

High-Dimensional Change-Point Estimation: Combining Filtering with Convex Optimization

Yong Sheng Soh[†] and Venkat Chandrasekaran^{†,‡} *

[†] Department of Computing and Mathematical Sciences

[‡] Department of Electrical Engineering

California Institute of Technology

Pasadena, CA 91125

January 8, 2015

Abstract

We consider change-point estimation in a sequence of high-dimensional signals given noisy observations. Classical approaches to this problem such as the filtered derivative method are useful for sequences of scalar-valued signals, but they have undesirable scaling behavior in the high-dimensional setting. However, many high-dimensional signals encountered in practice frequently possess latent low-dimensional structure. Motivated by this observation, we propose a technique for high-dimensional change-point estimation that combines the filtered derivative approach from previous work with convex optimization methods based on atomic norm regularization, which are useful for exploiting structure in high-dimensional data. Our algorithm is applicable in online settings as it operates on small portions of the sequence of observations at a time, and it is well-suited to the high-dimensional setting both in terms of computational scalability and of statistical efficiency. The main result of this paper shows that our method performs change-point estimation reliably as long as the product of the smallest-sized change (the Euclidean-norm-squared of the difference between signals at a change-point) and the smallest distance between change-points (number of time instances) is larger than a Gaussian width parameter that characterizes the low-dimensional complexity of the underlying signal sequence.

Keywords: High-dimensional time series; convex geometry; atomic norm thresholding; filtered derivative.

1 Introduction

Change-point estimation is the identification of abrupt changes or anomalies in a sequence of observations. Such problems arise in numerous applications such as product quality control, data segmentation, network analysis, and financial modeling; an overview of the change-point estimation literature can be found in [7, 23, 54, 67]. As in other inferential tasks encountered in contemporary settings, a key challenge underlying many modern change-point estimation problems is the increasingly large dimensionality of the underlying sequence of signals – that is, the signal at each location in the sequence is not scalar-valued but rather lies in a high-dimensional space. This challenge leads both to computational difficulties as well as to complications with obtaining statistical consistency

*Email: ysoh@caltech.edu, venkatc@caltech.edu

in settings in which one has access to a small number of observations (relative to the dimensionality of the space in which these observations live).

A prominent family of methods for estimating the locations of change-points in a sequence of noisy scalar-valued observations is based on the *filtered derivative* approach [4, 7, 8, 10, 11]. Broadly speaking, these procedures begin with an application of a low-pass filter to the sequence, followed by a computation of pairwise differences between successive elements, and finally the implementation of a thresholding step to estimate change-points. A large body of prior literature has analyzed the performance of this family of algorithms and its variants [4, 10, 11]. Unfortunately, as we describe in Section 3, the natural extension of this procedure to the high-dimensional setting leads to performance guarantees for reliable change-point estimation that require the underlying signal to remain unchanged for long portions of the sequence. Such requirements tend to be unrealistic in applications such as financial modeling and network analysis in which rapid transitions in the underlying phenomena trigger frequent changes in the associated signal sequences.

1.1 Our contributions

To alleviate these difficulties, modern signal processing methods for high-dimensional data – in a range of statistical inference tasks such as denoising [12, 27, 32], model selection [17, 46, 68], the estimation of large covariance matrices [13, 14], and others [19, 21, 28, 31, 55] – recognize and exploit the observation that signals lying in high-dimensional spaces typically possess low-dimensional structure. For example, images frequently admit sparse representations in an appropriately transformed domain [16, 19] while covariance matrices are well-approximated as low-rank matrices in many settings. The exploitation of low-dimensional structure in solving problems such as denoising leads to consistency guarantees that depend on the intrinsic low-dimensional “complexity” of the data rather than on the ambient (large) dimension of the space in which they live. A notable feature of several of these structure-exploiting procedures is that they are based on convex optimization methods, which can lead to tractable numerical algorithms for large-scale problems as well as to insightful statistical performance analyses. Motivated by these ideas, we propose a new approach for change-point estimation in high dimensions by integrating a convex optimization step into the filtered derivative framework (see Section 3). We prove that the resulting method provides reliable change-point estimation performance in high-dimensional settings, with guarantees that depend on the underlying low-dimensional structure in the sequence of observations rather than on their ambient dimension.

To illustrate our ideas and arguments concretely, we consider a setup in which we are given a sequence $\mathbf{Y}[t] \in \mathbb{R}^p$ for $t = 1, \dots, n$ of observations of the form:

$$\mathbf{Y}[t] = \mathbf{X}^*[t] + \boldsymbol{\varepsilon}[t]. \tag{1}$$

Here $\mathbf{X}^*[t] \in \mathbb{R}^p$ is the underlying signal and the noise is normally distributed $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{p \times p})$. The signal sequence $\mathcal{X} := \{\mathbf{X}^*[t]\}_{t=1}^n$ is assumed to be piecewise constant with respect to t . The set of change-points is denoted by $\tau^* \subset \{1, \dots, n\}$, i.e., $t \in \tau^* \Leftrightarrow \mathbf{X}^*[t] \neq \mathbf{X}^*[t+1]$, and the objective is to estimate the set τ^* . A central aspect of our setup is that each $\mathbf{X}^*[t]$ is modeled as having an efficient representation as a linear combination of a small number of elements from a known set \mathcal{A} of building blocks or atoms [12, 18, 19, 20, 21, 27, 28, 30, 55, 66]. This notion of signal structure includes widely studied models in which signals are specified by sparse vectors and low-rank matrices. It also encompasses several others such as low-rank tensors, orthogonal matrices, and permutation matrices. The convex optimization step in our approach exploits knowledge of the atomic set \mathcal{A} ; specifically, the algorithm described in Section 3.2 consists of a denoising operation in which the underlying signal is estimated from local averages of the sequence $\mathbf{Y}[t]$ using a proximal

operator based on the atomic norm associated to \mathcal{A} [12, 21, 27]. The main technical result of this paper is that the method we propose in Section 3 provides accurate estimates of the change-points τ^* with high probability under the condition:

$$\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2 \{\eta^2(\mathcal{X}) + \log n\}, \quad (2)$$

where Δ_{\min} denotes the size (in ℓ_2 -norm) of the smallest change among all change-points, T_{\min} denotes the smallest interval between successive change-points, and n is the number of observations. The quantity $\eta(\mathcal{X})$ captures the low-dimensional complexity in the signal sequence $\mathcal{X} := \{\mathbf{X}^*[t]\}_{t=1}^n$ via a Gaussian distance/width characterization, and it appears in our result due to the incorporation of the convex optimization step. In the high-dimensional setting, the parameter η^2 plays a crucial role as it reflects the underlying low-dimensional structure in the signal sequence of interest; as such it is usually much smaller than the ambient dimension p (we quantify the comparisons in Section 2). Indeed, directly applying the filtered derivative method without incorporating a convex optimization step that exploits the signal structure would lead to weaker performance guarantees in terms of p rather than η^2 .

The performance guarantee (2) highlights a number of tradeoffs in high-dimensional change-point estimation that result from using our approach. For example, the appearance of the term $\Delta_{\min}^2 T_{\min}$ on the left hand side of (2) implies that it is possible to compensate for one of these quantities being small if the other one is suitably large. Further, our algorithm also operates in a causal manner on small portions of the sequence at any given time rather than on the entire sequence simultaneously, and it is therefore useful in “online” settings. This feature of our method combined with the the result (2) leads to a more subtle tradeoff between the computational efficiency of the approach and the number of observations n ; specifically, our algorithm can be adapted to process larger datasets (e.g., settings in which observations are obtained via high-frequency sampling, leading to larger n) more efficiently without loss in statistical performance by employing a suitable form of convex relaxation based on the ideas discussed in [20]. We discuss these points in greater detail in Section 4.

1.2 Related work

A recent paper by Harchaoui and Lévy-Leduc [35] is closest in spirit to ours (in terms of providing change-point estimation guarantees of the form (2)); they describe a convex programming method based on total-variation minimization to detect changes in sequences of scalar-valued signals. In addition to the restriction to scalar-valued signals, the technique in [35] requires knowledge of the full sequence of observations in advance. As a result it is not directly applicable in high-dimensional and online settings unlike our proposed approach.

High-dimensional change-point estimation has received much attention in recent years based on different types of extensions of the scalar case. The diversity of these generalizations of the scalar setting is due to the wide range of applications in which change-point estimation problems arise, each with a unique set of considerations. For example, several papers [22, 29] investigate high-dimensional change-point estimation in settings in which the changes only occur in a small subset of components. Therefore, assumptions about low-dimensional structure are made with regards to the pattern of changes rather than in the signal itself at each time instance (as in our setup). Xie et al. [69] consider a high-dimensional change-point estimation problem in which the underlying signals are modeled as lying on a low-dimensional manifold; although this setup is similar to ours, their algorithmic approach is based on projections onto manifolds rather than on convex optimization, and the types of guarantees obtained in [69] are qualitatively quite different in comparison to (2). We also note recent work on high-dimensional change-point problems in which

the authors study the impact of random projections on the performance of classical algorithms such as the cumulative-sum method [5].

1.3 Paper outline

Section 2 gives the relevant background on structured signals that are concisely represented with respect to sets of elementary atoms as well as the analytical tools that are used in the remainder of the paper. In Section 3 we describe our algorithm for high-dimensional change-point estimation, and we state the main recovery guarantee of the procedure. In Section 4 we discuss the tradeoffs that result from using our approach, and their utility in adapting our algorithm to address challenges beyond high-dimensionality that arise in applications involving change-point estimation. We verify our theoretical results with numerical experiments on synthetic data in Section 5, and we conclude with brief remarks and further directions in Section 6. The proofs are given in the Appendix.

2 Background on Structured Signal Models

2.1 Efficient representations with respect to atomic sets

We outline a framework with roots in nonlinear approximation [6, 25, 39, 53] that generalizes several types of low-dimensional models considered in the literature such as sparse vectors and low-rank matrices [12, 13, 14, 19, 21, 28, 46, 50, 55].

Let $\mathcal{A} \subseteq \mathbb{R}^p$ be a compact set that specifies a collection of atoms. We say that a signal $\mathbf{X} \in \mathbb{R}^p$ has a concise representation with respect to \mathcal{A} if it admits a decomposition as a sum of a small number of atoms in \mathcal{A} , that is, we are able to write

$$\mathbf{X} = \sum_{i=1}^s c_i \mathbf{a}_i, \mathbf{a}_i \in \mathcal{A}, c_i \geq 0, \quad (3)$$

for some $s \ll p$. Sparse vectors and low-rank matrices are examples of low-dimensional representations that are expressible in this framework. Specifically, an atomic set for sparse vectors is the set of signed standard basis vectors $\mathcal{A} = \{\pm \mathbf{e}_i | 1 \leq i \leq p\}$, while a natural atomic set for low-rank matrices is set of all rank-one matrices with unit Euclidean norm. Other examples include binary vectors (e.g., in knapsack problems [45]), permutation matrices (in ranking problems [38]), low-rank tensors [41], and orthogonal matrices. Such classes of signals that have concise representations with respect to general atomic sets were studied in the context of linear inverse problems [21], and subsequently in the setting of statistical denoising [12, 20].

In comparison with alternative notions of low-dimensional structure, e.g., manifold models [69], which have been considered previously in the context of high-dimensional change-point estimation (and more generally in signal processing), the setup described here has the virtue that one can employ efficient algorithms for convex optimization methods and one can appeal to insights from convex geometry in developing and analyzing algorithms for high-dimensional change-point estimation. We discuss the relevant concepts in the next two subsections.

2.2 Minkowski functional and proximal operators

A key feature of our change-point estimation algorithm is the incorporation of a signal denoising step that exploits knowledge of the atomic set \mathcal{A} . To formally define the denoising operation, we consider the Minkowski functional $\|\cdot\|_{\mathcal{C}} : \mathbb{R}^p \mapsto [0, \infty]$

$$\|\mathbf{X}\|_{\mathcal{C}} := \inf\{t : \mathbf{X} \in t\mathcal{C}, t > 0\}, \quad (4)$$

defined with respect to a convex set $\mathcal{C} \subset \mathbb{R}^p$ such that $\mathcal{A} \subseteq \mathcal{C}$, as discussed in [21]. As \mathcal{C} is convex, the Minkowski functional $\|\cdot\|_{\mathcal{C}}$ is also convex. This function is also called the gauge function in the convex analysis literature [58]. For a given $\mathbf{Y} \in \mathbb{R}^p$ and a convex set \mathcal{C} , we consider denoisers specified in terms of the following *proximal operator*:

$$\hat{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\|_2^2 + \lambda \|\mathbf{X}\|_{\mathcal{C}}. \quad (5)$$

As $\|\cdot\|_{\mathcal{C}}$ is a convex function, this optimization problem is a convex program. To obtain a proximal operator that enforces signal structure in the denoising operation, the set \mathcal{C} is usually taken to be the tightest convex set containing the atomic set \mathcal{A} , i.e., $\mathcal{C} = \operatorname{conv}(\mathcal{A})$. With $\mathcal{C} = \operatorname{conv}(\mathcal{A})$, the resulting Minkowski functional is called the *atomic norm*¹ with respect to \mathcal{A} , and the associated proximal operator (5) is called *atomic norm thresholding* [12]. The atomic norm has been studied in the approximation theory literature for characterizing approximation rates associated with best k -term approximants [6, 25, 39, 53], and its convex-geometric properties were investigated in [21] in the context of ill-posed linear inverse problems. When $\mathcal{A} = \{\pm \mathbf{e}_i | 1 \leq i \leq p\}$ is the collection of signed standard basis vectors, the atomic norm with respect to \mathcal{A} is simply the ℓ_1 -norm in \mathbb{R}^p . Similarly, the atomic norm corresponding to unit-Euclidean-norm rank-one matrices is the matrix nuclear norm. More generally, one can define atomic norms associated to other types of structured objects such as permutation matrices, low-rank tensors, orthogonal matrices, and signed vectors; see [21] for a detailed list. Atomic norm thresholding naturally generalizes soft-thresholding based on the ℓ_1 -norm for sparse signals to a more general denoising operation for the types of structured signals described here.

One exception to the rule of thumb of choosing $\mathcal{C} = \operatorname{conv}(\mathcal{A})$ arises if the atomic norm is intractable to represent, e.g., the tensor nuclear norm [37]. That is, although these norms are convex functions, computing them may in general be computationally intractable. To overcome such difficulties, a natural approach described in [20, 21] is to consider Minkowski functionals of convex sets \mathcal{C} that contain \mathcal{A} and that are efficient to represent, i.e., further tractable convex relaxations of $\operatorname{conv}(\mathcal{A})$.

Finally, to avoid dealing with technicalities in degenerate cases, we assume throughout the remainder of the paper that the set $\operatorname{conv}(\mathcal{A}) \subset \mathbb{R}^p$ is a solid convex set containing the origin in its interior. Consequently, we have that $\|\mathbf{X}\|_{\mathcal{C}} < \infty$ for all $\mathbf{X} \in \mathbb{R}^p$.

2.3 Summary parameters in signal denoising

Next we describe the relevant convex-geometric concepts for analyzing the performance of proximal denoising operators. For $\mathbf{X} \in \mathbb{R}^p$, the *Gaussian distance* $\eta_{\mathcal{C}}(\mathbf{X})$ [31, 50] with respect to a norm $\|\cdot\|_{\mathcal{C}}$ is defined as

$$\eta_{\mathcal{C}}(\mathbf{X}) := \inf_{\lambda \geq 0} \left\{ \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, I_{p \times p})} [\operatorname{dist}(\mathbf{g}, \lambda \cdot \partial \|\mathbf{X}\|_{\mathcal{C}})] \right\}. \quad (6)$$

Here $\operatorname{dist}(\mathbf{g}, \partial \|\mathbf{X}\|_{\mathcal{C}}) := \inf_{\mathbf{w} \in \partial \|\mathbf{X}\|_{\mathcal{C}}} \|\mathbf{w} - \mathbf{g}\|_2$ denotes the distance of \mathbf{g} from the set $\partial \|\mathbf{X}\|_{\mathcal{C}}$, where $\partial \|\mathbf{X}\|_{\mathcal{C}}$ is the subdifferential of the function $\|\cdot\|_{\mathcal{C}}$ at the point \mathbf{X} [58]. We relate the Gaussian distance to the Gaussian width [33] in Appendix 7.1 by extending a result in [31].

The Gaussian distance $\eta_{\mathcal{C}}(\mathbf{X})$ is useful for characterizing the performance of the proximal denoising operator (5) [12, 20, 50]. Specifically, suppose $\hat{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{X}^* + \varepsilon - \mathbf{X}\|_2^2 + \lambda \|\mathbf{X}\|_{\mathcal{C}}$,

¹For (4) to formally define a norm, we would also need the set \mathcal{A} to be centrally symmetric. Nevertheless, the results in the remainder of the paper hold without this condition, so we use “norm” with an abuse of terminology.

then the error between $\hat{\mathbf{X}}$ and \mathbf{X}^* is bounded as [50]:

$$\|\mathbf{X}^* - \hat{\mathbf{X}}\|_2 \leq \text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial\|\mathbf{X}^*\|_C).$$

Taking expectations with respect to $\boldsymbol{\varepsilon}$ and subsequently optimizing the resulting bound with respect to λ yields the Gaussian distance (6). We prove a generalization of this result in Appendix 7.2, which is relevant to the analysis of the change-point estimation algorithm proposed in Section 3.2.

As we discuss in Section 3, the combination of the proximal denoising operator with a suitable filtering step leads to a change-point estimation procedure with performance guarantees in terms of $\eta_C(\mathbf{X})$ rather than \sqrt{p} . This point is significant because for many examples of structured signals that are encountered in practice, it is typically the case that $\eta_C(\mathbf{X}) \ll \sqrt{p}$. For example, if \mathbf{X} is an s -sparse vector in \mathbb{R}^p then proximal denoising via the ℓ_1 -norm gives $\eta_{\ell_1}(\mathbf{X}) \leq \sqrt{2s \log(p/s) + 3s/2 + 7}$ [21, 31, 59, 65]. Similarly, if \mathbf{X} is a rank- r matrix in $\mathbb{R}^{d \times d}$ then proximal denoising via the matrix nuclear norm gives $\eta_{\text{nuc}}(\mathbf{X}) \leq \sqrt{6rd} + 7$ [21, 31, 49, 56].

In order to state performance guarantees for a *sequence* of observations, we extend the definition of η_C to collections of vectors $\mathcal{X} = \{\mathbf{X}^*[1], \dots, \mathbf{X}^*[n]\}, \mathbf{X}^*(i) \in \mathbb{R}^p$ as follows:

$$\eta_C(\mathcal{X}) := \inf_{\lambda \geq 0} \max_{\mathbf{X}^*[t] \in \mathcal{X}} \left\{ \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, I_{p \times p})} [\text{dist}(\mathbf{g}, \lambda \cdot \partial\|\mathbf{X}^*[t]\|_C)] \right\}. \quad (7)$$

3 Convex Programming for Change-Point Estimation

In this section, we describe our algorithm for high-dimensional change-point estimation by combining the filtered derivative method with proximal denoising. We state the main theorem that characterizes the accuracy of the estimated set of change-points, and we outline the proof, with the full details given in the Appendix.

3.1 Motivation

We begin by highlighting some of the difficulties that arise in change-point estimation as a result of the high-dimensionality of the observations. In order to frame our discussion concretely, we consider the prominent and widely-employed class of change-point estimation techniques based on the filtered derivative algorithm [4, 8, 10, 11], although similar difficulties arise with other approaches as well. The *filtered derivative* method detects changes based on an application of a pairwise difference operator to the output of a suitable low-pass filter applied to the sequence of observations. For simplicity, we describe a particular low-pass filter that is given by the sample mean of the observations over a small window (again, elaborations on this scheme are possible, with qualitatively similar conclusions). Formally, consider the following sequence defined at time t by computing differences of sample means over windows of size θ :

$$\text{FD}_\theta[t] = -\frac{1}{\theta} \sum_{i=t-\theta+1}^t Y[i] + \frac{1}{\theta} \sum_{i=t+1}^{t+\theta} Y[i]. \quad (8)$$

Locations at which $\text{FD}_\theta[t]$ has large magnitude (i.e., above a suitably chosen threshold) are declared as change-points.

This approach is well-suited for settings with sequences of scalar-valued signals, i.e., each $\mathbf{X}^*[t]$ is scalar; see [4, 10, 11] for detailed analyses. However, if applied directly to the high-dimensional setting, the underlying sequence of signals $\mathbf{X}^*[t] \in \mathbb{R}^p$ is required to remain stationary over time scales on the order of p so that changes can be reliably estimated. This requirement is unfortunately

not realistic for practical purposes, e.g., in image processing applications one typically encounters $p \approx 10^6$. As such, it is desirable to develop an algorithm that detects changes in sequences of high-dimensional observations reliably even if the signal does not remain stationary over long time scales.

3.2 Our approach to high-dimensional change-point estimation

We base our method on the principle that more effective signal denoising by exploiting the low-dimensional structure underlying the sequence $\mathbf{X}^*[t]$ enables improved change-point estimation. The formal steps of our algorithm for obtaining an estimate $\hat{\tau}$ of τ^* are as follows:

1. **[Input]:** $\{\mathbf{Y}[t]\}_{t=1}^n$ the sequence of signal observations, a choice of parameters θ, γ, λ to be employed in the algorithm, and a specification of a convex set \mathcal{C} .
2. **[Filtering]:** Compute the sample means $\bar{\mathbf{Y}}[i] = \frac{1}{\theta} \sum_{t=i}^{i+\theta-1} \mathbf{Y}[t], 1 \leq i \leq n - \theta + 1$.
3. **[Denoising]:** Let $\hat{\mathbf{X}}[t], 1 \leq t \leq n - \theta + 1$ be the solutions to the following convex optimization problems:

$$\hat{\mathbf{X}}[t] = \underset{\mathbf{X} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|\bar{\mathbf{Y}}[t] - \mathbf{X}\|_2^2 + \lambda \|\mathbf{X}\|_{\mathcal{C}}. \quad (9)$$

4. **[Differencing]:** Compute $S[t] := \|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2$ for $\theta \leq t \leq n - \theta$.
5. **[Thresholding]:** For all t such that $S[t] < \gamma$, set $S[t] = 0$.
6. **[Output]:** $\hat{\tau} = \{\hat{t}_i\}$. Group time intervals such that consecutive non-zero entries in $S[t]$ are at most θ apart, and in each group select the index \hat{t} corresponding to the largest value of S as an estimate of the location of a change.

Observe that the proximal denoising step is applied before the differencing step. This particular integration of proximal denoising and the filtered derivative ensures that the differencing operator is applied to estimates $\hat{\mathbf{X}}[t]$ that are closer to the underlying signal $\mathbf{X}^*[t]$ than the raw averages $\bar{\mathbf{Y}}[t]$ (due to the favorable denoising properties of the proximal denoiser). As discussed in Theorem 3.1, this leads to improved change-point performance in comparison to a pure filtered derivative method. However, the analysis of our approach is complicated by the introduction of the proximal denoising step; we discuss this point in greater detail in Section 3.3.

The parameter θ determines the window over which we compute the sample mean, and it controls the resolution to which we estimate change-points. A larger value of θ allows the algorithm to detect small changes, although if θ is chosen too large, multiple change-points may be mistaken for a single change-point. A smaller choice of θ increases the resolution of the change-point estimates, but small changes cannot be reliably detected. The parameter γ specifies the threshold for declaring changes, and it governs the size of the change-points that can be reliably estimated. A small choice of γ allows the algorithm to detect smaller changes but it also increases the occurrence of false positives. Conversely, a larger value of γ reduces the number of false positives, but only those changes that are sufficiently large in magnitude may be detected by the algorithm (i.e., the number of false negatives may increase). In Theorem 3.1, we give precise guidelines for the choices of the parameters $(\theta, \gamma, \lambda)$ to guarantee reliable change-point estimation under suitable conditions via the method described above.

Theorem 3.1. *Consider a sequence of observations $\mathbf{Y}[t] = \mathbf{X}^*[t] + \boldsymbol{\varepsilon}[t], t = 1, \dots, n$, where each $\mathbf{X}^*[t] \in \mathbb{R}^p$ and each $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{p \times p})$ independently. Let $\tau^* \subset \{1, \dots, n\}$ be such that $t \in \tau^* \Leftrightarrow$*

$\mathbf{X}^*[t] \neq \mathbf{X}^*[t+1]$, let $\Delta_{\min} = \min_{t \in \tau^*} \|\mathbf{X}^*[t] - \mathbf{X}^*[t+1]\|_2$, let $T_{\min} = \min_{t_i, t_j \in \tau^*, t_i \neq t_j} |t_i - t_j|$, and let $\mathcal{X} = \{\mathbf{X}^*[1], \dots, \mathbf{X}^*[n]\}$. Suppose Δ_{\min} and T_{\min} satisfy

$$\Delta_{\min}^2 T_{\min} \geq 64\sigma^2 \{\eta_{\mathcal{C}}(\mathcal{X}) + r\sqrt{2\log n}\}^2 \quad (10)$$

for some $r > 1$ and some convex set \mathcal{C} , where $\eta_{\mathcal{C}}(\mathcal{X})$ is as defined in (7), and $\tau^* \subset \{T_{\min}/4, \dots, n - T_{\min}/4\}$. Suppose we apply our change-point estimation algorithm with any choice of parameters θ, γ , and λ satisfying

1. $T_{\min}/4 \geq \theta$,
2. $\Delta_{\min}/2 \geq \gamma \geq 2\frac{\sigma}{\sqrt{\theta}} \{\eta_{\mathcal{C}}(\mathcal{X}) + r\sqrt{2\log n}\}$, and
3. $\lambda = \frac{\sigma}{\sqrt{\theta}} \operatorname{argmin}_{\tilde{\lambda}} \max_{\mathbf{X}^*[t] \in \mathcal{X}} \left\{ \mathbb{E}_{g \sim \mathcal{N}(0, I_{p \times p})} [\operatorname{dist}(g, \tilde{\lambda} \cdot \partial \|\mathbf{X}^*[t]\|_{\mathcal{C}})] \right\}$.

Then the algorithm recovers an estimate of the change-points $\hat{\tau}$ satisfying

1. $|\hat{\tau}| = |\tau^*|$
2. $|\hat{t}_i - t_i^*| \leq \min\{(4r\sqrt{\log n}/\eta_{\mathcal{C}}(\mathcal{X}) + 4)\frac{\sigma\eta_{\mathcal{C}}(\mathcal{X})}{\Delta_{\min}}\sqrt{\theta}, \theta\}$ for all i , where \hat{t}_i and t_i^* are the i -th elements of $\hat{\tau}$ and τ^* when ordered sequentially,

with probability greater than $1 - 5n^{1-r^2}$.

Remark. If condition (10) is satisfied then the choice of $\theta = T_{\min}/4$ and $\gamma = \Delta_{\min}/2$ satisfies the requirements in Theorem 3.1.

As a concrete illustration of this theorem, if each element $\mathbf{X}^*[t] \in \mathbb{R}^p$, $t = 1, \dots, n$ of the signal sequence is a vector consisting of at most s nonzero entries, then our algorithm (with a proximal denoiser based on the ℓ_1 -norm) estimates change-points reliably under the condition $\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2 (s \log(\frac{p}{s}) + \log(n))$. Similarly, if each element $\mathbf{X}^*[t] \in \mathbb{R}^{d \times d}$, $t = 1, \dots, n$ is a matrix with rank at most r , then our algorithm (with a proximal denoiser now based on the nuclear norm) provides reliable change-point estimation performance under the condition $\Delta_{\min}^2 T_{\min} \gtrsim \sigma^2 (rd + \log(n))$.

3.3 Proof of Theorem 3.1

The proof broadly proceeds by bounding the probabilities of the following three events:

$$E_1 := \{S[t] \geq \gamma, \forall t \in \tau^*\} \quad (11)$$

$$E_2 := \{S[t] < \gamma, \forall t \in \tau_{\text{far}}\} \quad (12)$$

$$E_3 := \{\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2 > \|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2, \forall (t, \delta) \in \tau_{\text{buffer}}\}. \quad (13)$$

Here $\tau_{\text{far}} = \{i : \theta \leq i \leq n - \theta, |i - j| > \theta, j \in \tau^*\}$ and $\tau_{\text{buffer}} = \{(t_i^*, \delta) : t_i^* \in \tau^*, \theta \geq |\delta| > (4r\sqrt{\log n}/\eta_{\mathcal{C}}(\mathcal{X}) + 4)\frac{\sigma\eta_{\mathcal{C}}(\mathcal{X})}{\Delta_{\min}}\sqrt{\theta}\}$. Note that τ_{buffer} defines a non-empty set if $\theta > (4r\sqrt{\log n}/\eta_{\mathcal{C}}(\mathcal{X}) + 4)^2 \sigma^2 \eta_{\mathcal{C}}^2(\mathcal{X}) / \Delta_{\min}^2$. The event E_1 corresponds to the atomic-norm-thresholded derivative exceeding the threshold γ for all change-points, while event E_2 corresponds to the atomic-norm-thresholded derivative *not* exceeding the threshold γ in regions “far away” from the change-points. Bounding the probabilities of these two events is sufficient for a weaker recovery guarantee than is provided by Theorem 3.1, which is that any estimated change-point $\hat{t} \in \hat{\tau}$ will be within θ of an actual change-point $t^* \in \tau^*$. However, the selection of the *maximum* derivative in Step 6 of the algorithm

often leads to far more accurate estimates of the locations of change-points. To prove that this is the case, we consider the event E_3 corresponding to the atomic-norm-thresholded derivative at the change-point being larger than the atomic-norm-thresholded derivatives at other points that are still within a window of θ but outside a small buffer region around the change-point.

The next proposition gives bounds on the probabilities of the events E_1, E_2, E_3 :

Proposition 3.2. *Under the setup and conditions of Theorem 3.1, we have the following bounds:*

$$\mathbb{P}(E_1^c) \leq 2n^{1-r^2}, \quad \mathbb{P}(E_2^c) \leq 2n^{1-r^2}, \quad \mathbb{P}(E_3^c) \leq n^{1-r^2}. \quad (14)$$

The events E_1, E_2, E_3 are defined in (11), (12), and (13).

The proof of Proposition 3.2 is given in the Appendix, and it involves overcoming two difficulties. First, if the filtering operator is applied over a window containing a change-point in Step 2, the average $\bar{\mathbf{Y}}[t]$ is in effect a superposition of two structured signals corrupted by noise. This necessitates the analysis of the performance of a proximal denoiser applied to a noisy superposition of structured signals rather than to a single structured signal corrupted by noise. The second (more challenging) complication arises due to the fact that the differencing operator is applied to the result of a nonlinear mapping of the observations (via the proximal denoiser) rather than to just a linear average of the observations as in a standard filtered derivative framework. We address these difficulties by exploiting certain properties of the proximal operator such as its non-expansiveness and its robustness to perturbations. Assuming Proposition 3.2, the proof of Theorem 3.1 proceeds as follows:

Proof of Theorem 3.1 From Proposition 3.2 and the union bound we have that $\mathbb{P}(E_1 \cap E_2 \cap E_3) \geq 1 - 5n^{r^2-1}$. We condition on the event $E_1 \cap E_2 \cap E_3$ to complete the proof. Specifically, conditioning on E_1 ensures that we have $S[t] \geq \gamma$ for all $t \in \tau^*$ after Step 5. Conditioning on E_2 implies that all entries of S outside a window of θ from any change-point are set to 0 after Step 5. Hence, after Step 6 all non-zero entries of S that are within a window of at most θ around a change-point will have been grouped together (for all change-points), which implies that $|\hat{\tau}| = |\tau^*|$. Finally, conditioning on the event E_3 implies that $S[t]$ is larger than $S[t + \delta]$ for all δ such that $(4r\sqrt{\log n}/\eta_C(\mathcal{X}) + 4)\frac{\sigma\eta_C(\mathcal{X})}{\Delta_{\min}}\sqrt{\theta} < |\delta| \leq \theta$ and for all $t \in \tau^*$. Thus, our estimate of the change-point at $t \in \tau^*$ is at most $\min\left\{(4r\sqrt{\log n}/\eta_C(\mathcal{X}) + 4)\frac{\sigma\eta_C(\mathcal{X})}{\Delta_{\min}}\sqrt{\theta}, \theta\right\}$ away, which concludes the proof. \square

3.4 Signal reconstruction

Based on the estimated set of change-points $\hat{\tau}$, it is straightforward to obtain good reconstructions of the signal away from the change-points via proximal denoising (5). This result follows from the analysis in [12, 50], but we state and prove it here for completeness:

Proposition 3.3. *Suppose that the assumptions of Theorem 3.1 hold. Let $t_1, t_2 \in \hat{\tau}$ be two consecutive estimates of change-points, let $\lambda' = \frac{\sigma}{\sqrt{t_2 - t_1 - 2\theta}} \operatorname{argmin}_{\tilde{\lambda}} \max_{\mathbf{X}^*[t] \in \mathcal{X}} \{\mathbb{E}_{g \sim \mathcal{N}(0, I_{p \times p})} [\operatorname{dist}(g, \tilde{\lambda} \cdot \partial \|\mathbf{X}^*[t]\|_C)]\}$, and let $\bar{\mathbf{Y}} = \frac{1}{t_2 - t_1 - 2\theta} \sum_{t=t_1+\theta+1}^{t_2-\theta} \mathbf{Y}[t]$. Denote the solution of the proximal denoiser (5) applied to $\bar{\mathbf{Y}}$ as $\bar{\mathbf{X}}$:*

$$\bar{\mathbf{X}} := \operatorname{argmin}_{\mathbf{X}} \frac{1}{2} \|\bar{\mathbf{Y}} - \mathbf{X}\|_2^2 + \lambda' \|\mathbf{X}\|_C.$$

Letting $\bar{\mathbf{X}}$ be the estimate of the signal $\mathbf{X}^*[t]$ over the interval $\{t_1 + \theta + 1, \dots, t_2 - \theta\}$, we have that

$$\|\bar{\mathbf{X}} - \mathbf{X}^*[t]\|_2^2 \leq \frac{2\sigma^2}{t_2 - t_1 - 2\theta} \{\eta_C(\mathcal{X})^2 + s^2\}$$

with probability greater than $1 - 4n^{1-r^2} - \exp(-s^2/2)$, for all $t_1 + \theta + 1 \leq t \leq t_2 - \theta$.

The proof of Proposition 3.3 is given in the Appendix. The quantities Δ_{\min} and T_{\min} determine the accuracy of the change-point estimates from Theorem 3.1. Proposition 3.3 demonstrates that, in addition to these quantities, the *duration of stationarity* of the signal also determines the accuracy of signal reconstruction away from the change-points.

4 Tradeoffs in High-Dimensional Change-Point Estimation

Data analysis in practice involves a range of challenges beyond the high-dimensionality of the observations that motivated our development in this paper. For example, in change-point estimation in financial time series, one is typically faced with additional difficulties arising from the extremely rapid rate at which the data are acquired and the requirement that the data be processed in an “online” fashion, i.e., the change-point estimation procedure must process the incoming data in “real time.” In some settings rapid variations in an underlying phenomenon trigger frequent changes in the sequence of observations, while in other cases small changes in a signal can be difficult to detect when severely corrupted by noise. In this section, we describe adaptations of the algorithm proposed in Section 3 to handle some of these challenges. Specifically, Theorem 3.1 suggests a number of performance tradeoffs that can be obtained in change-point estimation problems by employing suitable variations of our algorithm. We demonstrate the utility of these modifications in addressing some of the difficulties mentioned above, which highlights the versatility of our approach.

4.1 Change-point frequency and size tradeoffs

The appearance of the term $\Delta_{\min}^2 T_{\min}$ in (10) suggests an explicit relation between the minimum time span between changes, the minimum size of a change, and estimation accuracy. To illustrate the tradeoffs between Δ_{\min} and T_{\min} clearly, we fix the complexity parameter $\eta_{\mathcal{C}}(\mathcal{X})$ and the number of observations n in this discussion. As a consequence, the quantity $\Delta_{\min}^2 T_{\min}$ in (10) can be interpreted as a *resolution* on the types of changes that can be detected. Specifically, Theorem 3.1 guarantees reliable estimation of changes whenever $\Delta_{\min}^2 T_{\min}$ is sufficiently large even if one of Δ_{\min} or T_{\min} may be small. There are two separate regimes where this observation has interesting implications.

In settings in which Δ_{\min} is small, i.e., there are change-points where the size of the change is small, Theorem 3.1 guarantees that all the changes can be detected reliably so long as the distance between change-points (T_{\min}) is sufficiently large. In order to accomplish this, one is required to choose a sufficiently small threshold parameter γ (for Step 5 of the procedure) and a suitably large averaging window θ (for Step 2) with $\theta \lesssim T_{\min}$, in accordance with the requirements of Theorem 3.1. By smoothing over large windows of size θ and subsequently applying a proximal denoiser, even small-sized changes can be detected as long as the averaging window does not include multiple change-points (hence the condition that $\theta \lesssim T_{\min}$). The downside with choosing a large value for the parameter θ is that we do not resolve the locations of the change-points well; in particular, we estimate the locations of each of the change-points to within a resolution of about $\sqrt{\theta}$. However, detecting small-sized changes in a sequence corrupted by noise necessitates the computation of averages over large windows in Step 2 of our algorithm in order to distinguish genuine changes from spurious ones. Therefore, the low resolution to which we estimate the locations of change-points is the price to pay for estimating the number of change-points exactly in settings in which some of the changes may be small in size.

In a similar manner, if changes occur frequently in a signal sequence, i.e., the distance between change-points T_{\min} is small, Theorem 3.1 guarantees that all the changes can be detected reliably

if the size of each change Δ_{\min} is sufficiently large. In such cases, the averaging window parameter θ must be chosen to be sufficiently small while the threshold parameter γ must be appropriately large with $\gamma \lesssim \Delta_{\min}$, as prescribed in Theorem 3.1. The choice of a small value for θ ensures that we do not smooth the observation sequence over windows that contain multiple change-points in Step 2 of our method. However, this restriction of the averaging window size implies that the proximal denoiser in Step 3 is applied to the average of a small number of observations, which negatively impacts its performance. This limitation underlies the choice of a large value for the threshold parameter γ in Step 5 of the algorithm, which ensures that spurious changes resulting from denoising over small windows do not impact the performance of our algorithm. Consequently, the size of each change must be sufficiently large (as required by the condition that $\gamma \lesssim \Delta_{\min}$) so that the change-points can be reliably estimated from a few noisy observations. Thus, the size of each change must be sufficiently large in settings with frequent changes so that the number of change-points can be reliably estimated.

4.2 Computational complexity and sample size tradeoffs

In change-point estimation tasks arising in many contemporary problem domains (e.g., financial time series), one is faced with a twin set of challenges: (a) the number of observations n may be quite large due to the increasingly higher frequencies at which data are acquired (this is in addition to the high dimensionality of each observation), and (b) the requirement that these large datasets be processed online or in real time. Consequently, as the number of observations per unit time grows, it is crucial that we adopt a *simpler* algorithmic strategy – i.e., a method requiring a smaller number of computational steps per observation – so that the overall computational complexity of our algorithm does not grow with the number of observations. In this section we describe a convex relaxation approach to adapt the algorithm described in Section 3 to achieve a tradeoff between the number of observations and the overall computational complexity of our procedure; in particular, we demonstrate that in certain change-point estimation problems one can even achieve an overall *reduction* in computational runtime as the number of observations grows. Our method is based on the ideas presented in [20, 21] in the context of statistical denoising and of linear inverse problems; here we demonstrate the utility of those insights in change-point estimation. We note that other researchers have also explored the idea of trading off computational resources and sample size in various inferential problems such as binary classifier learning [15, 24, 60, 61], in sparse principal component analysis [3, 9, 40], in model selection [1], and in linear regression [62].

A modified change-point estimation algorithm For ease of analysis and exposition, we consider a modification of our change-point estimation procedure from Section 3. Specifically, Step 6 of our algorithm is simplified so it only groups time indices corresponding to the nonzero entries of the thresholded derivative values, with consecutive time indices in a group at most θ apart (i.e., without further choosing the maximum element from each group). Thus, our algorithm only produces windows that localize change-points instead of returning precise estimates of change-points. The reason for restricting our attention to such a simplification is that the additional operation of choosing the maximum element in Step 6 of the original algorithm leads to unnecessary complications that are not essential to the point of the discussion in this section. The performance analysis of this modified algorithm follows from Theorem 3.1, and we record this result next:

Corollary 4.1. *Under the same setup and conditions as in Theorem 3.1, suppose that we perform the modified change-point estimation algorithm – that is, Step 6 is simplified to only return groups of times indices, where consecutive time indices in a group are at most θ apart. Then we have*

with probability greater than $1 - 4n^{1-r^2}$ that (i) there are exactly τ^* groups, and (ii) the j 'th group $g_j \subset \{\theta, \dots, n - \theta\}$ is such that $|t_j^* - \tilde{t}| \leq \theta$ for all $\tilde{t} \in g_j$.

In order to concretely illustrate tradeoffs between the number of observations and the overall computational runtime, we focus on the following stylized change-point estimation problem. Consider a continuous-time piecewise constant signal $\mathbf{X}^*(T) \in \mathbb{R}^p$, $T \in (0, 1]$ defined as:

$$\mathbf{X}^*(T) = \mathbf{X}^{*(i)}, \quad T \in (T_i, T_{i+1}].$$

That is, the signal $\mathbf{X}^*(T)$ takes on the value $\mathbf{X}^{*(i)} \in \mathbb{R}^p$ identically for the entire time interval $T \in (T_i, T_{i+1}]$ for $i = 1, \dots, k$. Here $i = 1, \dots, k$ and the time indices $\{T_i\}_{i=1}^{k+1}$ are such that $0 = T_1 \leq \dots \leq T_{k+1} = 1$. Suppose we have two collections of noisy observations obtained by sampling the signal $\mathbf{X}^*(T)$ at equally-spaced points $\frac{1}{n}$ apart and $\frac{1}{kn}$ apart for some positive integers $k > 1$ and n (we assume that $n \gg \frac{1}{T_{i+1} - T_i}$ for all i):

$$\begin{aligned} \mathbf{Y}^{(1)}[t] &= \mathbf{X}^* \left(\frac{t}{n} \right) + \epsilon[t], \quad t = 1, \dots, n \\ \mathbf{Y}^{(2)}[t] &= \mathbf{X}^* \left(\frac{t}{kn} \right) + \tilde{\epsilon}[t], \quad t = 1, \dots, kn. \end{aligned}$$

Here $\epsilon[t], \tilde{\epsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{p \times p})$ are i.i.d. Gaussian noise vectors. In words, $\mathbf{Y}^{(2)}[t]$ is a k -times more rapidly sampled version of $\mathbf{X}^*(T)$ than $\mathbf{Y}^{(1)}[t]$. As a result, the sequence $\mathbf{Y}^{(1)}[t]$ consists of n observations and the sequence $\mathbf{Y}^{(2)}[t]$ consists of kn observations. Consequently, it may seem that estimating the change-points in the sequence $\mathbf{Y}^{(2)}[t]$ requires at least as many computational resources as the estimation of change-points in $\mathbf{Y}^{(1)}[t]$. However, when viewed from the prism of Corollary 4.1 and Theorem 3.1, the sequence $\mathbf{Y}^{(2)}[t]$ is in some sense more favorable than $\mathbf{Y}^{(1)}[t]$ for change-point estimation – specifically, if the minimum distance between successive change-points underlying the sequence $\mathbf{Y}^{(1)}[t]$ is T_{\min} , then the minimum distance between successive change-points in $\mathbf{Y}^{(2)}[t]$ is kT_{\min} , i.e., larger by a factor of k . (Note that Δ_{\min} for both sequences remains the same.)

Let $\mathcal{X} = \{\mathbf{X}^{*(1)}, \dots, \mathbf{X}^{*(k)}\}$. Applying the modified change-point estimation algorithm described above to the sequence $\mathbf{Y}^{(1)}[t]$ with parameters $\theta_1 = T_{\min}/4, \gamma_1 = \Delta_{\min}/2$ and with a proximal denoising step based on a convex set \mathcal{C} , we obtain reliable localizations of the change-points under the condition:

$$\Delta_{\min}^2 T_{\min} \geq 64\sigma^2 \{ \eta_{\mathcal{C}}(\mathcal{X}) + r\sqrt{2 \log n} \}^2,$$

That is, we localize the change-points to a window of size $\theta_1 = T_{\min}/4$. Now suppose we apply the modified change-point algorithm to the sequence $\mathbf{Y}^{(2)}[t]$ (note that this sequence is of length kn) with parameters $\theta_2 = kT_{\min}/4, \gamma_2 = \Delta_{\min}/2$ and with a proximal denoising step based on the same convex set \mathcal{C} . In this case, we reliably localize each change-point in $\mathbf{Y}^{(2)}[t]$ to a window of size $\theta_2 = kT_{\min}/4$ under the following condition:

$$\Delta_{\min}^2 (kT_{\min}) \geq 64\sigma^2 \{ \eta_{\mathcal{C}}(\mathcal{X}) + r\sqrt{2 \log n + 2 \log k} \}^2, \quad (15)$$

The quality of the output in both cases is the same – identifying changes in $\mathbf{Y}^{(1)}[t]$ to a resolution of $\theta_1 = T_{\min}/4$ is comparable to identifying changes in $\mathbf{Y}^{(2)}[t]$ to a resolution of $\theta_2 = kT_{\min}/4$, because $\mathbf{Y}^{(2)}[t]$ is a k -times more rapidly sampled version of the continuous-time signal $\mathbf{X}^*(T)$ in comparison to $\mathbf{Y}^{(1)}[t]$. On the computational side, our algorithm involves roughly n applications of the proximal denoiser based on the set \mathcal{C} in the case of $\mathbf{Y}^{(1)}[t]$, and about kn applications of the

same proximal denoiser in the case of $\mathbf{Y}^{(2)}[t]$. Therefore, the overall runtime is higher in the case of $\mathbf{Y}^{(2)}[t]$ than in the case of $\mathbf{Y}^{(1)}[t]$.

Notice that the left-hand-side of the condition (15) goes up by a factor of k . We exploit this increased gap between the two sides of the inequality in (15) to obtain a smaller overall computational runtime for estimating changes in the sequence $\mathbf{Y}^{(2)}[t]$ than for estimating changes in the sequence $\mathbf{Y}^{(1)}[t]$. The key insight underlying our approach, borrowing from the ideas developed in [20, 21], is that we can employ a computationally cheaper proximal denoiser when applying our algorithm to the sequence $\mathbf{Y}^{(2)}[t]$. Specifically, for many interesting classes of structured signals, one can replace the proximal denoising operation with respect to the convex set \mathcal{C} in step (3) of our algorithm with a proximal operator corresponding to a *relaxation* $\mathcal{B} \subset \mathbb{R}^p$ of the set \mathcal{C} , i.e., \mathcal{B} is a convex set such that $\mathcal{C} \subset \mathcal{B}$. For suitable relaxations \mathcal{B} of the set \mathcal{C} , the proximal denoiser associated to \mathcal{B} is more efficient to compute than the proximal denoiser with respect to \mathcal{C} , and further $\eta_{\mathcal{C}}(\mathcal{X}) < \eta_{\mathcal{B}}(\mathcal{X})$. The reason for the second property is that, under appropriate conditions, the subdifferentials with respect to the gauge functions $\|\cdot\|_{\mathcal{B}}, \|\cdot\|_{\mathcal{C}}$ satisfy the condition that $\partial\|\mathbf{X}^*\|_{\mathcal{B}} \subset \partial\|\mathbf{X}^*\|_{\mathcal{C}}$ at signals of interest $\mathbf{X}^* \in \mathcal{X}$. We refer the reader to [20] for further details, and more generally, to the convex optimization literature [34, 42, 52, 57, 63] for various constructions of families of tractable convex relaxations.

Going back to the sequence $\mathbf{Y}^{(2)}[t]$, we can employ a proximal denoiser based on *any* tractable convex relaxation \mathcal{B} of the set \mathcal{C} as long as the following condition (a modification of (15)) for reliable change-point estimation is satisfied:

$$\Delta_{\min}^2(kT_{\min}) \geq 64\sigma^2\{\eta_{\mathcal{B}}(\mathcal{X}) + r\sqrt{2\log n + 2\log k}\}^2.$$

Indeed, if this condition is satisfied, we can still localize changes to a resolution of $kT_{\min}/4$, i.e., the same quality of performance as before with a proximal denoiser with respect to the set \mathcal{C} . However, the computational upshot is that the number of operations required to estimate change-points in $\mathbf{Y}^{(2)}[t]$ using the modified proximal denoising step is roughly kn applications of the proximal denoiser based on the relaxations \mathcal{B} . The contrast to n applications of a proximal denoiser based on \mathcal{C} for estimating change-points in the sequence $\mathbf{Y}^{(1)}[t]$ can be significant if computing the proximal denoiser with respect to \mathcal{B} is much more tractable than computing the denoiser with respect to \mathcal{C} .

We give an example in which such convex relaxations can lead to reduced computational runtime as the number of observations increases. We refer the reader to [20] for further illustrations in the context of statistical denoising, which can also be translated to provide interesting examples in a change-point estimation setting. Specifically, suppose that the underlying signal set $\mathcal{X} = \{\mathbf{a}\mathbf{a}^T | \mathbf{a} \in \{-1, +1\}^d\}$, i.e., the signal at each instant in time is a rank-one matrix formed as an outer product of signed vectors. In this case, a natural candidate for a set \mathcal{C} is the set of $d \times d$ *correlation matrices*, which is also called the *elliptope* in the convex optimization literature [26]. One can show that each application of a proximal denoiser with respect to \mathcal{C} requires $\mathcal{O}(d^{4.5})$ operations [36]. The $d \times d$ nuclear norm ball (scaled to contain all $d \times d$ matrices with nuclear norm at most d), which we denote as \mathcal{B} , is a relaxation of the set \mathcal{C} of correlation matrices. Interestingly, the distance $\eta_{\mathcal{B}}(\mathcal{X})$ is only a constant times larger (independent of the dimension d) than $\eta_{\mathcal{C}}(\mathcal{X})$ [20]. However, each application of a proximal denoiser with respect to \mathcal{B} requires only $\mathcal{O}(d^3)$ operations. In summary, even if the increased sampling factor k in our setup is larger than a constant (independent of the dimension d), one can obtain an overall reduction in computational runtime from about $\mathcal{O}(nd^{4.5})$ operations to about $\mathcal{O}(knd^3)$ operations.

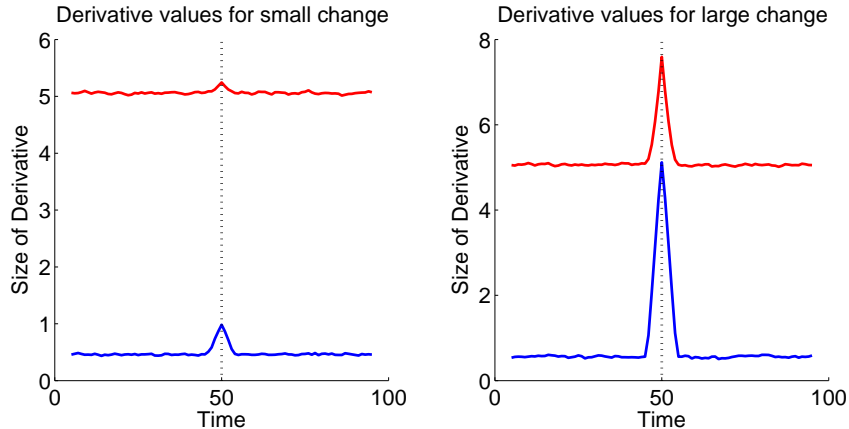


Figure 1: Experiment contrasting our algorithm (in blue) with the filtered derivative approach (in red): the left sub-plot corresponds to a small-sized change and the right sub-plot corresponds to a large-sized change.

5 Numerical Results

We illustrate the performance of our change-point estimation algorithm with two numerical experiments on synthetic data.

A contrast between our approach and the filtered derivative. The objective of the first experiment is to demonstrate the improved performance of our algorithm from Section 3.2 in comparison to the classical filtered derivative approach in a stylized problem setup. Recall that the filtered derivative method is equivalent to omitting the proximal denoising step in our algorithm, i.e., $\hat{\mathbf{X}}[t] = \bar{\mathbf{Y}}[t]$ in Step 3 of our algorithm.

We consider a signal sequence $\mathbf{X}^*[t] \in \mathbb{R}^{200 \times 200}$, $t = 1, \dots, 100$, consisting of exactly one change-point at time $t = 50$. Let $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathbb{R}^{200}$ be vectors with unit Euclidean-norm and direction chosen uniformly at random and independently. The signal $\mathbf{X}^*[t]$ is a 200×200 matrix equal to $\mathbf{u}^{(1)}\mathbf{v}^{(1) \prime}$ before the change-point and $\mathbf{u}^{(2)}\mathbf{v}^{(2) \prime}$ after the change-point, and the observations are $\mathbf{Y}[t] = \mathbf{X}^*[t] + \boldsymbol{\varepsilon}[t]$, $t = 1, \dots, 100$, where $\boldsymbol{\varepsilon}[t] \sim \mathcal{N}(0, \sigma^2 I_{200^2 \times 200^2})$ with $\sigma = 0.04$. Given this sequence of observations, we apply our algorithm with parameters $\lambda = 0.4$ and $\theta = 5$ (and with proximal denoising based on the nuclear-norm), and the filtered derivative algorithm with $\theta = 5$. The corresponding derivative values from our algorithm and the filtered derivative algorithm are given in the left sub-plot of Figure 1. We repeat the same experiment with the modification that the vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}$ now have Euclidean-norm equal to 2, thus leading to a larger-sized change relative to the noise. The corresponding derivative values from our algorithm and the filtered derivative algorithm are given in the right sub-plot in Figure 1.

One observation is that the derivative values are generally larger with the filtered derivative algorithm than with our approach; this is primarily due to the lack of a denoising step, as a larger amount of noise leads to greater derivative values. More crucially, however, the *relative difference* in the derivative values near a change-point and away from a change-point is much larger with our algorithm than with the filtered derivative method. This is also a consequence of the inclusion of the proximal denoising step in our algorithm and the lack of a similar denoising operation in the filtered derivative approach. By suppressing the impact of noise via proximal denoising, our approach identifies smaller-sized changes more reliably than a standard filtered derivative method without a denoising step (see the sub-plot on the left in Figure 1).

Run	θ	λ	γ
1	10	1	15
2	10	2	9
3	30	1	8
4	30	2	6

Figure 2: Table of parameters employed in our change-point estimation algorithm in synthetic experiment with sparse vectors.

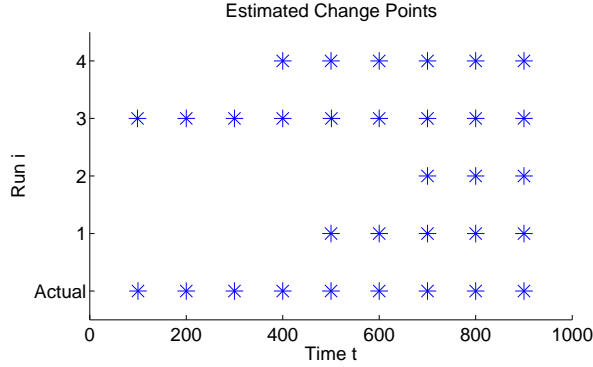


Figure 3: Plot of estimated change-points: the locations of the actual change-points are indicated in the bottom row.

Estimating change-points in sequences of sparse vectors. In our second experiment, we investigate the variation in the performance of our algorithm by choosing different sets of parameters θ, λ, γ . We consider a signal sequence $\mathbf{X}^*[t] \in \mathbb{R}^{1000}$, $t = 1, \dots, 1000$ consisting of sparse vectors. Specifically, we begin by generating 10 sparse vectors $\mathbf{S}^{(k)} \in \mathbb{R}^{1000}$, $k = 1, \dots, 10$ as follows: for each $k = 1, \dots, 10$, the vector $\mathbf{S}^{(k)}$ is a random sparse vector consisting of 30 nonzero entries (the locations of these entries are chosen uniformly at random and independently of k), with each nonzero entry being set to 1.2^{k-1} . We obtain the signal sequence $\mathbf{X}^*[t]$ from the $\mathbf{S}^{(k)}$'s by setting $\mathbf{X}^*[t] = \mathbf{S}^{(k)}$ for $t \in \{100(k-1) + 1, 100k\}$. In words, the signal sequence $\mathbf{X}^*[t]$ consists of 10 equally-sized blocks of length 100, and within each block the signal is identically equal to a sparse vector consisting of 30 nonzero entries. The magnitudes of the nonzero entries of $\mathbf{X}^*[t]$ in the latter blocks are larger than those in the earlier blocks. The observations are $\mathbf{Y}[t] = \mathbf{X}^*[t] + \varepsilon[t]$, $t = 1, \dots, 1000$, where each $\varepsilon[t] \sim \mathcal{N}(0, \sigma^2 I_{1000 \times 1000})$ with σ chosen to be 2.5. We then apply our proposed algorithm using the four choices of parameters listed in Figure 2, with a proximal operator based on the ℓ_1 -norm. The estimated sets of change-points are given in Figure 3, and the derivative values corresponding to Step 4 of our algorithm are given in Figure 4.

First, note that the algorithm generally detects smaller sized changes with larger values of θ and smaller values of γ (corresponding to Runs 3 and 4 from Figure 2), i.e., the averaging window size is larger in Step 2 of our algorithm and the threshold is smaller in Step 4. Next, consider the graph of derivative values in Figure 4. The estimated locations of change-points correspond to peaks in Figure 4, so the algorithm can be interpreted as selecting a subset of peaks that are sufficiently separated (Step 6). We note that a smaller choice of θ leads to sharper peaks (and hence, smaller-sized groups in Step 6), while a larger choice of θ leads to wider peaks (correspondingly, larger-size groups in Step 6).

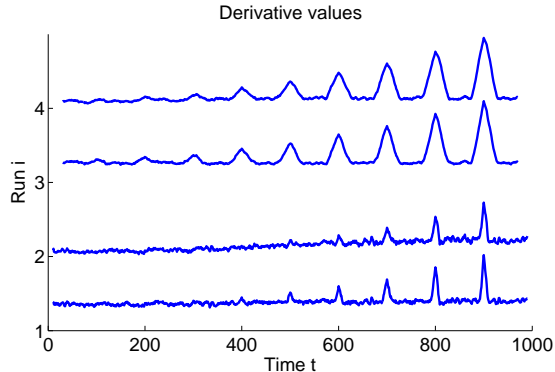


Figure 4: Experiment with sparse vectors: graphs of derivative values corresponding to different parameters choices from Figure 2.

6 Conclusions

We propose an algorithm for high-dimensional change-point estimation that blends the filtered derivative method with a convex optimization step that exploits low-dimensional structure in the underlying signal sequence. We prove that our algorithm reliably estimates change-points provided the product of the square of the size of the smallest change (measured in ℓ_2 -norm) and the smallest distance between changes is larger than Gaussian distance/width quantity η^2 , which characterizes the low-dimensional complexity in the signal sequence. The dependence on η^2 is a result of the integration of the convex optimization step (based on proximal denoising).

The change-point literature also consists of extensive investigations of *quickest change detection* problems [44, 51, 54, 64, 67], which are qualitatively somewhat different than the setup considered in our work. In those settings one is given access to observations sequentially, and the objective is to correctly declare when a change-point occurs in the shortest time possible (i.e., minimize the expected delay) subject to false alarm rate constraints. Building on the algorithmic ideas described in this paper, it would be of interest to design computationally and statistically efficient techniques for high-dimensional quickest change detection problems by exploiting structure in the underlying signal sequence.

Acknowledgments

This work was supported in part by the following sources: National Science Foundation Career award CCF-1350590, Air Force Office of Scientific Research grant FA9550-14-1-0098, an Okawa Research Grant in Information and Telecommunications, and an A*STAR (Agency for Science, Technology and Research, Singapore) Fellowship. Yong Sheng Soh would like to thank Michael McCoy for useful discussions, and Atul Ingle for pointing out a typographical error in a preliminary version of this paper.

References

- [1] A. Agarwal, P. L. Bartlett, and J. C. Duchi. Oracle Inequalities for Computationally Adaptive Model Selection. *CoRR*, abs/1208.0129, 2012.

- [2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living On The Edge: Phase Transitions in Convex Programs with Random Data. *Information and Inference*, 2014.
- [3] A. A. Amini and M. J. Wainwright. High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.
- [4] J. Antoch and M. Hušková. *Procedures for the Detection of Multiple Changes in Series of Independent Observations*. Contributions to Statistics. Physica-Verlag HD, 1994.
- [5] J. A. D. Aston and C. Kirch. Change Points in High Dimensional Settings. *CoRR*, abs/1409.1771, 2014.
- [6] A. R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.
- [7] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Prentice Halls, 1993.
- [8] A. Benveniste and M. Basseville. Detection of Abrupt Changes in Signals and Dynamical Systems: Some statistical aspects. *Lecture Notes in Control and Information Sciences*, 62:143–155, 1984.
- [9] Q. Berthet and P. Rigollet. Optimal Detection of Sparse Principal Components in High Dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.
- [10] P. R. Bertrand. A Local Method for Estimating Change Points: the “Hat-Function”. *Statistics: A Journal of Theoretical and Applied Statistics*, 34(3):215–235, 2000.
- [11] P. R. Bertrand, M. Fhima, and A. Guillin. Off-Line Detection of Multiple Change Points by the Filtered Derivative with p -Value Method. *Sequential Analysis*, 30:172–207, 2011.
- [12] B. N. Bhaskar, G. Tang, and B. Recht. Atomic Norm Denoising with Applications to Line Spectral Estimation. *CoRR*, abs/1204.0562, 2012.
- [13] P. J. Bickel and E. Levina. Covariance Regularization by Thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.
- [14] P. J. Bickel and E. Levina. Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics*, 36(1):199–227, 2008.
- [15] A. Birnbaum and S. Shalev-Shwartz. Learning Halfspaces with the Zero-One Loss: Time-Accuracy Tradeoffs. In F. Pereira, C.J.C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 926–934. Curran Associates, Inc., 2012.
- [16] A. M. Bruckstein, D. L. Donoho, and M. Elad. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review*, 51(1):34–81, 2009.
- [17] E. J. Candès and Y. Plan. Near-Ideal Model Selection by ℓ_1 Minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.
- [18] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

- [19] E. J. Candès, J. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [20] V. Chandrasekaran and M. I. Jordan. Computational and Statistical Tradeoffs via Convex Relaxation. *Proceedings of the National Academy of Sciences*, 110(13):1181–1190, 2013.
- [21] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
- [22] H. Cho and P. Fryzlewicz. Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014.
- [23] Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, 1971.
- [24] S. Decatur, O. Goldreich, and D. Ron. Computational Sample Complexity. *SIAM Journal on Computing*, 29:854–879, 1998.
- [25] R. A. DeVore and V. N. Temlyakov. Some Remarks on Greedy Algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996.
- [26] M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.
- [27] D. L. Donoho. De-noising by Soft-Thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.
- [28] D. L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [29] F. Enikeeva and Z. Harchaoui. High-Dimensional Change-Point Detection with Sparse Alternatives. *CoRR*, abs/1312.1900, 2013.
- [30] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Department of Electrical Engineering, Stanford University, 2002.
- [31] R. Foygel and L. Mackey. Corrupted Sensing: Novel Guarantees for Separating Structured Signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.
- [32] M. Gavish and D. L. Donoho. Optimal Shrinkage of Singular Values. *CoRR*, abs/1405.7511, 2014.
- [33] Y. Gordon. On Milman’s Inequality and Random Subspaces which Escape Through a Mesh in \mathbb{R}^n . In *Geometric Aspects of Functional Analysis*, volume 1317 of *Lecture Notes in Mathematics*, pages 84–106. Springer Berlin Heidelberg, 1988.
- [34] J. Gouveia, P. Parrilo, and R. Thomas. Theta Bodies for Polynomial Ideals. *SIAM Journal on Optimization*, 20:2097–2118, 2010.
- [35] Z. Harchaoui and C. Lévy-Leduc. Multiple Change-Point Estimation with a Total Variation Penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

- [36] N. J. Higham. Computing the Nearest Correlation Matrix – A Problem from Finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002.
- [37] C. J. Hillar and L.-H. Lim. Most Tensor Problems are NP-Hard. *Journal of the ACM*, 60(6):45:1–45:39, 2013.
- [38] S. Jagabathula and D. Shah. Inferring Rankings Using Constrained Sensing. *IEEE Transactions on Information Theory*, 57(11):7288–7306, 2011.
- [39] L. K. Jones. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *The Annals of Statistics*, 20(1):608–613, 1992.
- [40] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax Localization of Structural Information in Large Noisy Matrices. In *Neural Information Processing Systems*, 2011.
- [41] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2009.
- [42] J. B. Lasserre. Global Optimization with Polynomials and the Problem of Moments. *SIAM Journal on Optimization*, 11:796–817, 2001.
- [43] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical surveys and monographs*. American Mathematical Society, 2001.
- [44] G. Lorden. Procedures for Reacting to a Change in Distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.
- [45] O. L. Mangasarian and B. Recht. Probability of Unique Integer Solution to a System of Linear Equations. *European Journal of Operational Research*, 214(1):27 – 30, 2011.
- [46] N. Meinhausen and P. Bühlmann. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.
- [47] G. Minty. On the Monotonicity of the Gradient of a Convex Function. *Pacific Journal of Mathematics*, 14(1):243–247, 1964.
- [48] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.
- [49] S. Oymak and B. Hassibi. Tight Recovery Thresholds and Robustness Analysis for Nuclear Norm Minimization. In *IEEE International Symposium on Information Theory*, pages 2323 – 2327, 2011.
- [50] S. Oymak and B. Hassibi. On a Relation between the Minimax Risk and the Phase Transitions of Compressed Recovery. In *50th Annual Allerton Conference on Communication, Control, and Computing*, pages 1018–1025, 2012.
- [51] E. A. Page. Continuous Inspection Schemes. *Biometrika*, 41:100–115, 1954.
- [52] P. A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology, 2000.

- [53] G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz")*, pages 1–12, 1981.
- [54] H. V. Poor and O. Hadjiladis. *Quickest Detection*. Cambridge University Press, 2008.
- [55] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.
- [56] B. Recht, W. Xu, and B. Hassibi. Null Space Conditions and Thresholds for Rank Minimization. *Mathematical Programming*, 127(1):175–202, 2011.
- [57] J. Renegar. Hyperbolic Programs and their Derivative Relaxations. *Foundations of Computational Mathematics*, 6:59–79, 2006.
- [58] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [59] M. Rudelson and R. Vershynin. Sparse Reconstruction by Convex Relaxation: Fourier and Gaussian Measurements. In *40th Annual Conference on Information Sciences and Systems*, pages 207–212, 2006.
- [60] R. Servedio. Computational Sample Complexity and Attribute-Efficient Learning. *Journal of Computer and Systems Sciences*, 60:161–178, 2000.
- [61] S. Shalev-Shwartz, O. Shamir, and E. Tromer. Using More Data to Speed Up Training Time. In *Conference on Artificial Intelligence and Statistics*, 2012.
- [62] D. Shender and J. Lafferty. Computation-Risk Tradeoffs for Covariance-Thresholded Regression. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 28(3):756–764, 2013.
- [63] H. D. Sherali and W. P. Adams. A Hierarchy of Relaxations between the Continuous and Convex Hull Representations for Zero-One Programming Problems. *SIAM Journal on Discrete Mathematics*, 3:411–430, 1990.
- [64] A. N. Shiryaev. On Optimum Methods in Quickest Detection Problems. *Theory of Probability and its Applications*, 8(1):22–46, 1963.
- [65] M. Stojnic. Various Thresholds for ℓ_1 -Optimization in Compressed Sensing. *CoRR*, abs/0907.3666, 2009.
- [66] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed Sensing Off the Grid. *IEEE Transactions on Information Theory*, 59(11):7465–7490, 2013.
- [67] V. V. Veeravalli and T. Banerjee. Quickest Change Detection. *CoRR*, abs/1210.5552, 2012.
- [68] M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5), 2009.
- [69] Y. Xie, J. Huang, and R. Willett. Change-Point Detection for High-Dimensional Time Series With Missing Data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013.

7 Appendix

The Appendix is divided as follows. In Section 7.1 we describe the relation between the Gaussian distance and the Gaussian width. Next, in Section 7.2 we analyze the denoising properties of proximal operators. Finally, in Section 7.3 we prove the main results (from Section 3) of this paper. As described at the end of Section 2.2, we reiterate that the assumption that the set $\text{conv}(\mathcal{A}) \subset \mathbb{R}^p$ contains the origin in its interior holds throughout the Appendix.

7.1 Relationship between Gaussian distance and Gaussian width

The *Gaussian width* of a set $S \subseteq \mathbb{R}^p$ is defined as [33]:

$$\omega(S) := \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I_{p \times p})} \left[\sup_{\mathbf{z} \in S} \mathbf{g}^T \mathbf{z} \right].$$

The next definition that we need in order to relate the Gaussian distance and the Gaussian width is the *skewness* $\kappa_{\mathcal{C}}(\mathbf{X})$ of a norm $\|\cdot\|_{\mathcal{C}}$ at a point \mathbf{X} :

$$\kappa_{\mathcal{C}}(\mathbf{X}) := \frac{\|\Pi_{\partial\|\mathbf{X}\|_{\mathcal{C}}}(\mathbf{0})\|_2}{\|\Pi_{\text{aff.hull.}(\partial\|\mathbf{X}\|_{\mathcal{C}})}(\mathbf{0})\|_2},$$

where Π denotes the Euclidean projection and aff.hull. denotes the affine hull. The quantity κ has a natural geometric interpretation: since the subdifferential $\partial\|\mathbf{X}\|_{\mathcal{C}}$ corresponds to a face of the dual norm ball $\mathcal{C}^* = \{\mathbf{X} : \|\mathbf{X}\|_{\mathcal{C}}^* \leq 1\}$, the parameter $\kappa_{\mathcal{C}}(\mathbf{X})$ measures the skewness of the face $\partial\|\mathbf{X}\|_{\mathcal{C}}$. It is clear from this interpretation that $\kappa_{\mathcal{C}}(\mathbf{X}) = 1$ for all $\mathbf{X} \in \mathbb{R}^p$ whenever the unit ball with respect to the dual norm is suitably symmetric. Examples of such convex sets include the ℓ_1 -norm ball, the nuclear-norm ball, the ℓ_∞ -norm ball and the spectral-norm ball.

Our final definition relates to yet another convex-geometric concept. The tangent cone $T_{\mathcal{C}}(\mathbf{X})$ at a point $\mathbf{X} \in \mathbb{R}^p$ with respect to the unit ball of the $\|\cdot\|_{\mathcal{C}}$ -norm (i.e., the convex set \mathcal{C}) when $\|\mathbf{X}\|_{\mathcal{C}} = 1$ is:

$$T_{\mathcal{C}}(\mathbf{X}) := \text{cone}\{\mathbf{Z} - \mathbf{X} : \mathbf{Z} \in \mathbb{R}^p, \|\mathbf{Z}\|_{\mathcal{C}} \leq \|\mathbf{X}\|_{\mathcal{C}}\}.$$

For general unnormalized nonzero points $\mathbf{X} \in \mathbb{R}^p$, the tangent cone with respect to \mathcal{C} is $T_{\mathcal{C}}(\mathbf{X}/\|\mathbf{X}\|_{\mathcal{C}})$.

The next proposition relates the Gaussian distance and the Gaussian width. The result relaxes a “weak decomposability” assumption in [31, Prop.1]. We denote the Euclidean sphere in \mathbb{R}^p by \mathbb{S}^{p-1} .

Proposition 7.1. *The Gaussian distance is bounded above by the Gaussian width as follows:*

$$\eta_{\mathcal{C}}(\mathbf{X}) \leq \omega(T_{\mathcal{C}}(\mathbf{X}) \cap \mathbb{S}^{p-1}) + 3\kappa_{\mathcal{C}}(\mathbf{X}) + 4.$$

Proposition 7.1 is useful because it relates Theorem 3.1 and Proposition 3.3 with previously computed bounds on Gaussian widths [21, 31].

The proof of Proposition 7.1 requires two short lemmas.

Lemma 7.2. *Suppose $\mathbf{X} \neq \mathbf{0}$. Define \mathcal{H} to be the affine hull of $\partial\|\mathbf{X}\|_{\mathcal{C}}$ and $\mathbf{w}_0 := \Pi_{\mathcal{H}}(\mathbf{0})$. Then $\langle \mathbf{w} - \mathbf{w}_0, \mathbf{w}_0 \rangle = 0$ for all $\mathbf{w} \in \partial\|\mathbf{X}\|_{\mathcal{C}}$ and $\|\mathbf{w}_0\|_2 > 0$.*

Proof. Since $\mathbf{X} \neq \mathbf{0}$, the subdifferential $\partial\|\mathbf{X}\|_{\mathcal{C}}$ is a proper face of the dual norm ball $\mathcal{C}^* = \{\mathbf{Y} : \|\mathbf{Y}\|_{\mathcal{C}}^* \leq 1\}$. Also, since $\mathbf{w}_0 - \mathbf{0}$ is orthogonal to \mathcal{H} , we have $\langle \mathbf{w}_0 - \mathbf{0}, \mathbf{w} - \mathbf{w}_0 \rangle = 0$ for all $\mathbf{w} \in \mathcal{H}$. In particular, this holds for all $\mathbf{w} \in \partial\|\mathbf{X}\|_{\mathcal{C}}$. By the assumption that the unit-norm ball has a non-empty interior (see reminder at the beginning of the Appendix), we have that $\mathbf{0} \in \text{int}(\mathcal{C})$, which implies that $\mathbf{0} \in \text{int}(\mathcal{C}^*)$. Consequently, \mathcal{H} does not contain $\mathbf{0}$ and thus $\mathbf{w}_0 \neq \mathbf{0}$. This implies that $\|\mathbf{w}_0\|_2 > 0$. \square

Lemma 7.3. Let $\mathbf{X} \in \mathbb{R}^p$ be an arbitrary nonzero vector. Define $\lambda^* : \mathbb{R}^p \mapsto \mathbb{R}$ as the function

$$\begin{aligned}\lambda^*(\mathbf{g}) &:= \operatorname{argmin}_{\lambda \geq 0} \operatorname{dist}(\mathbf{g}, \lambda \cdot \partial\|\mathbf{X}\|_c) \\ &= \operatorname{argmin}_{\lambda \geq 0} \operatorname{dist}^2(\mathbf{g}, \lambda \cdot \partial\|\mathbf{X}\|_c).\end{aligned}$$

Let \mathcal{H} be the affine hull of $\lambda \cdot \partial\|\mathbf{X}\|_c$. Then λ^* is $\frac{1}{\operatorname{dist}(\mathbf{0}, \mathcal{H})}$ -Lipschitz.

Proof. Let $\mathbf{g}_1, \mathbf{g}_2$ be arbitrary vectors in \mathbb{R}^p . Since $\|\mathbf{X}\|_c < \infty$, the subdifferential $\partial\|\mathbf{X}\|_c$ is a closed convex set [58]. Hence we may let $\mathbf{w}_{\mathbf{g}_1}$ be the point in $\partial\|\mathbf{X}\|_c$ such that $\|\mathbf{w}_{\mathbf{g}_1} - \mathbf{g}_1\|_2 = \operatorname{dist}(\mathbf{g}_1, \lambda^*(\mathbf{g}_1) \cdot \partial\|\mathbf{X}\|_c)$. Define $\mathbf{w}_{\mathbf{g}_2}$ similarly. Let $\mathbf{w}_0 = \Pi_{\mathcal{H}}(\mathbf{0})$ so that

$$\begin{aligned}\|\lambda^*(\mathbf{g}_2)\mathbf{w}_{\mathbf{g}_2} - \lambda^*(\mathbf{g}_1)\mathbf{w}_{\mathbf{g}_1}\|_2 &= \|(\lambda^*(\mathbf{g}_2) - \lambda^*(\mathbf{g}_1))\mathbf{w}_0 + (\lambda^*(\mathbf{g}_2)(\mathbf{w}_{\mathbf{g}_2} - \mathbf{w}_0) + \lambda^*(\mathbf{g}_1)(\mathbf{w}_0 - \mathbf{w}_{\mathbf{g}_1}))\|_2 \\ &\geq \langle (\lambda^*(\mathbf{g}_2) - \lambda^*(\mathbf{g}_1))\mathbf{w}_0 + (\lambda^*(\mathbf{g}_2)(\mathbf{w}_{\mathbf{g}_2} - \mathbf{w}_0) + \lambda^*(\mathbf{g}_1)(\mathbf{w}_0 - \mathbf{w}_{\mathbf{g}_1})), \mathbf{w}_0 \rangle \frac{1}{\|\mathbf{w}_0\|_2} \\ &= \|\mathbf{w}_0\|_2 |\lambda^*(\mathbf{g}_2) - \lambda^*(\mathbf{g}_1)|,\end{aligned}\tag{16}$$

where the last equality follows from Lemma 7.2. Recall that projection onto a nonempty, closed convex set is nonexpansive, and thus we have $\|\mathbf{g}_2 - \mathbf{g}_1\|_2 \geq \|\Pi_{\cup_{\lambda \geq 0} \{\lambda \cdot \partial\|\mathbf{X}\|_c\}}(\mathbf{g}_2) - \Pi_{\cup_{\lambda \geq 0} \{\lambda \cdot \partial\|\mathbf{X}\|_c\}}(\mathbf{g}_1)\|_2 = \|\Pi_{\lambda^*(\mathbf{g}_2) \cdot \partial\|\mathbf{X}\|_c}(\mathbf{g}_2) - \Pi_{\lambda^*(\mathbf{g}_1) \cdot \partial\|\mathbf{X}\|_c}(\mathbf{g}_1)\|_2 = \|\lambda^*(\mathbf{g}_2)\mathbf{w}_{\mathbf{g}_2} - \lambda^*(\mathbf{g}_1)\mathbf{w}_{\mathbf{g}_1}\|_2 \geq \|\mathbf{w}_0\|_2 |\lambda^*(\mathbf{g}_2) - \lambda^*(\mathbf{g}_1)|$. \square

Proof of Proposition 7.1

Our proof is a minor modification of the proof of [31, Prop.1.]. Let \mathcal{H} be the affine hull of $\partial\|\mathbf{X}\|_c$ and $\mathbf{w}_0 = \Pi_{\mathcal{H}}(\mathbf{0})$. From Lemma 7.3, we have λ^* is $\frac{1}{\|\mathbf{w}_0\|_2}$ -Lipschitz function. Hence by [43, Theorem 5.3], we have $|\lambda^*(\varepsilon) - \mathbb{E}[\lambda^*(\bar{\varepsilon})]| \leq c$ for $\bar{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I_{p \times p})$ with probability greater than $1 - 2 \exp(-c\|\mathbf{w}_0\|_2^2/2)$. Suppressing the dependence on $\bar{\varepsilon}$, consider the event $E_c := \{|\lambda^*(\varepsilon) - \mathbb{E}[\lambda^*]| \leq c\}$, and condition on this event. Define $\mathbf{w}_1 := \Pi_{\partial\|\mathbf{X}\|_c}(\mathbf{0})$ so that $\|\mathbf{w}_1\|_2/\|\mathbf{w}_0\|_2 = \kappa(\mathbf{X})$. Let $\mathbf{w}_\varepsilon \in \partial\|\mathbf{X}\|_c$ be such that $\|\mathbf{w}_\varepsilon - \varepsilon\|_2 = \operatorname{dist}(\varepsilon, \lambda^*(\varepsilon) \cdot \partial\|\mathbf{X}\|_c)$. One has that $\frac{\lambda^*(\varepsilon)}{\mathbb{E}[\lambda^*] + c} \mathbf{w}_\varepsilon + \frac{\mathbb{E}[\lambda^*] + c - \lambda^*(\varepsilon)}{\mathbb{E}[\lambda^*] + c} \mathbf{w}_1$ is a convex combination of \mathbf{w}_1 and \mathbf{w}_ε (as we condition on E_c), and hence it belongs to $\partial\|\mathbf{X}\|_c$. Then

$$\begin{aligned}\operatorname{dist}(\varepsilon, (\mathbb{E}[\lambda^*] + c) \cdot \partial\|\mathbf{X}\|_c) &\stackrel{(i)}{\leq} \|\varepsilon - (\lambda^*(\varepsilon)\mathbf{w}_\varepsilon + (\mathbb{E}[\lambda^*] + c - \lambda^*(\varepsilon))\mathbf{w}_1)\|_2 \\ &\stackrel{(ii)}{\leq} \operatorname{dist}(\varepsilon, \lambda^*(\varepsilon) \cdot \partial\|\mathbf{X}\|_c) + \|(\mathbb{E}[\lambda^*] + c - \lambda^*(\varepsilon))\mathbf{w}_1\|_2 \\ &\stackrel{(iii)}{\leq} \operatorname{dist}(\varepsilon, \lambda^*(\varepsilon) \cdot \partial\|\mathbf{X}\|_c) + 2c\kappa(\mathbf{X})\|\mathbf{w}_0\|_2\end{aligned}\tag{17}$$

where (i) is a consequence of $\frac{\lambda^*(\varepsilon)}{\mathbb{E}[\lambda^*] + c} \mathbf{w}_\varepsilon + \frac{\mathbb{E}[\lambda^*] + c - \lambda^*(\varepsilon)}{\mathbb{E}[\lambda^*] + c} \mathbf{w}_1 \in \partial\|\mathbf{X}\|_c$, (ii) follows from the triangle inequality, and (iii) follows from the definition of $\kappa(\mathbf{X})$ and our conditioning on the event E_c . Define the function $m : \mathbb{R}^p \mapsto \mathbb{R}$

$$m(\varepsilon) = \operatorname{dist}(\varepsilon, (\mathbb{E}[\lambda^*] + c) \cdot \partial\|\mathbf{X}\|_c) - \operatorname{dist}(\varepsilon, \lambda^*(\varepsilon) \cdot \partial\|\mathbf{X}\|_c).$$

Since $m(\varepsilon)$ is the difference of two 1-Lipschitz functions and hence 2-Lipschitz, we have the concentration inequality $\mathbb{P}(m < \mathbb{E}[m] - r) \leq \exp(-r^2/8)$. By setting $r = \sqrt{8 \log(1/(1 - 2 \exp(-c\|\mathbf{w}_0\|_2^2/2)))}$ we have $\exp(-r^2/8) = 1 - 2 \exp(-c\|\mathbf{w}_0\|_2^2/2)$. From (17) the event $\{m(\varepsilon) \leq 2c\kappa(\mathbf{X})\|\mathbf{w}_0\|_2\}$ holds with probability greater than $1 - 2 \exp(-c\|\mathbf{w}_0\|_2^2/2)$. Hence it must be the case that

$$\mathbb{E}[m(\varepsilon)] \leq 2\kappa(\mathbf{X})c\|\mathbf{w}_0\|_2 + \sqrt{8 \log(1/(1 - 2 \exp(-c\|\mathbf{w}_0\|_2^2/2)))}.\tag{18}$$

Define $N := \cup_{\lambda \geq 0} \{\lambda \cdot \partial \|\mathbf{X}\|_c\}$. We have

$$\begin{aligned}
\eta_c(\mathbf{X}) &= \inf_{\lambda} \{\mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c)]\} \\
&\leq \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, (\mathbb{E}[\lambda^*] + c) \cdot \partial \|\mathbf{X}\|_c)] \\
&= \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \lambda^*(\boldsymbol{\varepsilon}) \cdot \partial \|\mathbf{X}\|_c)] + \mathbb{E}[m(\boldsymbol{\varepsilon})] \\
&\stackrel{(i)}{=} \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, N)] + \mathbb{E}[m(\boldsymbol{\varepsilon})] \\
&\stackrel{(ii)}{\leq} \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, N)] + 2\kappa(\mathbf{X})c\|\mathbf{w}_0\|_2 + \sqrt{8 \log(1/(1 - 2 \exp(-(c\|\mathbf{w}_0\|)^2/2)))} \\
&\stackrel{(iii)}{\leq} \{\mathbb{E}[\text{dist}^2(\boldsymbol{\varepsilon}, N)]\}^{1/2} + 2\kappa(\mathbf{X})c\|\mathbf{w}_0\|_2 + \sqrt{8 \log(1/(1 - 2 \exp(-(c\|\mathbf{w}_0\|)^2/2)))} \\
&\stackrel{(iv)}{\leq} \omega(T_{\mathcal{C}}(\mathbf{X}) \cap \mathbb{S}^{p-1}) + 1 + 2\kappa(\mathbf{X})c\|\mathbf{w}_0\|_2 + \sqrt{8 \log(1/(1 - 2 \exp(-(c\|\mathbf{w}_0\|)^2/2)))},
\end{aligned}$$

where (i) follows from the definition of λ^* , (ii) follows from (18), (iii) follows Jensen's Inequality, and (iv) follows from [2, Proposition 10.1]. We obtain the desired bound by setting $c = 1.5/\|\mathbf{w}_0\|_2$. \square

7.2 Analysis of proximal denoising operators

The first result describes a useful monotonicity property of convex functions [47].

Lemma 7.4 (Monotonicity, [47]). *Let f be a convex function. Let $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^p$. Then for any $\mathbf{Z}_i \in \partial f(\mathbf{X}_i), i = 1, 2$, we have*

$$\langle \mathbf{Z}_1 - \mathbf{Z}_2, \mathbf{X}_1 - \mathbf{X}_2 \rangle \geq 0.$$

Our second result applies this monotonicity property to show that the error of proximal denoising operators is robust to small changes in the underlying signal \mathbf{X}^* . Notice that our proposition also describes the performance of proximal denoisers for *combinations* of structured signals corrupted by noise. This is relevant in our subsequent analysis because the proximal denoiser is applied to averages computed near change-points.

Proposition 7.5 (Robustness). *Suppose $\hat{\mathbf{X}} = \text{argmin}_{\mathbf{X}} \frac{1}{2} \|\mathbf{X}^* + \boldsymbol{\varepsilon} - \mathbf{X}\|_2^2 + f(\mathbf{X})$ for some convex function f and*

Case 1 (Convex combination of two structured signals): $\mathbf{X}^ = \mu \mathbf{X}_0^* + (1 - \mu) \mathbf{X}_1^*$ for some $0 \leq \mu \leq 1$ is a convex combination of two signals \mathbf{X}_0^* and \mathbf{X}_1^* . Then*

$$\mathbb{E}[\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2] \leq (1 - \mu)\|\mathbf{X}_0^* - \mathbf{X}_1^*\|_2 + \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{X}_0^*))].$$

In particular when $\mu = 1$ there is no mixture. In this special case the error bound simplifies to

$$\mathbb{E}[\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2] \leq \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{X}_0^*))].$$

Case 2 (Small perturbation to a structured signal): $\mathbf{X}^ = \mathbf{X}_0^* + \boldsymbol{\Delta}$. Then*

$$\mathbb{E}[\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2] \leq \|\boldsymbol{\Delta}\|_2 + \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \partial f(\mathbf{X}_0^*))].$$

Here the expectations are with respect to $\boldsymbol{\varepsilon}$.

Proof. We only prove Case 1 since Case 2 follows from a change of variables. We begin by fixing an $\boldsymbol{\varepsilon}$. From the optimality conditions, we have $\mu \mathbf{X}_0^* + (1 - \mu) \mathbf{X}_1^* + \boldsymbol{\varepsilon} - \hat{\mathbf{X}} \in \partial \|\hat{\mathbf{X}}\|_c$. Let $\mathbf{Z}_0 = \text{argmin}_{\mathbf{Z} \in \partial f(\mathbf{X}_0^*)} \|\mathbf{Z} - \boldsymbol{\varepsilon}\|$. From the monotonicity property in Lemma 7.4 we have

$$\langle \mu \mathbf{X}_0^* + (1 - \mu) \mathbf{X}_1^* + \boldsymbol{\varepsilon} - \hat{\mathbf{X}} - \mathbf{Z}_0, \hat{\mathbf{X}} - \mathbf{X}_0^* \rangle \geq 0.$$

Rearranging terms and applying the Cauchy-Schwarz inequality, we obtain

$$(1 - \mu)\|\mathbf{X}_0^* - \mathbf{X}_1^*\|_2\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2 + \|\mathbf{Z}_0 - \boldsymbol{\varepsilon}\|_2\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2 \geq \|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2^2.$$

Finally, we divide through by $\|\mathbf{X}_0^* - \hat{\mathbf{X}}\|_2$ and take expectations on both sides with respect to $\boldsymbol{\varepsilon}$ to obtain the desired result. \square

The final result concerns a Lipschitz property of proximal operators. Demonstrating such a property allows us to subsequently appeal to concentration of measure results [43].

Lemma 7.6 (Proximal operators are non-expansive, Section 5 of [48]). *Suppose f is a convex function. Let $\hat{\mathbf{X}}(\boldsymbol{\varepsilon})$ be the optimal solution of the following optimization problem*

$$\hat{\mathbf{X}}(\boldsymbol{\varepsilon}) = \operatorname{argmin}_{\mathbf{X}} \frac{1}{2}\|\mathbf{X}^* + \boldsymbol{\varepsilon} - \mathbf{X}\|_2^2 + f(\mathbf{X}). \quad (19)$$

Then $\|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|_2 \geq \|\hat{\mathbf{X}}(\boldsymbol{\varepsilon}_1) - \hat{\mathbf{X}}(\boldsymbol{\varepsilon}_2)\|_2$.

Corollary 7.7. *Fix an $\mathbf{X}^* \in \mathbb{R}^p$. Define the function $h : \mathbb{R}^p \mapsto \mathbb{R}$ as $h(\boldsymbol{\varepsilon}) = \|\hat{\mathbf{X}}(\boldsymbol{\varepsilon}) - \mathbf{X}^*\|_2$, where $\hat{\mathbf{X}}(\boldsymbol{\varepsilon})$ is defined in (19). Then the function h is 1-Lipschitz.*

Proof. By applying the triangle inequality twice one has $\|\hat{\mathbf{X}}(\boldsymbol{\varepsilon}_1) - \hat{\mathbf{X}}(\boldsymbol{\varepsilon}_2)\|_2 \geq \|\hat{\mathbf{X}}(\boldsymbol{\varepsilon}_1) - \mathbf{X}^*\|_2 - \|\mathbf{X}^* - \hat{\mathbf{X}}(\boldsymbol{\varepsilon}_2)\|_2$. The result follows from an application of Lemma 7.6. \square

7.3 Proofs of results from Section 3

In this section we prove Proposition 3.2 (our precursor to Theorem 3.1) and Proposition 3.3. To simplify notation, we denote $\eta_{\mathcal{C}}(\mathcal{X})$ by η in this section. First we establish a tertiary result that is useful for obtaining a sharper bound on the accuracy of the locations of the estimated change-points.

Proposition 7.8. *Fix an $\mathbf{X}^* \in \mathbb{R}^p$. Let $\hat{\mathbf{X}}_0$ and $\hat{\mathbf{X}}_1$ be the optimal solutions to $\hat{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X}} \frac{1}{2}\|\mathbf{X}^* + \boldsymbol{\varepsilon} - \mathbf{X}\|_2^2 + f(\mathbf{X})$ for $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_1$, respectively. Define the function $j : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$, $j(\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_1) := \|\hat{\mathbf{X}}_0 - \hat{\mathbf{X}}_1\|_2$. Then j is $\sqrt{2}$ -Lipschitz.*

Proof. Let $\{\hat{\mathbf{X}}_0^1, \hat{\mathbf{X}}_1^1\}$ and $\{\hat{\mathbf{X}}_0^2, \hat{\mathbf{X}}_1^2\}$ be the optimal solutions corresponding to the two instantiations $(\boldsymbol{\varepsilon}_0^1, \boldsymbol{\varepsilon}_1^1)$ and $(\boldsymbol{\varepsilon}_0^2, \boldsymbol{\varepsilon}_1^2)$ of the vectors $(\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_1)$. From Lemma 7.6, we have $\|\hat{\mathbf{X}}_0^1 - \hat{\mathbf{X}}_0^2\|_2 \leq \|\boldsymbol{\varepsilon}_0^1 - \boldsymbol{\varepsilon}_0^2\|_2$ and $\|\hat{\mathbf{X}}_1^1 - \hat{\mathbf{X}}_1^2\|_2 \leq \|\boldsymbol{\varepsilon}_1^1 - \boldsymbol{\varepsilon}_1^2\|_2$. By applying the triangle inequality, we have $\|\hat{\mathbf{X}}_0^1 - \hat{\mathbf{X}}_1^1\|_2 \leq \|\hat{\mathbf{X}}_0^1 - \hat{\mathbf{X}}_0^2\|_2 + \|\hat{\mathbf{X}}_0^2 - \hat{\mathbf{X}}_1^2\|_2 + \|\hat{\mathbf{X}}_1^2 - \hat{\mathbf{X}}_1^1\|_2$. Then

$$\begin{aligned} \|\|\hat{\mathbf{X}}_0^1 - \hat{\mathbf{X}}_1^1\|_2 - \|\hat{\mathbf{X}}_0^2 - \hat{\mathbf{X}}_1^2\|_2\| &\leq \|\hat{\mathbf{X}}_0^1 - \hat{\mathbf{X}}_0^2\|_2 + \|\hat{\mathbf{X}}_1^2 - \hat{\mathbf{X}}_1^1\|_2 \\ &\leq \|\boldsymbol{\varepsilon}_0^1 - \boldsymbol{\varepsilon}_0^2\|_2 + \|\boldsymbol{\varepsilon}_1^1 - \boldsymbol{\varepsilon}_1^2\|_2 \\ &\leq \sqrt{2}\|(\boldsymbol{\varepsilon}_0^1, \boldsymbol{\varepsilon}_1^1) - (\boldsymbol{\varepsilon}_0^2, \boldsymbol{\varepsilon}_1^2)\|_2. \end{aligned}$$

Hence, j is $\sqrt{2}$ -Lipschitz. \square

Proof of Proposition 3.2 We divide the proof into three parts corresponding to the three events of interest.

Part one $[\mathbb{P}(E_1^c) \leq 2n^{1-r^2}]$: For each change-point $t \in \tau^*$, define the following event $E_{1,t} : \{S_t \geq \gamma\}$. Clearly, $E_1^c = \bigcup_{t \in \tau^*} E_{1,t}^c$. We will prove that $\mathbb{P}(E_{1,t}^c) \leq 2n^{-r^2}$. By taking a union bound over all $t \in \tau^*$, we have

$$\mathbb{P}(E_1^c) = \mathbb{P}\left(\bigcup_{t \in \tau^*} E_{1,t}^c\right) \leq \sum_{t \in \tau^*} \mathbb{P}(E_{1,t}^c) \leq 2|\tau^*|n^{-r^2} \leq 2n^{1-r^2}.$$

We now prove that $\mathbb{P}(E_{1,t}^c) \leq 2n^{-r^2}$. Conditioning on the event $E_{1,t}^c$, and by the triangle inequality, we have

$$\begin{aligned} \gamma &> \|\hat{\mathbf{X}}[t-\theta+1] - \hat{\mathbf{X}}[t+1]\|_2 \\ &\geq -\|\mathbf{X}^*[t-\theta+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2 + \|\mathbf{X}^*[t+1] - \mathbf{X}^*[t-\theta+1]\|_2 - \|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2. \end{aligned}$$

Since $\|\mathbf{X}^*[t+1] - \mathbf{X}^*[t-\theta+1]\|_2 \geq \Delta_{\min} \geq 2\gamma$, one of the two events $\{\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t]\|_2 \geq \gamma/2\}$ or $\{\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2 \geq \gamma/2\}$ must occur. Also, since $t \in \tau^*$, we have $|t-t'| \geq \theta$ for all $t' \in \tau^* \setminus \{t\}$. Hence the signal is constant over the time instances $\{t-\theta+1, \dots, t\}$ and $\{t+1, \dots, t+\theta\}$. By applying Proposition 7.5, we have the inequalities $\mathbb{E}[\|\mathbf{X}^*[t-\theta+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] \leq \frac{\sigma}{\sqrt{\theta}}\eta$ and $\mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2] \leq \frac{\sigma}{\sqrt{\theta}}\eta$. Thus

$$\begin{aligned} \mathbb{P}(E_{1,t}^c) &\leq \mathbb{P}(\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t]\|_2 \geq \gamma/2) + \mathbb{P}(\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2 \geq \gamma/2) \\ &\leq \mathbb{P}(\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t]\|_2 \geq \mathbb{E}[\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t]\|_2] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}) \\ &\quad + \mathbb{P}(\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2 \geq \mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}) \\ &\leq 2\exp(-(r\sqrt{2\log n})^2/2) = 2n^{-r^2} \end{aligned}$$

where the last inequality follows from Corollary 7.7 and from [43, Theorem 5.3].

Part two [$\mathbb{P}(E_2^c) \leq 2n^{1-r^2}$]: We prove that $\mathbb{P}(E_2^c) \leq 2n^{1-r^2}$ in essentially the same manner in which we showed that $\mathbb{P}(E_1^c) \leq 2n^{1-r^2}$. For all $t \in \tau_{\text{far}}$, define $E_{2,t}$ as the event $E_{2,t} := \{\|\hat{\mathbf{X}}[t-\theta+1] - \hat{\mathbf{X}}[t+1]\|_2 \leq \gamma\}$. Then $E_2^c = \bigcup_{t \in \tau_{\text{far}}} E_{2,t}^c$. We will start by proving that $\mathbb{P}(E_{2,t}^c) \leq 2n^{-r^2}$.

By applying the triangle inequality and conditioning on the event $E_{2,t}^c$ holding for some $t \in \tau_{\text{far}}$, we have $\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t+1]\|_2 + \|\mathbf{X}^*[t+1] - \hat{\mathbf{X}}[t+1]\|_2 > \|\hat{\mathbf{X}}[t-\theta+1] - \hat{\mathbf{X}}[t+1]\|_2 > \gamma$. Consequently, one of the two events $\{\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t+1]\|_2 \geq \gamma/2\}$ or $\{\|\mathbf{X}^*[t+1] - \hat{\mathbf{X}}[t+1]\|_2 \geq \gamma/2\}$ must hold. Since $t \in \tau_{\text{far}}$, we have $|t-t^*| > \theta$ for all $t^* \in \tau^*$, and thus the signal is constant over the time instances $\{t-\theta+1, \dots, t+\theta\}$. By Proposition 7.5, we have $\mathbb{E}[\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t-\theta+1]\|_2] \leq \frac{\sigma}{\sqrt{\theta}}\eta$ and $\mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2] \leq \frac{\sigma}{\sqrt{\theta}}\eta$. This implies that we have that at least one of the following two events $\{\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t-\theta+1]\|_2 \geq \mathbb{E}[\|\hat{\mathbf{X}}[t-\theta+1] - \mathbf{X}^*[t-\theta+1]\|_2] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}\}$ or $\{\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2 \geq \mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2] + r\sqrt{\sigma^2/\theta}\sqrt{2\log n}\}$ holds.

From Corollary 7.7 and from [43, Theorem 5.3], we have that the probability of either event (corresponding to these two inequalities) occurring is less than $2\exp(-(r\sqrt{2\log n})^2/2) = 2n^{-r^2}$. Thus

$$\mathbb{P}(E_2^c) = \mathbb{P}\left(\bigcup_{t \in \tau_{\text{far}}} E_{2,t}^c\right) \leq \sum_{t \in \tau_{\text{far}}} \mathbb{P}(E_{2,t}^c) \leq 2|\tau_{\text{far}}|n^{-r^2} \leq 2n^{1-r^2},$$

as required.

Part three [$\mathbb{P}(E_3^c) \leq n^{1-r^2}$]: Let us now consider the event E_3 . To simplify notation, we define $l := 4r\sqrt{\log n}/\eta$. To prove this part of the proposition, we show a slightly stronger result $\mathbb{P}(E_3^c) \leq 4\theta|\tau^*|\exp(-l^2\eta^2/16)$. Since $\theta|\tau^*| \leq n/4$, our bound would imply that $\mathbb{P}(E_3^c) \leq n^{1-r^2}$.

For all pairs $(t, \delta) \in \tau_{\text{buffer}}$, define the event $E_{3,t,\delta} = \{\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2 > \|\hat{\mathbf{X}}[t+1 + \delta] - \hat{\mathbf{X}}[t-\theta+1 + \delta]\|_2\}$. Then $E_3^c = \bigcup_{(t,\delta) \in \tau_{\text{buffer}}} E_{3,t,\delta}^c$. We start by proving the following bound

$$\mathbb{P}(E_{3,t,\delta}^c) \leq 2\exp(-l^2\eta^2/16)$$

for all pairs (t, δ) in τ_{buffer} . Fix one such pair and let Δ_t denote the magnitude of the change at

$t \in \tau^*$. From the triangle inequality and Proposition 7.5 we have that

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] \\ & \geq -\mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \mathbf{X}^*[t+1]\|_2] \\ & \quad + \mathbb{E}[\|\mathbf{X}^*[t+1] - \mathbf{X}^*[t-\theta+1]\|_2] - \mathbb{E}[\|\mathbf{X}^*[t-\theta+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] \\ & \geq \Delta_t - 2\sqrt{\sigma^2/\theta}\eta. \end{aligned}$$

Suppose that $\delta \geq 0$. By similarly applying the triangle inequality and Proposition 7.5 we have

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] \\ & \leq \mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \mathbf{X}^*[t+1]\|_2] + \mathbb{E}[\|\mathbf{X}^*[t+1] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] \\ & \leq (1-\delta/\theta)\Delta_t + 2\sqrt{\sigma^2/\theta}\eta. \end{aligned}$$

A similar set of computations will show that $\mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] \leq (1+\delta/\theta)\Delta_t + 2\sqrt{\sigma^2/\theta}\eta$ for $\delta < 0$. Combining these inequalities and using the range of values of δ we have

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] - \mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] & \geq \frac{|\delta|}{\theta}\Delta_t - 4\frac{\sigma}{\sqrt{\theta}}\eta \\ & \geq l\frac{\sigma}{\sqrt{\theta}}\eta. \end{aligned} \quad (20)$$

Then

$$\begin{aligned} \mathbb{P}(E_{3,t,\delta}^c) &= \mathbb{P}(\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2 > \|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2) \\ & \stackrel{(i)}{\leq} \mathbb{P}\left(\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2 - \|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2 \right. \\ & \quad \left. + \mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] - \mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] \geq \frac{l\sigma}{\sqrt{\theta}}\eta\right) \\ & \stackrel{(ii)}{\leq} \mathbb{P}\left(\mathbb{E}[\|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2] - \|\hat{\mathbf{X}}[t+1] - \hat{\mathbf{X}}[t-\theta+1]\|_2 \geq \frac{l\sigma}{2\sqrt{\theta}}\eta\right) \\ & \quad + \mathbb{P}\left(\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2 - \mathbb{E}[\|\hat{\mathbf{X}}[t+1+\delta] - \hat{\mathbf{X}}[t-\theta+1+\delta]\|_2] \right. \\ & \quad \left. \geq \frac{l\sigma}{2\sqrt{\theta}}\eta\right) \\ & \stackrel{(iii)}{\leq} 2\exp(-l^2\eta^2/16), \end{aligned}$$

where (i) follows from (20), (ii) follows from the triangle inequality, and (iii) follows from Proposition 7.8 and from [43, Theorem 5.3]. Since $E_3^c = \bigcup_{(t,\delta) \in \tau_{\text{buffer}}} E_{3,t,\delta}^c$, we have via a union bound

$$\mathbb{P}(E_3^c) \leq \sum_{(t,\delta) \in \tau_{\text{buffer}}} \mathbb{P}(E_{3,t,\delta}^c) \leq 2|\tau_{\text{buffer}}| \exp(-l^2\eta^2/16) \leq 4\theta|\tau^*| \exp(-l^2\eta^2/16).$$

This concludes the proof of Proposition 3.2. \square

Before proving Proposition 3.3 we require a short lemma.

Lemma 7.9. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_{p \times p})$. Then*

$$\text{dist}^2(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c) \leq 2(\mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c)])^2 + 2\sigma^2 t^2$$

with probability greater than $1 - 2 \exp(-t^2/2)$.

Proof. The mapping $\boldsymbol{\varepsilon} \mapsto \text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c)$ is nonexpansive and hence 1-Lipschitz. Using Theorem 5.3 from [43], we have

$$\text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c) \leq \mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c)] + t\sigma \tag{21}$$

with probability greater than $1 - \exp(-t^2/2)$. By conditioning on the event corresponding to the inequality (21), we apply the arithmetic-geometric-mean inequality and conclude that

$$\text{dist}^2(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c) \leq 2(\mathbb{E}[\text{dist}(\boldsymbol{\varepsilon}, \lambda \cdot \partial \|\mathbf{X}\|_c)])^2 + 2t^2\sigma^2$$

with probability greater than $1 - \exp(-t^2/2)$. □

Proof of Proposition 3.3 It follows from the proof of Proposition 3.2 that the event $E_1 \cap E_2$ holds with probability greater than $1 - 4n^{1-r^2}$. Conditioning on the event that $E_1 \cap E_2$ holds, the reconstructed signal is constant over the interval $\{t_1 + \theta, \dots, t_2 - \theta\}$. The result then follows from an application of Lemma 7.9 and a union bound. □