \rfloor$. Suppose that $d(S(\widetilde{\theta}) \,|\, S_0) > r_n$. Then, by definition, there is a changepoint $t_i \in S_0$ such that no changepoints of $\widetilde{\theta}$ are within r_n of t_i, which means that $\widetilde{\theta}_j$ is constant over $j \in \{t_i - r_n + 1, \ldots, t_i + r_n\}$. Denote

$$z = \widetilde{\theta}_{t_i - r_n + 1} = \ldots = \widetilde{\theta}_{t_i} = \widetilde{\theta}_{t_i+1} = \ldots = \widetilde{\theta}_{t_i + r_n}.$$

We then form the lower bound

$$\frac{1}{n} \sum_{j=t_i-r_n+1}^{t_i+r_n} \big(\widetilde{\theta}_j - \theta_{0,j}\big)^2 = \frac{r_n}{n}\big(z - \theta_{0,t_i}\big)^2 + \frac{r_n}{n}\big(z - \theta_{0,t_i+1}\big)^2 \ge \frac{r_n H_n^2}{2n} > \frac{C}{4} R_n, \tag{12}$$

where the first inequality holds because $(x-a)^2 + (x-b)^2 \ge (a-b)^2/2$ for all x (the quadratic in x here is minimized at $x = (a+b)/2$), and the second because $r_n = \lfloor CnR_n/H_n^2 \rfloor$. Therefore, we see that $d(S(\widetilde{\theta}) \,|\, S_0) > r_n$ implies

$$\|\widetilde{\theta} - \theta_0\|_n^2 \ge \frac{1}{n} \sum_{j=t_i-r_n+1}^{t_i+r_n} \big(\widetilde{\theta}_j - \theta_{0,j}\big)^2 > \frac{C}{4} R_n,$$

which implies

$$\mathbb{P}\Big(d\big(S(\widetilde{\theta}) \,|\, S_0\big) > r_n\Big) \le \mathbb{P}\bigg(\|\widetilde{\theta} - \theta_0\|_n^2 > \frac{C}{4} R_n\bigg) \le \epsilon,$$

for all $n \ge N$, completing the proof. $\square$

Remark 11 (**Conditions on W_n, H_n**)**.** The condition that $nR_n/H_n^2 = o(W_n)$ in Theorem 8 is not strong. Consider the simple case in which $H_n = \Omega(1)$ and $W_n = \Theta(n)$. This condition reduces to $R_n = o(1)$, requiring only that the estimator $\widetilde{\theta}$ in question be consistent. The theorem then gives the bound $d(S(\widetilde{\theta}) \,|\, S_0) = O_{\mathbb{P}}(nR_n)$ on the screening distance obtained by $\widetilde{\theta}$.

Remark 12 (**Generic setting: no particular assumptions on data model, or estimator**)**.** Importantly, Theorem 8 assumes no data model whatsoever, and treats $\widetilde{\theta}$ as a generic estimator of θ_0. Of course, through the statement $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$, one sees that $\widetilde{\theta}$ is random, constructed from data that depends on θ_0, but no specific data model is required, nor are any specific properties of $\widetilde{\theta}$ (other than its ℓ_2 error rate). This flexibility allows for the result to be applied in any problem setting in which one has control of the estimation error of a piecewise constant parameter θ_0. Apart from the applications of Theorem 8 to the fused lasso estimator, where we consider data from a standard model as in (1), with θ_0 being the mean, and i.i.d. sub-Gaussian errors (see Corollaries 9 and 10 below), we could instead suppose that data is distributed according to, e.g., a Poisson model with natural parameter θ_0,

$$y_i \sim \mathrm{Pois}(e^{\theta_{0,i}}), \quad i = 1, \dots, n. \tag{13}$$

If we knew of an estimate $\widetilde{\theta}$ for θ_0 such that the estimation error $\|\widetilde{\theta} - \theta_0\|_n^2$ was analyzable, then we could use Theorem 8 to infer a bound on the screening distance between changepoints of θ_0 and $\widetilde{\theta}$. In this paper, we do not describe particular applications of Theorem 8 beyond the sub-Gaussian error model in (1), since we are not aware of estimation error guarantees outside of this model. However, establishing ℓ_2 estimation error rates for models like (13) (which may be used to describe say copy number data in genetics), and interpreting the resulting changepoint approximation guarantees would be an interesting topic for future work. (For model (13), and other likelihood-based models with a piecewise constant parameter θ_0, we suspect that the fused lasso provides a basis for a good estimator: simply replace the squared error loss in (2) by the negative log likelihood.)

We present two different corollaries of Theorem 8 for the fused lasso. The first is given by using Theorem 1 and the associated rate $R_n = n^{-2/3} C_n$ in the weak sparsity case, and the second is given by using Theorem 4 and the associated rate $R_n = (\log n \log\log n)/n$ in the strong sparsity case. The proofs are immediate and are hence omitted.

Corollary 9 (**Fused lasso screening result, weak sparsity setting**)**.** *Assume the conditions of Theorem 1, so that $\mathrm{TV}(\theta_0) \le C_n$ for a sequence C_n. Also assume that $H_n = \omega(n^{1/6} C_n^{1/3}/\sqrt{W_n})$. Let $\widehat{\theta}$ be the fused lasso estimate in (2), with $\lambda = \Theta(n^{1/3} C_n^{-1/3})$. Then*

$$d\big(\widehat{S} \,|\, S_0\big) = O_{\mathbb{P}}\bigg(\frac{n^{1/3} C_n^{2/3}}{H_n^2}\bigg).$$

Corollary 10 (**Fused lasso screening result, strong sparsity setting**)**.** *Assume the conditions of Theorem 2, so that $s_0 = O(1)$ and $W_n = \Theta(n)$. Also assume that $H_n = \omega(\sqrt{(\log n \log\log n)/n})$. Let $\widehat{\theta}$ be the fused lasso estimate in (2), with $\lambda = \Theta(\sqrt{n})$. Then*

$$d\big(\widehat{S} \,|\, S_0\big) = O_{\mathbb{P}}\bigg(\frac{\log n \log\log n}{H_n^2}\bigg).$$

Remark 13 (**Conditions on W_n, H_n**)**.** We have rewritten the condition that $nR_n/H_n^2 = o(W_n)$ in Theorem 8 as $H_n = \omega(n^{1/6} C_n^{1/3}/\sqrt{W_n})$ in Corollary 9, and $H_n = \omega(\sqrt{(\log n \log\log n)/n})$ in Corollary 10 (note that in the latter, we are assuming that $W_n = \Theta(n)$).

Remark 14 (**Screening under weak sparsity**)**.** Corollary 9 handles a difficult setting in which the number of changepoints s_0 in θ_0 can grow quickly with n, and yet it still provides a reasonable bound on the screening distance $d(S(\widehat{\theta}) \,|\, S_0)$ provided that C_n is not too large (i.e., $\mathrm{TV}(\theta_0)$ is not growing too quickly), or H_n is large enough (i.e., the minimum signal gap in θ_0 is large enough). As an example, suppose that $s_0 = \Theta(n^{1/6})$, and the changepoints in θ_0 are evenly spread out, so that $W_n = \Theta(n^{5/6})$. Then Corollary 9 implies, provided that $H_n = \omega(n^{-1/4} C_n^{1/3})$,

$$d(\widehat{S} \,|\, S_0) = O_{\mathbb{P}}\left(\frac{n^{1/3} C_n^{2/3}}{H_n^2}\right) = o_{\mathbb{P}}(n^{5/6}),$$

so for each true changepoint, there is at least one estimated changepoint that is much closer to it than all of the other true changepoints (each of which is at least a distance $W_n = \Theta(n^{5/6})$ away). From the condition $H_n = \omega(n^{-1/4} C_n^{1/3})$, and the fact that we must always have $C_n \geq s_0 H_n$ (recall $s_0 = \Theta(n^{1/6})$), we can be more explicit here about the allowable ranges for H_n, C_n: combining the last two relationships gives $C_n = \omega(n^{-1/8})$, and then $H_n = \omega(n^{-7/24})$. Hence, the minimum signal gap requirement here is very reasonable, allowing H_n to shrink to 0, just not too quickly (this is far from a trivial regime, e.g., with $H_n = \omega(\sqrt{\log n})$, when simple thresholding of pairwise differences achieves perfect recovery, as shown in Sharpnack et al. (2012)).

Remark 15 (**Comparison to Theorem 3**)**.** Corollary 10 provides a similar conclusion to that in Harchaoui & Levy-Leduc (2010), restated in Theorem 3: in a strong sparsity setting, the fused lasso has a well-controlled screening distance, only slightly larger than $\log n / H_n^2$. However, we note that Corollary 10 guarantees this screening bound under a natural choice for the tuning parameter λ, known to provide good ℓ_2 estimation performance (see Theorem 4), whereas Theorem 3 implicitly requires λ to be very small, which seems unnatural (see Remark 5).

Remark 16 (**Changepoint detection limit**)**.** The restriction that $H_n = \omega(\sqrt{(\log n \log\log n)/n})$ in Corollary 10 is very close to the optimal limit of $H_n = \omega(1/\sqrt{n})$ for changepoint detection: Duembgen & Walther (2008) showed that in Gaussian changepoint model with a single elevated region, and $W_n = \Theta(n)$, there is no test for detecting a changepoint that has asymptotic power 1 unless $H_n = \omega(1/\sqrt{n})$. See also Chan & Walther (2013).

4.2 Post-processing for approximate changepoint recovery

We study a procedure for post-processing the estimated changepoints in $\widetilde{\theta}$, in such a way that aims to eliminate changepoints of $\widetilde{\theta}$ that lie far away from changepoints of θ_0. Our procedure is based on convolving $\widetilde{\theta}$ with a filter that resembles the mother Haar wavelet. Consider

$$F_i(\widetilde{\theta}) = \frac{1}{b_n} \sum_{j=i+1}^{i+b_n} \widetilde{\theta}_j - \frac{1}{b_n} \sum_{j=i-b_n+1}^{i} \widetilde{\theta}_j, \quad \text{for } i = b_n, \ldots, n - b_n, \tag{14}$$

for an integral bandwidth $b_n > 0$. Our result in this subsection asserts that, by evaluating the filter $F_i(\widetilde{\theta})$ at all locations $i = b_n, \ldots, n - b_n$, and retaining only locations at which the filter value is large (in magnitude), we can approximately recovery the changepoints of θ_0, in the Hausdorff metric.

Theorem 11 (**Generic recovery result**)**.** *Let $\theta_0 \in \mathbb{R}^n$ be a piecewise constant vector, and $\widetilde{\theta} \in \mathbb{R}^n$ be an estimator that satisfies the error bound $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. Consider the following procedure: we evaluate the filter in (14) with bandwidth b_n at all locations $i = b_n, \ldots, n - b_n$, and we keep only the locations whose filter value is greater than or equal to a threshold level τ_n, in magnitude. Denote the resulting "filtered" set by*

$$S_F(\widetilde{\theta}) = \Big\{ i \in \{b_n, \ldots, n - b_n\} : |F_i(\widetilde{\theta})| \geq \tau_n \Big\}. \tag{15}$$

If the bandwidth and threshold values satisfy $b_n = \omega(n R_n / H_n^2)$, $2b_n \leq W_n$, and $\tau_n / H_n \to \rho \in (0, 1)$ as $n \to \infty$, then we have

$$\mathbb{P}\Big(d_H\big(S_F(\widetilde{\theta}), S_0\big) \leq b_n \Big) \to 1 \quad \text{as } n \to \infty,$$

where $d_H(\,\cdot,\cdot\,)$ is the Hausdorff distance, as defined in (7).

Proof. The proof is not complicated conceptually, but requires some careful bookkeeping. Also, we make use of a few key lemmas whose details will be given later. Fix $\epsilon > 0$ and $C > 0$. Let $N_1 > 0$ be an integer such that for all $n \geq N_1$,

$$\mathbb{P}\Big(\|\widetilde{\theta} - \theta_0\|_n^2 > CR_n\Big) \leq \frac{\epsilon}{2}.$$

Set $\epsilon = \min\{\rho, 1-\rho\}/2$. As $b_n = \omega(nR_n/H_n^2)$, there is an integer $N_2 > 0$ such that for all $n \geq N_2$,

$$\frac{2CnR_n}{b_n} \leq (0.99\epsilon H_n)^2.$$

As $\tau_n/H_n \to \rho \in (0,1)$, there is an integer $N_3 > 0$ such that for all $n \geq N_3$,

$$(\rho - \epsilon)H_n \leq \tau_n \leq (\rho + \epsilon)H_n.$$

Set $N = \max\{N_1, N_2, N_3\}$, and take $n \geq N$. Note that $\epsilon \leq \rho - \epsilon$ and $\rho + \epsilon \leq 1 - \epsilon$ by construction, and thus by the last two displays,

$$\sqrt{\frac{2CnR_n}{b_n}} < \tau_n < H_n - \sqrt{\frac{2CnR_n}{b_n}}. \tag{16}$$

Now observe

$$\mathbb{P}\Big(d_H\big(S_F(\widetilde{\theta}), S_0\big) > b_n\Big) \leq \mathbb{P}\Big(d\big(S_F(\widetilde{\theta}) \,|\, S_0\big) > b_n\Big) + \mathbb{P}\Big(d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) > b_n\Big). \tag{17}$$

We focus on bounding each term on the right-hand side above separately. For the first term on the right-hand side in (17), observe that if $F_{t_i}(\widetilde{\theta}) \geq \tau_n$ for all $t_i \in S_0$, then $d(S_F(\widetilde{\theta}) \,|\, S_0) \leq b_n$. By the contrapositive,

$$\begin{aligned}
\mathbb{P}\Big(d\big(S_F(\widetilde{\theta}) \,|\, S_0\big) > b_n\Big) &\leq \mathbb{P}\Big(|F_{t_i}(\widetilde{\theta})| < \tau_n \text{ for some } t_i \in S_0\Big) \\
&\leq \mathbb{P}\bigg(|F_{t_i}(\widetilde{\theta})| < H_n - \sqrt{\frac{2CnR_n}{b_n}} \text{ for some } t_i \in S_0\bigg),
\end{aligned} \tag{18}$$

where in the second line we used the upper bound on τ_n in (16). Suppose that $\|\widetilde{\theta} - \theta_0\|_n^2 \leq CR_n$; then, for $t_i \in S_0$, Lemma 13 tells us how small $|F_{t_i}(\widetilde{\theta})|$ can be made with this error bound in place. Specifically, define

$$a = (\underbrace{-1/b_n, \ldots, -1/b_n}_{b_n \text{ times}}, \underbrace{1/b_n, \ldots, 1/b_n}_{b_n \text{ times}}) \quad \text{and} \quad c = (\theta_{0,t_i - b_n + 1}, \ldots, \theta_{0,t_i + b_n}),$$

and also $r = \sqrt{CnR_n}$. Then Lemma 13 implies the following: if $\|\widetilde{\theta} - \theta_0\|_n^2 \leq CR_n$, then

$$|F_{t_i}(\widetilde{\theta})| \geq |a^\top c| - r\|a\|_2 \geq |\theta_{0,t_i+1} - \theta_{0,t_i}| - \sqrt{\frac{2CnR_n}{b_n}} \geq H_n - \sqrt{\frac{2CnR_n}{b_n}}.$$

Therefore, continuing on from (18),

$$\begin{aligned}
\mathbb{P}\Big(d\big(S_F(\widetilde{\theta}) \,|\, S_0\big) > b_n\Big) &\leq \mathbb{P}\bigg(|F_{t_i}(\widetilde{\theta})| < H_n - \sqrt{\frac{2CnR_n}{b_n}} \text{ for some } t_i \in S_0\bigg) \\
&\leq \mathbb{P}\Big(\|\widetilde{\theta} - \theta_0\|_n^2 > CR_n\Big) \\
&\leq \frac{\epsilon}{2}.
\end{aligned}$$

It suffices to consider the second term in (17), and show that this is also bounded by $\epsilon/2$. Note that

$$\begin{aligned}\mathbb{P}\Big(d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) > b_n\Big) &\leq \mathbb{P}\Big(|F_i(\widetilde{\theta})| \geq \tau_n \text{ at some } i \text{ such that } \theta_{0,i-b_n+1} = \ldots = \theta_{0,i+b_n}\Big) \\ &\leq \mathbb{P}\bigg(|F_i(\widetilde{\theta})| > \sqrt{\frac{2CnR_n}{b_n}} \text{ at some } i \text{ such that } \theta_{0,i-b_n+1} = \ldots = \theta_{0,i+b_n}\bigg).\end{aligned} \tag{19}$$

In the second inequality we used the lower bound on τ_n in (16). Similar to the previous argument, suppose that $\|\widetilde{\theta} - \theta_0\|_n^2 \leq CR_n$; for any location i in consideration in (19), Lemma 12 tells us how large $|F_i(\widetilde{\theta})|$ can be made with this error bound in place. Defining

$$a = (\underbrace{-1/b_n, \ldots, -1/b_n}_{b_n \text{ times}}, \underbrace{1/b_n, \ldots, 1/b_n}_{b_n \text{ times}}) \quad \text{and} \quad c = (\theta_{0,i-b_n+1}, \ldots, \theta_{0,i+b_n}),$$

and $r = \sqrt{CnR_n}$, as before, the lemma says the following: if $\|\widetilde{\theta} - \theta_0\|_n^2 \leq CR_n$, then

$$|F_i(\widetilde{\theta})| \leq |a^\top c| + r\|a\|_2 = \sqrt{\frac{2CnR_n}{b_n}}.$$

Hence, continuing on from (19),

$$\begin{aligned}\mathbb{P}\Big(d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) > b_n\Big) &\leq \mathbb{P}\bigg(|F_i(\widetilde{\theta})| > \sqrt{\frac{2CnR_n}{b_n}} \text{ at some } i \text{ such that } \theta_{0,i-b_n+1} = \ldots = \theta_{0,i+b_n}\bigg) \\ &\leq \mathbb{P}\Big(\|\widetilde{\theta} - \theta_0\|_n^2 > CR_n\Big) \\ &\leq \frac{\epsilon}{2},\end{aligned}$$

completing the proof. □

Remark 17 (Comparison to Theorem 8). Though they are stated differently, the rates in Theorems 8 and 11 for approximate changepoint screening and recovery, respectively, are comparable. To see this, note that the conclusion in the latter implies

$$\mathbb{P}\Big(d\big(S(\widetilde{\theta}) \,|\, S_0\big) \leq d_n\Big) \to 1 \quad \text{as } n \to \infty,$$

for any sequence $d_n = \omega(nR_n/H_n^2)$, which is in line with Theorem 11. (The original conclusion that $d(S(\widetilde{\theta}) \,|\, S_0) = O_{\mathbb{P}}(nR_n/H_n^2)$ is a somewhat stronger statement, though the difference is not major.)

Remark 18 (Generic setting: no particular assumptions on data model, or estimator). To emphasize a similar point to that in Remark 12, Theorem 11 does not use a specific data model, and considers any estimator $\widetilde{\theta}$ for which we have ℓ_2 error control, $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. This makes it a very flexible and broadly applicable result. When the data comes from a model as in (1), where θ_0 is the mean and we have i.i.d. sub-Gaussian errors, we can apply Theorem 11 to the fused lasso, given our knowledge of its ℓ_2 error rate (see Corollaries 14 and 15 below). It could also be applied, under the same data model, to many other estimators whose ℓ_2 error rates are known (such as the Potts estimator, and unbalanced Haar wavelets). Moreover, it could be useful under different data models, like the Poisson model in (13), as it would provide approximate recovery guarantees for any method with a fast enough ℓ_2 estimation error rate. (Note that the post-processing step using the filter (14) does not itself require assumptions about the data.) We do not consider such extensions in the current paper, but they suggest interesting directions for future work.

The proof of Theorem 11 relied on two lemmas, that we state below. Their proofs are based on simple arguments in convex analysis and deferred until Appendix C.

Lemma 12. *Given $a, c \in \mathbb{R}^m$, $r \geq 0$, the optimal value of the (nonconvex) optimization problem*

$$\max_{x \in \mathbb{R}^m} \; |a^\top x| \;\text{ subject to }\; \|x - c\|_2 \leq r \tag{20}$$

is $|a^\top c| + r\|a\|_2$.

Lemma 13. *Given $a, c \in \mathbb{R}^m$, $r \geq 0$ such that $|a^\top c| - r\|a\|_2 \geq 0$, the optimal value of the (convex) optimization problem*

$$\min_{x \in \mathbb{R}^n} \; |a^\top x| \;\text{ subject to }\; \|x - c\|_2 \leq r \tag{21}$$

is $|a^\top c| - r\|a\|_2$.

We finish this subsection with two corollaries of Theorem 11 for the fused lasso estimator, in the weak and strong sparsity cases. The proofs are immediate and are thus omitted.

Corollary 14 (**Fused lasso recovery result, weak sparsity setting**)**.** *Assume the conditions of Theorem 1, so that $\mathrm{TV}(\theta_0) \leq C_n$ for a sequence C_n. Let $\widehat{\theta}$ be the fused lasso estimate in (2), with $\lambda = \Theta(n^{1/3} C_n^{-1/3})$, and consider applying the filter in (14) to $\widehat{\theta}$, as described in Theorem 11, to produce a filtered set $\widehat{S}_F = S_F(\widehat{\theta})$. If the bandwidth and threshold satisfy $b_n = \lfloor n^{1/3} C_n^{2/3} \nu_n^2 / H_n^2 \rfloor \leq W_n/2$ for a sequence $\nu_n \to \infty$, and $\tau_n / H_n \to \rho \in (0,1)$, then*

$$\mathbb{P}\bigg(d_H(\widehat{S}_F, S_0) \leq \frac{n^{1/3} C_n^{2/3} \nu_n^2}{H_n^2} \bigg) \to 1 \quad \text{as } n \to \infty.$$

Corollary 15 (**Fused lasso recovery result, strong sparsity setting**)**.** *Assume the conditions of Theorem 2, so that $s_0 = O(1)$ and $W_n = \Theta(n)$. Let $\widehat{\theta}$ denote the fused lasso estimate in (2), with $\lambda = \Theta(\sqrt{n})$, and consider applying the filter in (14) to $\widehat{\theta}$, as in Theorem 11, to produce a filtered set $\widehat{S}_F = S_F(\widehat{\theta})$. If the bandwidth and threshold values satisfy $b_n = \lfloor (\log n \log\log n) \nu_n^2 / H_n^2 \rfloor \leq W_n/2$ for a sequence $\nu_n \to \infty$, and $\tau_n / H_n \to \rho \in (0,1)$, then*

$$\mathbb{P}\bigg(d_H(\widehat{S}_F, S_0) \leq \frac{(\log n \log\log n) \nu_n^2}{H_n^2} \bigg) \to 1 \quad \text{as } n \to \infty.$$

Remark 19 (**Recovery under weak sparsity, comparison to BS**)**.** Corollary 14 considers a challenging setting in which the number of changepoints s_0 in θ_0 could be growing quickly with n, and the only control that we have is $\mathrm{TV}(\theta_0) \leq C_n$. We draw a comparison here to known results on binary segmentation (BS). Corollary 14 on the (filtered) fused lasso and Theorem 3.1 in Fryzlewicz (2014) on the BS estimator $\widehat{\theta}^{\mathrm{BS}}$, each under appropriate conditions on W_n, H_n, state that

$$d_H(\widehat{S}_F, S_0) \leq \frac{n^{1/3} C_n^{2/3} \log n}{H_n^2} \quad \text{versus} \quad d_H\big(S(\widehat{\theta}^{\mathrm{BS}}), S_0\big) \leq c \frac{n \log n}{H_n^2} \quad \text{respectively}, \tag{22}$$

where $c > 0$ is a constant, and both bounds hold with probability approaching 1. The result on $\widehat{S}_F$ is obtained by choosing $\nu_n = \sqrt{\log n}$ and then $b_n = \lfloor n^{1/3} C_n^{2/3} \log n / H_n^2 \rfloor$ in Corollary 14. Examining (22), we see that, when C_n scales more slowly than n, Theorem 11 provides the stronger result: the term $n^{1/3} C_n^{2/3}$ will be smaller than n, and thus the bound on $d_H(\widehat{S}_F, S_0)$ will be sharper than that on $d_H(S(\widehat{\theta}^{\mathrm{BS}}), S_0)$.

But we must also examine the specific restrictions that each result in (22) places on s_0, W_n, H_n. Consider the simplification $W_n = \Theta(n/s_0)$, corresponding to a case in which the changepoints in θ_0 are spaced evenly apart. For Corollary 14, starting with the condition $n^{1/3} C_n^{2/3} \log n / H_n^2 \leq W_n/2$, plugging in the relationship $C_n \geq s_0 H_n$, and rearranging to derive a lower bound on the minimum signal gap, gives $H_n = \Omega(s_0^{5/4} n^{-1/2} \log^{3/4} n)$. If $s_0 = \Theta(n^{2/5})$, then we see that the minimum signal gap requirement becomes $H_n = \Omega(\log^{3/4} n)$, which is growing with n and is thus too stringent to be interesting (recall, as discussed previously, that Sharpnack et al. (2012) showed simple thresholding of pairwise differences achieves perfect recovery when $H_n = \omega(\sqrt{\log n})$). Hence, to accommodate signals for which H_n remains constant or even shrinks with n, we must restrict the number of jumps in θ_0 according to $s_0 = O(n^{2/5-\delta})$, for any fixed $\delta > 0$. Meanwhile, inspection of Assumption 3.2 in Fryzlewicz (2014) reveals that his Theorem 3.1 requires $s_0 = O(n^{1/4-\delta})$, for any $\delta > 0$, in order to handle signals such that H_n remains constant or shrinks with n. In short, the (effectively) allowable range for s_0 is larger for Theorem 11 than for Theorem 3.1 in Fryzlewicz (2014). Even when we look within their common range, Theorem 11 places weaker conditions on H_n. As an example, consider $s_0 = \Theta(n^{1/6})$ and $W_n = \Theta(n^{5/6})$. The fused lasso result in (22) requires $H_n = \Omega(n^{-7/24} \log^{4/3} n)$, and the BS result in (22) requires $H_n = \Omega(n^{-1/6+\delta})$, for any $\delta > 0$. Finally, to reiterate, the fused lasso result in (22) gives a better Hausdorff recovery bound when C_n is small compared to n; at the extreme end, this is better by a full factor of $n^{2/3}$, when $C_n = O(1)$.

While the post-processed fused lasso looks favorable compared to BS, based on its approximate changepoint recovery properties in the weak sparsity setting, we must be clear that the analyses for other methods—wild binary segmentation (WBS), the simultaneous multiscale changepoint estimator (SMUCE), and tail-greedy unbiased Haar (TGUH) wavelets—are still much stronger in this setting. Such methods have Hausdorff recovery bounds that are only possible for the post-processed fused lasso (at least, using our current analysis technique) when we assume strong sparsity. This is discussed next.

Remark 20 (Recovery under strong sparsity, comparison to other methods). When $s = O(1)$ and $W_n = \Theta(n)$, Corollary 10 shows that the post-processed fused lasso estimator delivers a Hausdorff bound of

$$d_H(\widehat{S}_F, S_0) \le \frac{\log^2 n}{H_n^2}, \tag{23}$$

on the set $\widehat{S}_F$ of filtered changepoints, with probability approaching 1. This is obtained by choosing (say) $\nu_n = \sqrt{\log n / \log\log n}$ and $b_n = \lfloor \log^2 n / H_n^2 \rfloor \le W_n/2$ in the corollary. The effective restriction on the minimum signal gap is thus $H_n = \Omega(\log n/\sqrt{n})$, which is quite reasonable, as $H_n = \omega(1/\sqrt{n})$ is needed for any method to detect a changepoint with probability tending to 1 (recall Remark 16). Several other methods—the Potts estimator (Boysen et al. 2009), binary segmentation (BS) and wild binary segmentation (WBS) (Fryzlewicz 2014), the simultaneous multiscale changepoint estimator (SMUCE) (Frick et al. 2014), and tail-greedy unbiased Haar wavelets (TGUH) (Fryzlewicz 2016)—all admit Hausdorff recovery bounds that essentially match (23), under similarly weak restrictions on H_n. But, it should be noted that the latter three methods—WBS, SMUCE, and TGUH—continue to enjoy these same sharp Hausdorff bounds *outside of* the strong sparsity setting, namely, their analyses do not require that $s_0 = O(1)$ and $W_n = \Theta(n)$, and instead just place weak restrictions on the allowed combinations of W_n, H_n (e.g., the analysis of WBS in Fryzlewicz (2014) only requires $W_n H_n^2 \ge \log n$). These analyses (and those for all previously described estimators) are more refined than that given in Corollary 15: they are based on specific properties of the estimator in question. The corollary, on the other hand, follows from Theorem 8, which uses a completely generic analysis that only assumes knowledge of the ℓ_2 error rate.

4.3 Post-processing on a reduced set

Recall that the strategy studied in Theorem 11 was to apply the Haar filter in (14) at each location $i = b_n, \dots, n - b_n$ and then check for large absolute values. Computationally, this not expensive—it only requires $O(n)$ operations—but there is an undesirable feature of this strategy with respect to practical usage. Writing the original number of estimated changepoints as $\widetilde{s} = |S(\widetilde{\theta})|$, it is possible in practice for the size of the filtered set $S_F(\widetilde{\theta})$ in (15) to be much larger than $\widetilde{s}$, if the bandwidth and threshold parameters are not set appropriately. Indeed, as the filter is being applied at $n - 2b_n$ locations, it is possible for the filtered set to have precisely this many elements.

Here we propose a modified strategy that runs the filter on (at most) $3\widetilde{s}+2$ changepoints, and then as usual, keeps only changepoints whose absolute filter values are large. This modified strategy has essentially same the theoretical guarantee of approximate changepoint recovery as the original "exhaustive" strategy from Section 4.2, but enjoys the practical advantage that, no matter how the bandwidth and threshold parameters are chosen, the final set of detected changepoints is bounded in size by 3 times the number of changepoints in $\widetilde{\theta}$ (plus 2, to be precise). Before stating the main result of this subsection, we introduce a "candidate" set for locations for changepoints,

$$I_C(\widetilde{\theta}) = \Big\{ i \in \{b_n, \dots, n-b_n\} : i \in S(\widetilde{\theta}), \text{ or } i + b_n \in S(\widetilde{\theta}), \text{ or } i - b_n \in S(\widetilde{\theta}) \Big\} \cup \{b_n, n - b_n\}. \quad (24)$$

These are estimated changepoints, locations that are at a distance b_n from estimated changepoints, or boundary points.

Theorem 16 (**Generic recovery result, reduced post-processing**). *Assume the conditions of Theorem 11, but consider a modified strategy in which we only evaluate the filter in (14) at locations in the candidate set $I_C(\widetilde{\theta})$ in (24), and define a "reduced" set of filtered points based on the locations whose filter value is at least τ_n,*

$$S_R(\widetilde{\theta}) = \Big\{ i \in I_R(\widetilde{\theta}) : |F_i(\widetilde{\theta})| \geq \tau_n \Big\}. \quad (25)$$

Then, subject to the same conditions on b_n, τ_n as in Theorem 11, we have

$$\mathbb{P}\Big(d_H\big(S_R(\widetilde{\theta}), S_0\big) \leq 2b_n \Big) \to 1 \quad \text{as } n \to \infty.$$

Proof. We will show that

$$\Big\{ d_H\big(S_F(\widetilde{\theta}), S_0\big) \leq b_n \Big\} \subseteq \Big\{ d_H\big(S_R(\widetilde{\theta}), S_0\big) \leq 2b_n \Big\}, \quad (26)$$

Since the left-hand side occurs with probability tending to 1, by Theorem 11, so will the right-hand side. To show the desired containment, recall that, by the definition of Hausdorff distance,

$$\Big\{ d_H\big(S_F(\widetilde{\theta}), S_0\big) \leq b_n \Big\} = \Big\{ d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) \leq b_n \Big\} \cap \Big\{ d\big(S_F(\widetilde{\theta}) \,|\, S_0\big) \leq b_n \Big\}. \quad (27)$$

Inspecting the first term on the right-hand side of (27), we observe

$$\Big\{ d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) \leq b_n \Big\} \subseteq \Big\{ d\big(S_0 \,|\, S_F(\widetilde{\theta})\big) \leq 2b_n \Big\} \subseteq \Big\{ d\big(S_0 \,|\, S_R(\widetilde{\theta})\big) \leq 2b_n \Big\}, \quad (28)$$

where the last containment holds as $S_R(\widetilde{\theta}) \subseteq S_F(\widetilde{\theta})$. Inspecting the second term on the right-hand side of (27), we use Lemma 17 which states that for each $j \in \{b_n, \dots, n-b_n\}$, there exists $i \in I_C(\widetilde{\theta})$ such that $|i-j| \leq b_n$ and $|F_i(\widetilde{\theta})| \geq |F_j(\widetilde{\theta})|$. Using this, we see

$$\begin{aligned} \Big\{ d\big(S_F(\widetilde{\theta}) \,|\, S_0\big) \leq b_n \Big\} &= \Big\{ \text{for all } \ell \in S_0, \text{ there exists } j \in S_F(\widetilde{\theta}) \text{ such that } |\ell - j| \leq b_n \Big\} \\ &\subseteq \Big\{ \text{for all } \ell \in S_0, \text{ there exists } i \in I_C(\widetilde{\theta}) \text{ such that } |\ell - i| \leq 2b_n \Big\} \\ &= \Big\{ d\big(S_R(\widetilde{\theta}) \,|\, S_0\big) \leq 2b_n \Big\}. \end{aligned} \quad (29)$$

Above, we have used Lemma 17 for the containment in the second line. Combining (27), (28), and (29), we have established (26), as desired. ☐

The proof of Theorem 16 relied on the following lemma. Its proof can be found in Appendix D.

Lemma 17. *Let $I_C(\widetilde{\theta})$ be the candidate set defined in (24). For every location $j \in \{b_n, \dots, n-b_n\}$ where $|F_j(\widetilde{\theta})| > 0$, there exists a location $i \in I_C(\widetilde{\theta})$ such that $|i-j| \leq b_n$ and $|F_i(\widetilde{\theta})| \geq |F_j(\widetilde{\theta})|$.*

5 Implementation considerations and experiments

We develop a data-driven procedure to determine the threshold level τ_n of the filter in (14), used to derive a post-processed set of changepoints $S_F(\widetilde{\theta})$ from an estimate $\widetilde{\theta}$, as described in (15) in Theorem 11. We also present a number of simulation results to support and complement the theoretical developments in this paper.

A data-driven procedure for choosing τ_n. Let $\mathcal{A}(\cdot)$ denote a fitting algorithm that, applied to data y, outputs an estimate $\widetilde{\theta}$ of θ_0 (e.g., $\mathcal{A}(y)$ could be the minimizer in (2), so that its output is the fused lasso estimate). In Algorithm 1, we present a heuristic but intuitive method for choosing the threshold level τ_n, based on (entrywise) permutations of the residual vector $y - \widetilde{\theta}$. Aside from the choice of fitting algorithm $\mathcal{A}(\cdot)$, we must specify a number of permutations B to be explored, a bandwidth b_n for the filter in (14), and a quantile level $q \in (0, 1)$. The intuition behind Algorithm 1 is to set τ_n large enough to suppress "false positive" changepoints $100 \cdot q\%$ of the time (according to the permutations). This is revisited later, in the discussion of the simulation results.

Some example settings: we may choose $\mathcal{A}(\cdot)$ to be the fused lasso estimator, where the tuning parameter λ is selected to minimize 5-fold cross-validation (CV) error, $B = 100$, and $q = 0.95$. The choice of bandwidth b_n is more subtle, and unfortunately, there is no specific answer that works for all problems.[4] But, the theory in the last section provides some general guidance: e.g., for problems in which we believe there are a small number of changepoints (i.e., $s_0 = O(1)$) of reasonably large magnitude (i.e., $H_n = \Omega(1)$), Theorem 11 instructs us to choose a bandwidth that grows faster than $\log n \log \log n$, so, choosing b_n to scale as $\log^2 n$ would suffice. We will use this scaling, as well as the above suggestions for $\mathcal{A}(\cdot)$, B, and q in all coming experiments, unless otherwise specified.

Algorithm 1 Permutation-based approach for choosing τ_n

---

0. Input a fitting algorithm $\mathcal{A}(\cdot)$, number of permutations B, bandwidth b_n, and quantile level $q \in (0, 1)$.
1. Compute $\widetilde{\theta} = \mathcal{A}(y)$. Let $\widetilde{S} = S(\widetilde{\theta})$ denote the changepoints, and $r = y - \widetilde{\theta}$ the residuals.
2. For each $b = 1, \ldots, B$, repeat the following steps:
   (a) Let $r^{(b)}$ be a randomly-chosen permutation of r, and define auxiliary data $y^{(b)} = \widetilde{\theta} + r^{(b)}$.
   (b) Rerun the fitting algorithm on the auxiliary data to yield $\widetilde{\theta}^{(b)} = \mathcal{A}(y^{(b)})$.
   (c) Apply the filter in (14) to $\widetilde{\theta}^{(b)}$ (with the specified bandwidth b_n), and record the largest magnitude $\widehat{\tau}^{(b)}$ of the filter values at locations greater than b_n away from $\widetilde{S}$. Formally,
   $$\widehat{\tau}^{(b)} = \max_{\substack{i \in \{b_n, \ldots, n-b_n\}: \\ d(\widetilde{S} | \{i\}) > b_n}} \big| F_i(\widetilde{\theta}^{(b)}) \big|.$$
3. Output $\widehat{\tau}_n$, the level q quantile of the collection $\widehat{\tau}^{(b)}$, $b = 1, \ldots, B$.

After running Algorithm 1 to compute $\widehat{\tau}_n$, the idea is to proceed with the full filter $S_F(\widetilde{\theta})$ or the reduced filter $S_R(\widetilde{\theta})$, applied at the level $\tau_n = \widehat{\tau}_n$, to the estimate $\widetilde{\theta}$ computed on the original data y at hand. In the experiments that follow, we use the reduced filter, though similar conclusions would hold with the full filter.

[4] We note that in some situations, problem-specific intuition can yield a reasonable choice of bandwidth b_n. Also, it should be possible to extend Algorithm 1 to choose both τ_n and b_n, but we do not pursue this, for simplicity.

Simulation setup. In our experiments, we use the following simulation setup. For a given n, the mean parameter $\theta_0 \in \mathbb{R}^n$ is defined to have $s_0 = 5$ equally-sized segments, with levels 0, 2, 4, 1, 4, from left to right. Data $y \in \mathbb{R}^n$ is generated around θ_0 using i.i.d. $N(0, 4)$ noise. Lastly, the sample size n is varied between 100 and 10,000, equally-spaced on a log scale. Figure 2 shows example data sets with $n = 774$ and $n = 10,000$.

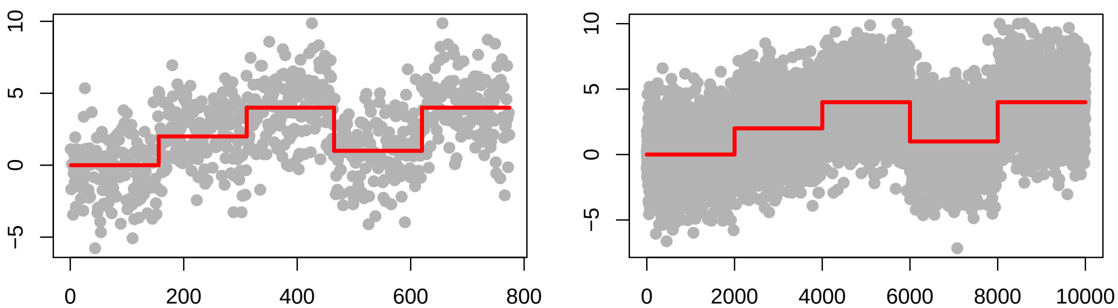

Figure 2: *An example from our simulation setup for $n = 774$ (left) and $n = 10,000$ (right), where in each panel, the mean θ_0 is plotted in red, and the data points in gray.*

Evaluation of the filter. We demonstrate that the filter in (14), with $b_n = \lfloor 0.25 \log^2 n \rfloor$, can be effective at reducing the Hausdorff distance between estimated and true changepoint sets. We first illustrate the use of the filter in a single data example with $n = 774$, in Figure 3. As we can see, the fused lasso originally places a spurious jump around location 250, but this jump is eliminated when we apply the filter, provided that we set the threshold to be (say) $\tau_n = 0.5$.

Figure 4 now reports the results from applying the filter in problems of sizes between $n = 100$ and $n = 10,000$, using 50 trials for each n. We consider three different sets of changepoint estimates: $\widehat{S} = S(\widehat{\theta})$, the original changepoints from fused lasso estimate $\widehat{\theta}$ tuned with 5-fold CV tuning; $S_R(\widehat{\theta})$, the changepoints after applying the reduced filter as described in Theorem 16 to $\widehat{\theta}$, with τ_n chosen by Algorithm 1; and $S_O(\widehat{\theta})$, an oracle set of changepoints given by trying a wide range of τ_n values and choosing the value that minimizes the Hausdorff distance after filtering (this assumes knowledge of S_0, and is infeasible in practice). These are labeled as "original", "data-driven", and "oracle" in the figure, respectively. As we can see from the left and middle panels, the Hausdorff distance achieved by the original changepoint set grows nearly linearly with n, but after applying the reduced filter, the Hausdorff distance becomes very small, provided that n is larger than 1000 or so. Empirically, the Hausdorff distance associated with the filtered set appears to grow very slowly with n, nearly constant (slower than the the $\log n \log \log n$ rate guaranteed by Corollary 15). The right panel shows that our data-driven choices of τ_n are not substantially different from those made by the oracle.

Screening distances, false positives. Figure 5 examines the outcomes from varying the filter threshold τ_n in between 0 and 2, and then applying the reduced filter to produce $S_R(\widehat{\theta})$. The results are aggregated over 500 trials when $n = 774$ (i.e., 500 data instances drawn from the simulation setup), and the screening distance $d(S_R(\widehat{\theta}) \,|\, S_0)$ and "precision distance" $d(S_0 \,|\, S_R(\widehat{\theta}))$ are plotted with τ_n. The former increases with τ_n, and the latter decreases; recall, the Hausdorff distance is the maximum of the two. We see that threshold levels from 0.5 to 1 yield a small Hausdorff distance.

The left panel of Figure 6 shows the same results, but with the screening distance on the x-axis, and the precision distance on the y-axis. The red dot marks the screening distance and precision distance achieved by the data-driven rule from Algorithm 1, using $B = 150$ permutations. This lies basically at the "elbow" of the curve, just as we would desire. The middle panel of the figure plots the proportion of false positive detections (out of the 500 repetitions total) on the x-axis, versus the

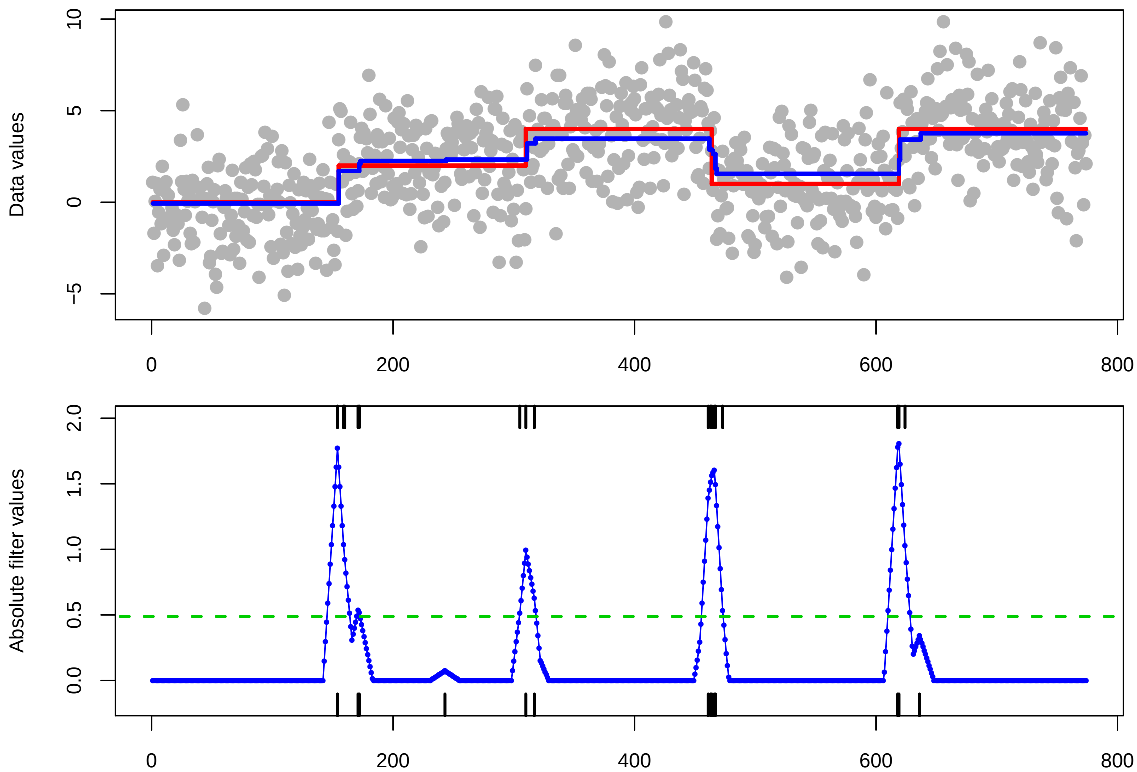

Figure 3: *In the top plot, an example with $n = 774$ is shown from our simulation setup, where the data y is drawn in gray, the mean θ_0 in red, and the fused lasso estimate $\widehat{\theta}$ in blue. In the bottom plot, the filter values $F_i(\widehat{\theta})$, $i = 1, \ldots, n$ are drawn in blue, and the threshold τ_n is drawn as a horizontal green line. Changepoints before and after filtering are marked by short black lines along the bottom and top x-axes, respectively.*

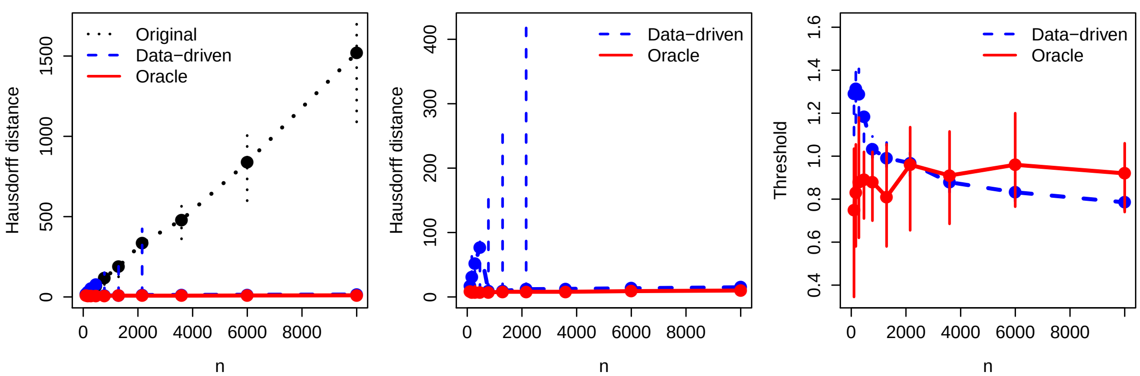

Figure 4: *In the left panel, the Hausdorff distances between original changepoints, filtered changepoints with a data-driven threshold, and filtered changepoints with an oracle threshold, are plotted (in black, blue, and red, respectively). The results are aggregated across 50 trial runs for each sample size n; the solid dots display the median values, and the vertical segments display the interquartile ranges (25th to 75th percentiles). The middle panel zooms in on the Hausdorff distances for the data-driven and oracle filtering procedures, and the right panel displays the choices of τ_n for these procedures.*

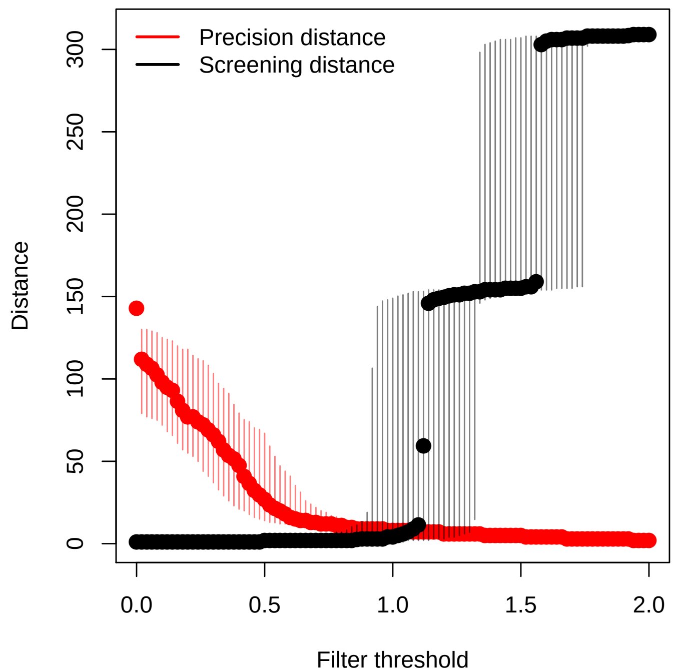

Figure 5: *The screening distance $d(S_R(\widehat{\theta})\,|\,S_0)$ (in black) and the precision distance $d(S_0\,|\,S_R(\widehat{\theta}))$ (in red) are shown as functions of the threshold τ_n used for the filtered set. These were aggregated over 500 trials, in which $n = 774$; the dots display the median values, and the vertical segments are drawn from the 25th to 75th percentiles. We can see a substantial jump in the screening distance once a bit after $\tau_n = 1$ and again after $\tau_n = 1.5$, where the median value is close to one of the quartiles. This is due to $S_R(\widehat{\theta})$ suppressing all of the estimated changepoints near a particular true changepoint, at these critical values of τ_n.*

proportion of true positive detections on the y-axis. Here, note, we define a false positive detection to be the event that *any* estimated changepoint is more than b_n away from all true changepoints, or simply, the event that $d(S_0\,|\,S_R(\widehat{\theta})) > b_n$, and a true positive detection to be the event that *all* true changepoints have estimated changepoints at most b_n away, or simply, $d(S_R(\widehat{\theta})\,|\,S_0) \le b_n$. Therefore, to be perfectly concrete, the x-axis and y-axis are displaying a certain type of false positive and true positive rates (FPR and TPR), defined as

$$\text{FPR} = \frac{\#\text{ trials in which } d(S_0\,|\,S_R(\widehat{\theta})) > b_n}{\#\text{ of trials}} \quad \text{and} \quad \text{TPR} = \frac{\#\text{ trials in which } d(S_R(\widehat{\theta})\,|\,S_0) \le b_n}{\#\text{ of trials}}.$$

The red dot again marks the FPR and TPR achieved by the data-driven rule in Algorithm 1 for choosing the threshold, about 0.26 and 0.7, respectively. We might expect here, having set $q = 0.95$ in Algorithm 1, to see a FPR close to 0.05 (because the choice of threshold in Algorithm 1 precisely controls the FPR at 0.05 over the permutations encountered in the procedure). However, this is not the case on in our simulation, and the actual FPR is higher. This phenomenon is not specific to the quantile choice of $q = 0.95$, as shown in the right panel of Figure 6. For a varying quantile level q in between 0 and 1, we ran Algorithm 1, used the corresponding threshold for our filter, and measured the FPR achieved by the filtered changepoint set. As we can see, the actual FPR is generally higher than $1 - q$.

Fused lasso fast ℓ_2 error rate, under strong sparsity. We finish by examining the (squared) ℓ_2 error $\|\widehat{\theta} - \theta_0\|_n^2$ as it scales with n, when the fused lasso estimate $\widehat{\theta}$ in (2) is tuned appropriately.

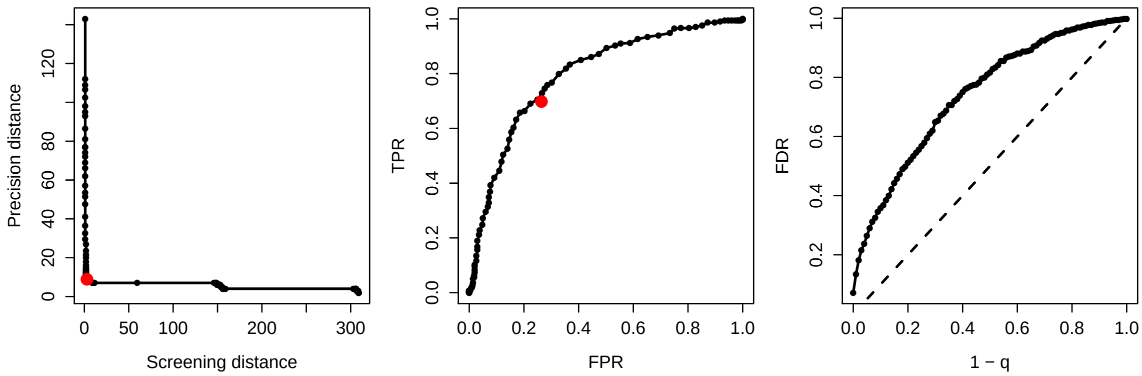

Figure 6: *The left panel plots the precision distance $d(S_0 \,|\, S_R(\widehat{\theta}))$ and the screening distance $d(S_R(\widehat{\theta}) \,|\, S_0)$, as the threshold τ_n is varied from 0 to 2. Shown are the median distances over 500 trials for a problem with $n = 774$. The red dot marks the precision and screening distances achieved by the data-driven threshold level chosen by Algorithm 1 with $q = 0.95$. The middle panel shows the same, but with true positive rate (TPR) against false positive rate (FPR). The right panel shows the achieved FPR against $1 - q$, as the input quantile level q is varied in Algorithm 1.*

For different sample sizes ranging from $n = 100$ to $n = 10,000$, we generated 50 example data sets from the same setup described previously, and on each data set, computed the fused lasso estimate $\widehat{\theta}$ with 5-fold CV to select the tuning parameter λ. Figure 7 reports the median value of λ, and the median achieved ℓ_2 error rate $\|\widehat{\theta} - \theta_0\|_n^2$, over the 50 trials, as functions of n. The results support the theoretical conclusion in Theorem 4, as the achieved ℓ_2 error rate scales at about the rate $(\log n \log\log n)/n$. Also, since $s_0 = O(1)$, the results support the theoretical result that λ scales with $\sqrt{n}$.

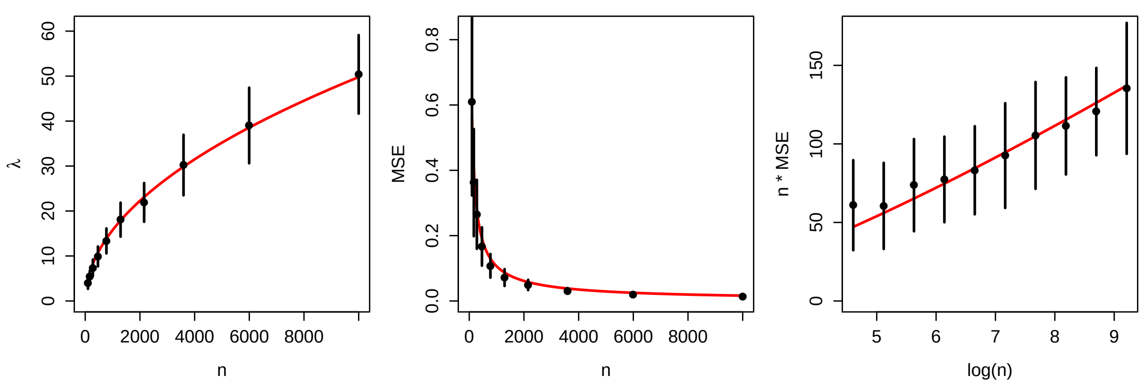

Figure 7: *The left panel shows the value of λ chosen to minimize 5-fold CV error over the fused lasso path, aggregated over repetitions in our simulation setup, as the sample size n varies. This scales approximately as $\sqrt{n}$, which is drawn as a red curve (with a best-fitting constant). The middle panel shows the corresponding (squared) ℓ_2 estimation error $\|\widehat{\theta} - \theta_0\|_n^2$, again aggregated over repetitions, as n varies. The scaling is about $(\log n \log\log n)/n$ (red curve). The right panel plots the median achieved values of $n\|\widehat{\theta} - \theta_0\|_n^2$ against $\log n$; this looks close to linear (red line), which provides empirical support to the claim that the fused lasso error rate is indeed $(\log n \log\log n)/n$ instead of $\log^2 n/n$ (as this would have appeared as a quadratic trend in the right panel). In each panel, the vertical bars denote ± 1 standard deviations.*

6 Extensions

We study screening properties that are implied by ℓ_2 error properties in two related problems: first, piecewise linear segmentation, and then, segmentation on graphs.

6.1 Piecewise linear segmentation

We now consider data from a model as in (1) but where $\theta_{0,i}$, $i = 1, \ldots, n$ is a piecewise linear (rather than a piecewise constant) sequence. The main estimator of interest is *linear trend filtering* (Steidl et al. 2006, Kim et al. 2009, Tibshirani 2014), which can be seen as an extension of the fused lasso that penalizes second-order (rather than first-order) differences:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-2} |\theta_i - 2\theta_{i+1} + \theta_{i+2}|, \tag{30}$$

for a tuning parameter $\lambda \geq 0$. Several other estimators are available in the piecewise linear segmentation problem, but given its ties to the fused lasso (and our focus on the fused lasso thus far), we focus on linear trend filtering in particular.

In terms of detection, we are now interested in the locations of nonzero second-order differences, i.e., the "knots", which mark the changes in slope across the entries of a parameter $\theta \in \mathbb{R}^n$:

$$S_2(\theta) = \left\{ i \in \{2, \ldots, n-1\} : \theta_i \neq \frac{\theta_{i-1} + \theta_{i+1}}{2} \right\}.$$

We use the abbreviations $S_{0,2} = S_2(\theta_0)$ and $\widehat{S}_2 = S_2(\widehat{\theta})$. We again write $S_{0,2} = \{t_1, \ldots, t_{s_0}\}$, where $2 \leq t_1 < \ldots < t_{s_0} < n$ and $s_0 = |S_{0,2}|$, and for convenience $t_0 = 0$, $t_{s_0+1} = n$. We also carry forward analogous definitions for W_n, H_n:

$$W_n = \min_{i=0,1\ldots,s_0} (t_{i+1} - t_i) \quad \text{and} \quad H_n = \min_{i \in S_{0,2}} |\theta_{0,i-1} - 2\theta_{0,i} + \theta_{0,i+1}|. \tag{31}$$

Lastly, we define the discrete second-order total variation operator, acting on a vector $x \in \mathbb{R}^n$, by

$$\mathrm{TV}_2(x) = \sum_{i=2}^{n-1} |x_{i-1} - 2x_i + x_{i+1}|.$$

The following describes the ℓ_2 estimation error of linear trend filtering, under weak sparsity.

Theorem 18 (**Trend filtering error rate, weak sparsity setting, Theorem 10 of Mammen & van de Geer 1997**). *Assume the data model in (1), with errors ϵ_i, $i = 1, \ldots, n$ i.i.d. from a sub-Gaussian distribution as in (8). Also assume that $\mathrm{TV}_2(\theta_0) \leq C_n$, for a sequence C_n. Then for $\lambda = \Theta(n^{1/5} C_n^{-3/5})$, the linear trend filtering estimate $\widehat{\theta}$ in (30) satisfies*

$$\|\widehat{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(n^{-4/5} C_n^{2/5}).$$

Remark 21 (**Consistency, optimality**). The lemma shows that linear trend filtering is consistent when $C_n = o(n^2)$. When $C_n = O(1)$, its error rate is $n^{-4/5}$, which is in fact minimax optimal as θ_0 varies over the class of signals having bounded second-order total variation, i.e., $\theta_0 \in \{\theta \in \mathbb{R}^n : \mathrm{TV}_2(\theta) \leq C\}$ for a constant $C > 0$ (Donoho & Johnstone 1998). As in the fused lasso case, we refer the reader to Tibshirani (2014) for explanations of the above theorem and this minimax result, in notation that is more consistent with that of the current paper.

Remark 22 (**Strong sparsity, higher polynomial degrees**)**.** Results for linear trend filtering in the strong sparsity setting, i.e., one in which s_0 is assumed to be bounded (so that we are estimating a piecewise linear function with few knots) are not currently available, to the best of our knowledge. However, we suspect that the achieved error rate here will be close to the "parametric" $1/n$ rate, as in Theorems 2 and 4, on the fused lasso. It is also worth noting that the extension of trend filtering to fit piecewise polynomials of higher degrees (i.e., higher than 1, as in the current piecewise linear case) is covered in Tibshirani (2014), where ℓ_2 estimation error rates (under weak sparsity) are also derived. For simplicity, we do not consider the general piecewise polynomial setting in our study of approximate screening, below, though such an extension should be possible.

Now we give our generic approximate screening result, analogous to that in Theorem 8.

Theorem 19 (**Generic screening result, piecewise linear segmentation**)**.** *Let $\theta_0 \in \mathbb{R}^n$ be a piecewise linear vector, and $\widetilde{\theta} \in \mathbb{R}^n$ be an estimator satisfying the error bound $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. Assume that $n^{1/3} R_n^{1/3} H_n^{-2/3} = o(W_n)$, where, recall, W_n, H_n are as defined in (31). Then*

$$d\big(S_2(\widetilde{\theta}) \,|\, S_{0,2}\big) = O_{\mathbb{P}}\bigg(\frac{n^{1/3} R_n^{1/3}}{H_n^{2/3}}\bigg).$$

Proof. The proof follows that of Theorem 8 closely, but differs in the lower bound asserted in (12). As before, given any $\epsilon > 0$, $C > 0$, as know that for some integer $N_1 > 0$ and all $n \geq N_1$,

$$\mathbb{P}\bigg(\|\widetilde{\theta} - \theta_0\|_n^2 > \frac{C^3}{4} R_n\bigg) \leq \epsilon.$$

We also know that for some integer $N_2 > 0$ and $n \geq n_2$, it holds that $Cn^{1/3} R_n^{1/3} H_n^{-2/3} \leq W_n$. Let $N = \max\{N_1, N_2\}$, take $n \geq N$, and let $r_n = \lfloor Cn^{1/3} R_n^{1/3} H_n^{-2/3} \rfloor$. Suppose that $d(S_2(\widetilde{\theta}) \,|\, S_{0,2}) > r_n$. Then there is a knot $t_i \in S_{0,2}$ such that there are no knots in $\widetilde{\theta}$ within r_n of t_i, which means that $\widetilde{\theta}_j$ displays a linear trend over the entire segment $j \in \{t_i - r_n, \ldots, t_i + r_n\}$. Hence

$$\frac{1}{n} \sum_{j=t_i-r_n}^{t_i+r_n} \big(\widetilde{\theta}_j - \theta_{0,j}\big)^2 \geq \frac{13 r_n^3 ((\theta_{0,t_i+1} - \theta_{0,t_i}) - (\theta_{0,t_i} - \theta_{0,t_i-1}))^2}{24n} \geq \frac{13 r_n^3 H_n^2}{24n} > \frac{C^3}{4} R_n. \tag{32}$$

Here, the first inequality holds due to Lemma 20, the second holds by definition of H_n, and the third by definition of r_n. We see that $d(S_2(\widetilde{\theta}) \,|\, S_{0,2}) > r_n$ implies the estimation error exceeds $(C^3/4) R_n$, an event that we know occurs with probability at most ϵ, completing the proof. ☐

The proof of Theorem 19 relied on the next lemma, to construct the key lower bound (32). The lemma characterizes how well a piecewise linear function can be approximated by a linear one, and is proved in Appendix E.

Lemma 20. *Let $f(x)$ be a piecewise linear function, defined over $x = -r, \ldots, r$, by*

$$f(x) = \begin{cases} a_1 x & \text{for } x \geq 0 \\ a_2 x & \text{for } x < 0 \end{cases}.$$

Let $\widetilde{a}x + \widetilde{b}$ be the optimal linear function for estimating $f(x)$, according to squared error loss, i.e.,

$$(\widetilde{a}, \widetilde{b}) = \operatorname*{argmin}_{a,b \in \mathbb{R}} \sum_{x=-r}^{r} \big(f(x) - ax - b\big)^2.$$

Then

$$\sum_{x=-r}^{r} \big(f(x) - \widetilde{a}x - \widetilde{b}\big)^2 \geq (a_2 - a_1)^2 \frac{13 r^3}{24}.$$

By combining Theorems 18 and 19, we have the following approximate screening result for linear trend filtering. The proof is omitted.

Corollary 21 (**Trend filtering screening result, weak sparsity setting**). *Assume the conditions in Theorem 18, thus $TV_2(\theta_0) \le C_n$ for a sequence C_n. Also assume $H_n = \omega(n^{1/10} C_n^{1/5} W_n^{-3/2})$. Let $\widehat{\theta}$ denote the linear trend filtering estimate in (30) with $\lambda = \Theta(n^{1/5} C_n^{-2/3})$. Then*

$$d\big(\widehat{S}_2 \,|\, S_{0,2}\big) = O_{\mathbb{P}}\bigg(\frac{n^{1/15} C_n^{2/15}}{H_n^{2/3}}\bigg).$$

Remark 23 (**Knot screening under weak sparsity**). To give an example of a challenging case that can be accommodated by Corollary 21, consider a setting in which θ_0 has $s_0 = \Theta(\sqrt{n})$ knots, evenly spread apart, so that $W_n = \Theta(\sqrt{n})$. Then, provided $H_n = \omega(n^{-13/20} C_n^{1/5})$, Corollary 21 says

$$d\big(\widehat{S}_2 \,|\, S_{0,2}\big) = O_{\mathbb{P}}\bigg(\frac{n^{1/15} C_n^{2/15}}{H_n^{2/3}}\bigg) = o_{\mathbb{P}}(\sqrt{n}),$$

so that each true knot has a detected knot that is much closer to it than all other true knots. Note that $C_n \ge s_0 H_n$, and combining this with the requirement on H_n reveals the implicit requirement $C_n = \omega(n^{-3/16})$, which in turn implies that $H_n = \omega(n^{-11/16})$. This seems to be a weak requirement on the minimum nonzero change in slopes that is present in θ_0.

6.2 Changepoint detection on a graph

We depart from the 1-dimensional setting considered throughout the paper thus far, and study the model (1) in a case where the mean parameter has components $\theta_{0,i}$, $i = 1, \dots, n$ that correspond to nodes $V = \{1, \dots, n\}$ of a graph G, with edges $E = \{e_1, \dots, e_m\}$. Note that, for each $\ell = 1, \dots, m$, we may write $e_\ell = (i, j)$ for some nodes i, j (and all edges are to be considered undirected, so that (i, j) and (j, i) are equivalent). Moreover, the mean θ_0 is assumed to behave in a piecewise constant fashion over the graph, which means that there are clusters of nodes over which θ_0 admits constant values, or, equivalently, $\theta_{0,i} = \theta_{0,j}$ for many edges $(i, j) \in E$. For estimation of θ_0, we focus on the *graph fused lasso* or *graph-based total variation denoising* (Tibshirani et al. 2005, Hoefling 2010, Tibshirani & Taylor 2011, Sharpnack et al. 2012), defined by

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^n}{\mathrm{argmin}} \; \frac{1}{2} \sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{(i,j) \in E} |\theta_i - \theta_j|, \tag{33}$$

for a tuning parameter $\lambda \ge 0$. When G is a 1d chain graph (i.e., $E = \{(1,2), (2,3), \dots, (n-1, n)\}$), the estimator in (33) reduces to the "usual" 1d fused lasso estimator in (2).

In the current graph-based setting, the "changepoints" of interest are actually edges for which the corresponding nodes display differing values, under a vector $\theta \in \mathbb{R}^n$:

$$S_G(\theta) = \big\{(i,j) \in E : \theta_i \neq \theta_j\big\}.$$

We use the abbreviations $S_{0,G} = S_G(\theta_0)$ and $\widehat{S}_G = S_G(\widehat{\theta})$. For an edge $(i, j) \in E$, let P_{ij} denote the set of paths in G centered around (i, j), and embedded within two constant clusters of nodes, i.e.,

$$P_{ij} = \Big\{ \big\{(i_\ell, i_{\ell-1}) \dots, (i_1, i), (i,j), (j, j_1), \dots, (j_{\ell-1}, j_\ell)\big\} \text{ for any } \ell = 1, \dots, n : \\ \theta_{0,i} = \theta_{0,i_1} = \dots = \theta_{0,i_\ell} \text{ and } \theta_{0,j} = \theta_{0,j_1} = \dots = \theta_{0,j_\ell} \Big\}.$$

We now define W_n, H_n over the graph G, in an analogous fashion to our notions in the 1d setting,

$$W_n = \min_{(i,j)\in S_{0,G}} \max_{p\in P_{ij}} \frac{|p|-1}{2} \quad \text{and} \quad H_n = \min_{(i,j)\in S_{0,G}} |\theta_{0,i} - \theta_{0,j}|, \tag{34}$$

where we write $|p|$ for the number of edges that form a path p. Note that, as defined, $2W_n + 1$ is the minimax length of any path centered around a changepoint in S_0; in other words, by construction, for each changepoint $(i,j) \in S_0$, there exists a path of at least W_n edges embedded entirely within a cluster on either side of (i,j). When W_n is small, this is indicative of one of the constant clusters of nodes in θ_0 being small in size. For $x \in \mathbb{R}^n$, we define its graph-based discrete total variation to be

$$\mathrm{TV}_G(x) = \sum_{(i,j)\in E} |x_i - x_j|.$$

Finally, we must precisely define our screening distance metric in the graph-based setting. For any two edges $e_1 = (i_1, j_1), e_2 = (i_2, j_2) \in E$, let $d_G(e_1, e_2)$ denote the length of the shortest path that starts either i_1 or j_1, and ends at either i_2 or j_2. For sets $A, B \in E$, we define the screening distance

$$d_G(A|B) = \max_{e_1=(i_1,j_1)\in B} \; \min_{e_2=(i_2,j_2)\in A} \; d_G(e_1, e_2).$$

Hence, if $d_G(A|B) = k$, then for any edge in B, there is a path of at most k edges starting from this edge, and ending at an edge in A.

Between Wang et al. (2016) and Hutter & Rigollet (2016), various estimation error rates are available for the graph fused lasso. These results take on different forms, depending on the assumptions placed on θ_0 and on the graph G. Below we recite a result from Hutter & Rigollet (2016) in the case that G is a 2d grid graph. Here $\widehat{\theta}$ in (33) is called the *2d fused lasso* estimate.

Theorem 22 (2d fused lasso error rate, weak and strong sparsity settings, Corollary 5 of Hutter & Rigollet 2016). *Assume the data model in (1), with errors ϵ_i, $i = 1, \ldots, n$ i.i.d. from $N(0, \sigma^2)$. Assume that G is a 2d grid graph, with n nodes (hence the 2d grid is of dimension $\sqrt{n} \times \sqrt{n}$). Write $s_0 = |S_{0,G}|$, and $\mathrm{TV}_G(\theta_0) \le C_n$, for some nondecreasing sequence C_n. Then for $\lambda = \Theta(\log n)$, the 2d fused lasso estimate $\widehat{\theta}$ in (33) satisfies*

$$\|\widehat{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}\bigg(\min\{s_0, C_n\} \frac{\log^2 n}{n} \bigg).$$

Remark 24 (Consistency, optimality). According to the theorem, the 2d fused lasso estimator is consistent when either $s_0 = o(n/\log^2 n)$ or $C_n = o(n/\log^2 n)$. Theorem 22 covers both the weak and strong sparsity cases (since it allows us to draw conclusions involving either s_0 or C_n). In the case of weak sparsity, the $C_n \log^2 n/n$ rate achieved by the 2d fused lasso was recently shown to be essentially minimax optimal (differing only by log factors), over the class of signals having bounded total variation over the 2d grid G, i.e., $\theta_0 \in \{\theta \in \mathbb{R}^n : \mathrm{TV}_G(\theta) \le C_n\}$, by Sadhanala et al. (2016).

Remark 25 (Other graphs). Basically the same result as in Theorem 22 holds for 3d and higher-dimensional grids (except with one fewer log factor) (Hutter & Rigollet 2016). Estimation error rates for various types of random graphs, the complete graph, and star graphs are derived in Wang et al. (2016), Hutter & Rigollet (2016). For simplicity, we do not consider any of these cases when we give an application of our generic graph screening result, below; however, given the availability of ℓ_2 rates, we remark that results over different graph models (over than the 2d grid) are certainly possible, and are just a matter of plugging in the proper rates in the proper settings.

Here is our generic graph screening result, analogous to that in Theorem 8 for the 1d chain.

Theorem 23 (**Generic screening result, changepoint detection on a graph**). *Let $\theta_0 \in \mathbb{R}^n$ be piecewise constant over a graph G, and $\widetilde{\theta} \in \mathbb{R}^n$ be an estimator that satisfies $\|\widetilde{\theta} - \theta_0\|_n^2 = O_{\mathbb{P}}(R_n)$. Assume that $nR_n/H_n^2 = o(W_n)$, where, recall W_n, H_n are as defined in (34). Then*

$$d_G\big(S_G(\widetilde{\theta}) \,|\, S_{0,G}\big) = O_{\mathbb{P}}\bigg(\frac{nR_n}{H_n^2}\bigg).$$

Proof. The proof is again very similar to the proof of Theorem 8. Fix $\epsilon > 0$, $C > 0$, and let $N_1 > 0$ be an integer such that, for $n \geq N_1$,

$$\mathbb{P}\bigg(\|\widetilde{\theta} - \theta_0\|_n^2 > \frac{C}{4} R_n\bigg) \leq \epsilon.$$

Let $N_2 > 0$ be an integer such that, for $n \geq N_2$, we have $CnR_n/H_n^2 \leq W_n$. Let $N = \max\{N_1, N_2\}$, take $n \geq N$, and define $r_n = \lfloor CnR_n/H_n^2 \rfloor$. Suppose that $d_G(S_G(\widetilde{\theta}) \,|\, S_{0,G}) > r_n$. By definition, there exists a changepoint $(i,j) \in S_0$ such that no changepoints in $\widetilde{\theta}$ are within r_n of (i,j), which in our distance metric, means that $\widetilde{\theta}_k$ is constant over all nodes k that are r_n away from i or j. Construct an arbitrary path p centered around edge (i,j) with $2r_n + 1$ edges

$$p = \Big\{(i_{r_n}, i_{r_n - 1}), \ldots, (i_1, i), (i,j), (j, j_1) \ldots, (j_{r_n - 1}, j_{r_n})\Big\},$$

where $\theta_{0,i} = \theta_{0,i_1} = \ldots = \theta_{0,i_{r_n}}$ and $\theta_{0,j} = \theta_{0,j_1} = \ldots = \theta_{0,j_{r_n}}$. (This is possible because $r_n \leq W_n$.) Denote

$$z = \widetilde{\theta}_{i_{r_n}} = \ldots = \widetilde{\theta}_i = \widetilde{\theta}_j = \ldots = \widetilde{\theta}_{j_{r_n}}.$$

Also let $I(p) = \{i_{r_n - 1}, \ldots, i_1, i, j, j_1, \ldots, j_{r_n - 1}\}$ denote the internal nodes of the path p. Then

$$\frac{1}{n} \sum_{k \in I(p)} \big(\widetilde{\theta}_k - \theta_{0,k}\big)^2 = \frac{r_n}{n}\big(z - \theta_{0,i}\big)^2 + \frac{r_n}{n}\big(z - \theta_{0,j}\big)^2 \geq \frac{r_n H_n^2}{2n} > \frac{C}{4} R_n,$$

where the first inequality holds because, as argued before, $(x-a)^2 + (x-b)^2 \geq (a-b)^2/2$ for all x, and the second by definition of r_n. Invoking the assumed ℓ_2 error rate for $\widetilde{\theta}$ completes the proof. □

Combining Theorems 22 and 23 gives the next and final result, whose proof is omitted.

Corollary 24 (**2d fused lasso screening result, weak and strong sparsity settings**). *Assume the conditions in Theorem 22, so that G is a 2d grid, and $s_0 = |S_{0,G}|$, $\mathrm{TV}_G(\theta_0) \leq C_n$. Also assume that $H_n = \omega(\min\{\sqrt{s_0}, \sqrt{C_n}\}\log n/\sqrt{W_n})$. Let $\widehat{\theta}$ denote the 2d fused lasso estimate in (33), with the choice of tuning parameter $\lambda = \Theta(\log n)$. Then*

$$d\big(\widehat{S}_G \,|\, S_{0,G}\big) = O_{\mathbb{P}}\bigg(\min\{s_0, C_n\}\frac{\log^2 n}{H_n^2}\bigg).$$

Remark 26 (**Screening over a 2d grid**). Consider, as a concrete example, a case in which θ_0 is piecewise constant with just 2 pieces or clusters, over the 2d grid G (of dimension, recall, $\sqrt{n} \times \sqrt{n}$). Then $s_0 = |S_{0,G}|$ reflects the length of the boundary separating the 2 pieces. In the typical case (in which the 2 pieces are of roughly equal size, and both have volume proportional to n), this scales as $s_0 = \Theta(\sqrt{n})$. Moreover, in the typical case, the length W_n of the longest path on either side of this boundary also scales as $W_n = \Theta(\sqrt{n})$. Here, $C_n \geq s_0 H_n$, which is larger than s_0 unless H_n is quite small ($H_n \leq 1$). Thus when H_n is large ($H_n > 1$), we can use Corollary 24 to conclude that

$$d\big(\widehat{S}_G \,|\, S_{0,G}\big) = O_{\mathbb{P}}\bigg(\frac{\sqrt{n}\log^2 n}{H_n^2}\bigg).$$

Roughly speaking, this says if we were to place a "tube" of radius $B_n = O_{\mathbb{P}}(\sqrt{n}\log^2 n/H_n^2)$ around the boundary edges $S_{0,G}$, containing edges that are at distance of at most B_n from $S_{0,G}$, then each changepoint in $S_{0,G}$ has a corresponding detected changepoint in $\widehat{S}_G$ lying inside this tube. Seeing as the entire grid itself is of dimension $\sqrt{n} \times \sqrt{n}$, this statement is not really interesting unless H_n is fairly large, say $H_n = \Theta(n^{1/8}\log n)$. Then $B_n = O_{\mathbb{P}}(n^{1/4})$, giving a reasonably tight tube around the boundary $S_{0,G}$.

An illustration of the true changepoints versus those detected by the 2d fused lasso, in a simple simulated 2d image example, is given in Figure 8. See the figure caption for details.

Figure 8: *An example on a 2d grid graph, of dimension* 100×100 *(i.e.,* $n = 10,000$*). The left panel displays the mean* θ_0*; the middle panel the data* y*, whose entries were generated by adding i.i.d.* $N(0,1)$ *noise to* θ_0*; and the right panel displays the 2d fused lasso estimate, whose tuning parameter* λ *was chosen to minimize the* ℓ_2 *estimation error in retrospect. The mean only has two constant pieces, taking values 0 and 1, denoted by dark gray and light gray colors, respectively, in the left panel. A consistent color scale is used throughout the three plots. The true changepoint set* $S_{0,G}$ *is drawn in red, and the estimated changepoint set* $\widehat{S}_G$ *in blue. In this example,* $d_G(\widehat{S}_G \,|\, S_{0,G}) = 1$*.*

7 Discussion

We have derived a new ℓ_2 error bound for the fused lasso in a strong sparsity setting, which, to the best of our knowledge, yields the sharpest available rate in this setting. We have also undertaken a detailed study of the manner in which ℓ_2 error bounds for generic estimators $\widetilde{\theta}$ can be used to prove changepoint screening results for $\widetilde{\theta}$, and after simple post-processing, changepoint recovery results for $\widetilde{\theta}$. As a prime example, we have derived new changepoint screening and recovery results for the fused lasso estimator, in various settings, based solely on its ℓ_2 error guarantees, in these settings. To reiterate, our general technique for analyzing changepoint screening and recovery properties is not specific to the fused lasso, and is potentially much more broadly applicable, as it only assumes knowledge of the ℓ_2 error rate of the estimator $\widetilde{\theta}$ in question. This could be applied even outside of the typical Gaussian data model.

We have also presented extensions to the piecewise linear segmentation and graph changepoint detection problems, as well as detailed simulations. The code to run all our simulations is located at https://github.com/linnylin92/fused_lasso, and relies on the R package `genlasso`.

References

Boysen, L., Kempe, A., Liebscher, V., Munk, A. & Wittich, O. (2009), 'Consistencies and rates of convergence of jump-penalized least squares estimators', *The Annals of Statistics* **37**(1), 157–183.

Chan, H. P. & Walther, G. (2013), 'Detection with the scan and the average likelihood ratio', *Statistica Sinica* **23**(1), 409–428.

Dalalyan, A. S., Hebiri, M. & Lederer, J. (2014), 'On the prediction performance of the lasso', *arXiv preprint arXiv:1402.1700* .

Davies, L. & Kovac, A. (2001), 'Local extremes, runs, strings and multiresolution', *Annals of Statistics* **21**(1), 1–65.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**(3), 425–455.

Donoho, D. L. & Johnstone, I. M. (1998), 'Minimax estimation via wavelet shrinkage', *Annals of Statistics* **26**(8), 879–921.

Duembgen, L. & Walther, G. (2008), 'Multiscale inference about a density', *The Annals of Statistics* **36**(4), 1758–1785.

Frick, K., Munk, A. & Sieling, H. (2014), 'Multiscale change point inference', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(3), 495–580.

Fryzlewicz, P. (2007), 'Unbalanced Haar technique for nonparametric function estimation', *Journal of the American Statistical Association* **102**(480), 1318–1327.

Fryzlewicz, P. (2014), 'Wild binary segmentation for multiple change-point detection', *The Annals of Statistics* **42**(6), 2243–2281.

Fryzlewicz, P. (2016), 'Tail-greedy bottom-up data decompositions and fast multiple change-point detection'. Available at http://stats.lse.ac.uk/fryzlewicz/tguh/tguh.pdf.

Harchaoui, Z. & Levy-Leduc, C. (2010), 'Multiple change-point estimation with a total variation penalty', *Journal of the American Statistical Association* **105**(492), 1480–1493.

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

Hutter, J.-C. & Rigollet, P. (2016), 'Optimal rates for total variation denoising', *arXiv preprint arXiv:1603.09388* .

Kim, S.-J., Koh, K., Boyd, S. & Gorinevsky, D. (2009), 'ℓ_1 trend filtering', *SIAM Review* **51**(2), 339–360.

Mammen, E. & van de Geer, S. (1997), 'Locally adaptive regression splines', *The Annals of Statistics* **25**(1), 387–413.

Qian, J. & Jia, J. (2012), 'On pattern recovery of the fused lasso', *arXiv preprint arXiv:1211.5194* .

Rinaldo, A. (2009), 'Properties and refinements of the fused lasso', *Annals of Statistics* **37**(5), 2922–2952.

Rojas, C. R. & Wahlberg, B. (2014), 'On change point detection using the fused lasso method', *arXiv preprint arXiv:1401.5408* .

Rudin, L., Osher, S. & Faterni, E. (1992), 'Nonlinear total variation based noise removal algorithms', *Physica D: Nonlinear Phenomena* **60**(1–4), 259–268.

Sadhanala, V., Wang, Y.-X. & Tibshirani, R. J. (2016), 'Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers', *arXiv preprint arXiv:1605.08400* .

Sharpnack, J., Rinaldo, A. & Singh, A. (2012), Sparsistency of the edge lasso over graphs, *in* 'Proceedings of the 15th International Conference on Artificial Intelligence and Statistics', pp. 1028–1036.

Steidl, G., Didas, S. & Neumann, J. (2006), 'Splines in higher order TV regularization', *International Journal of Computer Vision* **70**(3), 214–255.

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

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

van de Geer, S. (1990), 'Estimating a regression function', *Annals of Statistics* **18**(2), 907–924.

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press.

Venkatraman, E. S. (1992), Consistency results in multiple change-point problems, PhD thesis, Department of Statistics, Stanford University.

Vostrikova, L. (1981), 'Detecting 'disorder' in multidimensional random processes', *Soviet Mathematics Doklady* **24**, 55–59.

Wang, Y.-X., Sharpnack, J., Smola, A. & Tibshirani, R. J. (2016), 'Trend filtering on graphs', *Journal of Machine Learning Research* . To appear.

A Proof of Theorem 4

Here and henceforth, we write $N(r, S, \|\cdot\|)$ to denote the covering number of a set S in a norm $\|\cdot\|$, i.e., the smallest number of $\|\cdot\|$-balls of radius r needed to cover S. We call $\log N(r, S, \|\cdot\|)$ the log covering or entropy number. Recall that we define the scaled norm $\|\cdot\|_n = \|\cdot\|_2/\sqrt{n}$.

In the proof of Theorem 4, we will rely on the following result from van de Geer (1990) (which is derived closely from Dudley's chaining for sub-Gaussian processes).

Theorem 25 (**Theorem 3.3 of van de Geer 1990**). *Assume that $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^n$ has i.i.d. components drawn from a sub-Gaussian distribution, as in (8). Consider a set $\mathcal{X} \subseteq \mathbb{R}^n$ with $\|x\|_n \le 1$ for all $x \in \mathcal{X}$, and let $\mathcal{K}(\cdot)$ be a continuous function upper bounding the $\|\cdot\|_n$ entropy of $\mathcal{X}$, i.e., $\mathcal{K}(r) \ge \log N(r, \mathcal{X}, \|\cdot\|_n)$. Then there are constants $C_1, C_2, C_3, C_4 > 0$ (depending only on M, σ, the parameters of the underlying sub-Gaussian distribution in (8)), such that for all $t > C_1$, with*

$$t > C_2 \int_0^{t_0} \sqrt{\mathcal{K}(r)}\, dr,$$

where $t_0 = \inf\{r : \mathcal{K}(r) \le C_3 t^2\}$, we have

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} \frac{|\epsilon^\top x|}{\sqrt{n}} > t \right) \le 2\exp(-C_4 t^2).$$

Now we give the proof of Theorem 4.

Proof of Theorem 4. The proof is given in two parts, one in which we bound $\|\widehat{\delta}\|_2^2$ and the other in which we bound $\|\widehat{x}\|_2^2$. Recall that $\widehat{\delta} = P_0(\widehat{\theta} - \theta_0)$ and $\widehat{x} = P_1\widehat{\theta}$. Each part begins with a different "basic inequality", established by comparing the fused lasso objective at different points. First, we define the following events,

$$\Omega_0 = \left\{ \sup_{z \in \mathcal{M}} \frac{|\epsilon^\top z|}{\|z\|_2} \le \gamma c_I \sqrt{(\log s_0 + \log\log n) s_0 \log n} \right\}, \tag{35}$$

$$\Omega_1 = \left\{ \sup_{w \in \mathcal{R}^\perp} \frac{|\epsilon^\top w|}{\|D_{-S_0} w\|_1^{1/2} \|w\|_2^{1/2}} \le \gamma c_R (n s_0)^{1/4} \right\}, \tag{36}$$

$$\Omega_2 = \left\{ \sup_{\delta \in \mathcal{R}} \frac{|\epsilon^\top \delta|}{\|\delta\|_2} \le \gamma c_S \sqrt{s_0} \right\}, \tag{37}$$

where $\gamma > 1$ is parameter free to vary in our analysis, $c_I, c_R > 0$ are the constants in Lemmas 6, 7, and $c_S > 0$ is a constant to be determined below. Focusing on the third event, we will lower bound its probability by applying Theorem 25 to $\mathcal{X} = \mathcal{R} \cap \{\delta : \|\delta\|_n \le 1\}$. Note that

$$\log N(r, \mathcal{R} \cap \{\delta : \|\delta\|_n \le 1\}, \|\cdot\|_n) \le (s_0 + 1)\log(3/r),$$

as $\mathcal{R}$ is (s_0+1)-dimensional, and it is well-known that in $\mathbb{R}^d$, the number of balls of radius r that are needed to cover the unit ball is at most $(3/r)^d$. The quantity t_0 in Theorem 25 may be taken to be $t_0 = \inf\{r : (s_0+1)\log(3/r) \le C_3 C_1^2\} = 3\exp(-C_3 C_1^2/(s_0+1))$. The restrictions on t are hence $t > C_1$, as well as

$$t > C_2 \int_0^{t_0} \sqrt{(s_0+1)\log(3/r)}\, dr.$$

But, writing $\mathrm{erf}(\cdot)$ for the error function,

$$C_2 \int_0^{t_0} \sqrt{(s_0+1)\log(3/r)}\, dr = (\sqrt{s_0+1}) \cdot 3C_2 \left[r\sqrt{\log\frac{1}{r}} - \frac{1}{2}\mathrm{erf}\left(\sqrt{\log\frac{1}{r}}\right)\right]\Bigg|_0^{t_0/3} \le C_2\sqrt{s_0},$$

where the constant $C_2 > 0$ is adjusted to be larger, as needed. Let us define $c_S = \max\{C_1, C_2\}$ and $C_S = C_4$. Then we have by Theorem 25, for $t = \gamma c_S \sqrt{s_0}$ and any $\gamma > 1$,

$$1 - 2\exp(-C_S \gamma^2 c_S^2 s_0) \le \mathbb{P}\left(\sup_{\delta \in \mathcal{R}} \frac{|\epsilon^\top \delta|}{\sqrt{n}\|\delta\|_n} \le \gamma c_S \sqrt{s_0}\right) = \mathbb{P}\left(\sup_{\delta \in \mathcal{R}} \frac{|\epsilon^\top \delta|}{\|\delta\|_2} \le \gamma c_S \sqrt{s_0}\right) = \mathbb{P}(\Omega_2). \quad (38)$$

Controlling $\widehat{\delta}$. Comparing the objective in (2) at $\widehat{\theta} = P_0\widehat{\theta} + P_1\widehat{\theta} = \theta_0 + \widehat{\delta} + \widehat{x}$ and at $\theta_0 + \widehat{x}$, we have

$$\|\widehat{\delta} + \widehat{x} - \epsilon\|_2^2 + \lambda\|D\widehat{\theta}\|_1 \le \|\widehat{x} - \epsilon\|_2^2 + \lambda\|D(\theta_0 + \widehat{x})\|_1,$$

and by rearranging terms we obtain our basic inequality,

$$\|\widehat{\delta}\|_2^2 \le 2\widehat{\delta}^\top \epsilon + \lambda\Big(\|D(\theta_0 + \widehat{x})\|_1 - \|D\widehat{\theta}\|_1\Big), \quad (39)$$

which follows from the fact that $\widehat{\delta}^\top \widehat{x} = 0$ (as they lie in orthogonal subspaces). Furthermore, since $\widehat{\theta} = \theta_0 + \widehat{\delta} + \widehat{x}$,

$$\|D(\theta_0 + \widehat{x})\|_1 - \|D\widehat{\theta}\|_1 \le \|D\widehat{\delta}\|_1 = \|D_{S_0}\widehat{\delta}\|_1,$$

where we used the triangle inequality, and the fact that $D_{-S_0}\widehat{\delta} = 0$. So from our basic inequality in (39), we have that

$$\|\widehat{\delta}\|_2^2 \le 2\widehat{\delta}^\top \epsilon + \lambda\|D_{S_0}\widehat{\delta}\|_1,$$

and dividing by $\|\widehat{\delta}\|_2$, we get

$$\|\widehat{\delta}\|_2 \le 2\frac{|\widehat{\delta}^\top \epsilon|}{\|\widehat{\delta}\|_2} + \lambda\frac{\|D_{S_0}\widehat{\delta}\|_1}{\|\widehat{\delta}\|_2}.$$

Now observe that

$$\|D_{S_0}\widehat{\delta}\|_1 = \sum_{i=1}^{s_0} |\widehat{\delta}_{t_{i+1}} - \widehat{\delta}_{t_i}| \le 2\sum_{i=1}^{s_0+1} |\widehat{\delta}_{t_i}| \le 2\sqrt{(s_0+1)\sum_{i=1}^{s_0+1} \widehat{\delta}_{t_i}^2} \le 4\sqrt{s_0 \sum_{i=1}^{s_0+1} \frac{t_i - t_{i-1}}{W_n}\widehat{\delta}_{t_i}^2} = 4\sqrt{\frac{s_0}{W_n}}\|\widehat{\delta}\|_2.$$

The second inequality used Cauchy-Schwartz, and the last equality used that $\widehat{\delta}$ is piecewise constant on the blocks $B_0, \ldots, B_{s_0}$, as $\widehat{\delta} \in \mathcal{R} = \mathrm{span}\{\mathbb{1}_{B_0}, \ldots, \mathbb{1}_{B_{s_0}}\}$. Hence, on the event Ω_2, we have

$$\|\widehat{\delta}\|_2 \le 2\gamma c_S\sqrt{s_0} + 4\lambda\sqrt{\frac{s_0}{W_n}}. \quad (40)$$

Controlling $\widehat{x}$. We can establish our next basic inequality by comparing the objective in (2) at $\widehat{\theta}$ and $\theta_0 + \widehat{\delta}$,

$$\|\widehat{x} + \widehat{\delta} - \epsilon\|_2^2 + \lambda\|D\widehat{\theta}\|_1 \le \|\widehat{\delta} - \epsilon\|_2^2 + \lambda\|D(\theta_0 + \widehat{\delta})\|_1,$$

or, rearranged,

$$\begin{aligned}\|\widehat{x}\|_2^2 &\le 2\epsilon^\top \widehat{x} + \lambda\Big(\|D_{S_0}(\theta_0 + \widehat{\delta})\|_1 - \|D_{S_0}\widehat{\theta}\|_1 - \|D_{-S_0}\widehat{x}\|_1\Big) \\ &\le 2\epsilon^\top \widehat{x} + \lambda\Big(\|D_{S_0}\widehat{x}\|_1 - \|D_{-S_0}\widehat{x}\|_1\Big), \quad (41)\end{aligned}$$

where the first line used $\widehat{x}^\top\widehat{\delta} = 0$ and $D_{-S_0}\theta_0 = D_{-S_0}\widehat{\delta} = 0$, and the second used $\widehat{\theta} = \theta_0 + \widehat{\delta} + \widehat{x}$ and the triangle inequality.

Decompose $\widehat{x} = \widehat{z} + \widehat{w}$, where $\widehat{z} \in \mathcal{M}$ is the lower interpolant to $\widehat{x}$, as defined in Lemma 5, and $\widehat{w} = \widehat{x} - \widehat{z}$ is the remainder. Combining the basic inequality in (41) with (9) and (10) from Lemma 5,

$$\begin{aligned}\|\widehat{x}\|_2^2 &\le 2\epsilon^\top \widehat{z} + 2\epsilon^\top \widehat{w} + \lambda\Big(\|D_{S_0}\widehat{z}\|_1 - \|D_{-S_0}\widehat{z}\|_1 - \|D_{-S_0}\widehat{w}\|_1\Big) \\ &\le 2\epsilon^\top \widehat{z} + 4\lambda\sqrt{\frac{s_0}{W_n}}\|\widehat{z}\|_2 + 2\epsilon^\top \widehat{w} - \lambda\|D_{-S_0}\widehat{w}\|_1. \qquad (42)\end{aligned}$$

On the event Ω_0 in (35)

$$\epsilon^\top \widehat{z} \le \gamma c_I \sqrt{(\log s_0 + \log\log n)s_0 \log n}\|\widehat{z}\|_2.$$

Further, on the event Ω_1 in (36), since $P_1\widehat{w} \in \mathcal{R}^\perp$, $\|D_{-S_0}P_1\widehat{w}\|_1 = \|D_{-S_0}\widehat{w}\|_1$, and $\|P_1\widehat{w}\|_2 \le \|\widehat{w}\|_2$,

$$\epsilon^\top P_1\widehat{w} \le \gamma c_R (ns_0)^{1/4}\|D_{-S_0}\widehat{w}\|_1^{1/2}\|\widehat{w}\|_2^{1/2},$$

Also, on the event Ω_2 in (37), since $P_0\widehat{w} \in \mathcal{R}$,

$$\epsilon^\top P_0\widehat{w} \le \gamma c_S\sqrt{s_0}\|\widehat{w}\|_2.$$

Hence, on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$, combining the last three displays with (42),

$$\begin{aligned}\|\widehat{x}\|_2^2 \le 2\Bigg(\gamma c_I\sqrt{(\log s_0 + \log\log n)s_0\log n} + 2\lambda\sqrt{\frac{s_0}{W_n}}\Bigg)\|\widehat{z}\|_2 + 2\gamma c_S\sqrt{s_0}\|\widehat{w}\|_2 + \\ 2\gamma c_R(ns_0)^{1/4}\|D_{-S_0}\widehat{w}\|_1^{1/2}\|\widehat{w}\|_2^{1/2} - \lambda\|D_{-S_0}\widehat{w}\|_1. \qquad (43)\end{aligned}$$

Consider the first case in which $2\gamma c_R(ns_0)^{1/4}\|D_{-S_0}\widehat{w}\|_1^{1/2}\|\widehat{w}\|_2^{1/2} \ge \lambda\|D_{-S_0}\widehat{w}\|_1$. Then

$$\|D_{-S_0}\widehat{w}\|_1 \le 4\bigg(\frac{\gamma c_R}{\lambda}\bigg)^2\sqrt{ns_0}\|\widehat{w}\|_2,$$

and from (43), on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$,

$$\|\widehat{x}\|_2 \le 2\gamma c_I\sqrt{(\log s_0 + \log\log n)s_0\log n} + 4\lambda\sqrt{\frac{s_0}{W_n}} + 2\gamma c_S\sqrt{s_0} + \frac{4\gamma^2 c_R^2\sqrt{ns_0}}{\lambda}. \qquad (44)$$

where in the above we used (11). In the case that $2\gamma c_R(ns_0)^{1/4}\|D_{-S_0}\widehat{w}\|_1^{1/2}\|\widehat{w}\|_2^{1/2} < \lambda\|D_{-S_0}\widehat{w}\|_1$, we have from (43), on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$,

$$\|\widehat{x}\|_2 \le 2\gamma c_I\sqrt{(\log s_0 + \log\log n)s_0\log n} + 4\lambda\sqrt{\frac{s_0}{W_n}} + 2\gamma c_S\sqrt{s_0}.$$

Therefore, the bound (44) always holds on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$.

Putting it all together. As $\|\widehat{\theta} - \theta_0\|_2 \le \|\widehat{x}\|_2 + \|\widehat{\delta}\|_2$, combining (44) and (40), we see that

$$\|\widehat{\theta} - \theta_0\|_2 \le 4\gamma c_S\sqrt{s_0} + 8\lambda\sqrt{\frac{s_0}{W_n}} + 2\gamma c_I\sqrt{(\log s_0 + \log\log n)s_0\log n} + \frac{4\gamma^2 c_R^2\sqrt{ns_0}}{\lambda},$$

on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$. We see that there exists a constant $c > 0$, such that for large enough n, and any $\gamma > 1$,

$$\|\widehat{\theta} - \theta_0\|_2^2 \le \gamma^4 c s_0\Bigg((\log s_0 + \log\log n)\log n + \frac{\lambda^2}{W_n} + \frac{n}{\lambda^2}\Bigg), \qquad (45)$$

on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$. Furthermore, using the union bound along with Lemmas 6, 7, and (38), we find that

$$\begin{aligned}\mathbb{P}\big((\Omega_0 \cap \Omega_1 \cap \Omega_2)^c\big) \le 2\exp\big(-C_I\gamma^2 c_I^2(\log s_0 + \log\log n)\big) + \\ 2\exp(-C_R\gamma^2 c_R^2\sqrt{s_0}) + 2\exp(-C_S\gamma^2 c_S^2 s_0) \le \exp(-C\gamma^2),\end{aligned}$$

for an appropriately defined constant $C > 0$. Optimizing the bound in (45) to choose the tuning parameter λ yields $\lambda = (nW_n)^{1/4}$. Plugging this in gives the final result. □

B Proofs of Lemmas 5, 6, 7

Proof of Lemma 5. We give an explicit construction of a lower interpolant $z \in \mathcal{M}$ to x, given the changepoints $0 = t_0 < \ldots < t_{s_0+1} = n$. We will use the notation $a_+ = \max\{0, a\}$. For $i = 0, \ldots, s_0$, define $z^{(i+)} \in \mathbb{R}^{t_{i+1}-t_i}$ by setting $g_i^+ = \mathrm{sign}(x_{t_i})$ and

$$z_j^{(i+)} = g_i^+ \cdot \min\Big\{(g_i^+ x_{t_i+1})_+, \ldots, (g_i^+ x_{t_i+j})_+\Big\}, \quad j = 1, \ldots, t_{i+1} - t_i.$$

Similarly, define $z^{(i-)} \in \mathbb{R}^{t_{i+1}-t_i}$ by setting $g_i^- = \mathrm{sign}(x_{t_{i+1}-1})$ and

$$z_j^{(i-)} = g_i^- \cdot \min\Big\{(g_i^- x_{t_i+j})_+, \ldots, (g_i^- x_{t_{i+1}})_+\Big\}, \quad j = 1, \ldots, t_{i+1} - t_i.$$

Note that $z_1^{(i+)} = x_{t_i+1}$ and $z_{t_{i+1}-t_i}^{(i-)} = x_{t_{i+1}}$; also, $\{|z_j^{(i+)}|\}_{j=1}^{t_{i+1}-t_i}$ is a nonincreasing sequence, and $\{|z_j^{(i-)}|\}_{j=1}^{t_{i+1}-t_i}$ is nondecreasing. Furthermore,

$$\mathrm{sign}\big(z_1^{(i+)}\big) \cdot \mathrm{sign}\big(z_j^{(i+)}\big) \geq 0 \quad \text{and} \quad \mathrm{sign}\big(z_{t_{i+1}-t_i}^{(i-)}\big) \cdot \mathrm{sign}\big(z_j^{(i-)}\big) \geq 0, \qquad j = 1, \ldots, t_{i+1} - t_i.$$

Lastly, notice that there exists a point $j' \in 1, \ldots, t_{i+1} - t_i - 1$ (not necessarily unique) such that

$$\min_{k \in \{1,\ldots,t_{i+1}-t_i\}} \big|z_k^{(i+)}\big| = \big|z_{j'+1}^{(i+)}\big| = \big|z_j^{(i+)}\big|, \qquad j = j'+1, \ldots, t_{i+1} - t_i, \tag{46}$$

$$\min_{k \in \{1,\ldots,t_{i+1}-t_i\}} \big|z_k^{(i-)}\big| = \big|z_{j'}^{(i-)}\big| = \big|z_j^{(i-)}\big|, \qquad j = 1, \ldots, j'. \tag{47}$$

We construct by $z_{t_i+j} = z_j^{(i+)}$ for $j = 1, \ldots, j'$, and $z_{t_i+j} = z_j^{(i-)}$ for $j = j'+1, \ldots, t_{i+1} - t_i$. Letting $t_i' = t_i + j'$ and repeating this process for $i = 0, \ldots, s_0$, we have constructed $z \in \mathcal{M}$.

We now verify the claimed properties for the constructed lower interpolant z. For $i = 0, \ldots, s_0$, and any $j = 1, \ldots, t_{i+1} - t_i$, we have

$$\mathrm{sign}(z_j^{(i+)}) \cdot \mathrm{sign}(x_{t_i+j}) \geq 0, \tag{48}$$

$$|z_j^{(i+)}| \leq |x_{t_i+j}|, \tag{49}$$

Further, for any $j = 1, \ldots, t_{i+1} - t_i - 1$,

$$\mathrm{sign}\Big((Dz^{(i+)})_j\Big) \cdot \mathrm{sign}\big((Dx)_{t_i+j}\big) \geq 0, \tag{50}$$

$$\Big|(Dz^{(i+)})_j\Big| \leq \big|(Dx)_{t_i+j}\big|. \tag{51}$$

To see why (50) holds, note that $\mathrm{sign}(Dz^{(i+)})_j \in \{-1, 0\}$, $(Dz^{(i+)})_j < 0$ imply $(D(g_i^+ x)_+)_{t_i+j} < 0$. To see why (51) holds, if $(Dz^{(i+)})_j \neq 0$, then we know that

$$|z_{j+1}^{(i+)} - z_j^{(i+)}| \leq \Big| \min\Big\{(g_i^+ x_{t_i+j+1})_+, (g_i^+ x_{t_i+j})_+\Big\} - (g_i^+ x_{t_i+j})_+ \Big| \leq |x_{t_i+j+1} - x_{t_i+j}|,$$

where we used the observation that $|\min\{a, b\} - b| \geq |\min\{a, b, c\} - \min\{b, c\}|$.

It can be shown by nearly equivalent steps that $z^{(i-)}$, and z both satisfy properties analogous to (48)–(51). Using (48) and (49) on z gives (11). Using (50) and (51) on z gives (9) (note that if $\mathrm{sign}(a) = \mathrm{sign}(b)$ and $|a| > |b|$, then $|a| = |b| + |a - b|$). Because $z_{t_i+1} = x_{t_i+1}$ and $z_{t_{i+1}} = x_{t_{i+1}}$ for all $i = 0, \ldots, s_0$, we have the equality in (10) (since $D_{t_i} z = z_{t_i+1} - z_{t_i} = x_{t_i+1} - x_{t_i} = D_{t_i} x$).

Finally, for each $i = 0, \ldots, s_0$, define $t_i'' = t_i'$ if $|z_{t_i'}| \geq |z_{t_i'+1}|$ and $t_i'' = t_i' + 1$ otherwise. Observe that by (46) and (47), it holds that $|z_{t_i''}| = \min_{j=1,\ldots,t_{i+1}-t_i} |z_{t_i+j}|$. The inequality in (10) is finally established by the following chain of inequalities:

$$\begin{aligned}\|D_{S_0}z\|_1 &= \sum_{i=1}^{s_0} |z_{t_i+1} - z_{t_i}| \le \sum_{i=1}^{s_0} |z_{t_i+1}| + |z_{t_i}| \\ &= \sum_{i=1}^{s_0} \big(|z_{t_i+1}| - |z_{t''_i}|\big) + \big(|z_{t_i}| - |z_{t''_{i-1}}|\big) + |z_{t''_{i-1}}| + |z_{t''_i}| \\ &\le \|D_{-S_0}z\|_1 + 2\sum_{i=0}^{s_0} |z_{t''_i}| \le \|D_{-S_0}z\|_1 + 4\sqrt{\frac{s_0}{W_n}}\|z\|_2,\end{aligned}$$

where in the second inequality, we used $|a| - |c| \le |a-c| \le |a-b| + |b-c|$, and in the last inequality, we used the above property of $z_{t''_i}$ and

$$\sum_{i=0}^{s_0} |z_{t''_i}| \le 2\sqrt{s_0}\sqrt{\sum_{i=0}^{s_0} |z_{t''_i}|^2} \le 2\sqrt{s_0 \sum_{i=0}^{s_0} \frac{t_{i+1}-t_i}{W_n} z_{t''_i}^2} \le 2\sqrt{\frac{s_0}{W_n}}\|z\|_2.$$

This completes the proof. $\square$

Proof of Lemma 6. We consider $\epsilon \in \mathbb{R}^n$, an i.i.d. sub-Gaussian vector as referred to in the statement of the lemma, and arbitrary $z \in \mathcal{M}$. In this proof, we will also consider $E(t)$ and $Z(t)$, real-valued functions over $[0,n]$, constructed so that $E(t) = \epsilon_{\lceil t \rceil}$ for all t (i.e., $E(t)$ is a step function), $Z(t) = z_t$ for $t = 1,\dots,n$, and $Z(t)$ is smooth and monotone over $(t_i, t'_i]$ and $(t'_i, t_{i+1}]$ for $i = 0,\dots,s_0$. These functions will also satisfy the boundary conditions $E(0) = \epsilon_1$ and $Z(0) = z_1$.

Let $F(t) = \int_0^t E(u)\,du$. As ϵ is random, $E(t)$ and $F(t)$ are also random. It can be shown that there exists constants $c_I, C_I > 0$ such that for any $\gamma > 1$,

$$\begin{aligned}\mathbb{P}\left(\frac{|F(t)-F(t_i)|}{\sqrt{|t-t_i|}} \le \gamma c_I \sqrt{\log s_0 + \log\log n},\ \text{for } t \in (t_i, t_{i+1}],\ i = 0,\dots,s_0\right) \\ \ge 1 - 2\exp\big(-C_I\gamma^2 c_I^2(\log s_0 + \log\log n)\big). \quad (52)\end{aligned}$$

So as not to distract from the main flow of ideas, we now proceed to prove Lemma 6, and we later provide a proof of (52). Let Ω_3 denote the event in consideration on the left-hand side of (52). By integration by parts,

$$\int_{t_i}^{t'_i} E(t)Z(t)\,dt = Z(t'_i)(F(t'_i) - F(t_i)) - \int_{t_i}^{t'_i} Z'(t)(F(t) - F(t_i))\,dt$$

where $Z'(t) = \frac{d}{dt}Z(t)$. Thus, on the event Ω_3,

$$\left|\int_{t_i}^{t'_i} E(t)Z(t)\,dt\right| \le \gamma c_I \sqrt{\log s_0 + \log\log n}\left(|Z(t'_i)|\sqrt{t'_i - t_i} + \left|\int_{t_i}^{t'_i} Z'(t)\sqrt{t-t_i}\,dt\right|\right), \quad (53)$$

since Z' does not change sign within the intervals $(t_i, t'_i], (t'_i, t_{i+1}]$ (as $z \in \mathcal{M}$). For n large enough, we can upper bound the last term in (53) as follows

$$\left|\int_{t_i}^{t'_i} Z'(t)\sqrt{t-t_i}\,dt\right| = \left|\int_{t_i}^{t_i+n^{-1}} Z'(t)\sqrt{t-t_i}\,dt\right| + \left|\int_{t_i+n^{-1}}^{t'_i} Z'(t)\sqrt{t-t_i}\,dt\right|. \quad (54)$$

Using integration by parts and the triangle inequality, on the second term in (54),

$$\left|\int_{t_i+n^{-1}}^{t'_i} Z'(t)\sqrt{t-t_i}\,dt\right| = |Z(t'_i)|\sqrt{t'_i - t_i} + \left|\frac{Z(t_i+n^{-1})}{\sqrt{n}}\right| + \frac{1}{2}\left|\int_{t_i+n^{-1}}^{t'_i} \frac{Z(t)}{\sqrt{t-t_i}}\,dt\right|. \quad (55)$$

By Cauchy-Schwartz on the last term in (55),

$$\left|\int_{t_i+n^{-1}}^{t_i'} \frac{Z(t)}{\sqrt{t-t_i}}\,dt\right| \le \left(\int_{t_i+n^{-1}}^{t_i'} Z(t)^2\,dt\right)^{1/2} \left(\int_{t_i+n^{-1}}^{t_i'} \frac{1}{t-t_i}\,dt\right)^{1/2}$$
$$\le \left(\int_{t_i+n^{-1}}^{t_i'} Z(t)^2\,dt\right)^{1/2} \sqrt{2\log n}. \tag{56}$$

Now examining the first term in (54),

$$\left|\int_{t_i}^{t_i+n^{-1}} Z'(t)\sqrt{t-t_i}\,dt\right| \le n^{-1/2}\left|\int_{t_i}^{t_i+n^{-1}} Z'(t)\,dt\right| = \frac{|Z(t_i+n^{-1})-Z(t_i)|}{\sqrt{n}}.$$

But because we only require Z to be a piecewise monotonic and smooth interpolant then we are at liberty to make $Z(t_i+n^{-1}) = Z(t_i)$, forcing this term to be 0. In order to bound $Z(t_i')$, notice that because $|Z(t)|$ is non-increasing over the interval $(t_i, t_i']$ we have that

$$Z(t_i')^2|t_i'-t_i| \le \int_{t_i}^{t_i'} Z(t)^2\,dt. \tag{57}$$

Combining (53)–(57), we have that on the event Ω_3,

$$\left|\int_{t_i}^{t_i'} E(t)Z(t)\,dt\right| \le \alpha_n\left(2+\sqrt{\frac{\log n}{2}}\right)\left(\int_{t_i}^{t_i'} Z(t)^2\,dt\right)^{1/2} + \alpha_n\frac{|Z(t_i)|}{\sqrt{n}}. \tag{58}$$

where we have abbreviated $\alpha_n = \gamma c_I\sqrt{\log s_0 + \log\log n}$. Through nearly identical steps we can show that on the event Ω_3,

$$\left|\int_{t_i'}^{t_{i+1}} E(t)Z(t)\,dt\right| \le \alpha_n\left(2+\sqrt{\frac{\log n}{2}}\right)\left(\int_{t_i'}^{t_{i+1}} Z(t)^2\,dt\right)^{1/2} + \alpha_n\frac{|Z(t_{i+1})|}{\sqrt{n}}. \tag{59}$$

Therefore

$$\left|\int_0^n E(t)Z(t)\,dt\right| \le \sum_{i=0}^{s_0}\left(\left|\int_{t_i}^{t_i'} E(t)Z(t)\,dt\right| + \left|\int_{t_i'}^{t_{i+1}} E(t)Z(t)\,dt\right|\right)$$
$$\le \alpha_n\sqrt{2s_0+2}\left(2+\sqrt{\frac{\log n}{2}}\right)\left(\int_0^n Z(t)^2\,dt\right)^{1/2} + 2\alpha_n\frac{\|z\|_1}{\sqrt{n}}, \tag{60}$$

where in the second line we applied (58), (59), and the Cauchy-Schwartz inequality. Because we can choose $Z(t)$ to be arbitrarily close to $z_{\lceil t\rceil}$ over all t, the integral $(\int_0^n Z(t)^2\,dt)^{1/2}$ is approaching $\|z\|_2$ and $\int_0^n E(t)Z(t)\,dt$ is approaching $\epsilon^\top z$. Furthermore, because $\|z\|_1 \le \sqrt{n}\|z\|_2$, the first term in (60) dominates. Hence on the event Ω_3, we have established yet

$$|\epsilon^\top z| \le \gamma c_I\sqrt{(\log s_0 + \log\log n)s_0\log n}\,\|z\|_2,$$

where the constant c_I is adjusted to be larger, as needed. Noting that the event Ω_3 does not depend on z, the result follows. $\square$

Proof of claim (52). We will construct a covering for $\mathcal{V} = \cup_{i=0}^{s_0}\mathcal{V}_i$, where for each $i = 0,\dots,s_0$,

$$\mathcal{V}_i = \left\{\sqrt{\frac{n}{|A|}}\mathbb{1}_A : A = \{t_i,\dots,t\},\ t = t_i+1,\dots,n\right\} \cup \left\{\sqrt{\frac{n}{|A|}}\mathbb{1}_A : A = \{t,\dots,t_i\},\ t = 1,\dots,t_i-1\right\}.$$

Note that our scaling is such that, for any $a = \sqrt{n/|A|}\mathbb{1}_A$, where $A \subseteq \{1,\dots,n\}$, we have $\|a\|_n = 1$. Further, for any other $b = \sqrt{n/|B|}\mathbb{1}_B$, where $B \subseteq \{1,\dots,n\}$, we have

$$\|a-b\|_n^2 = \frac{|A\cap B|}{(\sqrt{|A|}-\sqrt{|B|})^2} + \frac{|A\setminus B|}{|A|} + \frac{|B\setminus A|}{|B|} = 2\bigg(1 - \frac{|A\cap B|}{\sqrt{|A||B|}}\bigg), \tag{61}$$

We first construct a covering for each set $\mathcal{V}_i$, $i = 0,\dots,s_0$, restricting our attention to a radius $0 < r < \sqrt{2}$. Let $\alpha = \lceil (1-r^2/2)^{-2} \rceil$, and consider the set

$$\mathcal{C}_i = \left\{ \sqrt{\frac{n}{|A|}}\mathbb{1}_A : A = \big\{t_i,\dots,\min\{t_i+\alpha^j, n\}\big\},\ j = 1,\dots,\lceil \log n/\log\alpha \rceil \right\} \\ \cup\ \left\{ \sqrt{\frac{n}{|A|}}\mathbb{1}_A : A = \big\{\max\{t_i-\alpha^j, 1\},\dots,t_i\big\},\ j = 1,\dots,\lceil \log n/\log\alpha \rceil \right\}.$$

Here, the set $\mathcal{C}_i$ has at most $2\lceil \log n/\log\alpha \rceil \le 4\log n/\log\alpha$ elements, and by (61), balls of radius r around elements in $\mathcal{C}_i$ cover the set $\mathcal{V}_i$. This establishes that

$$N(r, \mathcal{V}_i, \|\cdot\|_n) \le \frac{-2\log n}{\log(1-r^2/2)}. \tag{62}$$

For a radius $0 < r < \sqrt{2}$, the covering number for $\mathcal{V} = \cup_{i=0}^{s_0}\mathcal{V}_i$ can be obtained by just taking a union of the covers in (62) over $i = 0,\dots,s_0$, giving

$$N(r, \mathcal{V}, \|\cdot\|_n) \le \sum_{i=0}^{s_0} N(r, \mathcal{V}_i, \|\cdot\|_n) \le 2(s_0+1)\bigg(\frac{-\log n}{\log(1-r^2/2)}\bigg). \tag{63}$$

Using (61) once more, the diameter of the set $\mathcal{V}$ is $\sqrt{2}$, hence if $r \ge 1/\sqrt{2}$, then we need only 1 ball to cover $\mathcal{V}$. Combining this fact with (63), we obtain

$$N(r, \mathcal{V}, \|\cdot\|_n) \le \begin{cases} 2(s_0+1)\bigg(\dfrac{-\log n}{\log(1-r^2/2)}\bigg) & \text{if } 0 < r < 1/\sqrt{2} \\ 1 & \text{if } r \ge 1/\sqrt{2} \end{cases}. \tag{64}$$

Now let us apply Theorem 25, with $\mathcal{X} = \mathcal{V}$. First, we remark that the quantity t_0 in Theorem 25 may be taken to be $t_0 = 1/\sqrt{2}$. The bounds on t in the theorem are $t > C_1$, as well as

$$t > C_2 \int_0^{1/\sqrt{2}} \sqrt{\log\bigg(2(s_0+1)\frac{-\log n}{\log(1-r^2/2)}\bigg)}\, dr.$$

Next, we know that the right-hand side above is upper bounded by

$$C_2 \int_0^{1/\sqrt{2}} \left[\sqrt{\log\big(2(s_0+1)\log n\big)} + \sqrt{\log\bigg(\frac{-1}{\log(1-r^2/2)}\bigg)} \right] dr \\ = C_2\sqrt{\frac{\log\big(2(s_0+1)\log n\big)}{2}} + C_2\sqrt{2}\int_0^{1/2} \sqrt{\log\left(\frac{1}{\log\big(\frac{1}{1-x^2}\big)}\right)}\, dx.$$

One can verify that the the integral in the second term above converges to a finite constant (upper bounded by 1 in fact). Thus the entire expression above is upper bounded by $C_2\sqrt{\log s_0 + \log\log n}$, where the constant $C_2 > 0$ is adjusted to be larger, as needed. Therefore, letting $c_I = \max\{C_1, C_2\}$, we may restrict our attention to $t > c_I \sqrt{\log s_0 + \log\log n}$ in Theorem 25, and letting $C_I = C_4$, the conclusion reads, for $t = \gamma c_I$ and $\gamma > 1$,

$$\mathbb{P}\bigg(\sup_{a \in \mathcal{V}} \frac{\epsilon^\top a}{\sqrt{n}} > \gamma c_I \sqrt{\log s_0 + \log\log n} \bigg) \leq 2 \exp\big(- C_I \gamma^2 c_I^2 (\log s_0 + \log\log n) \big).$$

Recalling the form of $a = \sqrt{n/|A|} \mathbb{1}_A \in \mathcal{V}$, the above may be rephrased as

$$\mathbb{P}\bigg(\frac{\sum_{j=t_i}^t \epsilon_j}{\sqrt{|t - t_i|}} > \gamma c_I \sqrt{\log s_0 + \log\log n}, \text{ for } t = 1, \ldots, n, \ i = 0, \ldots, s_0 \bigg) \\ \leq 2 \exp\big(- C_I \gamma^2 c_I^2 (\log s_0 + \log\log n) \big). \quad (65)$$

Finally, consider the following event

$$\Omega_4 = \left\{ \frac{|F(t) - F(t_i)|}{\sqrt{|t - t_i|}} \leq \gamma c_I \sqrt{\log s_0 + \log\log n}, \text{for } t = 1, \ldots, n, \ i = 0, \ldots, s_0 \right\}.$$

Recalling that $E(t) = \epsilon_{\lceil t \rceil}$ for all $t \in [0, 1]$, we have $F(t) = \int_0^t E(u)\, du = \sum_{j=0}^t \epsilon_j$ for $t = 1, \ldots, n$. In (65), we have thus shown $\mathbb{P}(\Omega_4) \geq 1 - 2 \exp(-C_I \gamma^2 c_I^2 (\log s_0 + \log\log n))$. Note that $|F(t) - F(t_i)|$ is piecewise linear with knots at $t = 1, \ldots, n$ and $\sqrt{|t - t_i|}$ is concave in between these knots, so if $|F(t) - F(t_i)| / \sqrt{|t - t_i|} \leq \gamma c_I \sqrt{\log s_0 + \log\log n}$ for $t = 1, \ldots, n$, then the same bound must hold over all $t \in [0, n]$. This shows that $\Omega_4 \supseteq \Omega_3$, where Ω_3 is the event in question in the left-hand side of (52); in other words, we have verified (52). ☐

For the proof of Lemma 7, we will need the following result from van de Geer (1990).

Lemma 26 (**Lemma 3.5 of van de Geer 1990**)**.** *Assume the conditions in Theorem 25, and additionally, assume that for some $\zeta \in (0, 1)$ and $K > 0$,*

$$\mathcal{K}(r) \leq K r^{-2\zeta},$$

where, recall, $\mathcal{K}(r)$ is a continuous function upper bounding the entropy number $\log N(r, \mathcal{X}, \|\cdot\|_n)$. Then there exists constants C_0, C_1 (depending only on M, σ in (8)) such that for any $t \geq C_0$,

$$\mathbb{P}\bigg(\sup_{x \in \mathcal{X}} \frac{|\epsilon^\top x|}{\sqrt{n} \|x\|_n^{1-\zeta}} > t\sqrt{K} \bigg) \leq \exp(-C_1 t^2 K).$$

Proof of Lemma 7. Recall that for $i = 0, \ldots, s_0$, we define $B_i = \{t_i + 1, \ldots, t_{i+1}\}$. For $i = 0, \ldots, s_0$, also define $n_i = |B_i|$, the scaled norm $\|\cdot\|_{n_i} = \|\cdot\|_2 / \sqrt{n_i}$, and

$$\mathcal{X}_i = \Big\{ w^{(i)} \in \mathbb{R}^{n_i} : (\mathbb{1}^{(i)})^\top w^{(i)} = 0, \ \|D^{(i)} w^{(i)}\|_1 \leq 1, \ \|w^{(i)}\|_{n_i} \leq 1 \Big\}.$$

Here, we write $\mathbb{1}^{(i)} \in \mathbb{R}^{n_i}$ for the vector of all 1s, and $D^{(i)} \in \mathbb{R}^{(n_i - 1) \times n}$ for the difference operator, as in (5) but of smaller dimension. The set $\mathcal{X}_i$ is the discrete total variation space in $\mathbb{R}^{n_i}$, where all elements are centered and have scaled norm at most 1. From well-known results on entropy bounds for total variation spaces (e.g., from Lemma 11 and Corollary 12 of Wang et al. (2016)), we have

$$\log N(r, \mathcal{X}_i, \|\cdot\|_{n_i}) \leq \frac{C}{r},$$

for a universal constant $C > 0$. Hence we may apply Lemma 26 with $\mathcal{X} = \mathcal{X}_i$ and $\zeta = 1/2$: for the random variable

$$M_i = \sup \left\{ \frac{|\epsilon_{B_i}^\top w^{(i)}|}{\sqrt{n_i} \|w^{(i)}\|_{n_i}^{1/2}} : w^{(i)} \in \mathcal{X}_i \right\},$$

we may take $t = \gamma C_0$ in the lemma, for any $\gamma > 1$, and conclude that

$$\mathbb{P}\Big(M_i > \gamma C_0 \sqrt{C}\Big) \le \exp(-C_1 \gamma^2 C_0^2 C).$$

Notice that we may rewrite M_i as

$$M_i = \sup\left\{ \frac{|\epsilon_{B_i}^\top w^{(i)}|}{n_i^{1/4} \|D^{(i)} w^{(i)}\|_1^{1/2} \|w^{(i)}\|_2^{1/2}} : w^{(i)} \in \mathbb{R}^{n_i},\ (\mathbb{1}^{(i)})^\top w^{(i)} = 0 \right\},$$

and therefore

$$\mathbb{P}\Bigg(\sup_{w^{(i)} \in \mathbb{R}^{n_i},\, (\mathbb{1}^{(i)})^\top w^{(i)} = 0} \frac{|\epsilon_{B_i}^\top w^{(i)}|}{\|D^{(i)} w^{(i)}\|_1^{1/2} \|w^{(i)}\|_2^{1/2}} > \gamma C_0 \sqrt{C} n_i^{1/4} \Bigg) \le \exp(-C_1 \gamma^2 C_0^2 C).$$

Using the union bound,

$$\mathbb{P}\left(\sup_{\substack{w^{(i)} \in \mathbb{R}^{n_i},\, (\mathbb{1}^{(i)})^\top w^{(i)} = 0 \\ i = 0,\ldots,s_0}} \frac{|\epsilon_{B_i}^\top w^{(i)}|}{\|D^{(i)} w^{(i)}\|_1^{1/2} \|w^{(i)}\|_2^{1/2}} > \gamma C_0 \sqrt{C} n_i^{1/4} \right) \le (s_0 + 1) \exp(-C_1 \gamma^2 C_0^2 C).$$

Define the constants $c_R = \max\{C_0 \sqrt{C}, 1\}$ and $C_R = \max\{C_1/2, 1\}$. Then this ensures that we have $2 C_R \gamma^2 c_R^2 \sqrt{s_0} \ge \log(s_0 + 1)$ for any $\gamma > 1$ and any s_0, thus

$$\mathbb{P}\left(\sup_{\substack{w^{(i)} \in \mathbb{R}^{n_i},\, (\mathbb{1}^{(i)})^\top w^{(i)} = 0 \\ i = 0,\ldots,s_0}} \frac{|\epsilon_{B_i}^\top w^{(i)}|}{\|D^{(i)} w^{(i)}\|_1^{1/2} \|w^{(i)}\|_2^{1/2}} > \gamma c_R (n_i s_0)^{1/4} \right) \le \exp(-C_R \gamma^2 c_R^2 \sqrt{s_0}).$$

The proof is completed by noting the following: if $w \in \mathcal{R}^\perp$, then $(\mathbb{1}^{(i)})^\top w_{B_i} = 0$ for all $i = 0, \ldots, s_0$, and so on the event in consideration in the last display,

$$\begin{aligned}
|\epsilon^\top w| &\le \sum_{i=0}^{s_0} |\epsilon_{B_i}^\top w_{B_i}| \le \gamma c_R s_0^{1/4} \sum_{i=0}^{s_0} n_i^{1/4} \|D^{(i)} w_{B_i}\|_1^{1/2} \|w_{B_i}\|_2^{1/2} \\
&\le \gamma c_R s_0^{1/4} \left(\sum_{i=0}^{s_0} \|D^{(i)} w_{B_i}\|_1 \right)^{1/2} \left(\sum_{i=0}^{s_0} n_i^{1/2} \|w_{B_i}\|_2 \right)^{1/2} \\
&= \gamma c_R s_0^{1/4} \|D_{-S_0} w\|_1^{1/2} \left(\sum_{i=0}^{s_0} n_i^{1/2} \|w_{B_i}\|_2 \right)^{1/2} \\
&\le \gamma c_R s_0^{1/4} \|D_{-S_0} w\|_1^{1/2} \left(\sum_{i=0}^{s_0} \|w_{B_i}\|_2^2 \right)^{1/4} \left(\sum_{i=0}^{s_0} n_i \right)^{1/4} \\
&= \gamma c_R s_0^{1/4} \|D_{-S_0} w\|_1^{1/2} \|w\|_2^{1/2} n^{1/4},
\end{aligned}$$

by two successive uses of Cauchy-Schwartz. □

C Proofs of Lemmas 12, 13

Both proofs follow from standard techniques in convex analysis.

Proof of Lemma 12. We first consider the convex optimization problem

$$\min_{x \in \mathbb{R}^m} \; a^\top x \;\; \text{subject to} \;\; \|x - c\|_2 \le r, \tag{66}$$

whose Lagrangian may be written as, for a dual variable $\lambda \ge 0$,

$$L(x, \lambda) = a^\top x + \lambda(\|x - c\|_2^2 - r^2).$$

The stationarity condition is $a + \lambda(x - c) = 0$, thus $x = c - a/\lambda$. By primal feasibility, $\|x - c\|_2 \le r$, we see that we can take $\lambda = \|a\|_2/r$, which gives a solution $x = c - ra/\|a\|_2$. The optimal value in (67) is therefore $a^\top x = a^\top c - r\|a\|_2$. By the same logic, the optimal value of the convex problem

$$\max_{x \in \mathbb{R}^m} \; a^\top x \;\; \text{subject to} \;\; \|x - c\|_2 \le r \tag{67}$$

is $a^\top c + r\|a\|_2$. Now we can read off the optimal value of (20) from those of (66), (67): its optimal value is

$$\max\big\{ -\big(a^\top c - r\|a\|_2^2\big), \; a^\top c + r\|a\|_2 \big\} = |a^\top c| + r\|a\|_2^2,$$

completing the proof. ☐

Proof of Lemma 13. The proof is nearly immediate from the proof of Lemma 12, above. Notice that the optimal value of (21) is lower bounded by that of (66), which we already know is $a^\top c - r\|a\|_2^2$. But when the latter is nonnegative, this is also the optimal value of (21). Repeating the argument with $-a$ in place of a gives the result as stated in the lemma. ☐

D Proof of Lemma 17

To facilitate the proof, we define the concept of a *local maximum* among the absolute filter values: a location i is a local maximum if its absolute filter value $|F_i(\widetilde{\theta})|$ is be greater than or equal to the absolute values at neighboring locations, and strictly greater than at least one of these values (where the boundary points are treated as having just one neighboring location). Specifically, a location i must satisfy one of the following conditions

$$|F_{i-1}(\widetilde{\theta})| < |F_i(\widetilde{\theta})|, \; |F_{i+1}(\widetilde{\theta})| \le |F_i(\widetilde{\theta})|, \qquad \text{if } i \in \{b_n + 1, \dots, n - b_n - 1\}, \tag{68}$$

$$|F_{i-1}(\widetilde{\theta})| \le |F_i(\widetilde{\theta})|, \; |F_{i+1}(\widetilde{\theta})| < |F_i(\widetilde{\theta})|, \qquad \text{if } i \in \{b_n + 1, \dots, n - b_n - 1\}, \tag{69}$$

$$|F_{i+1}(\widetilde{\theta})| < |F_i(\widetilde{\theta})| \qquad \text{if } i = b_n, \tag{70}$$

$$|F_{i-1}(\widetilde{\theta})| < |F_i(\widetilde{\theta})| \qquad \text{if } i = n - b_n. \tag{71}$$

Let $L(\widetilde{\theta})$ denote the set of local maximums derived from the filter with bandwidth b_n, i.e., the set of locations i satisfying one of the four conditions (68)–(71).

We first show that $L(\widetilde{\theta}) \subseteq I_C(\widetilde{\theta})$. Fix $i \in L(\widetilde{\theta})$. The boundary cases, $i = b_n$ or $i = n - b_n$, are handled directly by the definition of $I_C(\widetilde{\theta})$. Hence, we may assume that $i \in \{b_n + 1, \dots, n - b_n - 1\}$, and without a loss of generality,

$$|F_i(\widetilde{\theta})| > |F_{i-1}(\widetilde{\theta})| \quad \text{and} \quad |F_i(\widetilde{\theta})| \ge |F_{i+1}(\widetilde{\theta})|,$$

as well as $F_i(\widetilde{\theta}) > 0$. This means that

$$F_i(\widetilde{\theta}) > |F_{i-1}(\widetilde{\theta})| \quad \text{and} \quad F_i(\widetilde{\theta}) \ge |F_{i+1}(\widetilde{\theta})|,$$

which of course implies

$$F_i(\widetilde{\theta}) > F_{i-1}(\widetilde{\theta}) \quad \text{and} \quad F_i(\widetilde{\theta}) \ge F_{i+1}(\widetilde{\theta}).$$

Applying the definition of the filter in (14) gives

$$\Big(\sum_{j=i+1}^{i+b_n}\widetilde{\theta}_j-\sum_{j=i-b_n+1}^{i}\widetilde{\theta}_j\Big)-\Big(\sum_{j=i}^{i+b_n-1}\widetilde{\theta}_j-\sum_{j=i-b_n}^{i-1}\widetilde{\theta}_j\Big)>0$$
$$\Big(\sum_{j=i+1}^{i+b_n}\widetilde{\theta}_j-\sum_{j=i-b_n+1}^{i}\widetilde{\theta}_j\Big)-\Big(\sum_{j=i+2}^{i+b_n+1}\widetilde{\theta}_j-\sum_{j=i-b_n+2}^{i+1}\widetilde{\theta}_j\Big)\geq 0,$$

or, after simplification,

$$\widetilde{\theta}_{i+b_n}-2\widetilde{\theta}_i+\widetilde{\theta}_{i-b_n}>0 \quad \text{and} \quad -\widetilde{\theta}_{i+b_n+1}+2\widetilde{\theta}_{i+1}-\widetilde{\theta}_{i-b_n+1}\geq 0.$$

Adding the above two equations together, we get

$$-\big(\widetilde{\theta}_{i+b_n+1}-\widetilde{\theta}_{i+b_n}\big)+2\big(\widetilde{\theta}_{i+1}-\widetilde{\theta}_i\big)-\big(\widetilde{\theta}_{i-b_n+1}-\widetilde{\theta}_{i-b_n}\big)>0,$$

which implies at least one of the three bracketed pairs of terms must be nonzero, i.e., a changepoint must occur at one of the locations i, $i+b_n$, or $i-b_n$. The proves that $L(\widetilde{\theta})\subseteq I_C(\widetilde{\theta})$.

Now we show the intended statement. Let $j\in\{b_n,\ldots,n-b_n\}$, and $i\in L(\widetilde{\theta})$ be in the direction of ascent from j with respect to $F(\widetilde{\theta})$, where $j\leq i$, without a loss of generality (for the case $i<j$, replace $\ell+b_n$ below by $\ell-b_n$). That is, the location i is a local maximum where

$$|F_j(\widetilde{\theta})|\leq|F_{j+1}(\widetilde{\theta})|\leq\ldots\leq|F_{i-1}(\widetilde{\theta})|\leq|F_i(\widetilde{\theta})|. \tag{72}$$

If $|i-j|\leq b_n$, then we have the desired result, due to (72). If $|i-j|>b_n$, then there must be at least one location $\ell\in S(\widetilde{\theta})$ such that $|\ell-j|\leq b_n$. (To see this, note that if $\widetilde{\theta}_{j-b_n+1}=\ldots=\widetilde{\theta}_{j+b_n}$, then $F_j(\widetilde{\theta})=0$.) Thus, at least one of $\ell,\ell+b_n$ lies in between j and i, and then again (72) implies the result, completing the proof.

E Proof of Lemma 20

Our optimization problem may be rewritten as

$$(\widetilde{a},\widetilde{b})=\underset{a,b\in\mathbb{R}}{\operatorname{argmin}}\ b^2+\sum_{x=1}^{r}\Big((ax+b-a_1x)^2+(ax-b-a_2x)^2\Big).$$

Taking a derivative of the criterion with respect to b and setting this equal to 0 gives

$$0=b+\sum_{x=1}^{r}(ax+b-a_1x-ax+b+a_2x),$$

i.e., we see that the optimal value is

$$\widetilde{b}=(a_1-a_2)\frac{\sum_{x=1}^{r}x}{2r+1}=(a_1-a_2)\frac{r(r+1)}{2(2r+1)}.$$

Taking a derivative of the criterion with respect to a and setting this equal to 0 gives

$$0=b+\sum_{x=1}^{r}(ax+b-a_1x+ax-b-a_2x),$$

i.e., we see that the optimal value is

$$\widetilde{a}=\frac{a_1+a_2}{2}.$$

Plugging in $\widetilde{a}, \widetilde{b}$ into the criterion, and abbreviating $c_r = r(r+1)/(2(2r+1))$, we can compute the optimal criterion value:

$$
\begin{aligned}
(a_1 - a_2)^2 c_r^2 + \sum_{x=1}^{r} \left[\left(\frac{a_2 - a_1}{2} x + (a_2 - a_1) c_r \right)^2 + \left(\frac{a_1 - a_2}{2} x - (a_1 - a_2) c_r \right)^2 \right] \\
&= (a_1 - a_2)^2 c_r^2 + \frac{1}{2} (a_1 - a_2)^2 \sum_{x=1}^{r} x^2 + 2r(a_1 - a_2)^2 c_r^2 + 2(a_1 - a_2)^2 c_r \sum_{x=1}^{r} x \\
&= (a_1 - a_2)^2 \left(\frac{r^2(r+1)^2}{4(2r+1)^2} + \frac{r(r+1)(2r+1)}{12} + \frac{r^3(r+1)^2}{2(2r+1)^2} + \frac{r^2(r+1)^2}{2(2r+1)} \right) \\
&\geq (a_1 - a_2)^2 \left(\frac{r^2}{16} + \frac{r(r+1)(2r+1)}{12} + \frac{r^3}{8} + \frac{r^2(r+1)}{4} \right) \\
&\geq (a_1 - a_2)^2 r^3 \left(\frac{1}{6} + \frac{1}{8} + \frac{1}{4} \right) \\
&= (a_1 - a_2)^2 \frac{13 r^3}{24}.
\end{aligned}
$$

□