Robust Mean Change Point Testing in High-dimensional Data with Heavy Tails

Mengchu Li, Yudong Chen, Tengyao Wang and Yi Yu

Abstract

We study mean change point testing problems for high-dimensional data, with exponentially- or polynomially-decaying tails. In each case, depending on the ℓ_0-norm of the mean change vector, we separately consider dense and sparse regimes. We characterise the boundary between the dense and sparse regimes under the above two tail conditions for the first time in the change point literature and propose novel testing procedures that attain optimal rates in each of the four regimes up to a poly-iterated logarithmic factor. To be specific, when the error distributions possess exponentially-decaying tails, a near-optimal CUSUM-type statistic is considered. As for polynomially-decaying tails, admitting bounded α-th moments for some $\alpha \geq 4$, we introduce a median-of-means-type test statistic that achieves a near-optimal testing rate in both dense and sparse regimes. Our investigation in the even more challenging case of $2 \leq \alpha < 4$, unveils a new phenomenon that the minimax testing rate has no sparse regime, i.e. testing sparse changes is information-theoretically as hard as testing dense changes. Finally, we consider various extensions where we also obtain near-optimal performances, including testing against multiple change points, allowing temporal dependence as well as fewer than two finite moments in the data generating mechanisms. We also show how sub-Gaussian rates can be achieved when an additional minimal spacing condition is imposed under the alternative hypothesis.

Index Terms

Change points, Heavy-tailed error, Minimax testing, High-dimensional data, Robustness

I. INTRODUCTION

IN this paper, we study change point testing problems when the observations are corrupted by heavy-tailed errors. To be specific, consider the 'signal plus noise' model

$$X_t = \theta_t + E_t \in \mathbb{R}^p, \qquad t = 1, \ldots, n,$$

where X_t represents the p-variate observation at time t, θ_t the signal and E_t the error term. Writing $X := (X_1, \ldots, X_n) \in \mathbb{R}^{p \times n}$, and similarly for θ and E, we express the model in matrix form as:

$$X = \theta + E, \tag{1}$$

where X, θ and E are all $p \times n$ matrices. We start by assuming the entries of E are independent random variables with zero mean and unit variance and we denote the distribution of E as $P_e \in \mathcal{Q}$. We are interested in understanding the fundamental difficulty of testing whether the columns of θ undergo a change at some unknown location when the class $\mathcal{Q}$ contains heavy-tailed distributions. Focusing on the single change point alternative hypothesis for now, our goal can be formalised as testing

$$\mathrm{H}_0 : \theta \in \Theta_0(p, n) \quad \text{vs.} \quad \mathrm{H}_1 : \theta \in \Theta(p, n, s, \rho) := \bigcup_{t_0=1}^{n-1} \Theta^{(t_0)}(p, n, s, \rho), \tag{2}$$

with

$$\Theta_0(p, n) := \{\theta : \theta_t = \mu \text{ for all } t = 1, \ldots, n, \text{ for some } \mu \in \mathbb{R}^p\} \tag{3}$$

and

$$\Theta^{(t_0)}(p, n, s, \rho) := \left\{ \theta : \theta_t = \mu_1 \text{ for } t = 1, \ldots, t_0, \ \theta_t = \mu_2 \text{ for } t = t_0 + 1, \ldots, n, \right.$$
$$\left. \text{for some } \mu_1, \mu_2 \in \mathbb{R}^p \text{ s.t. } \|\mu_1 - \mu_2\|_0 \leq s, \ \frac{t_0(n - t_0)}{n} \|\mu_1 - \mu_2\|_2^2 \geq \rho^2 \right\}. \tag{4}$$

For any $v = \big(v(1), \ldots, v(d)\big)^\top \in \mathbb{R}^d$, we denote $\|v\|_0 := \sum_{i=1}^d \mathbb{1}_{\{v(i) \neq 0\}}$ and $\|v\|_2 := \big\{\sum_{i=1}^d v(i)^2\big\}^{1/2}$. To put it in words, we use $\Theta_0(p, n)$ to denote the space of signals without a change point, and $\Theta^{(t_0)}(p, n, s, \rho)$ to denote the space of signals with a change at location t_0 of entry-wise sparsity level s and (normalised) signal strength ρ. The multiplicative factor $t_0(n - t_0)n^{-1}$ of $\|\mu_1 - \mu_2\|_2^2$ can be regarded as the effective sample size of the problem. It reflects the fact that the difficulty of testing change point is related to where the change happens.

Change point analysis as a broad topic has received increasing attention in recent years. Various models [1]–[6] are considered in the literature focusing on different tasks, including testing the existence of change points, estimating their locations and quantifying the uncertainty of the proposed estimators. From a theoretical point of view, many of the problems studied are shown to exhibit a phase transition phenomenon, i.e. a change point can only be reliably tested or accurately localised when its signal strength, measured in some problem-dependent way, exceeds some threshold. It is, therefore, crucial to understand the boundary of this phase transition behaviour. For the testing problem that we are concerned with here, the key quantity is the minimax testing rate, $v_{\mathcal{Q}}^*(p, n, s)$, defined below. For a given θ and $E \sim P_e$, we write $\mathbb{P}_{\theta, P_e}$ the probability measure of the data X generated from (1) and $\mathbb{E}_{\theta, P_e}$ the corresponding expectation operator.

Definition 1 (Minimax testing rate). *Let Φ denote the set of all measurable test functions $\phi : \mathbb{R}^{p \times n} \to \{0, 1\}$. Consider the minimax testing error*

$$\mathcal{R}_{\mathcal{Q}}(\rho) := \inf_{\phi \in \Phi} \mathcal{R}_{\mathcal{Q}}(\rho, \phi) := \inf_{\phi \in \Phi} \left\{ \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_{\theta, P_e}(\phi) + \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta(p,n,s,\rho)} \mathbb{E}_{\theta, P_e}(1 - \phi) \right\}.$$

For a fixed $\varepsilon \in (0, 1/2)$, we say that $v_{\mathcal{Q}}^*(p, n, s)$ is the minimax testing rate if $\mathcal{R}_{\mathcal{Q}}(\rho) \leq \varepsilon$ when $\rho^2 \geq C v_{\mathcal{Q}}^*(p, n, s)$, and $\mathcal{R}_{\mathcal{Q}}(\rho) \geq 1/2$ when $\rho^2 \leq c v_{\mathcal{Q}}^*(p, n, s)$, where $c, C > 0$ are constants depending only on ε and $\mathcal{Q}$.

Note that in Definition 1, C is allowed to depend on ε. Since the primary goal of the paper is to characterise the minimal size of the signal, in terms of various model parameters, where the testing problem starts to become feasible, we will treat ε as a constant throughout the rest of the paper.

A minimax testing rate is previously studied in [2] under model (1), where the entries of noise matrix E are assumed to be independent standard normal random variables. It is shown that

$$v_{N^\otimes(0,1)}^*(p, n, s) = \left\{ \sqrt{p \log \log(8n)} \wedge \left[s \log \left\{ eps^{-2} \log \log(8n) \right\} \right] \right\} \vee \log \log(8n), \tag{5}$$

where $N^\otimes(0, 1)$ denotes the joint distribution of all pn independent $N(0, 1)$ entries in E. Our main contribution, presented in Section I-A, is to characterise the impact of heavy-tailed distributions on the minimax testing rate. More specifically, we consider two classes of error distributions.

Definition 2 ($\mathcal{G}_{\alpha, K}$ class of distributions). *For $K > 0$ and $\alpha \in (0, 2]$, let $\mathcal{G}_{\alpha, K}$ denote the class of distributions on $\mathbb{R}$ such that for any $P \in \mathcal{G}_{\alpha, K}$ and random variable $W \sim P$, it holds that*

$$\mathbb{E}(W) = 0, \quad \mathbb{E}(W^2) = 1 \quad and \quad \mathbb{E}\big(\exp\{|W/K|^\alpha\}\big) \leq 2.$$

The $\mathcal{G}_{\alpha, K}$ class consists of sub-Weibull distributions of order α with mean 0, variance 1 and the Orlicz ψ_α-norm upper bounded by K (see Definitions 6 and 7). By Proposition 29(a), they possess exponentially-decaying tails, as $\mathbb{P}(|W| \geq x) \leq 2e^{-(x/K)^\alpha}$, for $x > 0$.

Definition 3 ($\mathcal{P}_{\alpha, K}$ class of distributions). *For $K > 0$ and $\alpha \geq 2$, let $\mathcal{P}_{\alpha, K}$ denote the class of distributions on $\mathbb{R}$ such that for any $P \in \mathcal{P}_{\alpha, K}$ and random variable $W \sim P$, it holds that*

$$\mathbb{E}(W) = 0, \quad \mathbb{E}(W^2) = 1 \quad and \quad \mathbb{E}\big(|W/K|^\alpha\big) \leq 1.$$

In words, each distribution within this class has its α-th moment bounded above by $K^\alpha < \infty$ and possesses a polynomially-decaying tail. This is typically much heavier than an exponentially-decaying tail and thus poses a much bigger statistical challenge.

We study the minimax rate of testing $v_{\mathcal{Q}}^*(p, n, s)$ defined in Definition 1 for $\mathcal{Q} = \mathcal{G}_{\alpha, K}^\otimes$ and $\mathcal{Q} = \mathcal{P}_{\alpha, K}^\otimes$, respectively. Let $\mathcal{G}_{\alpha, K}^\otimes$ and $\mathcal{P}_{\alpha, K}^\otimes$ denote the class of joint distributions of all the entries in the error matrix $E \in \mathbb{R}^{p \times n}$, when each entry of E independently follows a distribution on $\mathbb{R}$ that belongs to the class $\mathcal{G}_{\alpha, K}$ and $\mathcal{P}_{\alpha, K}$, respectively. Throughout the paper, we treat K and α as constants.

A. Main results and outline

Our main results, developed in Sections II and III, characterise the minimax testing rates under both exponentially- and polynomial-decaying tails. In particular, we establish that in the case of exponentially-decaying tails

$$v_{\mathcal{G}_{\alpha,K}^\otimes}^*(p, n, s) \asymp L \min\{\sqrt{p}, \, s \log^{2/\alpha}(ep/s)\} + \log \log(8n) \tag{6}$$

for some $L \in \left[1, \sqrt{\log \log(8n)}\right]$, and in the case of polynomial-decaying tails

$$v^*_{\mathcal{P}^{\otimes}_{\alpha,K}}(p, n, s) \asymp L \min\{p^{\frac{2}{\alpha} \vee \frac{1}{2}}, s(p/s)^{\frac{2}{\alpha}}\} + \log \log(8n) \tag{7}$$

for some $L \in [1, \log \log(8n)]$. Upper and lower bounds with explicit dependence on iterated logarithmic factors are detailed in Section II-C and section III-C.

Note that as the level of sparsity s increases from 1 to p, there is a changeover in the dominating term in both (6) and (7). Depending on whether directly involving s in the final rates, we refer to the resulting regimes the sparse and dense regimes, with their transition boundaries $s^*_{\mathcal{G}}$ and $s^*_{\mathcal{P}}$ determined by $s^*_{\mathcal{G}} \log^{2/\alpha}(ep/s^*_{\mathcal{G}}) = \sqrt{p}$ for $\mathcal{G}^{\otimes}_{\alpha,K}$ and $s^*_{\mathcal{P}}(p/s^*_{\mathcal{P}})^{\frac{2}{\alpha}} = p^{\frac{2}{\alpha} \vee \frac{1}{2}}$ for $\mathcal{P}^{\otimes}_{\alpha,K}$.

The transition boundaries are demonstrated in Figure 1. When $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$, the minimax testing rate transition occurs at $s^*_{\mathcal{P}} = p^{1/2 - 1/(\alpha - 2)}$ when $\alpha \geq 4$. When $\alpha \in [2, 4)$, there is essentially no sparse regime, since in this range of α, there does not exist any $s \in [1, p]$ such that $s(p/s)^{\alpha/2} < p^{2/\alpha}$. This observation implies that testing sparse change is information-theoretically as hard as testing dense changes when $\alpha \in [2, 4]$, as the minimax testing rate is independent of s. When $P_e \in \mathcal{G}^{\otimes}_{\alpha,K}$, the transition boundary takes a simpler form of $s^*_{\mathcal{G}} \asymp \sqrt{p} \log^{-2/\alpha}(ep)$ for $\alpha \in (0, 2]$.
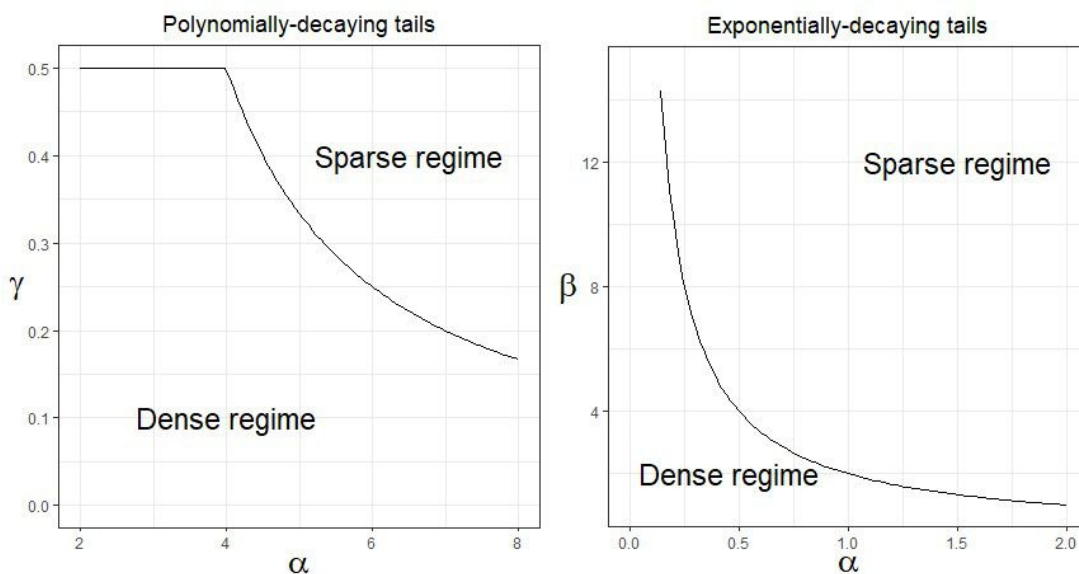

Fig. 1. Minimax testing rate transition boundaries between dense and sparse regimes when the distribution of the error matrix belongs to $\mathcal{P}^{\otimes}_{\alpha,K}$ (left panel) and $\mathcal{G}^{\otimes}_{\alpha,K}$ (right panel). The left panel plots the curve $\gamma(\alpha) = (\alpha - 2)^{-1} \wedge 1/2$ for $\alpha \in [2, \infty)$, and the two regimes are separated by $s^*_{\mathcal{P}} = p^{1/2 - \gamma}$. The right panel plots the curve $\beta(\alpha) = 2/\alpha$ for $\alpha \in (0, 2]$, and the two regimes are separated by $s^*_{\mathcal{G}} \asymp \sqrt{p} \log^{-\beta}(ep)$.

The upper bounds on (6) and (7) are each obtained by analysing two different testing procedures separately, targeting at dense and sparse regimes. In practice, the level of sparsity is usually unknown and we address the issue of adaptation to sparsity in Section IV. We show that there is no additional cost of adaptation in achieving the optimal minimax testing rates. Finally, in Section VI, we consider extensions in four interesting directions, including (1) testing against multiple change points, (2) accounting for temporal dependence among observations, (3) addressing the case where the noise matrix entries have fewer than two finite moments, and (4) examining the situation where an additional minimal spacing condition is imposed between the potential change point and the boundary time points. Generally speaking, we develop near-optimal procedures under these more general settings and demonstrate an interesting phenomenon in (4)—tests can achieve sub-Gaussian performances under heavy-tailed noise assumptions if it is known that the potential change point is away from the endpoints by a small distance that only depends on the model parameters p, n and s through logarithmic terms.

To highlight our contributions relative to the existing literature, we note that in previous works on robust mean change point testing problems [7], [8], change point locations are required to be comparable to the length of time series in order to achieve near-optimal guarantees. Our results, in comparison, cover a much more general parameter space, where the change point locations may be arbitrarily close to the boundary. Compared to recent works on optimal mean change point testing problems without robustness [2], [6], our results allow for general classes of distributions beyond Gaussian and sub-Gaussian cases and quantify the costs of heavy-tailedness. Finally, compared to relevant recent works on robust estimation in sequence models [9], we investigate the more challenging case where noise entries have fewer than four finite moments, and unveil a new phenomenon on the effects of sparsity that was previously unknown even in sequence models. More in-depth discussions on these works can be found in Section I-B.

B. Relation to existing literature

Many real-world data such as financial returns and macroeconomic variables exhibit heavy-tail phenomena, which often violate the convenient sub-Gaussian/exponential assumptions adopted by data analysts. Statistical procedures that mitigate the effects of heavy-tailed and/or contaminated data, therefore, have been sought after in practice, see [10] for more in-depth discussions. In the realm of change point analysis, one line of recent works [4], [5], [11] consider change point models with exponentially-decaying heavy-tailed noise and study the performance of non-robust algorithms that perform well under sub-Gaussian noise assumptions. Theoretical results therein all require stronger assumptions on the strength of change points compared to the setting under sub-Gaussian assumptions. One motivation for our work is thus to investigate to what extent ideas from robust statistics are useful in analysing change points within high-dimensional heavy-tailed data streams.

Another line of work develops algorithms with robust components for change point analysis. In particular, in the univariate mean change setting, [12] propose to swap the commonly used ℓ_2-loss with other loss functions, including the biweight and Huber loss functions to enhance robustness against heavy-tailed errors in localising change points. [13] deploy a robust mean estimator with a scanning window idea to estimate multiple change point locations under a more general Huber contamination framework. Their results show that, in terms of the minimax detection boundary, there is essentially no cost of relaxing the sub-Gaussian assumption to more flexible finite moment assumptions. Robust change point analysis methodologies have also been proposed in other contexts including change point detection in stump models [14], high-dimensional linear models [15] and functional time series [16], as well as detecting covariance changes [17] and distributional changes [18]. There is also work exploring rank-based methods and focusing on univariate time series data [19]–[23]. We remark that some works focus on robust online change point detection [24]–[26], which is different from the offline version that we study here[1].

Closer to our high-dimensional mean change point setting, [7] and [8] both consider the testing problem (2) and propose robust methodology targeting at sparse and dense changes, respectively. [7] formulate the problem as testing location parameter change, which in contrast to our model, allows the noise distribution to have mean parameter being infinite. Their methodology involves a U-statistic with an anti-symmetric and bounded kernel, followed by an ℓ_∞ aggregation. The power analysis of their proposed test (cf. Theorem 3.3 therein) along with subsequent remarks provide finite sample results showing that their test is able to detect the change point when it is sufficiently away from the boundary. In particular, their Remark 4 suggests that detection is only possible for local alternative when the change point location satisfies

$$t_0 \wedge (n - t_0) \geq c\sqrt{n \log(np)},$$

for some absolute constant $c > 0$. In comparison, our results hold for the parameter space $\Theta(p, n, s, \rho)$ that covers all possible locations of change points. Moreover, as discussed in Remark 5 therein, their procedure achieves the sparse regime rate in $v^*_{N \otimes (0,1)}(p, n, s)$ up to a poly-logarithmic factor in n and p only when $t_0 = cn$ for some fixed constant $c \in (0, 1)$. [8] consider the same mean change point testing problem as ours but without sparsity constraints, while allowing for a form of weak spatial dependence across coordinates. In terms of methodology, they also utilise a robustified U-statistic and combine it with the self-normalisation technique. They derive the limiting distributions of the proposed test under the sequential asymptotics. It is discussed in Remark 2 therein that, asymptotically, their test achieves the dense rate $v^*_{N \otimes (0,1)}(p, n, p)$ up to a logarithmic factor in n, when the change point location satisfies $t_0 = cn$ for some fixed constant $c \in (0, 1)$.

In comparison to the results in [7] and [8], our results are non-asymptotic and reveal that when considering the whole parameter space $\Theta(p, n, s, \rho)$, where the change point locations may be arbitrarily close to the boundary, the fundamental difficulty of the testing problem changes drastically. In particular, the heavy-tailed distributions manifest a strong impact on the minimax testing rates and one can no longer achieve the Gaussian-like minimax testing rates, especially in the sparse regime. Moreover, our results are generally sharper in the sense that we characterise the minimax testing rates up to a factor of at most $\log \log(8n)$.

Lastly, we mention two recent works [9] and [2], that are technically related to ours. [9] study the sparse sequence models where

$$Y_i = \theta_i + \sigma \xi_i, \quad i = 1, \ldots, p.$$

The noise random variables ξ_i's are i.i.d. with some distribution belonging to either $\mathcal{G}_{\alpha,K}$ or $\mathcal{P}_{\alpha,K}$, and the signal θ is assumed to be ℓ_0-sparse with sparsity s. They provide minimax rates for estimating $\|\theta\|_2$ among other results (cf. Table 1 therein) under these two noise classes. Our results recover theirs when n is of constant order and provide a link between these two problems, while significantly generalising to the arbitrary n case. To achieve the minimax estimation rates, [9] first estimate θ via a penalised least squares estimator $\hat{\theta}$ in the sparse regime, and use $\|\hat{\theta}\|_2$ as an estimator for $\|\theta\|_2$. We adopt a different yet more intuitive hard-thresholding methodology in extracting information from sparse changes. Moreover, their upper bound rate under $\mathcal{P}_{\alpha,K}$ requires the assumption of bounded fourth moments, i.e. $\alpha \geq 4$. We investigate the more challenging case when $\alpha \in (2, 4)$, as well as $\alpha \leq 2$ in Section VI, and unveil a previously unknown phase transition behaviour even when n is of constant order.

---

[1] In an online change point analysis problem, one monitors the change points while collecting data. In the offline context, the change point analysis is conducted retrospectively.

[2] study the same testing problem (2) as ours under the Gaussian noise assumption while also considering spatial and temporal dependence. Their proposed testing procedure computes CUSUM-type statistics [27] at each location on a dyadic grid. This also serves as the starting point of various procedures in our work. By comparing the results in Table I with the rate $v^*_{N^{\otimes}(0,1)}(p,n,s)$ derived by [2] under the Gaussianity assumption, we show that the heavy-tailed errors mainly affect the difficulty of testing sparse changes, whereas in the $\mathcal{P}^{\otimes}_{\alpha,K}$ case with $\alpha \in [2,4)$, the dense rate also changes dramatically. In the special case of $p = s = 1$, our results (both upper and lower bounds in all cases) reduce to $\log\log(8n)$, which is the same rate as $v^*_{N^{\otimes}(0,1)}(1,n,1)$. This shows that, in the univariate setting, there is no extra cost of allowing for heavy-tailed errors in testing change point compared to Gaussian errors.

C. Notation

We introduce the notation used throughout the paper. Let $\mathbb{Z}^+$ denote the set of positive integers. For $d \in \mathbb{Z}^+$, write $[d] := \{1,\ldots,d\}$. Let $\lceil\cdot\rceil$, $\lfloor\cdot\rfloor$ and $\Gamma(\cdot)$ denote the ceiling, floor and Gamma functions, respectively. Given $a,b \in \mathbb{R}$, denote $a \wedge b := \min(a,b)$ and $a \vee b := \max(a,b)$. For a set S, use $\mathbb{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a vector $v = \left(v(1),\ldots,v(d)\right)^\top \in \mathbb{R}^d$, define $\|v\|_1 := \sum_{i=1}^d |v(i)|$ and $\|v\|_\infty := \max_{i\in[d]} |v(i)|$. For two vectors $v,w \in \mathbb{R}^d$, we use $\langle v,w \rangle$ to denote their inner product. For a matrix $A = (A_{ij})_{i\in[d_1],j\in[d_2]} = \left(A_j(i)\right)_{i\in[d_1],j\in[d_2]} \in \mathbb{R}^{d_1 \times d_2}$, denote the Frobenius norm $\|A\|_{\mathrm{F}} := \left(\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} A_{ij}^2\right)^{1/2}$, the operator norm $\|A\|_2 := \max_{v\in\mathbb{R}^{d_2},v\neq 0} \|Av\|_2/\|v\|_2$, the two-to-infinity norm $\|A\|_{2\to\infty} := \max_{v\in\mathbb{R}^{d_2},v\neq 0} \|Av\|_\infty/\|v\|_2$ and the max norm $\|A\|_{\max} := \max_{i\in[d_1],j\in[d_2]} |A_{ij}|$. For two probability measures P and Q on a measurable space $(\mathcal{X},\mathcal{A})$, denote the total variation distance between them as $\mathrm{TV}(P,Q) := \sup_{A\in\mathcal{A}} |P(A) - Q(A)|$. If, in addition, P and Q are absolute continuous with respect to some base measure λ, then define the squared Hellinger distance between them as $\mathrm{H}^2(P,Q) := \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \lambda(\mathrm{d}x)$, where p and q are the Radon–Nikodym derivatives of P and Q with respect to λ respectively. When the distribution is clear from the context, let $\mathbb{P}$, $\mathbb{E}$ and Var be probability, expectation and variance operators respectively. Finally, we write $a \gtrsim b$ if $a \geq C_1 b$, write $a \lesssim b$ if $a \leq C_2 b$, and write $a \asymp b$ if $C_3 b \leq a \leq C_4 b$, for some constants $C_1, C_2, C_3, C_4 > 0$ that depend only on α, K, and ε, which are treated as constants throughout this work.

II. TESTING UNDER SUB-WEIBULL NOISE DISTRIBUTIONS

In this section, we consider the entries of the noise matrix E to be independent random variables and each follows a distribution belonging to the class $\mathcal{G}_{\alpha,K}$; see Definition 2. Recall the definitions of minimax testing rates, $v^*_{\mathcal{G}_{\alpha,K}}(p,n,s)$, the worst case testing error of a given test ϕ, $\mathcal{R}_\mathcal{G}(\rho,\phi)$, and the minimax testing error $\mathcal{R}_\mathcal{G}(\rho)$, from Definition 1. For notational simplicity, we use $\mathcal{G}$ in place of $\mathcal{G}^{\otimes}_{\alpha,K}$.

As mentioned in Section I-A, we shall establish an upper bound on $\mathcal{R}_\mathcal{G}(\rho)$ by developing two testing procedures, targeting at dense and sparse change signals. We provide the details of these two testing procedures with corresponding theoretical guarantees leading to the dense and sparse rates in Sections II-A and II-B. The minimum between the two rates serves as an upper bound on $\mathcal{R}_\mathcal{G}(\rho)$ and we prove and discuss its optimality in Section II-C.

A. Testing for dense signals

To derive the dense rate, we consider the testing procedure that is used in [2]. Consider $\mathcal{T} := \left\{1,2,4,\ldots,2^{\lfloor \log_2(n/2)\rfloor}\right\}$ and a CUSUM-type statistic

$$Y_t := \frac{\sum_{i=1}^t X_i - \sum_{i=1}^t X_{n+1-i}}{\sqrt{2t}}.$$

We define our test as

$$\phi_{\mathcal{G},\mathrm{dense}} := \mathbb{1}_{\{\max_{t\in\mathcal{T}} A_t > r\}}, \tag{8}$$

where

$$A_t := \sum_{j=1}^p \left\{Y_t^2(j) - 1\right\} \tag{9}$$

and $r > 0$ is the detection threshold specified in Theorem 1. Note that it suffices to test for a change point over the dyadic grid $\mathcal{T}$ since for any true change point location $t_0 \in [n-1]$ under the alternative, there exists some $t \in \mathcal{T}$ such that $t \leq t_0 \leq 2t$, approximating the true change location up to a constant factor. The logarithmic size of $\mathcal{T}$ is the main reason behind the appearance of the $\log\log(8n)$ terms in our bounds below. The following theorem establishes the theoretical guarantee of the test $\phi_{\mathcal{G},\mathrm{dense}}$.

Theorem 1. *Let $0 < \alpha \leq 2$ and $K > 0$. For any $\varepsilon \in (0,1)$, there exist constants $C_1, C_2 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{G},\mathrm{dense}}$ defined in* (8) *with*

$$r = C_1\left(\sqrt{p\log\log(8n)} + \log\log(8n)\right)$$

satisfies

$$\mathcal{R}_\mathcal{G}(\rho, \phi_{\mathcal{G}, \text{dense}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_2 v_{\mathcal{G}, \text{dense}}^{\text{U}}$, *where*

$$v_{\mathcal{G}, \text{dense}}^{\text{U}} := \sqrt{p \log \log(8n)} + \log \log(8n).$$

Note that this simple test actually achieves the same rate in the dense regime as $v_{N^{\otimes}(0,1)}^*$ defined in (5), even though the noise distributions possess heavier tails than Gaussian and sub-Gaussian distributions. One key technical ingredient is a careful analysis of the probability of the Type I error using Lemma 35 instead of a crude union bound.

B. Testing for sparse signals

To derive the sparse rate, we employ a sample-splitting testing procedure similar to that proposed in Section 5.1 of [28]. Intuitively, we first use half of the data to identify coordinates that exhibit strong signals of change, and then use the other half to aggregate the selected 'signal' coordinates. Such a methodology is applicable for testing potential change locations $t \in \mathcal{T} \setminus \{1\}$ and we deal with testing the special case of $t = 1$ separately.

To be specific, for $t \in \mathcal{T} \setminus \{1\}$, we define a sample-splitting version of (9) that

$$\phi_{\mathcal{G}, \text{sparse}} := \mathbb{1}_{\{\max_{t \in \mathcal{T} \setminus \{1\}} A_{t,a} > r\}} \vee \mathbb{1}_{\{A_{1,a} > r_1\}}, \tag{10}$$

with

$$Y_{t,1} := \frac{\sum_{i=1}^{t/2} X_{2i-1} - \sum_{i=1}^{t/2} X_{n+2-2i}}{\sqrt{t}}, \qquad Y_{t,2} := \frac{\sum_{i=1}^{t/2} X_{2i} - \sum_{i=1}^{t/2} X_{n+1-2i}}{\sqrt{t}} \tag{11}$$

and

$$A_{t,a} := \begin{cases} \sum_{j=1}^{p} \{Y_{t,1}^2(j) - 1\} \mathbb{1}_{\{|Y_{t,2}(j)| \geq a\}}, & t \geq 2, \\ \sum_{j=1}^{p} \{Y_t^2(j) - 1\} \mathbb{1}_{\{|Y_t(j)| \geq a\}}, & t = 1, \end{cases} \tag{12}$$

where a, r, r_1 are specified in Theorem 2.

Theorem 2. *Let* $0 < \alpha \leq 2$ *and* $K > 0$. *For any* $\varepsilon \in (0,1)$, *there exist constants* $C_1, C_2, C_3, C_4 > 0$ *depending only on* α, K *and* ε, *such that the test* $\phi_{\mathcal{G}, \text{sparse}}$ *defined in* (10) *with*

$$a = C_1 \left[\log^{1/\alpha}(ep/s) + s^{-1/2} \log^{1/2} \{\log(8n)\} \right],$$

$$r = C_2 \left(\sqrt{s \log \log(8n)} + \log \log(8n) \right) \quad \text{and} \quad r_1 = C_3 s \log^{2/\alpha}(ep/s),$$

satisfies that

$$\mathcal{R}_\mathcal{G}(\rho, \phi_{\mathcal{G}, \text{sparse}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_4 v_{\mathcal{G}, \text{sparse}}^{\text{U}}$, *where*

$$v_{\mathcal{G}, \text{sparse}}^{\text{U}} := s \log^{2/\alpha}(ep/s) + \log \log(8n).$$

The idea of selecting coordinates via hard-thresholding has been widely used and in particular, in the change point context, considered in [2] and [28] under the Gaussian noise assumption. Our use of sample-splitting prompts the independence between the coordinate selection step and the ℓ_2-aggregation step. It simplifies the analysis while achieving the optimal testing rate, as we will show in Section II-C.

C. Minimax optimality

We derive lower bounds on the minimax testing rate and discuss the optimality of our testing procedures in this section. First, note that the theoretical guarantees in Theorems 1 and 2 hold for any sparsity level s. In other words, for any given $s \in [1, p]$, we can simultaneously run the two testing procedures described in Sections II-A and II-B and take $\phi_{\mathcal{G}, \text{dense}} \vee \phi_{\mathcal{G}, \text{sparse}}$ as our test. This leads to an upper bound

$$v_{\mathcal{G}, \text{dense}}^{\text{U}} \wedge v_{\mathcal{G}, \text{sparse}}^{\text{U}} = \left\{ s \log^{2/\alpha}(ep/s) \wedge \sqrt{p \log \log(8n)} \right\} + \log \log(8n) \tag{13}$$

on the minimax testing rate $v_{\mathcal{G}_{\alpha, K}^{\otimes}}^*(p, n, s)$. The following theorem presents a corresponding lower bound.

Theorem 3. *Let* $0 < \alpha \leq 2$, $K \geq K_\alpha$ *and* $s \geq c$, *for some absolute constant* $c \geq 1$ *and some constant* $K_\alpha > 0$ *depending only on* α. *There exists some constant* $c' > 0$ *depending only on* α *and* K, *such that* $\mathcal{R}_\mathcal{G}(\rho) \geq 1/2$ *whenever* $\rho^2 \leq c' v_\mathcal{G}^{\text{L}}$, *where*

$$v_\mathcal{G}^{\text{L}} := \left\{ s \log^{2/\alpha}(ep/s) \wedge \sqrt{p \{\log \log(8n)\}^{\omega_1}} \right\} + \log \log(8n)$$

and $\omega_1 = \mathbb{1}_{\{s > \sqrt{p \log \log(8n)}\}}$.

By combining the lower bound $v_{\mathcal{G}}^{\mathrm{L}}$ in Theorem 3 and the upper bound in (13), we are able to quantify the minimax test rate up to a factor of $\sqrt{\log\log(8n)}$ and conclude that our testing procedures are minimax rate optimal up to an iterated logarithmic factor. Moreover, a closer look into the sparse and dense regimes, as defined in Section I-A, reveals that our results exactly quantify $v_{\mathcal{G}_{\alpha,K}^{\otimes}}^{*}(p,n,s)$ in almost all regimes of sparsity.

Both the upper and lower bounds consist of a minimum of two terms, directly involving the sparsity level s and not. As s grows, the final rate undergoes a phase transition, namely from a sparse regime to a dense one. To better understand the transition, we focus on the relationship between p and s, setting the boundary between the sparse and the dense regime to be the solution to $s_{\mathcal{G}}^{*}\log^{2/\alpha}(ep/s_{\mathcal{G}}^{*}) = \sqrt{p}$, i.e.

$$s_{\mathcal{G}}^{*} \asymp \frac{\sqrt{p}}{\log^{2/\alpha}(ep)}. \tag{14}$$

We summarise this phenomenon in Table I. From the table, it is clear that our upper and lower bounds match exactly in the entire sparse regime (i.e. $s < s_{\mathcal{G}}^{*}$) and the majority region of the dense regime (i.e. $s > \sqrt{p\log\log(8n)}$). The $\sqrt{\log\log(8n)}$ gap between the upper and lower bounds, only exists in the region $s_{\mathcal{G}}^{*} \leq s \leq \sqrt{p\log\log(8n)}$ within the dense regime. Closing such gap is challenging and a similar gap exists even when each entry of the noise matrix follows a sub-Gaussian, yet Gaussian, distribution; see Section 7.1 of [29], where it is suggested that a procedure exploring the exact distribution of the noise might be required to close this gap.

TABLE I
BOUNDS ON THE MINIMAX TESTING RATES IN THE SUB-WEIBULL NOISE DISTRIBUTION CLASS $\mathcal{G}_{\alpha,K}^{\otimes}$, WHERE $\omega_1 = \mathbb{1}_{\left\{s > \sqrt{p\log\log(8n)}\right\}}$. UPPER BOUNDS ARE OBTAINED IN THEOREMS 1 AND 2. LOWER BOUNDS ARE OBTAINED IN THEOREM 3.

		Upper bound	Lower bound
$\mathcal{G}_{\alpha,K}^{\otimes}$	Dense	$\sqrt{p\log\log(8n)} + \log\log(8n)$	$\sqrt{p\{\log\log(8n)\}^{\omega_1}} + \log\log(8n)$
	Sparse	$s\log^{2/\alpha}(ep/s) + \log\log(8n)$	$s\log^{2/\alpha}(ep/s) + \log\log(8n)$

To highlight the effects of sub-Weibull distributions on the minimax test rate, we note that allowing heavier tails does not affect the minimax testing rate when $s > \sqrt{p\log\log(8n)}$, relative to the results under Gaussian noise assumptions; see (5). However, the tail behaviour, quantified by the parameter α, does affect the minimax rate in the sparse regime and hence the transition boundary. Specifically, as α decreases (i.e. the tail becomes heavier), the sparse rate increases, meaning that it is fundamentally more difficult to detect sparse changes as the tail of the noise distribution becomes heavier. As a prelude to Section VI-D, we also show that a modification of the test $\phi_{\mathcal{G},\mathrm{sparse}}$ in Section II-B can achieve a sparse rate that is independent of α, if a different alternative hypothesis is considered, where the change point is known to be at least $\log(ep/s) + s^{-1}\log\log(8n)$ away from the end points 1 and n.

III. TESTING UNDER FINITE MOMENT NOISE DISTRIBUTIONS

In this section, we consider the case when $P_e \in \mathcal{P}_{\alpha,K}^{\otimes}$, or equivalently, we assume that the distribution of each entry in the noise matrix E has only finite α-th moments, for some constant $\alpha \geq 2$, see Definition 3. Compared to the $\mathcal{G}_{\alpha,K}$ class of distributions considered in Section II, where standard CUSUM-type testing procedures already achieve near-optimal minimax testing rates, the $\mathcal{P}_{\alpha,K}$ class of distributions include a much wider range of noise distributions, e.g. t distributions and centred Pareto distributions. As a result, it poses a much larger statistical challenge. New approaches to tackle the testing problem are thus required.

Similar to Section II, we write the worst case testing error as $\mathcal{R}_{\mathcal{P}}(\rho,\phi)$ and the minimax testing error as $\mathcal{R}_{\mathcal{P}}(\rho)$. We again assume the sparsity level to be known and derive the dense and sparse testing rates separately in Sections III-A and III-B.

A. Testing for dense signals

We consider a testing procedure built on the median-of-means-type statistics. For $i \leq n/2$, we denote $Z_i := (X_i - X_{n-i+1})/\sqrt{2}$. For $t \in \mathcal{T}$, we split $\{Z_1, \ldots, Z_t\}$ into G_t groups of equal size (assuming that t is always a multiple of chosen G_t for simplicity) that

$$\mathcal{Z}_{t,1}, \mathcal{Z}_{t,2}, \ldots, \mathcal{Z}_{t,G_t},$$

where each group contains $t/G_t \geq 1$ elements and the number of groups G_t is specified later in (18). Set $V_{t,g} \in \mathbb{R}^p$ with

$$V_{t,g}(j) := \overline{Z}_{t,g}^2(j) - \frac{G_t}{t}, \quad g \in [G_t], \tag{15}$$

where $\overline{Z}_{t,g} \in \mathbb{R}^p$ is the sample mean of the g-th group. This quantity $V_{t,g}$ can be thought as a scaled version of the statistic A_t defined in (9), but computed using only a subset of the data. To achieve robustness against heavy-tailed errors, we consider the following median-of-means-type statistic

$$A_t^{\mathrm{MoM}} := t \cdot \mathrm{median}\left(\sum_{j=1}^{p} V_{t,1}(j), \sum_{j=1}^{p} V_{t,2}(j), \ldots, \sum_{j=1}^{p} V_{t,G_t}(j)\right). \tag{16}$$

Our test is denoted as

$$\phi_{\mathcal{P},\mathrm{dense}} := \mathbb{1}_{\left\{\max_{t \in \mathcal{T}} A_t^{\mathrm{MoM}}/r_t > 1\right\}}, \tag{17}$$

with the detection threshold r_t specified in (18). Before presenting the theoretical guarantee of the test $\phi_{\mathcal{P},\mathrm{dense}}$ in Theorem 4, we first briefly explain the significance of median-of-means-type statistics and the novelty of our procedure.

Median-of-means-type statistics like (16) have been applied in a wide range of statistical problems [30]–[34]. The most well-known and simplest form is its univariate mean estimation version. Suppose that we have i.i.d. data of sample size n with mean μ and variance σ^2. The median-of-means estimator $\hat{\mu}^{\mathrm{MoM}}$ is obtained by first partitioning the data into G groups of equal size, then calculating the sample mean within each group and finally computing the median of these G sample means. It is shown in Theorem 2 of [35] that, for $\delta \in (0,1)$, when the number of groups G is chosen to be at least $8\log(1/\delta)$, with probability at least $1 - \delta$, the estimator $\hat{\mu}^{\mathrm{MoM}} = \hat{\mu}^{\mathrm{MoM}}(\delta)$ satisfies that

$$|\hat{\mu}^{\mathrm{MoM}} - \mu| \leq \sigma\sqrt{\frac{32\log(1/\delta)}{n}}.$$

Thus, the median-of-means estimator can achieve sub-Gaussian performance in mean estimation under only the assumption of finite second moment.

However, in our context, the aforementioned methodology is not applicable for testing potential change point that is too close to the boundary, as we will not have enough data to split into the required number of groups to ensure good statistical guarantees. Therefore, for $t \in \mathcal{T}$ such that $t \leq \Delta$ with the threshold Δ specified in (18), we directly take the median of t statistics in (16), i.e. $G_t = t$. We now present the theoretical guarantee of the test $\phi_{\mathcal{P},\mathrm{dense}}$ in (17).

Theorem 4. *Assume $\alpha \geq 2$. For any $\varepsilon \in (0,1)$, there exist $C_1, C_2 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{P},\mathrm{dense}}$ defined in (17) with*

$$r_t = C_1 p^{(1/2)\vee(2/\alpha)} G_t, \quad G_t = t \wedge \Delta \quad and \quad \Delta = 2^{3+\lceil \log_2 \log\log(8n)\rceil}, \tag{18}$$

satisfies that

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\mathrm{dense}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_2 v_{\mathcal{P},\mathrm{dense}}^{\mathrm{U}}$, where

$$v_{\mathcal{P},\mathrm{dense}}^{\mathrm{U}} := p^{(2/\alpha)\vee(1/2)} \log\log(8n)$$

One challenge in our context is analysing the performance of the test $\phi_{\mathcal{P},\mathrm{dense}}$ when $\alpha \in [2,4]$. Since we compute a second-order statistic $V_{t,g}$ within each group g, standard variance-based analysis would require a bounded fourth moment condition on the distribution. However, through a more refined analysis, we extend our results to this more demanding case of $\alpha \in [2,4]$. In this setting, the dense testing rate $v_{\mathcal{P},\mathrm{dense}}^{\mathrm{U}}$ is affected by α, and we reveal a phase transition in the rate at $\alpha = 4$.

An even more challenging scenario arises when the distribution of each entry of E lacks a finite variance. In this case, an alternative test to $\phi_{\mathcal{P},\mathrm{dense}}$ is required, as the mean of the aforementioned second-order statistic $V_{t,g}$ is no longer guaranteed to be finite. We defer a detailed discussion of this setting to Section VI-C.

B. Testing for sparse signals

To derive the sparse rate, we employ a mean estimator satisfying a general condition detailed in Condition 1 in Appendix A5 to construct our test. There are potentially many choices of such a mean estimator, but one specific choice $\hat{\mu}^{\mathrm{RSM}}$ is given in [36]:

$$\hat{\mu}_{n,s}^{\mathrm{RSM}}(\{W_i\}_{i=1}^{n}; \eta) := \inf_{\mu \in \mathcal{L}_s} \sup_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} \left|u^\top \mu - 1\mathrm{DRobust}(\{u^\top W_i\}_{i=1}^{n}, \eta/(6ep/s)^s)\right|, \tag{19}$$

where $W_1, \ldots, W_n \in \mathbb{R}^p$ are input data, $\mathcal{L}_s := \{v \in \mathbb{R}^p : \|v\|_0 \leq s\}$ is the set of s-sparse vectors in $\mathbb{R}^p$, $\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})$ is a $(1/2)$-cover of the set of $2s$-sparse unit vectors with cardinality $|\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})| \leq (6ep/s)^s$ [37], and 1DRobust is a univariate robust mean estimator defined in Algorithm 2 of [36]. Other univariate robust mean estimators can be considered in place of 1DRobust, including the median-of-means and trimmed mean variants [38]. Notably, the estimator in (19) achieves a near-optimal statistical guarantee for sparse mean estimation (Corollary 11 of [36]), despite its high computational complexity, which scales exponentially in s.

We describe our test $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$ using $\hat{\mu}_{n,s}^{\mathrm{RSM}}(\{W_i\}_{i=1}^n; \eta)$. For $\tilde{\Delta}_1$ specified in (21) and for $t \leq \tilde{\Delta}_1$, we use the non-robust statistic $A_{t,a}$ as defined in (12). For $t \in \mathcal{T} \cap \{t > \tilde{\Delta}_1\}$, we construct the statistic from the ℓ_2-norm of this robust sparse mean estimator:

$$A_t^{\mathrm{RSM}} := t\|\hat{\mu}_{t,s,\eta_t}^{\mathrm{RSM}}\|_2^2 = t\|\hat{\mu}_{t,s}^{\mathrm{RSM}}(\{Z_i\}_{i=1}^t; \eta_t)\|_2^2.$$

With all the parameters $a, \tilde{\Delta}_1, \tilde{r}_t, r_t^{\mathrm{RSM}}$ and η_t specified later in (21), we define

$$\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}} := \mathbb{1}_{\{\max_{t \in \mathcal{T} \cap \{t \leq \tilde{\Delta}_1\}} A_{t,a}/\tilde{r}_t > 1\}} \vee \mathbb{1}_{\{\max_{t \in \mathcal{T} \cap \{t > \tilde{\Delta}_1\}} A_t^{\mathrm{RSM}}/r_t^{\mathrm{RSM}} > 1\}}. \tag{20}$$

The theoretical guarantee of $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$ is established as follows.

Theorem 5. *Assume $\alpha \geq 4$. For any $\varepsilon \in (0,1)$, there exist $C_1, C_2, C_3, C_4, C_5 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$ defined in (20) with*

$$a = C_1\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log\tilde{\Delta}_1)\big), \quad \tilde{r}_t = C_2\Big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + \sqrt{s\log\tilde{\Delta}_1}\,\mathbb{1}_{\{t>1\}}\Big),$$

$$\eta_t = \exp\Big\{s\log(ep/s) - \frac{t \wedge \tilde{\Delta}_2}{C_3}\Big\}, \quad r_t^{\mathrm{RSM}} = C_4(t \wedge \tilde{\Delta}_2), \tag{21}$$

$$\tilde{\Delta}_1 = C_3\big(s\log(ep/s) + \log(16/\varepsilon)\big) \quad \text{and} \quad \tilde{\Delta}_2 = C_3\big(s\log(ep/s) + \log(16\log(2n)/\varepsilon)\big),$$

satisfies that

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_5 v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{U}}$, where

$$v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{U}} := s(p/s)^{2/\alpha} + \log\log(8n). \tag{22}$$

The main reason we separate $\mathcal{T}$ into different regions in our test (20) is that (19) cannot be applied when t is too close to the boundary in order to achieve the required statistical performance. We, therefore, resort to the non-robust testing statistics $A_{t,a}$ for t close to boundaries.

A significant drawback of using the robust sparse mean estimator to construct our test, $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$, is its high computational cost, which scales exponentially with s. We defer the discussion of this issue and a two-component remedy that achieves the same rate, which is in fact optimal, in polynomial time to Section III-C3.

C. Minimax optimal testing using a polynomial-time procedure

Similar to Section II-C, we first derive lower bounds on the minimax testing rate under $\mathcal{P}_{\alpha,K}$ and examine the optimality of our testing procedures in both the dense and sparse regimes in Section III-C1. Then, in Sections III-C2 and III-C3, we address the computational intractability issue of $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$ by combining it with a median-of-means-type test, yielding a procedure that is both minimax optimal in the sparse regime and computationally feasible, with complexity polynomial in p and n.

1) Minimax optimality: For any given $s \in [1, p]$, by simultaneously running $\phi_{\mathcal{P},\mathrm{dense}}$ and $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}}$ when $\alpha \geq 4$ and only running $\phi_{\mathcal{P},\mathrm{dense}}$ when $2 \leq \alpha < 4$, we obtain an upper bound

$$\begin{cases} \{s(p/s)^{2/\alpha} \wedge \sqrt{p}\log\log(8n)\} + \log\log(8n), & \text{when } \alpha \geq 4, \\ p^{2/\alpha}\log\log(8n), & \text{when } 2 \leq \alpha < 4, \end{cases} \tag{23}$$

on the minimax testing rate $v_{\mathcal{G}_{\alpha,K}^{\otimes}}^*(p,n,s)$. This upper bound also implies the one presented in (7), i.e.

$$\Big\{s(p/s)^{2/\alpha} \wedge p^{\frac{2}{\alpha} \vee \frac{1}{2}}\Big\}\log\log(8n).$$

The following result provides a corresponding lower bound.

Theorem 6. *Let $\alpha \geq 2$, $K \geq K_\alpha$ and $s \geq c$, for some absolute constant $c \geq 1$ and some constant $K_\alpha > 0$ depending only on α. There exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_{\mathcal{P}}(\rho) \geq 1/2$ whenever $\rho^2 \leq c'v_{\mathcal{P}}^{\mathrm{L}}$, where*

$$v_{\mathcal{P}}^{\mathrm{L}} := \Big\{s(p/s)^{2/\alpha} \wedge p^{(2/\alpha)\vee(1/2)}(\log\log(8n))^{\omega_2}\Big\} + \log\log(8n)$$

and $\omega_2 = (1/2)\mathbb{1}_{\{s > \sqrt{p\log\log(8n)}\} \cap \{\alpha \geq 4\}}$.

Similar to the exponentially-decaying tail case, by combining the lower bound $v_{\mathcal{P}}^{\mathrm{L}}$ from Theorem 6 with the upper bound in (23), we quantify the minimax test rate up to a factor of $\log\log(8n)$ and conclude that our testing procedures are minimax rate near-optimal. Importantly, our results reveal a critical phenomenon that the minimax testing rate is independent of s when $\alpha \in [2, 4]$. As in this case, both (23) and $v_{\mathcal{P}}^{\mathrm{L}}$ reduce to $p^{2/\alpha}$, up to a factor of $\log\log(8n)$. In other words, knowing the change

is sparse does not make the testing problem easier, and it is fundamentally impossible to exploit the sparse structure of the change in pursuit of better results. We further discuss its consequence in the language of sparse and dense regimes below.

To understand the effect of sparsity, we again ignore the iterated logarithmic factor in n and focus on the relationship between s and p. The boundary between the dense and sparse regimes is obtained by determining which term dominates the upper/lower bound rate in the minimum operator. As defined in Section I-A, the boundary is

$$s_{\mathcal{P}}^* := p^{\frac{1}{2} - (\frac{1}{\alpha - 2} \wedge \frac{1}{2})}, \tag{24}$$

which satisfies $s_{\mathcal{P}}^*(p/s_{\mathcal{P}}^*)^{\frac{2}{\alpha}} = p^{\frac{2}{\alpha} \vee \frac{1}{2}}$. The dense and sparse regimes are those where the sparsity level s is directly involved in the rate or not. The characterisation is summarised in Table II. Notably, when $2 \leq \alpha \leq 4$, we always have

$$\frac{1}{\alpha - 2} \wedge \frac{1}{2} = \frac{1}{2},$$

which means that there is no sparse regime in this extremely heavy-tailed setting.

TABLE II
BOUNDS ON THE MINIMAX TESTING RATES UNDER FINITE MOMENT NOISE DISTRIBUTION CLASS $\mathcal{P}_{\alpha,K}^{\otimes}$ WITH $\alpha \geq 2$, WHERE $\omega_2 = (1/2)\mathbb{1}_{\left\{s > \sqrt{p \log \log(8n)}\right\} \cap \{\alpha \geq 4\}}$.

		Upper bound	Lower bound
$\mathcal{P}_{\alpha,K}^{\otimes}$	Dense ($\alpha \geq 2$)	$p^{(2/\alpha) \vee (1/2)} \log \log(8n)$	$p^{(2/\alpha) \vee (1/2)} (\log \log(8n))^{\omega_2} + \log \log(8n)$
	Sparse ($\alpha \geq 4$)	$s(p/s)^{2/\alpha} + \log \log(8n)$	$s(p/s)^{2/\alpha} + \log \log(8n)$

From the table, we observe that our upper and lower bounds match exactly across the entire sparse regime ($s < s_{\mathcal{P}}^*$) when it exists ($\alpha \geq 4$). In the dense regime, the upper and lower bounds are off by a factor of order at most $\log \log(8n)$. We briefly note that the upper bound in the special case of $\alpha = 2$ can be improved to $p + \log \log(8n)$, matching the lower bound in this case up to constants. This is achieved as a by-product when we consider noise distributions with no more than two finite moments in Section VI-C.

We now discuss the effects of α on the minimax testing rates. When both the dense and sparse regimes exist ($\alpha \geq 4$), we observe from Table II that the dense rate is not affected by α since $2/\alpha \leq 1/2$. Moreover, even compared to the dense rate $\sqrt{p \log \log(8n)}$ under Gaussian noise assumptions, the cost of heavy-tailedness is minimal. However, the sparse rates are completely different from their counterparts in Section II-C, implying a significant increase of difficulty in detecting sparse changes under heavy-tailed noises. We note that this difficulty can be largely mitigated if one assumes that the change point is away from the boundary, and we discuss this interesting extension in Section VI-D. Finally, as α further decreases to between 2 and 4, the sparse regime becomes empty, and the dense rates are also affected by α.

2) A median-of-means-type test: As mentioned in Section III-B, one challenge of using a robust estimator such as (19) to construct our testing procedure $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ is its computational intractability. This stems from the need to project data onto every $2s$-sparse unit vector or its covering set, causing the computational complexity to scale exponentially in s. This issue is, in fact, common in high-dimensional robust statistics.

From Table II, we observe that $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ is used to establish the upper bound rate only in the sparse regime $s < s_{\mathcal{P}}^* = p^{\frac{1}{2} - (\frac{1}{\alpha-2} \wedge \frac{1}{2})}$, where it achieves minimax optimality. A natural question, then, is whether a polynomial-time algorithm can also attain minimax optimality. To answer this, we first examine a computationally-efficient test that combines the median-of-means approach with a hard-thresholding step for coordinate selection.

Recall that $Z_i = (X_i - X_{n-i+1})/\sqrt{2}$ for $i \in [n/2]$. For $t \in \mathcal{T} \backslash \{1\}$, we split $\{Z_1, \ldots, Z_t\}$ into two halves: $\{Z_1, Z_3, \ldots, Z_{t-1}\}$ and $\{Z_2, Z_4, \ldots, Z_t\}$. We further split the first set into G_t groups of equal size, denoted as $\mathscr{Z}_{t,1,1}, \mathscr{Z}_{t,2,1}, \ldots, \mathscr{Z}_{t,G_t,1}$, with the number of groups G_t specified later in (29), and use $\overline{Z}_{t,g,1}$ to denote the sample mean of the g-th group. The set $\{Z_2, Z_4, \ldots, Z_t\}$ is reserved for selecting the signal coordinates as we did in Section II-B. Consider the statistic $V_{t,g,a} \in \mathbb{R}^p$ with

$$V_{t,g,a}(j) := \left(\overline{Z}_{t,g,1}^2(j) - \frac{2G_t}{t}\right)\mathbb{1}_{\{|Y_{t,2}(j)| \geq a\}}, \quad j \in [p], \tag{25}$$

where $Y_{t,2}(j)$ is defined in (11) and a is a selection threshold to be specified in (29). Our test statistic takes the same form as in the dense case that

$$A_{t,a}^{\text{MoM}} := \frac{t}{2} \cdot \text{median}\left(\sum_{j=1}^{p} V_{t,1,a}(j), \sum_{j=1}^{p} V_{t,2,a}(j), \ldots, \sum_{j=1}^{p} V_{t,G_t,a}(j)\right). \tag{26}$$

For $t = 1$, we cannot perform sample-splitting and therefore we deal with it separately by considering

$$A_{1,a}^{\text{MoM}} := A_{1,a} = \sum_{j=1}^{p} (Z_1^2(j) - 1)\mathbb{1}_{\{|Z_1(j)| \geq a\}}. \tag{27}$$

Finally, the test is given by

$$\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}} A_{t,a}^{\text{MoM}}/r_t > 1\}}. \tag{28}$$

Proposition 7. *Assume* $\alpha \geq 4$. *For any* $\varepsilon \in (0,1)$, *there exist* $C_1, C_2, C_3 > 0$ *depending only on* α, K *and* ε, *such that* $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ *defined in* (28) *with*

$$a = C_1\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log(8n))\big), \quad r_t = C_2\big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + \sqrt{s}G_t\mathbb{1}_{\{t>1\}}\big),$$
$$G_t = (t \wedge \Delta)/2 \quad and \quad \Delta = 2^{4 + \lceil \log_2 \log\log(8n) \rceil}, \tag{29}$$

satisfies that

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_3 v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}}$, *where*

$$v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}} := s\big((p/s)^{2/\alpha} + \log\log(8n)\big).$$

We observe that the rate $v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}}$ is dominated by $v_{\mathcal{P},\text{sparse}}^{\text{U}} = s(p/s)^{2/\alpha} + \log\log(8n)$ in Theorem 5, meaning that using solely this MoM-type test $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ is not statistically optimal. However, the computational complexity of each step in constructing this test (defined in (25), (26), (27) and (28)) is polynomial in n and p, making it a feasible polynomial-time testing procedure.

It is also worth mentioning that in the hard-thresholding step, we simply use the non-robust quantity $Y_{t,2}$ to estimate the signal of each coordinate instead of its robust counterparts to avoid further complication of the procedure. If, however, we know that the change point is sufficiently far away from the endpoints, employing a robust procedure for coordinate selection can significantly improve the testing rate. This additional assumption requires modifying the alternative space in (2), leading to a different testing problem. A detailed discussion of this scenario is provided in Section VI-D.

3) An optimal polynomial-time test in the sparse regime: By comparing $v_{\mathcal{P},\text{sparse}}^{\text{U}}$ and $v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}}$, as established in Theorem 5 and Proposition 7, we observe that the improvement offered by $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ over $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ occurs only when $(p/s)^{2/\alpha} < \log\log(8n)$, in which case $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ attains the suboptimal rate $s\log\log(8n)$. Given the range of s in the sparse regime, we deduce that the computationally expensive $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ is only necessary when $p < \log^{\alpha-2}(\log(8n))$, whereas $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ can be used otherwise. We define the following combined testing procedure:

$$\phi_{\mathcal{P},\text{sparse}} := \begin{cases} \phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}, & \text{if } p < \log^{\alpha-2}(\log(8n)), \\ \phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}, & \text{otherwise.} \end{cases} \tag{30}$$

The corollary below confirms that this combined test runs in polynomial time in both n and p while achieving the optimal rate $v_{\mathcal{P},\text{sparse}}^{\text{U}}$ in the sparse regime.

Corollary 8. *Assume* $\alpha \geq 4$ *and* $s < s_{\mathcal{P}}^*$. *Consider the test* $\phi_{\mathcal{P},\text{sparse}}$ *defined in* (30), *with its two components* $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ *and* $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ *described in Sections III-B and III-C2, respectively. For any* $\varepsilon \in (0,1)$, *there exists a constant* $C > 0$ *depending only on* α, K, *and* ε, *such that* $\phi_{\mathcal{P},\text{sparse}}$, *with the parameters of its two components chosen according to* (21) *and* (29), *satisfies that*

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\text{sparse}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_1 v_{\mathcal{P},\text{sparse}}^{\text{U}}$, *with* $v_{\mathcal{P},\text{sparse}}^{\text{U}}$ *defined in* (22). *Moreover, the computational complexity of* $\phi_{\mathcal{P},\text{sparse}}$ *is polynomial in both* n *and* p.

We conclude this section by summarising in Table III the main features of each test constructed to achieve the upper bounds of the minimax testing rates in Tables I and II for $\mathcal{G}_{\alpha,K}$ and $\mathcal{P}_{\alpha,K}$ and under both dense and sparse signals.

TABLE III
A SUMMARY OF MAIN FEATURES OF EACH TEST CONSTRUCTED TO ACHIEVE THE UPPER BOUNDS OF THE MINIMAX TESTING RATES IN TABLES I AND II.

	Dense	Sparse
$\mathcal{G}_{\alpha,K}^{\otimes}$	CUSUM-type, ℓ_2 aggregation	CUSUM-type, thresholding, ℓ_2 aggregation
$\mathcal{P}_{\alpha,K}^{\otimes}$	MoM-type, ℓ_2 aggregation	Big p: MoM-type, thresholding, ℓ_2 aggregation Small p: Robust sparse mean estimator based test

When the noise has exponentially decaying tails, CUSUM-type statistics with ℓ_2 aggregation are sufficient to achieve near-optimal testing. In contrast, for polynomially decaying tails, robust methods such as median-of-means or robust sparse mean estimation are required to construct near-optimal tests. For both types of heavy-tailed noise, when testing sparse signals, a thresholding step is applied to identify the signal coordinates before aggregation. Finally, under finite-moment noise distributions, the two robust methods, chosen according to whether p is large or small, combine to yield an optimal and computationally efficient testing procedure.

IV. ADAPTATION TO SPARSITY

In Sections II and III, we have studied the change point testing problem under two types of heavy-tail assumptions on the error distributions: (1) exponentially-decaying/sub-Weibull tails and (2) polynomially-decaying/finite α-th moment assumption with $\alpha \geq 2$. The corresponding upper bound rates, e.g. $v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}$ and $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$, are currently achieved by testing procedures that take the sparsity level s as an input. In this section, we study the adaptation of these procedures to unknown sparsity levels.

First off, in the very heavy-tailed setting, i.e. each entry of E has only finite α-th moments for $\alpha \in [2, 4]$, there is no sparse regime; see the discussion in Section III-C1. The test $\phi_{\mathcal{P},\mathrm{dense}}$ defined in (17) with its parameters specified in Theorem 4 does not require the knowledge of the sparsity, and, therefore, the corresponding rate $v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}} = p^{2/\alpha} \log\log(8n)$ can already be achieved by an adaptive procedure.

We focus on the case where $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$ for $\alpha > 4$. Recall from Theorem 4 and Corollary 8 that the tests $\phi_{\mathcal{P},\mathrm{dense}}$ and $\phi_{\mathcal{P},\mathrm{sparse}}$ achieve the rates $v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}}$ and $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$, respectively, when the sparsity level is known. To address the scenario where sparsity is unknown, we propose the following adaptive testing procedure that combines these two tests:

$$\phi_{\mathcal{P},\mathrm{adaptive}} := \phi_{\mathcal{P},\mathrm{dense}} \vee \max_{s \in \mathcal{K}} \phi_{\mathcal{P},\mathrm{sparse},s}, \tag{31}$$

where the dependence of $\phi_{\mathcal{P},\mathrm{sparse}}$ on s is made explicit by writing it as $\phi_{\mathcal{P},\mathrm{sparse},s}$, and the set $\mathcal{K} := \{1, 2, 4, \ldots, 2^{\lceil \log_2(p) \rceil - 1}\}$ is a dyadic grid. The details of this test along with its parameter choices are provided in Appendix B1.

Theorem 9. *Assume $\alpha \geq 4$. For any $\varepsilon \in (0, 1)$, there exists a test $\phi_{\mathcal{P},\mathrm{adaptive}}$ of the form (31) that satisfies*

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\mathrm{adaptive}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C\big(v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}} \wedge v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}\big)$, with $C > 0$ being a constant depending only on α, K and ε.

Theorem 9 establishes that the adaptive test $\phi_{\mathcal{P},\mathrm{adaptive}}$ achieves the same rate as (23) without requiring the sparsity parameter as an input. Therefore, the discussion on the optimality in Section III-C1 also holds. When the errors have exponentially-decaying tails instead, a similar adaptive testing procedure $\phi_{\mathcal{G},\mathrm{adaptive}}$ can be constructed using (31) with $\phi_{\mathcal{G},\mathrm{dense}}$ and $\phi_{\mathcal{G},\mathrm{sparse}}$ instead and achieve the same rate as in (13). For brevity, we omit further details here.

V. ILLUSTRATIVE SIMULATIONS

In this section, we illustrate our main results, as presented in Section I-A via simulation studies. Specifically, we numerically verify the shrinkage of sparse regimes as the tail becomes heavier in both exponentially- and polynomially-decaying tail cases.

In our simulations, we set $p = 100$ and $n = 300$. The data matrices $X \in \mathbb{R}^{100 \times 300}$ are generated according to (1), where the entries of the noise matrix E are i.i.d. following some distribution F. Let t_k denotes the t-distribution with k degrees of freedom. We consider the following four choices for the distribution F:

- GG(2): standard normal distribution, which belongs to the class $\mathcal{G}_{2,K}$ for some K;
- GG(0.5): Generalised Gaussian distribution with mean 0, shape parameter 0.5, and scale parameter chosen so that the variance is 1; this belongs to the class $\mathcal{G}_{0.5,K}$ for some K;
- Nt(8): $\sqrt{3/4} \cdot t_8$, which has variance 1 and belongs to the class $\mathcal{P}_{8-\epsilon,K}$ for any $\epsilon > 0$[2];
- Nt(3): $\sqrt{1/3} \cdot t_3$, which has variance 1 and belongs to the class $\mathcal{P}_{3-\epsilon,K}$ for any $\epsilon > 0$.

We adopt the adaptive tests $\phi_{\mathcal{P},\mathrm{adaptive}}$ defined in (31), and $\phi_{\mathcal{G},\mathrm{adaptive}}$, the corresponding version for the class $\mathcal{G}_{\alpha,K}$. To calibrate the constants in our tests, we set θ to be the matrix with all entries being zero. In both the exponentially- and polynomially-decaying tail cases, the adaptive test consists of a dense test and a collection of sparse tests. For the dense test $\phi_{\mathcal{G},\mathrm{dense}}$ (or $\phi_{\mathcal{P},\mathrm{dense}}$), we calibrate the constant in the detection threshold r (or r_t) so that the empirical Type I error is below 0.025. For the sparse tests, we specify two constants that appear in a and r (or r_t), used in the thresholding and detection steps, respectively. We select these constants so that the empirical Type I error is below 0.025. Taken together, this ensures that the overall Type I error of the adaptive test is controlled at 0.05.

The power of our adaptive test is evaluated empirically on a grid of $(s, \mathtt{Signal})$ values, where $s \in \{1, \ldots, 50\}$ and $\mathtt{Signal}$ varies over a fine grid within a suitable interval, which may differ across the tail cases. For each fixed $(s, \mathtt{Signal})$ pair, the mean matrix $\theta = (\theta_1, \ldots, \theta_{300}) \in \mathbb{R}^{100 \times 300}$ is generated as follows. The change point location t_0 is chosen uniformly at random from $\{1, \ldots, 150\}$, and we set

$$\theta_t = 0 \in \mathbb{R}^{100}, \quad 1 \leq t \leq t_0,$$

and

$$\theta_t = \sqrt{\frac{n}{t_0(n - t_0)}} \frac{\mathtt{Signal}}{\sqrt{s}} (\underbrace{1, \ldots, 1}_{s}, \underbrace{0, \ldots, 0}_{p-s})^{\top}, \quad t_0 + 1 \leq t \leq n.$$

---

[2]A t distribution with ν degrees of freedom has $(\nu - \epsilon)$-th moment finite for any $\epsilon > 0$.

The power for each pair $(s, \mathtt{Signal})$ is calculated by repeating the process 500 times of generating θ, E, and thus X, and computing $\phi_{\mathcal{G},\mathrm{adaptive}}$ (or $\phi_{\mathcal{P},\mathrm{adaptive}}$) with the constants calibrated as described above. We plot two heatmaps showing the power of $\phi_{\mathcal{G},\mathrm{adaptive}}$, with $\mathtt{Signal}$ on the horizontal axis and sparsity s on the vertical axis, for GG(2) and GG(0.5), respectively, in Figure 2. Similarly, heatmaps of the power of $\phi_{\mathcal{P},\mathrm{adaptive}}$ for Nt(8) and Nt(3), respectively, are shown in Figure 3.
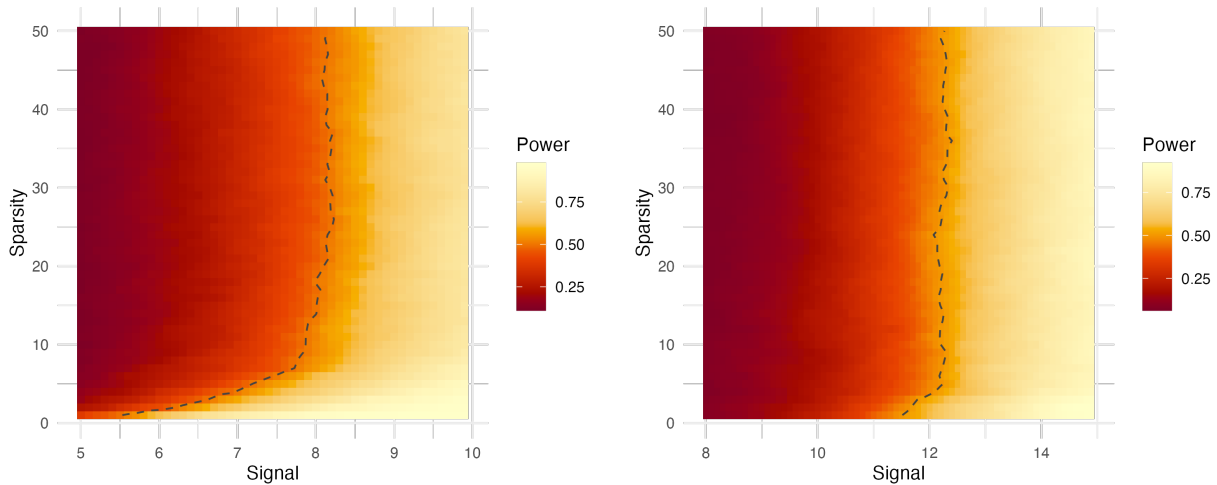

Fig. 2. Simulation results for noise GG(2) (left panel) and GG(0.5) (right panel). The dashed line represents the 0.5 contour of the power.
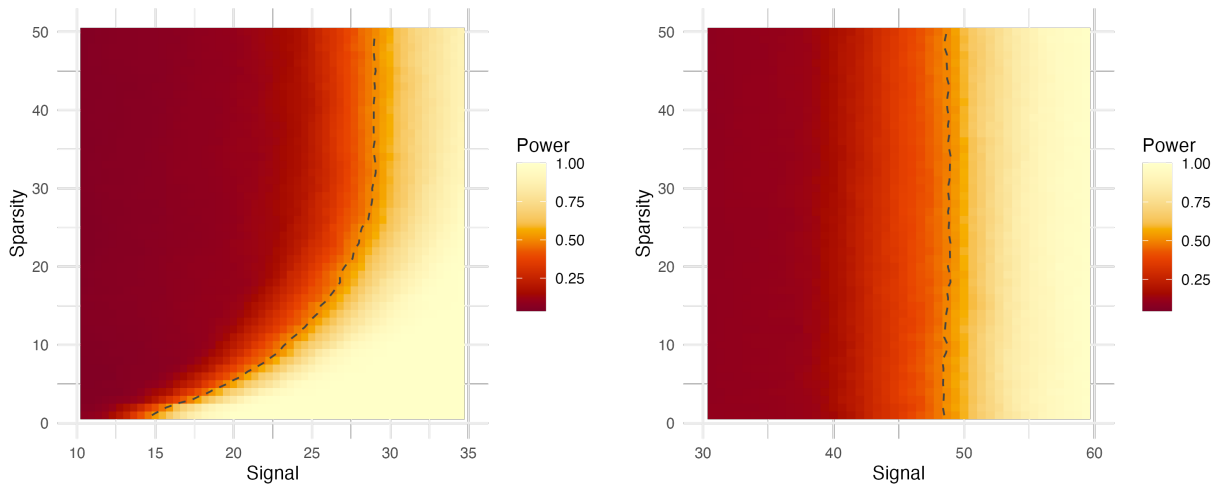

Fig. 3. Simulation results for noise Nt(8) (left panel) and Nt(3) (right panel). The dashed line represents the 0.5 contour of the power.

In both Figures 2 and 3, we observe that as the tails become heavier (moving from the left panel to the right panel), the power generally decreases, as indicated by the need for higher signal levels to achieve the same power. Moreover, the power for detecting sparse changes is much more affected than for dense changes as the tails become heavier. In Figure 2, when the noise is $\mathrm{GG}(2)$, there exists a range of sparsity levels where higher power can be achieved compared to dense changes at the same signal level. However, when the noise is $\mathrm{GG}(0.5)$, this advantage for sparse changes diminishes significantly. A similar pattern is seen in Figure 3. Moreover, as suggested by our theoretical results and now verified here, when the noise distribution has fewer than four finite moments, there is no sparsity level for which testing sparse changes is easier than testing dense changes.

VI. EXTENSIONS

In the previous sections, we characterised the minimax rates for testing a single mean change within independent high-dimensional data streams under two broad classes of heavy-tailed noise distributions. In this section, we consider several extensions, including (1) testing against multiple change points in Section VI-A, (2) accounting for temporal dependence among observations in Section VI-B, (3) addressing the case where noise matrix entries have fewer than two finite moments

in Section VI-C, and (4) examining a modified testing problem where the potential change point is known to be some distance away from the endpoints in Section VI-D.

A summary of the problem setups and notation is provided in Table IV. Specifically, within each section, we propose testing procedures ϕ and control their worst case testing error, which takes the form

$$\mathcal{R}_{\mathcal{Q},\#}(\rho,\phi) := \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in H_0} \mathbb{E}_{\theta,P_e}(\phi) + \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in H_1} \mathbb{E}_{\theta,P_e}(1-\phi),$$

where $\#$ denotes the problem-dependent quantities and ρ is some measure of change strength defined clearly in each alternative parameter space. Lower bounds on the minimax testing rates will be provided where available by considering $\inf_\phi \mathcal{R}_{\mathcal{Q},\#}(\rho,\phi)$ and, for brevity, we do not separately define the minimax testing rates for each problem.

TABLE IV
SUMMARY OF THE DIFFERENT SETTINGS IN SECTION VI.

	Alternative hypothesis H_1	Noise distribution $\mathcal{Q}$	Notation
Section VI-A	$\theta \in \Theta_{\mathrm{multi}}(p,n,\rho)$ (32)	$\mathcal{P}_{\alpha,K}^{\otimes}$	$\mathcal{R}_{\mathcal{P},\mathrm{multi}}$
Section VI-B	$\theta \in \Theta(p,n,p,\rho)$ (2)	$\mathcal{P}_{\mathrm{Temp}}$ (Definition 4)	$\mathcal{R}_{\mathcal{P}_{\mathrm{Temp}}}$
Section VI-C	$\theta \in \bigcup_{t_0=1}^{n-1} \Theta(p,n,\rho_{t_0})$ (37)	$\mathcal{W}_{\alpha}^{\otimes}$ (Definition 5)	$\mathcal{R}_{\mathcal{W}_{\alpha}^{\otimes}}$
Section VI-D	$\theta \in \Theta_{\mathrm{res}}(p,n,s,\rho)$ (40)	$\mathcal{P}_{\alpha,K}^{\otimes}$ & $\mathcal{G}_{\alpha,K}^{\otimes}$	$\mathcal{R}_{\mathcal{Q},\mathrm{res}}$

A. Testing against multiple change points

We consider the following testing problem between no change point and at least one change point:

$$H_0 : \theta \in \Theta_0(p,n) \quad \text{vs.} \quad H_1 : \theta \in \Theta_{\mathrm{multi}}(p,n,\rho) := \bigcup_{k \in \mathbb{Z}^+} \bigcup_{1 \leq \tau_1 < \ldots < \tau_k \leq n-1} \Theta^{(\tau_1,\ldots,\tau_k)}(p,n,\rho), \tag{32}$$

where $\Theta_0(p,n)$ is the same as in (3) and

$$\Theta^{(\tau_1,\ldots,\tau_k)}(p,n,\rho) := \Big\{ \theta : \theta_{\tau_{i-1}+1} = \ldots = \theta_{\tau_i} = \mu_i \text{ for all } i \in [k+1], \text{ for some } \mu_1,\ldots,\mu_{k+1} \in \mathbb{R}^p,$$

$$\text{where there exists } i \in [k] \text{ s.t. } \Delta_i \geq 4\log(n) \text{ and } \kappa_i^2 \Delta_i \geq \rho^2, \text{ with}$$

$$\Delta_i := \min\{\tau_i - \tau_{i-1}, \tau_{i+1} - \tau_i\}, \ \kappa_i := \|\mu_{i+1} - \mu_i\|_2 \text{ for } i \in [k],$$

$$\tau_0 = 0 \text{ and } \tau_{k+1} = n \text{ by convention} \Big\}.$$

The worst case testing error for a measurable test function ϕ is

$$\mathcal{R}_{\mathcal{Q},\mathrm{multi}}(\rho,\phi) := \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_{\theta,P_e}(\phi) + \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta_{\mathrm{multi}}(p,n,\rho)} \mathbb{E}_{\theta,P_e}(1-\phi).$$

Throughout this subsection, we consider the class of distributions $\mathcal{Q} = \mathcal{P}_{\alpha,K}^{\otimes}$ with $\alpha \geq 4$ and $K < \infty$.

Theorem 10. *Assume $n \geq 50$ and $\alpha \geq 4$. For any $\varepsilon \in (0,1)$, there exists a test $\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}$ that satisfies*

$$\mathcal{R}_{\mathcal{P},\mathrm{multi}}(\rho,\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}) \leq \varepsilon,$$

as long as

$$\rho^2 \geq C\sqrt{p}\log(n),$$

where $C > 0$ is a constant depending only on α, K and ε.

The details of the test $\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}$ is described in Appendix D1. At a high level, it exhibits a multi-scale nature by performing a collection of 'local tests' at all possible change locations with a dyadic grid of scales. Such multi-scale statistics are commonly used for detecting multiple change points [29], [39]. The robustness property is achieved by integrating the median-of-means methodology into the local tests. Another ideal feature of our test is that it is adaptive to the unknown number of change points k under the alternative hypothesis, i.e. it does not take k as an input.

In contrast to the single change point setting where we allow the potential change point to be arbitrarily close to the endpoints, here, we impose a mild minimum spacing condition $\Delta_i \geq 4\log(n)$ for some change point. However, this condition is not required for testing under Gaussian or sub-Gaussian assumptions [29]. In Appendix D1, we provide a heuristic example

illustrating why the heavy-tailed nature of the data necessitates a minimum spacing condition to achieve a rate that has logarithmic dependence on n.

The following proposition shows that the test $\phi_{\mathcal{P},\text{multi}}^{\text{MoM}}$ is minimax rate-optimal up to a factor of $\sqrt{\log(n)}$ when $\alpha \geq 4$.

Proposition 11. *Let $n \geq 72$, $\alpha \geq 4$ and $K \geq \sqrt{\alpha+1}$. Then it holds that $\inf_\phi \mathcal{R}_{\mathcal{P},\text{multi}}(\rho, \phi) \geq 1/2$ whenever*

$$\rho^2 \leq c\left(\sqrt{p\log(n)} + \log(n)\right),$$

for some absolute constant $c > 0$.

By comparing Theorem 10 and Proposition 11 for testing against multiple change points, with Theorems 4 and 6 for the single change point problem in Section III, we observe that the main difference is the presence of $\log(n)$ instead of $\log\log(n)$ in the rates. This distinction has been observed in the change point literature for univariate data under sub-Gaussian noise distributions [6]. Our results reaffirm this phenomenon even in the context of high dimensional data with heavy-tailed distributions.

B. Temporal dependence

We now consider temporal dependence in heavy-tailed observations. To isolate its impact, we focus on the case where $s = p$ (eliminating the effect of sparsity) and assume that the error matrix E has independent rows.

For simplicity, assume $n \geq 4$ to be an even number. The temporal dependence assumption to be introduced involves the interlaced α-mixing coefficient of our noise sequence $\{E_i\}_{i \in [n]}$:

$$\alpha^*(i) := \sup_{\substack{S,T \subseteq [n]: \\ \min_{s \in S, t \in T} |s-t| \geq i}} \sup_{\substack{A \in \sigma(E_j : j \in S), \\ B \in \sigma(E_j : j \in T)}} \left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right|. \tag{33}$$

Throughout this subsection, we consider the class of distributions $\mathcal{Q} = \mathcal{P}_{\text{Temp}}$, defined as follows:

Definition 4 ($\mathcal{P}_{\text{Temp}}$ class of distributions). *For $K > 0$, $\alpha \geq 2$ and $c_1, c_2 > 0$, let $\mathcal{P}_{\text{Temp}}$ denote the class of all distributions of the error matrix E satisfying the following conditions:*
1) *the matrix E has independent rows;*
2) *the marginal distribution of each entry in E belongs to $\mathcal{P}_{\alpha,K}$; and*
3) *the interlaced α-mixing coefficient satisfies that*

$$\alpha^*(i) \leq c_1 e^{-c_2 i}, \quad \text{for all } i \in [n-1]. \tag{34}$$

Theorem 12. *Assume $\alpha > 4$. For any $\varepsilon \in (0,1)$, there exists a test $\phi_{\mathcal{P}_{\text{Temp}}}$ that satisfies*

$$\mathcal{R}_{\mathcal{P}_{\text{Temp}}}(\rho, \phi_{\mathcal{P}_{\text{Temp}}}) \leq \varepsilon,$$

as long as

$$\rho^2 \geq C\, p^{1/2}\{\log\log(8n)\}\{\log\log\log(64n)\}^2,$$

where C is a constant depending only on α, K, c_1, c_2 and ε.

The test is essentially constructed in the same way as in Section III-A and we leave the detailed description to Appendix D2. We also note that a similar result can be obtained for the case of $2 < \alpha \leq 4$; see Remark 1 in Appendix D2. By comparing Theorem 12 with Theorem 4, we conclude that under certain weak dependence settings, described by the class $\mathcal{P}_{\text{Temp}}$, we can still achieve nearly the same rate as in the independence setting, up to a factor of squared triple logarithm.

To further interpret the main temporal dependence condition (34) required for $\mathcal{P}_{\text{Temp}}$, we note that (34) on the interlaced α-mixing coefficient (33) of the noise sequence is stricter than imposing the same condition on the (usual) α-mixing coefficient. The reason we choose this interlaced version of the coefficient is that our technique involves pairing observations (first with last, second with second last, etc.) to form Z_i's before grouping and averaging.

A sufficient condition for (34) is shown in Theorem 5.1(b) and Equation (1.12) of [40]. If each component series $\{E_i(j)\}_{i \in [n]}$ of the noise sequence satisfies

$$\sup_{\substack{S,T \subseteq [n]: \\ \min_{s \in S, t \in T} |s-t| \geq i}} \sup_{\substack{f \in L^2(\sigma(E_k(j):k \in S)) \\ g \in L^2(\sigma(E_k(j):k \in T))}} \text{Corr}(f,g) \leq c_1 e^{-c_2 i}, \tag{35}$$

then the α^*-mixing rate condition (34) is satisfied with constants c_1 and c_2. This fact implies that moving average processes essentially belong to the class $\mathcal{P}_{\text{Temp}}$. Specifically, consider a data-generating mechanism in which $\{E_i(j)\}_{i \in [n]}$, the j-th component series of the noise, is a moving average process of order q_j, $j \in [p]$. Assuming that there exists a constant $q_{\max}$ such that $\max_{j \in [p]} q_j \leq q_{\max}$, each univariate component series satisfies (35) with $c_1 = e^{q_{\max}}$ and $c_2 = 1$. We present a more detailed analysis of an MA(1) example in Appendix D2. For further discussion and examples satisfying (34) or (35), see [41].

C. Fewer than two finite moments

In Section III, we studied the minimax testing rate when $P_e \in \mathcal{P}_{\alpha,K}^{\otimes}$, for some constant $\alpha \geq 2$ and $K < \infty$. Now, we tackle the even more challenging case when each entry of the noise matrix does not necessarily have a finite variance. Given our main findings that when $\alpha \in [2,4)$, there is no sparse regime (c.f. Sections I-A and III-C), we do not need to consider sparse changes in this even more challenging setting. We first introduce the class of distributions $\mathcal{W}_\alpha$ that we work with.

Definition 5. *For $1 \leq \alpha \leq 2$, let $\mathcal{W}_\alpha$ denote the class of distributions on $\mathbb{R}^p$ such that for any $P \in \mathcal{W}_\alpha$ and random variable $W \sim P$, it holds that*

$$\mathbb{E}W = 0 \quad and \quad \mathbb{E}|\langle W, v \rangle|^\alpha \leq 1, \quad \forall v \in \mathbb{R}^p, \|v\|_2 = 1. \tag{36}$$

Let $\mathcal{W}_\alpha^{\otimes}$ denote the class of joint distributions of all entries in the error matrix E, where each column of E independently follows a distribution on $\mathbb{R}^p$ that belongs to the class $\mathcal{W}_\alpha$. It follows by Jensen's inequality that $\mathcal{P}_{2,1}^{\otimes} \subseteq \mathcal{W}_\alpha^{\otimes} \subseteq \mathcal{W}_{\alpha'}^{\otimes}$ for all $1 \leq \alpha' \leq \alpha \leq 2$.

We now specify the alternative parameter space. For a given sequence of $\{\rho_{t_0}\}_{t_0 \in [n-1]}$, let

$$\mathrm{H}_1 : \theta \in \bigcup_{t_0=1}^{n-1} \Theta(p, n, \rho_{t_0}),$$

where

$$\Theta(p, n, \rho_{t_0}) := \left\{ \theta : \theta_t = \mu_1 \text{ for } t = 1, \ldots, t_0, \ \theta_t = \mu_2 \text{ for } t = t_0 + 1, \ldots, n, \right.$$

$$\left. \text{for some } \mu_1, \mu_2 \in \mathbb{R}^p \text{ s.t. } \|\mu_1 - \mu_2\|_2^2 \geq \rho_{t_0}^2 \right\}. \tag{37}$$

Letting $\rho_{t_0} = \rho \sqrt{t_0^{-1}(n - t_0)^{-1}n}$, we see that the alternative parameter space (37) is a superset of $\Theta^{(t_0)}(p, n, p, \rho)$ defined in (4). When the noise variances are finite, the factor $t_0^{-1}(n - t_0)^{-1}n$ is the variance of the natural test statistic $t_0^{-1} \sum_{i=1}^{t_0} X_i - (n - t_0)^{-1} \sum_{i=t_0+1}^{n} X_i$, when $p = 1$. Therefore, the normalising factor $t_0(n - t_0)n^{-1}$ in (4) ensures a variance-stable signal strength parameter ρ across different change locations. In the absence of finite variance assumptions, as considered in this section, we opt out for such a specific normalising factor and use the general form ρ_{t_0}.

For a given sequence $\{\rho_{t_0}\}_{t_0 \in [n-1]}$, the worst case testing error for a given test ϕ is

$$\mathcal{R}_{\mathcal{W}_\alpha^{\otimes}}(\{\rho_{t_0}\}_{t_0 \in [n-1]}, \phi) := \sup_{P_e \in \mathcal{W}_\alpha^{\otimes}} \sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_{\theta, P_e}(\phi) + \sup_{P_e \in \mathcal{W}_\alpha^{\otimes}} \sup_{\theta \in \bigcup_{t_0=1}^{n-1} \Theta(p,n,\rho_{t_0})} \mathbb{E}_{\theta, P_e}(1 - \phi).$$

To establish an upper bound on the minimax testing rates, we develop a test $\phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}$, which has two components that separately target at whether the potential change point is sufficiently away from the boundary or not. When the potential change is away from the boundary, our test utilises on a robust mean estimator $\hat{\mu}^{\mathrm{RM}}$ from Algorithm 1-7 of [42]. Similar to the previous subsections in Section VI, we directly present the result, with the details of the test $\phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}$ deferred to Appendix D3.

Theorem 13. *Assume $1 \leq \alpha \leq 2$. For any $\varepsilon \in (0,1)$, if*

$$\rho_{t_0} \geq C\left\{ \sqrt{\frac{p}{t_0 \wedge (n - t_0)}} + \left(\frac{p}{t_0 \wedge (n - t_0)}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log\log(n)}{t_0 \wedge (n - t_0)}\right)^{\frac{\alpha-1}{\alpha}} \right\}, \tag{38}$$

for all $t_0 \in [n-1]$, where $C > 0$ is some constant that depends only on α and ε, then there exists a test $\phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}$ such that

$$\mathcal{R}_{\mathcal{W}_\alpha^{\otimes}}(\{\rho_{t_0}\}_{t_0 \in [n-1]}, \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}) \leq \varepsilon.$$

To interpret the requirement on ρ_{t_0} and compare it with the cases where at least two finite moments exist, we first note that for any $t_0 \in [n-1]$,

$$\frac{(n - t_0) \wedge t_0}{2} \leq \frac{t_0(n - t_0)}{n} \leq (n - t_0) \wedge t_0,$$

i.e. the normalising factor in (4) is of the same order as $(n - t_0) \wedge t_0$. Since $\|\mu_1 - \mu_2\|_2^2 \geq \rho_{t_0}^2$ under (37), we can rewrite (38) as

$$\frac{t_0(n - t_0)}{n}\|\mu_1 - \mu_2\|_2^2 \gtrsim p + p^{\frac{2\alpha-2}{\alpha}}\{t_0 \wedge (n - t_0)\}^{\frac{2-\alpha}{\alpha}} + \log^{\frac{2\alpha-2}{\alpha}}\log(n)\{t_0 \wedge (n - t_0)\}^{\frac{2-\alpha}{\alpha}}. \tag{39}$$

We observe that in the special case when $\alpha = 2$, the right hand side of (39) reduces to $p + \log\log(n)$. Theorem 13 shows that when the noise distribution belongs to $\mathcal{W}_2^{\otimes}$, the worst case testing error is controlled at ε, under this condition on the normalised signal strength. Since $\mathcal{P}_{2,1}^{\otimes} \subseteq \mathcal{W}_2^{\otimes}$, we conclude that the minimax testing rate under $\mathcal{P}_{2,1}^{\otimes}$ satisfies $v_{\mathcal{P}_{2,1}^{\otimes}}^*(p, n, s) \lesssim p + \log\log(8n)$ for all $s \in [p]$, which improves upon the result in Theorem 4 in the special case $\alpha = 2$. Combining this with Theorem 6 leads to an exact characterisation of the minimax testing rate under $\mathcal{P}_{2,1}^{\otimes}$, i.e. $v_{\mathcal{P}_{2,1}^{\otimes}}^*(p, n, s) \asymp p + \log\log(8n)$.

Investigating optimality under the more general alternative hypothesis (37) is conceptually more challenging, as it actually necessitates lower bound constructions that are valid at every potential change location t_0 in order to derive the optimal form of $\{\rho_{t_0}\}_{t_0 \in [n-1]}$. When $p = 1$, we show in Proposition 26 in Appendix D3 that $\inf_\phi \mathcal{R}_{\mathcal{W}_\alpha^\otimes}(\{\rho_{t_0}\}_{t_0 \in [n-1]}, \phi) \geq 1/2$, if

$$\rho_{t_0} \leq \{t_0 \wedge (n - t_0)\}^{-(\frac{\alpha-1}{\alpha})},$$

for some $t_0 \in [n-1]$, which confirms the optimality of our upper bound (38) up to logarithmic factors. Establishing optimality in general dimension p is significantly more challenging and we leave that for future investigation.

D. Change away from boundary

In this final extension, we aim to address whether it is possible to achieve the sub-Gaussian rate in the sparse regime, $s \log(ep/s) + \log \log(8n)$ (c.f. Table I with $\alpha = 2$), when it is known *a priori* that the change location is at least some distance away from the endpoints, 1 and n. We denote this minimum requirement on the distance from boundary by $t^{\mathrm{res}} \geq 1$, and modify the alternative space in (4) to include additional restrictions (we use sub/superscripts 'res' to denote this):

$$\mathrm{H}_1 : \theta \in \Theta_{\mathrm{res}}(p, n, s, \rho) := \bigcup_{t_0 = t^{\mathrm{res}}+1}^{n-t^{\mathrm{res}}} \Theta^{(t_0)}(p, n, s, \rho). \tag{40}$$

We separately consider the cases $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \geq 4$ and $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$ for $0 < \alpha \leq 2$. Notice that in both cases, the rates in the dense regime are not affected by the level of heavy-tailedness α, and they differ from the sub-Gaussian dense rate by at most a factor of $\log \log(n)$; see Tables I and II. Therefore, we are interested in whether the same phenomenon can happen in the sparse regime by considering a 'simpler' problem with a restricted alternative parameter space specified in (40). For a given test ϕ, we denote the worst case testing error, with (40) as the alternative hypothesis, by $\mathcal{R}_{\mathcal{Q},\mathrm{res}}(\rho, \phi)$.

For $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \geq 4$, we modify the median-of-means-type test proposed in Section III-C2 and provide the details of the new test $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$ in Appendix D4. Our new test relies on a more robust coordinate selection step compared to (25), which is made possible due to the additional assumption that the change location is away from boundary.

Theorem 14. *Let* $t^{\mathrm{res}} = 32\{\log(e^2 p/s) + s^{-1} \log \log(8n)\}$ *and* $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$ *with* $\alpha \geq 4$. *For any* $\varepsilon \in (0, 1)$, *there exists a test* $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$ *such that*

$$\mathcal{R}_{\mathcal{Q},\mathrm{res}}(\rho, \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$, *where*

$$v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}} := s\{\log(ep/s) + \log \log(8n)\},$$

and C *is a constant that only depends on* α, K *and* ε.

The rate in Theorem 14 offers a significant improvement over the original sparse rate, $s\{(p/s)^{2/\alpha} + \log \log(8n)\}$, achieved in Proposition 7. In particular, the rate $v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$ depends on the dimension p only through a logarithmic factor and is independent of α—the number of finite moments assumed for the noise variables. We emphasise that this improvement is achieved by restricting the original alternative parameter space in (4) to (40) with $t^{\mathrm{res}} = 32\{\log(e^2 p/s) + s^{-1} \log \log(8n)\}$, i.e. we only consider testing a potential change point that is not too close to the boundary. Finally, we observe that when $\log(ep/s) \geq \log \log(8n)$, with a very mild condition imposed by t^{res} that the change is away from the endpoints by at least the order of $\log(ep/s)$, the rate $v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$ matches the sparse sub-Gaussian rate for all $\alpha \geq 4$.

For $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$ with $0 < \alpha < 2$, a similar result can be shown with the test construction deferred to Appendix D4.

Theorem 15. *Let* $t^{\mathrm{res}} = 32\{\log(e^2 p/s) + s^{-1} \log \log(8n)\}$ *and* $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$ *with* $0 < \alpha < 2$. *For any* $\varepsilon \in (0, 1)$, *there exists a test* $\phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}}$ *such that*

$$\mathcal{R}_{\mathcal{Q},\mathrm{res}}(\rho, \phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C v_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}}$, *where*

$$v_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}} := s \log(ep/s) + \log \log(8n),$$

and C *is a constant that only depends on* α, K *and* ε.

Theorem 15 parallels Theorem 14 and demonstrates that $\phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}}$ achieves the sparse sub-Gaussian rate when testing against change that is at least the order of $\log(ep/s) + s^{-1} \log \log(8n)$ away from the boundary. We note that this idea of obtaining sub-Gaussian rates by assuming change is away from the boundary is also briefly explored in [7]. Compared to their result, our result offers a significant improvement on the requirement of how far a change needs to be away from the endpoints, in order to achieve sub-Gaussian rates; see Appendix D4 for detailed discussions. More generally, it would be interesting, from a lower bound perspective, to understand the smallest order of t^{res} such that it is possible to achieve sub-Gaussian rates. We leave this as an interesting future research direction.

VII. Discussion

In this paper, we have studied the problem of testing against a single mean change point for high-dimensional heavy-tailed data. We have characterised the minimax testing rates of this problem up to iterated logarithm factors in both the exponentially-decaying and polynomially-decaying tail cases. The transition boundary between the sparse and dense regimes occurs at $s_{\mathcal{G}}^* \asymp \sqrt{p} \log^{-2/\alpha}(ep)$ for $\mathcal{G}_{\alpha,K}$ with $\alpha \in (0,2]$. For $\mathcal{P}_{\alpha,K}$, the transition happens at $s_{\mathcal{P}}^* = p^{1/2-1/(\alpha-2)}$ when $\alpha \geq 4$ and there is no sparse regime when $\alpha \in [2,4)$. Our results also quantify the costs of heavy-tailed distributions in this problem by comparing to the previous results under Gaussian error assumption [2] and unveil a new phenomenon that the minimax testing rates of mean change point problem undergo a phase transition when the error distribution has finite fourth moment. There are several avenues for future research and we briefly discuss them below.

Spatial dependence. Throughout this paper, we have assumed independence across coordinates. To relax this independence assumption, one could allow for weak or strong coordinate-wise dependence via ρ-mixing, and employ alternative finite-sample analysis tools, as in, for example, [8]. Alternatively, for allowing a general covariance matrix Σ, if we assume that $\Sigma^{-1/2}E$ has independent components with all eigenvalues of Σ being of constant order, then at least in the dense case, all our theoretical results remain valid. We leave a thorough investigation into these two generalisations for future endeavours.

Adaptation to α. All of our proposed testing procedures require the knowledge of α, the tail decay index in the case of $\mathcal{G}_{\alpha,K}$ and the number of finite moments in the case of $\mathcal{P}_{\alpha,K}$, through the choices of parameters. Note that if we under-specify α, all of our theoretical guarantees still hold, albeit non-optimal rates achieved by the procedures. On the other hand, an over-specification of α would invalidate our results. In practice, practitioners, based on domain knowledge, usually have a conservative idea on how heavy the tails may be. There have been some recent works on distinguishing between exponentially-decaying and polynomially-decaying tails, e.g. [43], [44], and on estimating the tail index parameter for sub-Weibull distributions [45], which may be combined with our tests to obtain adaptivity. We leave this ambitious task for the future.

Acknowledgments

Appendix

The proofs of all theoretical results are presented in the Appendix. Appendix A contains proofs of upper bound results, including Theorem 1, Theorem 2, Theorem 4, Theorem 5, Proposition 7 and Corollary 8. Theorem 9 regarding the adaptive test is proved in Appendix B. The lower bound results, Theorem 3 and Theorem 6, are proved in Appendix C. Technical details of Section VI are collected in Appendix D. Appendix E contains auxiliary results.

A. Proofs of upper bound results in Sections II and III

Throughout the proofs in this subsection, we fix $P_e \in \mathcal{G}_{\alpha,K}^{\otimes}$ (resp. $\mathcal{P}_{\alpha,K}^{\otimes}$) and write $\mathbb{E}_\theta$ in place of $\mathbb{E}_{\theta,P_e}$ for the ease of notation. In every proof, we desire to control the two terms $\sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_\theta \phi$ ('**null term**') and $\sup_{\theta \in \Theta_1(p,n,s,\rho)} \mathbb{E}_\theta (1-\phi)$ ('**alternative term**') respectively. The values of the constants $C_1, C_2, \ldots$ vary from proof to proof. Note also that the order of the constants in each proof do not necessarily match that in the statement of the result, e.g. C_2 in the proof of Theorem 1 below corresponds to C_1 in the statement of Theorem 1.

1) Proof of Theorem 1: **Null term.** For any $\theta \in \Theta_0(p,n)$, we can write

$$Y_t = \frac{\sum_{i=1}^t (X_i - \theta_1) - \sum_{i=1}^t (X_{n+1-i} - \theta_1)}{\sqrt{2t}}.$$

Observe that $Y_t = (Y_t(1), \ldots, Y_t(p))^\top$ has independent components, each having mean 0 and variance 1. Moreover, each $X_i(j) - \theta_1(j)$ is a (centered) sub-Weibull random variable of order α belonging to the class $\mathcal{G}_{\alpha,K}$. Now, we consider the following block diagonal matrix $B \in \mathbb{R}^{2tp \times 2tp}$:

$$B := \begin{pmatrix} B^{\text{block}} & 0 & \cdots & 0 \\ 0 & B^{\text{block}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B^{\text{block}} \end{pmatrix},$$

where $B^{\text{block}} = (b_{ij})_{i,j \in [2t]} \in \mathbb{R}^{2t \times 2t}$ is defined as follows:

$$b_{ij} = \begin{cases} \frac{1}{2t} & \text{if } i = j, \\ \frac{1}{t} & \text{if } 1 \leq i \neq j \leq t \text{ or } t < i \neq j \leq 2t, \\ -\frac{1}{t} & \text{if } 1 \leq i \leq t < j \leq 2t \text{ or } 1 \leq j \leq t < i \leq 2t. \end{cases}$$

Let $U_i(j) := X_i(j) - \theta_1(j)$ for $i \in [n]$ and $j \in [p]$. Now, we can write

$$\sum_{j \in [p]} Y_t^2(j) = \sum_{j \in [p]} \frac{1}{2t} \left(\sum_{i=1}^{t} U_i(j) - \sum_{i=1}^{t} U_{n+1-i}(j) \right)^2 = \tilde{U}^\top B \tilde{U},$$

where $\tilde{U} \in \mathbb{R}^{2tp}$ has its first $2t$ coordinates as

$$(U_1(1), U_2(1), \ldots, U_t(1), U_{n+1-t}(1), \ldots, U_n(1))^\top,$$

and the remaining entries take the same form but with the coordinate index changing from 1 to p.

We calculate four different norms of matrix B:

$$\|B\|_{\mathrm{F}} = \sqrt{2tp \left(\frac{1}{4t^2} + \frac{2t-1}{t^2} \right)} \le 2\sqrt{p},$$

$$\|B\|_2 = \frac{1}{2t} + \frac{2t-1}{t} \le 2,$$

$$\|B\|_{2\to\infty} = \max_{i \in [2t]} \sqrt{\sum_{j \in [2t]} b_{ij}^2} \le \sqrt{\frac{2}{t}},$$

$$\|B\|_{\max} = 1/t.$$

For $\alpha \in [1, 2]$, we observe by Proposition 30 that $\mathcal{G}_{\alpha,K} \subseteq \mathcal{G}_{1,K'}$ for some constant $K' > 0$, depending only on K. Recall that $A_t = \sum_{j \in [p]} Y_t^2(j) - p$. Thus, for any $\alpha \in (0, 2]$, by applying Proposition 32, we have

$$\mathbb{P}_\theta(A_t > r) \le \exp\left\{ 1 - \left(\frac{r}{C_1\sqrt{p}} \right)^2 \right\} + \exp\left\{ 1 - \frac{r}{C_1} \right\} + \exp\left\{ 1 - \left(\frac{r\sqrt{t}}{C_1} \right)^{\frac{2\alpha}{2+\alpha} \wedge \frac{2}{3}} \right\}$$
$$+ \exp\left\{ 1 - \left(\frac{rt}{C_1} \right)^{\frac{\alpha}{2} \wedge \frac{1}{2}} \right\}, \tag{41}$$

where $C_1 > 0$ is some constant depending only on α and K from Proposition 32. Then, by union bounds (for all four terms) and Lemma 35 (for the last two terms), we obtain that for any $\theta \in \Theta_0(p, n)$ and $r \ge C_1 \Big\{ \big(2^{\frac{\alpha}{2+\alpha} \wedge \frac{1}{3}} - 1 \big)^{-\left(\frac{2+\alpha}{\alpha} \vee 3 \right)} \vee \big(2^{\frac{\alpha}{2} \wedge \frac{1}{2}} - 1 \big)^{-\left(\frac{2}{\alpha} \vee 2 \right)} \Big\}$,

$$\mathbb{E}_\theta \phi_{\mathcal{G},\mathrm{dense}} = \mathbb{P}_\theta \big(\max_{t \in \mathcal{T}} A_{t,0} > r \big) \le e \log_2(n) \exp\left\{ -\left(\frac{r}{C_1\sqrt{p}} \right)^2 \right\} + e \log_2(n) \exp\left\{ -\frac{r}{C_1} \right\}$$
$$+ e \sum_{t \in \mathcal{T}} \exp\left\{ -\left(\frac{r\sqrt{t}}{C_1} \right)^{\frac{2\alpha}{2+\alpha} \wedge \frac{2}{3}} \right\} + e \sum_{t \in \mathcal{T}} \exp\left\{ -\left(\frac{rt}{C_1} \right)^{\frac{\alpha}{2} \wedge \frac{1}{2}} \right\}$$
$$\le e \log_2(n) \exp\left\{ -\left(\frac{r}{C_1\sqrt{p}} \right)^2 \right\} + e \log_2(n) \exp\left\{ -\frac{r}{C_1} \right\}$$
$$+ 2e \exp\left\{ -\left(\frac{r}{C_1} \right)^{\frac{2\alpha}{2+\alpha} \wedge \frac{2}{3}} \right\} + 2e \exp\left\{ -\left(\frac{r}{C_1} \right)^{\frac{\alpha}{2} \wedge \frac{1}{2}} \right\}, \tag{42}$$

Thus, when

$$r \ge \left\{ C_1 \sqrt{p \log(8e\varepsilon^{-1} \log_2(n))} \right\} \vee \left\{ C_1 \log(8e\varepsilon^{-1} \log_2(n)) \right\}$$
$$\vee \left\{ C_1 \log^{\frac{2+\alpha}{2\alpha} \vee \frac{3}{2}}(16e\varepsilon^{-1}) \right\} \vee \left\{ C_1 \log^{\frac{2}{\alpha} \vee 2}(16e\varepsilon^{-1}) \right\},$$

each of the four terms in (42) can be upper bounded by $\varepsilon/8$. Equivalently, when

$$r \ge C_2 \big(\sqrt{p \log \log(8n)} + \log \log(8n) \big),$$

for some constant $C_2 > 0$, depending only on α, K and ε, we have $\mathbb{E}_\theta \phi_{\mathcal{G},\mathrm{dense}} \le \varepsilon/2$ for any $\theta \in \Theta_0(p, n)$.

Alternative term. For any $\theta \in \Theta(p, n, s, \rho)$, there exists some $t_0 \in [n]$, such that the mean change happens at time t_0, with $\frac{t_0(n-t_0)}{n}\|\mu_1 - \mu_2\|^2 \geq \rho^2$. We may assume without loss of generality that $t_0 \leq n/2$. By the definition of $\mathcal{T}$, there exists a unique $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0$. Note that then we can write

$$Y_{\tilde{t}} = \sqrt{\frac{\tilde{t}}{2}}(\mu_1 - \mu_2) + \frac{\sum_{i=1}^{\tilde{t}}(X_i - \mu_1) - \sum_{i=1}^{\tilde{t}}(X_{n+1-i} - \mu_2)}{\sqrt{2\tilde{t}}} =: \delta + Y_{\tilde{t}}', \tag{43}$$

where $\|\delta\|_2^2 \geq t_0\|\mu_1 - \mu_2\|_2^2/4 \geq \rho^2/4$. Note also that for all $j \in [p]$, we have $\mathbb{E}_\theta[Y_{\tilde{t}}'(j)] = 0$ and $\mathbb{E}[(Y_{\tilde{t}}'(j))^2] = 1$. By Proposition 29(b) and Lemma 36(a), we have $\mathbb{E}[(Y_{\tilde{t}}'(j))^4] \leq C_3$ for some constant $C_3 > 0$, depending only on α and K. When $\rho^2 \geq 8r \geq 8C_2(\sqrt{p \log \log(8n)} + \log \log(8n))$, we have by Chebyshev's inequality that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{G},\text{dense}}) \leq \mathbb{P}_\theta\left(\max_{t \in \mathcal{T}} \sum_{j=1}^{p} Y_t(j)^2 - p \leq \rho^2/8\right) \leq \mathbb{P}_\theta\left(\sum_{j=1}^{p}\left(Y_{\tilde{t}}(j)^2 - \delta(j)^2 - 1\right) \leq -\|\delta\|_2^2/2\right)$$

$$\leq \frac{4\sum_{j=1}^{p}\text{Var}_\theta(Y_{\tilde{t}}(j)^2)}{\|\delta\|_2^4} = \frac{4\sum_{j=1}^{p}\text{Var}_\theta(Y_{\tilde{t}}'(j)^2 + 2\delta(j)Y_{\tilde{t}}'(j))}{\|\delta\|_2^4}$$

$$\leq \frac{4\sum_{j=1}^{p}\{2\text{Var}_\theta(Y_{\tilde{t}}'(j)^2) + 8\delta(j)^2\text{Var}_\theta(Y_{\tilde{t}}'(j))\}}{\|\delta\|_2^4} \leq \frac{\sum_{j=1}^{p}\{8\mathbb{E}_\theta[Y_{\tilde{t}}'(j)^4] + 32\delta(j)^2\}}{\|\delta\|_2^4}$$

$$\leq \frac{8C_3 p + 32\|\delta\|_2^2}{\|\delta\|_2^4} \leq 128\left(\frac{C_3 p}{\rho^4} + \frac{1}{\rho^2}\right) \leq \frac{2C_3}{C_2^2 \log \log(8n)} + \frac{16}{C_2\sqrt{p \log \log(8n)}}, \tag{44}$$

where we have used the fact that $\text{Var}(X + Y) \leq 2(\text{Var}(X) + \text{Var}(Y))$ in the fourth inequality. Therefore, by having $C_2 > \max\{64/\varepsilon, \sqrt{8C_3/\varepsilon}\}$, we are guaranteed that $\mathbb{E}_\theta(1 - \phi_{\mathcal{G},\text{dense}}) \leq \varepsilon/2$ and the desired result follows.

2) Proof of Theorem 2: **Null term.** For any $\theta \in \Theta_0(p, n)$, we have by a union bound that

$$\mathbb{E}_\theta \phi_{\mathcal{G},\text{sparse}} \leq \mathbb{P}_\theta(A_{1,a} > r_1) + \sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta(A_{t,a} > r). \tag{45}$$

We first control the second term in (45). Recall the definition of $Y_{t,1}$ and $Y_{t,2}$ from (11) and denote $\mathcal{J}_{t,a} := \{j \in [p] : |Y_{t,2}(j)| \geq a\}$ for $t \in \mathcal{T}$ and $a \geq 0$. Note that $\mathcal{J}_{t,a}$ is a random set. Then,

$$\sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta(A_{t,a} > r) \leq \sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta\left(\sum_{j \in \mathcal{J}_{t,a}}(Y_{t,1}^2(j) - 1) > r\right)$$

$$= \sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{E}_\theta\left[\mathbb{P}_\theta\left(\sum_{j \in \mathcal{J}_{t,a}}(Y_{t,1}^2(j) - 1) > r \,\Big|\, \mathcal{J}_{t,a}\right)\right]$$

$$= \sum_{t \in \mathcal{T} \setminus \{1\}} \sum_{J \subseteq [p]}\left\{\mathbb{P}_\theta\left(\sum_{j \in J}(Y_{t,1}^2(j) - 1) > r\right)\mathbb{P}_\theta(\mathcal{J}_{t,a} = J)\right\}$$

$$\leq \sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t \in \mathcal{T} \setminus \{1\}} \sup_{J \subseteq [p]:|J| \leq s} \mathbb{P}_\theta\left(\sum_{j \in J}(Y_{t,1}^2(j) - 1) > r\right), \tag{46}$$

where the third line follows from the independence of $Y_{t,1}$ and $Y_{t,2}$. We now control the two terms in (46) respectively. Using Proposition 31 with $u_i = t^{-1/2}$ for $i = 1, \ldots, t/2$ and $u_i = -t^{-1/2}$ for $i = t/2 + 1, \ldots, t$, we obtain that for any $t \in \mathcal{T}$, $j \in [p]$ and $x \geq 0$

$$\mathbb{P}_\theta(|Y_{t,2}(j)| \geq x) \leq \exp\left\{1 - \min\left\{\left(\frac{x}{C_1}\right)^2, \left(\frac{x}{C_1\|u\|_{\beta(\alpha)}}\right)^\alpha\right\}\right\},$$

for some constant $C_1 \geq 1$ depending only on α and K. For $\alpha \leq 1$, we have $\|u\|_{\beta(\alpha)} = \|u\|_\infty = t^{-1/2}$ and for $1 < \alpha \leq 2$, we have $\|u\|_{\beta(\alpha)} = \|u\|_{\alpha/(\alpha-1)} = t^{1/2-1/\alpha}$. Thus

$$q_{t,a} := \mathbb{P}_\theta(|Y_{t,2}(j)| \geq a) \leq \exp\left\{1 - \min\left\{\left(\frac{a}{C_1}\right)^2, \left(\frac{a}{C_1 t^{(-\frac{1}{2}) \vee (\frac{1}{2} - \frac{1}{\alpha})}}\right)^\alpha\right\}\right\}. \tag{47}$$

For $0 < \alpha < 2$, by a binomial tail bound, equation (2.1) in Theorem 1 of [46], we have

$$\mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \left(\frac{q_{t,a}}{s/p}\right)^s\left(\frac{1 - q_{t,a}}{1 - s/p}\right)^{p-s} = \left(\frac{pq_{t,a}}{s}\right)^s\left(1 + \frac{s - pq_{t,a}}{p - s}\right)^{p-s}$$

$$\leq \left(\frac{pq_{t,a}}{s}\right)^s e^{s-pq_{t,a}} \leq \left(\frac{epq_{t,a}}{s}\right)^s. \tag{48}$$

Combining this with (47), we have

$$\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \sum_{t\in\mathcal{T}\setminus\{1\}} \left(\frac{epq_{t,a}}{s}\right)^s$$

$$\leq \log_2(n)\left(\frac{2e^2p}{s}\right)^s \exp\left\{-\frac{sa^2}{C_1^2}\right\} + \left(\frac{2e^2p}{s}\right)^s \sum_{t\in\mathcal{T}\setminus\{1\}} \exp\left\{-s\left(\frac{a^{\frac{2\alpha}{\alpha\wedge(2-\alpha)}}t}{C_1^{\frac{2\alpha}{\alpha\wedge(2-\alpha)}}}\right)^{\frac{\alpha\wedge(2-\alpha)}{2}}\right\}$$

$$\leq \log_2(n)\left(\frac{2e^2p}{s}\right)^s \exp\left\{-\frac{sa^2}{C_1^2}\right\} + 2\left(\frac{2e^2p}{s}\right)^s \exp\left\{-\frac{sa^\alpha}{C_1^\alpha}\right\}, \tag{49}$$

provided that $a \geq C_1\left(2^{\frac{\alpha\wedge(2-\alpha)}{2}} - 1\right)^{-1/\alpha}$, where we have used Lemma 35 in the last inequality. In fact, for $\alpha = 2$, by (47), the final bound in (49) remains valid for all $a \geq 0$. Now, the first term in the final bound above can be bounded by $\varepsilon/16$ when

$$a \geq C_1 s^{-1/2}\log^{1/2}(16\varepsilon^{-1}\log_2(n)) + C_1\log^{1/2}(2e^2p/s),$$

and the second term can be bounded by $\varepsilon/16$ when

$$a \geq C_1 s^{-1/\alpha}\log^{1/\alpha}(32\varepsilon^{-1}) + C_1\log^{1/\alpha}(2e^2p/s).$$

Thus, as long as we choose a to satisfy

$$a \geq C_2\left(\log^{1/\alpha}(ep/s) + s^{-1/2}\log^{1/2}(\varepsilon^{-1}\log(8n)) + s^{-1/\alpha}\log^{1/\alpha}(e\varepsilon^{-1})\right)$$

for some large enough $C_2 > 0$, depending only on α and K, we are guaranteed that

$$\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \varepsilon/8. \tag{50}$$

We now begin to bound the second term in (46). By replacing p with $|J|$ in (41), we have

$$\mathbb{P}_\theta\left(\sum_{j\in J}(Y_{t,1}^2(j) - 1) > r\right) \leq \exp\left\{1 - \left(\frac{r}{C_3\sqrt{|J|}}\right)^2\right\} + \exp\left\{1 - \frac{r}{C_3}\right\} + \exp\left\{1 - \left(\frac{r\sqrt{t}}{C_3}\right)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\right\}$$

$$+ \exp\left\{1 - \left(\frac{rt}{C_3}\right)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\right\},$$

where $C_3 > 0$ is some constant depending only on α and K from Proposition 32. Then, by the same technique as used in (42) in the proof of Theorem 1 (applying union bounds to the first two terms, and using Lemma 35 for the last two terms), we obtain that for all $r \geq C_3\left\{\left(2^{\frac{\alpha}{2+\alpha}\wedge\frac{1}{3}} - 1\right)^{-\left(\frac{2+\alpha}{\alpha}\vee 3\right)} \vee \left(2^{\frac{\alpha}{2}\wedge\frac{1}{2}} - 1\right)^{-\left(\frac{2}{\alpha}\vee 2\right)}\right\}$

$$\sum_{t\in\mathcal{T}\setminus\{1\}} \sup_{J\subseteq[p]:|J|\leq s} \mathbb{P}_\theta\left(\sum_{j\in J}(Y_{t,1}^2(j) - 1) > r\right)$$

$$\leq \sum_{t\in\mathcal{T}\setminus\{1\}}\left\{\exp\left\{1 - \left(\frac{r}{C_3\sqrt{s}}\right)^2\right\} + \exp\left\{1 - \frac{r}{C_3}\right\} + \exp\left\{1 - \left(\frac{r\sqrt{t}}{C_3}\right)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\right\}\right.$$

$$\left. + \exp\left\{1 - \left(\frac{rt}{C_3}\right)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\right\}\right\}$$

$$\leq e\log_2(n)\exp\left\{-\left(\frac{r}{C_3\sqrt{s}}\right)^2\right\} + e\log_2(n)\exp\left\{-\frac{r}{C_3}\right\} + \sum_{t\in\mathcal{T}\setminus\{1\}}\exp\left\{1 - \left(\frac{r\sqrt{t}}{C_3}\right)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\right\}$$

$$+ \sum_{t\in\mathcal{T}\setminus\{1\}}\exp\left\{1 - \left(\frac{rt}{C_3}\right)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\right\}$$

$$\leq e\log_2(n)\exp\left\{-\left(\frac{r}{C_3\sqrt{s}}\right)^2\right\} + e\log_2(n)\exp\left\{-\frac{r}{C_3}\right\} + 2e\exp\left\{-\left(\frac{r}{C_3}\right)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\right\}$$

$$+ 2e\exp\left\{-\left(\frac{r}{C_3}\right)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\right\}. \tag{51}$$

Thus, when

$$r \geq \left\{ C_3 \sqrt{s \log(32e\varepsilon^{-1} \log_2(n))} \right\} \vee \left\{ C_3 \log(32e\varepsilon^{-1} \log_2(n)) \right\}$$

$$\vee \left\{ C_3 \log^{\frac{2+\alpha}{2\alpha} \vee \frac{3}{2}}(64e\varepsilon^{-1}) \right\} \vee \left\{ C_1 \log^{\frac{2}{\alpha} \vee 2}(64e\varepsilon^{-1}) \right\},$$

each of the four terms in (51) can be upper bounded by $\varepsilon/32$. Equivalently, when

$$r \geq C_4 \left(\sqrt{s \log(\varepsilon^{-1} \log(8n))} + \log(\varepsilon^{-1} \log(8n)) + \log^{\frac{2}{\alpha} \vee 2}(e\varepsilon^{-1}) \right),$$

for some constant $C_4 > 0$, depending only on α and K, we are guaranteed

$$\sum_{t \in \mathcal{T} \setminus \{1\}} \sup_{J \subseteq [p]: |J| \leq s} \mathbb{P}_\theta \left(\sum_{j \in J} (Y_{t,1}^2(j) - 1) > r \right) \leq \varepsilon/8. \tag{52}$$

Finally, for the first term in (45), by Proposition 34(a), whenever $r_1 \geq C_4' s \log^{2/\alpha}(ep/s)$ for some sufficiently large $C_4' > 0$, depending on α, K and ε, we have

$$\mathbb{P}_\theta(A_{1,a} > r_1) \leq \varepsilon/4. \tag{53}$$

By combining (45), (46), (50), (52) and (53), we conclude that $\mathbb{E}_\theta \phi_{\mathcal{G},\mathrm{sparse}} \leq \varepsilon/2$ for all $\theta \in \Theta_0(p, n)$.

Alternative term. We use the same argument as at the beginning of the alternative part of the proof of Theorem 1. Recall that there exists a unique $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0$. We first consider the case $t_0 \geq 2$. This implies $\tilde{t} \geq 2$. Now, similar to (43), we can write

$$Y_{\tilde{t},1} = \frac{\sqrt{\tilde{t}}}{2}(\mu_1 - \mu_2) + \frac{\sum_{i=1}^{\tilde{t}/2}(X_{2i-1} - \mu_1) - \sum_{i=1}^{\tilde{t}/2}(X_{n-2i+1} - \mu_2)}{\sqrt{\tilde{t}}} =: \delta + Y'_{\tilde{t},1},$$

$$Y_{\tilde{t},2} = \frac{\sqrt{\tilde{t}}}{2}(\mu_1 - \mu_2) + \frac{\sum_{i=1}^{\tilde{t}/2}(X_{2i} - \mu_1) - \sum_{i=1}^{\tilde{t}/2}(X_{n-2i+2} - \mu_2)}{\sqrt{\tilde{t}}} =: \delta + Y'_{\tilde{t},2}.$$

The quantity $\delta := \sqrt{\tilde{t}}(\mu_1 - \mu_2)/2$ satisfies $\|\delta\|_2^2 \geq \rho^2/8$. Denote $\mathcal{S}_\delta := \{j \in [p] : \delta(j) \neq 0\}$ and $\mathcal{H}_{\delta,a} := \{j \in [p] : |\delta(j)| \geq 2a\}$. Note that these two sets are deterministic, while $\mathcal{J}_{\tilde{t},a} = \{j \in [p] : |Y_{\tilde{t},2}(j)| \geq a\}$ is random. Then, when $\rho^2 \geq 192(r + 2s) \log(8/\varepsilon)$, we have

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{G},\mathrm{sparse}}) \leq \mathbb{P}_\theta \left(\sum_{j \in \mathcal{J}_{\tilde{t},a}} (Y_{\tilde{t},1}^2(j) - 1) \leq r \right)$$

$$= \mathbb{P}_\theta \left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}^c} (Y_{\tilde{t},1}^2(j) - 1) + \sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} (Y_{\tilde{t},1}^2(j) - 1) \leq r \right)$$

$$\leq \mathbb{P}_\theta(|\mathcal{J}_{\tilde{t},a}| > 2s) + \mathbb{P}_\theta \left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} (Y_{\tilde{t},1}^2(j) - 1) \leq r + 2s \right)$$

$$\leq \mathbb{P}_\theta(|\mathcal{J}_{\tilde{t},a}| > 2s) + \mathbb{P}_\theta \left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)} \right)$$

$$+ \mathbb{P}_\theta \left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} (Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1) \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)} \right). \tag{54}$$

We now control the three terms in (54) respectively. By (50), we have

$$\mathbb{P}_\theta(|\mathcal{J}_{\tilde{t},a}| > 2s) \leq \mathbb{P}_\theta(|\mathcal{J}_{\tilde{t},a} \cap \mathcal{S}_\delta^c| > s) \leq \varepsilon/8. \tag{55}$$

For the second term, we observe that for all $j \in \mathcal{H}_{\delta,a}$

$$\mathbb{P}_\theta(j \notin \mathcal{J}_{\tilde{t},a}) = \mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) = \mathbb{P}_\theta(|\delta(j) + Y'_{\tilde{t},2}(j)| < a) \leq \mathbb{P}_\theta(|Y'_{\tilde{t},2}(j)| > |\delta(j)| - a)$$

$$\leq \exp\{1 - ((|\delta(j)| - a)/C_1)^\alpha\} \leq \frac{1}{256 \log(8/\varepsilon)}, \tag{56}$$

where the last two inequalities follow from (47) and the choice $a \geq C_1 \log^{1/\alpha}(700 \log(8/\varepsilon))$. Consequently,

$$\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\big(\delta(j)^2 \mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a}\}}\big) \leq \sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^4 \mathbb{P}_\theta(j \notin \mathcal{J}_{\tilde{t},a}) \leq \frac{\sum_{j=1}^p \delta(j)^4}{256 \log(8/\varepsilon)} \leq \frac{\|\delta\|_2^4}{256 \log(8/\varepsilon)}. \tag{57}$$

Moreover, when $\rho^2 \geq 64a^2 s$, we obtain

$$\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \geq \|\delta\|_2^2 - s(2a)^2 \geq \|\delta\|_2^2/2. \tag{58}$$

We first consider the case $\|\delta\|_2 \geq \sqrt{12 \log(8/\varepsilon)}\|\delta\|_\infty$. Then, by combining (56), (57), (58) and Bernstein's inequality, we have

$$
\begin{aligned}
\mathbb{P}_\theta\bigg(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \|\delta\|_2^2/8\bigg) &= \mathbb{P}_\theta\bigg(\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a}\}} < \|\delta\|_2^2/8\bigg) \\
&\leq \mathbb{P}_\theta\bigg(\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \big(\mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a}\}} - \mathbb{P}_\theta(j \in \mathcal{J}_{\tilde{t},a})\big) < -\|\delta\|_2^2/8\bigg) \\
&\leq \exp\bigg\{-\frac{\|\delta\|_2^4/64}{2\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\big(\delta(j)^2 \mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a}\}}\big) + \|\delta\|_\infty^2 \|\delta\|_2^2/12}\bigg\} \\
&\leq \max\bigg\{\exp\{-\log(8/\varepsilon)\}, \exp\bigg\{-\frac{\|\delta\|_2^2}{12\|\delta\|_\infty^2}\bigg\}\bigg\} \leq \varepsilon/8.
\end{aligned} \tag{59}
$$

If instead $\|\delta\|_\infty \leq \|\delta\|_2 < \sqrt{12 \log(8/\varepsilon)}\|\delta\|_\infty$, we assume that $|\delta(j^*)| = \|\delta\|_\infty$ for some $j^* \in \mathcal{H}_{\delta,a}$. Note that when $\rho^2 \geq 384 C_1^2 \log^{\frac{\alpha+2}{\alpha}}(8e/\varepsilon)$, we have $|\delta(j^*)| \geq 2C_1 \log^{1/\alpha}(8e/\varepsilon)$ and thus

$$
\begin{aligned}
\mathbb{P}_\theta\bigg(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)}\bigg) &\leq \mathbb{P}_\theta\bigg(\delta(j^*)^2 \mathbb{1}_{\{j^* \in \mathcal{J}_{\tilde{t},a}\}} < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)}\bigg) \\
&\leq \mathbb{P}_\theta(|Y_{\tilde{t},2}(j^*)| < a) \leq \exp\{1 - (|\delta(j^*)|/(2C_1))^\alpha\} \leq \varepsilon/8.
\end{aligned} \tag{60}
$$

For the third and final term in (54), we have by Chebyshev's inequality that

$$
\begin{aligned}
\mathbb{P}_\theta\bigg(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \big(Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1\big) &\leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)}\bigg) \\
&\leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\big((Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1)\mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}}\big)}{\|\delta\|_2^4/(576 \log^2(8/\varepsilon))} \leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\big(Y_{\tilde{t},1}^2(j)\big)}{\|\delta\|_2^4/(576 \log^2(8/\varepsilon))} \\
&\leq C_5 \log^2(8/\varepsilon)\bigg(\frac{s}{\rho^4} + \frac{1}{\rho^2}\bigg),
\end{aligned} \tag{61}
$$

where $C_5 \geq 1$ is a constant depending on α and K and the penultimate inequality follows from a similar argument to (44). Hence, when

$$\rho^2 \geq C_5 \varepsilon^{-1} \max\Big\{192(r+2s) \log(8/\varepsilon), 64a^2 s, 384 C_1^2 \log^{\frac{\alpha+2}{\alpha}}(8e/\varepsilon)\Big\},$$

we have by combining (54), (55), (59), (60) and (61) that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{G},\mathrm{sparse}}) \leq \varepsilon/4 + C_5 \log^2(8/\varepsilon)\bigg(\frac{s}{\rho^4} + \frac{1}{\rho^2}\bigg) \leq \varepsilon/2.$$

Finally, We consider the case that the mean change happens at $t_0 = 1$ instead. Recall that in this case we have $\tilde{t} = 1$. (54) remains true when $\rho^2 \geq 192(r_1 + 2s) \log(8/\varepsilon)$ if we redefine $\mathcal{J}_{\tilde{t}=1,a} := \{j \in [p] : |Y_1(j)| \geq a\}$. All three terms in (54) can be controlled in the same way as when $t_0 \geq 2$ and this completes the proof.

 3) Proof of Theorem 4: We first prove the result for $\alpha \geq 4$.
Null term. For any $\theta \in \Theta_0(p, n)$, we have $\mathbb{E}_\theta \overline{Z}_{t,g}(j) = 0$ and $\mathrm{Var}_\theta \overline{Z}_{t,g}(j) = G_t/t$ for every $t \in \mathcal{T}$, $g \in [G_t]$ and $j \in [p]$. Furthermore, from the class assumption $\mathbb{E}|E_i(j)|^\alpha \leq K^\alpha$, for all $i \in [n]$ and $j \in [p]$ and Jensen's inequality, we deduce $\mathbb{E}E_i(j)^4 \leq K^4$. We thus obtain, for all $i \leq n/2$ and $j \in [p]$

$$\mathbb{E}_\theta Z_i^4(j) = \mathbb{E}_\theta\bigg[\frac{X_i(j) - X_{n-i}(j)}{\sqrt{2}}\bigg]^4 = \frac{\mathbb{E}_\theta\big[E_i(j) - E_{n-i}(j)\big]^4}{4} \leq \frac{K^4 + 3}{2} =: C_1. \tag{62}$$

Then, by Chebyshev's inequality (or, alternatively, Lemma 37) and Lemma 36(a), with $r_t = C_2\sqrt{p}G_t$, we have for all $t \in \mathcal{T}$ and $g \in [G_t]$ that

$$\mathbb{P}_\theta\bigg(t\sum_{j=1}^p V_{t,g}(j) > r_t\bigg) = \mathbb{P}_\theta\bigg(\sum_{j=1}^p \bigg(\overline{Z}_{t,g}^2(j) - \frac{G_t}{t}\bigg) > \frac{C_2\sqrt{p}G_t}{t}\bigg) \leq \frac{t^2 \sum_{j=1}^p \mathbb{E}_\theta \overline{Z}_{t,g}^4(j)}{C_2^2 p G_t^2}$$

$$\leq \frac{3pt^2 C_1 (G_t/t)^2}{C_2^2 p G_t^2} \leq \frac{3C_1}{C_2^2} \leq \varepsilon/36, \tag{63}$$

where C_2 is chosen to satisfies $C_2 \geq \sqrt{108 C_1 \varepsilon^{-1}}$. We denote

$$\mathcal{B}_t := \left\{ g \in [G_t] : t \sum_{j=1}^p V_{t,g}(j) > r_t \right\}.$$

By (63) and the multiplicative Chernoff bound, e.g. Corollary 4.9 of [47], we have for $t \in \mathcal{T}$

$$\mathbb{P}_\theta (A_t^{\mathrm{MoM}} > r_t) \leq \mathbb{P}_\theta (|\mathcal{B}_t| \geq G_t/2) = \mathbb{P}_\theta \left(|\mathcal{B}_t| \geq \frac{\varepsilon G_t}{36} \left(1 + \left(\frac{18}{\varepsilon} - 1 \right) \right) \right)$$

$$\leq \exp\left\{ -\frac{\varepsilon G_t}{36} \left(\frac{18}{\varepsilon} \log\left(\frac{18}{\varepsilon} \right) - \frac{18}{\varepsilon} + 1 \right) \right\} \leq \exp\left\{ -\frac{G_t}{2} \log(6/\varepsilon) \right\}. \tag{64}$$

Thus, by (63), (64), the choices of G_t and Δ in (18) and a union bound, we conclude that

$$\mathbb{E}_\theta \phi_{\mathcal{P},\mathrm{dense}} \leq \mathbb{P}_\theta \left(\sum_{j=1}^p V_{t=1,1}(j) > r_{t=1} \right) + \sum_{t \in \mathcal{T} : 2 \leq t \leq \Delta} \mathbb{P}_\theta (A_t^{\mathrm{MoM}} > r_t) + \sum_{t \in \mathcal{T} : t > \Delta} \mathbb{P}_\theta (A_t^{\mathrm{MoM}} > r_t)$$

$$\leq \varepsilon/36 + \sum_{t \in \mathcal{T} : 2 \leq t \leq \Delta} (6/\varepsilon)^{-t/2} + \sum_{t \in \mathcal{T} : t > \Delta} (6/\varepsilon)^{-\Delta/2}$$

$$\leq \varepsilon/36 + \frac{(6/\varepsilon)^{-1}}{1 - (6/\varepsilon)^{-1}} + \log_2(n/2)(6/\varepsilon)^{-\Delta/2} \leq \varepsilon/36 + \varepsilon/5 + \varepsilon/5 < \varepsilon/2, \tag{65}$$

for all $\theta \in \Theta_0(p, n)$.

Alternative term. We again follow the argument in the first paragraph of the alternative term part of the proof of Theorem 1. In particular, recall that there exists a unique $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0$, where t_0 (without loss of generality $t_0 \leq n/2$) is the true mean change location. For all $i \leq n/2$, we denote

$$Z_i' := Z_i - \frac{\mu_1 - \mu_2}{\sqrt{2}} = \frac{(X_i - \mu_1) - (X_{n+1-i} - \mu_2)}{\sqrt{2}},$$

and correspondingly $\overline{Z}_{\tilde{t},g}' := \overline{Z}_{\tilde{t},g} - (\mu_1 - \mu_2)/\sqrt{2}$, for $g \in [G_{\tilde{t}}]$. It follows from the null term part of the proof that $\mathbb{E}_\theta \overline{Z}_{\tilde{t},g}'(j) = 0$, $\mathrm{Var}_\theta \overline{Z}_{\tilde{t},g}'(j) = G_{\tilde{t}}/\tilde{t}$ and $\mathbb{E}_\theta(Z_i'(j))^4 \leq C_1$, where C_1 is as in (62). When $\rho^2 \geq 16 C_2 \sqrt{p} \Delta$, we have

$$2\tilde{t} \|\mu_1 - \mu_2\|^2 \geq \frac{t_0(n - t_0)}{n} \|\mu_1 - \mu_2\|^2 \geq \rho^2 \geq 16 C_2 \sqrt{p} G_{\tilde{t}} = 16 r_{\tilde{t}},$$

since $G_{\tilde{t}} \leq \Delta$. Thus, for all $g \in [G_{\tilde{t}}]$, we have

$$\mathbb{P}_\theta \left(\tilde{t} \sum_{j=1}^p V_{\tilde{t},g}(j) \leq r_{\tilde{t}} \right) = \mathbb{P}_\theta \left(\sum_{j=1}^p \left(\left(\overline{Z}_{\tilde{t},g}'(j) + \frac{\mu_1(j) - \mu_2(j)}{\sqrt{2}} \right)^2 - \frac{G_{\tilde{t}}}{\tilde{t}} \right) \leq \frac{r_{\tilde{t}}}{\tilde{t}} \right)$$

$$= \mathbb{P}_\theta \left(\sum_{j=1}^p \left((\overline{Z}_{\tilde{t},g}'(j))^2 - \frac{G_{\tilde{t}}}{\tilde{t}} + \sqrt{2}(\mu_1(j) - \mu_2(j))\overline{Z}_{\tilde{t},g}'(j) \right) \leq \frac{r_{\tilde{t}}}{\tilde{t}} - \frac{\|\mu_1 - \mu_2\|_2^2}{2} \right)$$

$$\leq \mathbb{P}_\theta \left(\sum_{j=1}^p \left((\overline{Z}_{\tilde{t},g}'(j))^2 - \frac{G_{\tilde{t}}}{\tilde{t}} + \sqrt{2}(\mu_1(j) - \mu_2(j))\overline{Z}_{\tilde{t},g}'(j) \right) \leq -\frac{\rho^2}{16\tilde{t}} - \frac{\|\mu_1 - \mu_2\|_2^2}{4} \right). \tag{66}$$

By Chebyshev's inequality and Lemma 36(a), we obtain

$$\mathbb{P}_\theta \left(\sum_{j=1}^p \left((\overline{Z}_{\tilde{t},g}'(j))^2 - G_{\tilde{t}}/\tilde{t} \right) \leq -\frac{\rho^2}{16\tilde{t}} \right) \leq \frac{256(\tilde{t})^2 \sum_{j=1}^p \mathbb{E}_\theta (\overline{Z}_{\tilde{t},g}'(j))^4}{\rho^4} \leq \frac{768 C_1 p G_{\tilde{t}}^2}{\rho^4}, \tag{67}$$

and

$$\mathbb{P}_\theta \left(\sum_{j=1}^p \sqrt{2}(\mu_1(j) - \mu_2(j))\overline{Z}_{\tilde{t},g}'(j) \leq -\frac{\|\mu_1 - \mu_2\|_2^2}{4} \right) \leq \frac{32 G_{\tilde{t}} \|\mu_1 - \mu_2\|_2^2/\tilde{t}}{\|\mu_1 - \mu_2\|_2^4} \leq \frac{64 G_{\tilde{t}}}{\rho^2}.$$

Combining these with (66), as long as

$$\rho^2 \geq \max\left\{ 16 C_2 \sqrt{p} \Delta, 96 \sqrt{\frac{2 C_4}{\varepsilon}} \sqrt{p} \Delta, \frac{1536 \Delta}{\varepsilon} \right\},$$

we are guaranteed

$$\mathbb{P}_\theta\left(\tilde{t}\sum_{j=1}^p V_{\tilde{t},g}(j) \leq r_{\tilde{t}}\right) \leq \varepsilon/12.$$

If $\tilde{t} = 1$, then $G_{\tilde{t}} = 1$ and we immediately have

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{dense}}) \leq \mathbb{P}_\theta\left(A_{\tilde{t}}^{\text{MoM}} \leq r_{\tilde{t}}\right) = \mathbb{P}_\theta\left(\tilde{t}\sum_{j=1}^p V_{\tilde{t},1}(j) \leq r_{\tilde{t}}\right) \leq \varepsilon/12.$$

If $\tilde{t} \geq 2$, then $G_{\tilde{t}} \geq 2$ and we use the same binomial tail bound argument as in (64) to conclude that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{dense}}) \leq \mathbb{P}_\theta\left(A_{\tilde{t}}^{\text{MoM}} \leq r_{\tilde{t}}\right) \leq \exp\left\{-\frac{\varepsilon G_{\tilde{t}}}{12}\left(\frac{6}{\varepsilon}\log\left(\frac{6}{\varepsilon}\right) - \frac{6}{\varepsilon} + 1\right)\right\} \leq \left(\frac{2}{\varepsilon}\right)^{-1}.$$

This completes the proof for $\alpha \geq 4$. We now consider the case $\alpha < 4$. The proof is similar to above and we essentially replace Chebyshev's inequality wherever used by Lemma 37. We only highlight the difference for brevity.

Null term. Note that for all $t \in \mathcal{T}$ and $g \in [G_t]$, using Lemma 37 with $k = \alpha/2 < 2$ and $L = t/G_t$, we have with $r_t = C_2 p^{2/\alpha} G_t$ that

$$\mathbb{P}_\theta\left(t\sum_{j=1}^p V_{t,g}(j) > r_t\right) = \mathbb{P}_\theta\left(\frac{t}{G_t}\sum_{j=1}^p\left(\overline{Z}_{t,g}^2(j) - \frac{G_t}{t}\right) > C_2 p^{2/\alpha}\right) \leq \frac{\varepsilon}{36}, \tag{68}$$

for $C_2 \geq C_{\alpha/2}(36/\varepsilon)^{2/\alpha}$, where $C_{\alpha/2} > 0$ is the constant depending on α and K from Lemma 37. By substituting (63) with (68) and following the rest of the argument in the above proof, we prove that $\mathbb{E}_\theta\phi_{\mathcal{P},\text{dense}} \leq \varepsilon/2$ for all $\theta \in \Theta_0(p, n)$.

Alternative term. For all $g \in [G_{\tilde{t}}]$, again using Lemma 37 with $k = \alpha/2 < 2$ and $L = \tilde{t}/G_{\tilde{t}}$, we have

$$\mathbb{P}_\theta\left(\sum_{j=1}^p\left(\left(\overline{Z}_{\tilde{t},g}'(j)\right)^2 - G_{\tilde{t}}/\tilde{t}\right) \leq -\frac{\rho^2}{16\tilde{t}}\right) = \mathbb{P}_\theta\left(\frac{\tilde{t}}{G_{\tilde{t}}}\sum_{j=1}^p\left(\left(\overline{Z}_{\tilde{t},g}'(j)\right)^2 - G_{\tilde{t}}/\tilde{t}\right) \leq -\frac{\rho^2}{16G_{\tilde{t}}}\right) \leq \frac{\varepsilon}{24}, \tag{69}$$

for $\rho^2 \geq 24^{(2+\alpha)/\alpha}C_{\alpha/2}\varepsilon^{-2/\alpha}p^{2/\alpha}\Delta$, where $C_{\alpha/2}$ is, as above, a constant depending only on α and K. By substituting (67) with (69) and following the rest of argument in the above proof, we prove that as long as

$$\rho^2 \geq \max\left\{16C_2 p^{2/\alpha}\Delta, 24^{(2+\alpha)/\alpha}C_{\alpha/2}\varepsilon^{-2/\alpha}p^{2/\alpha}\Delta, \frac{1536\Delta}{\varepsilon}\right\},$$

we can control $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{dense}}) \leq \varepsilon/2$.

4) *Proof of Proposition 7:* **Null term.** For any $\theta \in \Theta_0(p, n)$, we have by a union bound that

$$\mathbb{E}_\theta\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}} \leq \mathbb{P}_\theta(A_{1,a} > r_1) + \sum_{t \in \mathcal{T}\backslash\{1\}}\mathbb{P}_\theta\left(A_{t,a}^{\text{MoM}} > r_t\right). \tag{70}$$

We first control the second term. Recall that $\mathcal{J}_{t,a} = \{j \in [p] : |Y_{t,2}(j)| \geq a\}$ for $t \in \mathcal{T} \backslash \{1\}$. For $J \subseteq [p]$, we denote

$$A_{t,*,J}^{\text{MoM}} := \frac{t}{2} \cdot \text{median}\left\{\sum_{j \in J}\left(\overline{Z}_{t,g,1}^2(j) - \frac{2G_t}{t}\right) : g \in [G_t]\right\}.$$

Note that $A_{t,a}^{\text{MoM}} = A_{t,*,\mathcal{J}_{t,a}}^{\text{MoM}}$. Using the same technique as (46) in the proof of Theorem 2, we have

$$\sum_{t \in \mathcal{T}\backslash\{1\}}\mathbb{P}_\theta(A_{t,a}^{\text{MoM}} > r_t) \leq \sum_{t \in \mathcal{T}\backslash\{1\}}\mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t \in \mathcal{T}\backslash\{1\}}\sup_{J \subseteq [p]:|J| \leq s}\mathbb{P}_\theta(A_{t,*,J}^{\text{MoM}} > r_t), \tag{71}$$

where s is the sparsity. From the assumption that $\mathbb{E}|E_i(j)|^\alpha \leq K^\alpha$, for all $i \in [n]$ and $j \in [p]$ and Jensen's inequality, we deduce that

$$\mathbb{E}_\theta|Z_i(j)|^\alpha = \frac{\mathbb{E}_\theta\left|E_i(j) - E_{n-i}(j)\right|^\alpha}{2^{\alpha/2}} \leq \frac{\mathbb{E}_\theta\left(|E_i(j)| + |E_{n-i}(j)|\right)^\alpha}{2^{\alpha/2}} \leq 2^{\alpha/2}K^\alpha.$$

Then, by Fuk–Nagaev inequality (Proposition 33), we have

$$q_{t,a} = \mathbb{P}_\theta(|Y_{t,2}(j)| \geq a) \leq 2\left(\frac{(\alpha+2)(K^\alpha 2^{\alpha/2}t/2)^{1/\alpha}}{\alpha a\sqrt{t/2}}\right)^\alpha + 2\exp\left\{-\frac{2a^2}{(\alpha+2)^2 e^\alpha}\right\}$$

$$\leq \frac{K^\alpha}{(a/3)^\alpha t^{\alpha/2-1}} + \exp\left\{1 - \frac{a^2}{2\alpha^2 e^\alpha}\right\}, \tag{72}$$

where we have used $\alpha \geq 4$ in the last inequality. Similar to (48) and (49), by a binomial tail bound, we have

$$\sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \sum_{t \in \mathcal{T} \setminus \{1\}} \left(\frac{epq_{t,a}}{s}\right)^s \leq \left(\frac{2epK^\alpha}{s(a/3)^\alpha}\right)^s + \log_2(n)\left(\frac{2e^2 p}{s}\right)^s \exp\left\{-\frac{sa^2}{2\alpha^2 e^\alpha}\right\}. \tag{73}$$

Thus, as long as we choose a to satisfy

$$a \geq C_1\left(\varepsilon^{-1}(p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\varepsilon^{-1}\log(8n))\right) \tag{74}$$

for some large enough $C_1 > 0$, depending only on α and K, we are guaranteed that

$$\sum_{t \in \mathcal{T} \setminus \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \frac{\varepsilon}{8}.$$

Furthermore, By setting $r_t = C_2\sqrt{s}G_t$ with a suffciently large $C_2 > 0$ and $\Delta = 2^{4 + \lceil \log_2 \log \log(8n) \rceil}$ and by following the argument from (63) to (65), we can upper bound the second term in (71) at $\varepsilon/8$ as well. Finally, to control the first term in (70), by Proposition 34(b), whenever $r_1 \geq C_1' s(p/s)^{2/\alpha}$ for sufficiently large $C_1' > 0$, depending on α, K and ε, we have $\mathbb{P}_\theta(A_{1,a} > r_1) \leq \varepsilon/4$. Hence, we conclude that $\mathbb{E}_\theta \phi_{\mathcal{P},\text{sparse}}^{\text{MoM}} \leq \varepsilon/2$ for all $\theta \in \Theta_0(p,n)$.

Alternative term. Recall the definitions of $\delta, \mathcal{S}_\delta$ and $\mathcal{H}_{\delta,a}$ from the alternative term part of the proof of Theorem 2:

$$\delta = \frac{\sqrt{\tilde{t}}}{2}(\mu_1 - \mu_2), \quad \mathcal{S}_\delta = \{j \in [p] : \delta(j) \neq 0\}, \quad \mathcal{H}_{\delta,a} = \{j \in [p] : |\delta(j)| \geq 2a\},$$

and the notation $\overline{Z}'_{\tilde{t},g} := \overline{Z}_{\tilde{t},g} - (\mu_1 - \mu_2)/\sqrt{2}$, for $g \in [G_{\tilde{t}}]$ introduced at the start of the alternative term part of the proof of Theorem 4. We first consider the case $t_0 \geq 2$, which implies $\tilde{t} \geq 2$. For $J \subseteq [p]$, we further denote

$$A_{t,*,J}^{\text{MoM}'} = \frac{t}{2} \cdot \text{median}\left\{\sum_{j \in J}\left(\overline{Z}_{t,g,1}^2(j) - \frac{(\mu_1(j) - \mu_2(j))^2}{2} - \frac{2G_t}{t}\right) : g \in [G_t]\right\}.$$

Observe that for $g \in [G_{\tilde{t}}]$

$$\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}\left(\overline{Z}_{\tilde{t},g,1}^2(j) - \frac{(\mu_1(j) - \mu_2(j))^2}{2} - \frac{2G_{\tilde{t}}}{\tilde{t}}\right) - \sum_{j \in \mathcal{J}_{\tilde{t},a}} V_{\tilde{t},g,a}(j)$$

$$= -\frac{\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}(\mu_1(j) - \mu_2(j))^2}{2} + \sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} V_{\tilde{t},g,a}(j) - \sum_{j \in \mathcal{J}_{\tilde{t},a}} V_{\tilde{t},g,a}(j)$$

$$\leq -\frac{\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}(\mu_1(j) - \mu_2(j))^2}{2} + \frac{2G_{\tilde{t}}|\mathcal{J}_{\tilde{t},a}|}{\tilde{t}}.$$

Then, on the event $\{|\mathcal{J}_{\tilde{t},a}| \leq 2s\} \cap \left\{\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 \geq \frac{\|\delta\|_2^2}{12\log(8/\varepsilon)}\right\}$, by Lemma 38, we deduce

$$A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}^{\text{MoM}'} \leq A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a}}^{\text{MoM}} - \frac{\|\delta\|_2^2}{12\log(8/\varepsilon)} + 2sG_{\tilde{t}},$$

and consequently, when $\rho^2 \geq 192C_2 s\Delta\log(8/\varepsilon)$, we have, with $C_2 \geq 2$, that

$$\frac{\|\delta\|_2^2}{24\log(8/\varepsilon)} \geq C_2 s\Delta \geq \max_{t \in \mathcal{T} \setminus \{1\}}\{r_t + 2sG_t\},$$

where the first inequality is due to $\|\delta\|_2^2 \geq \rho^2/8$ and the second inequality is due to the choice of $G_t = (t \wedge \Delta)/2$. Hence

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}) \leq \mathbb{P}_\theta\left(A_{\tilde{t},a}^{\text{MoM}} \leq r_{\tilde{t}}\right) = \mathbb{P}_\theta\left(A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a}}^{\text{MoM}} \leq r_{\tilde{t}}\right)$$

$$\leq \mathbb{P}_\theta\left(|\mathcal{J}_{\tilde{t},a}| > 2s\right) + \mathbb{P}_\theta\left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12\log(8/\varepsilon)}\right) + \mathbb{P}_\theta\left(A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}^{\text{MoM}'} \leq -\frac{\|\delta\|_2^2}{24\log(8/\varepsilon)}\right). \tag{75}$$

We control the three terms respectively. The arguments below mirror those made in the proof of Theorem 2 between (55) and (61) and we will omit details in places where the same reasoning is used in the last proof. First, it remains true that

$$\mathbb{P}_\theta\left(|\mathcal{J}_{\tilde{t},a}| > 2s\right) \leq \varepsilon/8.$$

By (72) and the choice $a \geq \{3K(512\log(8/\varepsilon))^{1/\alpha}\} \vee \{2\alpha e^{\alpha/2}\log^{1/2}(700\log(8/\varepsilon))\}$, we have for all $j \in \mathcal{H}_{\delta,a}$ that

$$\mathbb{P}_\theta(j \notin \mathcal{J}_{\tilde{t},a}) = \mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) \leq \frac{K^\alpha}{((|\delta(j)| - a)/3)^\alpha} + \exp\left\{1 - \frac{(|\delta(j)| - a)^2}{2\alpha^2 e^\alpha}\right\} \leq \frac{1}{256\log(8/\varepsilon)},$$

and thus again

$$\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\big(\delta(j)^2 \mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a}\}}\big) \leq \frac{\|\delta\|_2^4}{256 \log(8/\varepsilon)}.$$

At this point, we consider

$$\rho^2 \geq C_3 \max\Big\{192 C_2 s\Delta \log^2(8/\varepsilon), 64a^2 s, 3456 K^2 (16/\varepsilon)^{2/\alpha} \log(8/\varepsilon), 768\alpha^2 e^\alpha \log^2(16e/\varepsilon)\Big\},$$

with some $C_3 \geq 1$. Then, by repeating the argument in (58), (59) and (60), we obtain

$$\mathbb{P}_\theta\left(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)}\right) \leq \varepsilon/8.$$

We now bound the third and final term in (75). By Chebyshev's inequality, we deduce that for $g \in [G_t]$

$$\mathbb{P}\left(\frac{\tilde{t}}{2} \sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \left(\overline{Z}_{\tilde{t},g,1}^2(j) - \frac{(\mu_1(j) - \mu_2(j))^2}{2} - \frac{2G_{\tilde{t}}}{\tilde{t}}\right) \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)}\right)$$

$$\leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\left(\overline{Z}_{\tilde{t},g,1}^2(j)\right)}{\|\delta\|_2^4/(144 \tilde{t}^2 \log^2(8/\varepsilon))} \leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} 2\tilde{t}^2 \mathrm{Var}_\theta\left(\left(\overline{Z}_{\tilde{t},g,1}'(j)\right)^2\right) + \sum_{j \in \mathcal{H}_{\delta,a}} 16\tilde{t}\delta(j)^2 \mathrm{Var}_\theta\left(\overline{Z}_{\tilde{t},g,1}'(j)\right)}{\|\delta\|_2^4/(144 \log^2(8/\varepsilon))}$$

$$\leq C_4 \log^2(8/\varepsilon)\left(\frac{s\Delta^2}{\rho^4} + \frac{\Delta}{\rho^2}\right) \leq \varepsilon/48,$$

when $C_3 \geq 1$ is sufficiently large. The third inequality above follows from (67) and C_4 is a constant depending only on α and K. If $\tilde{t} = 2$, then $G_{\tilde{t}} = 1$ and we immediately have

$$\mathbb{P}_\theta\left(A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}^{\mathrm{MoM}'} \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)}\right) \leq \varepsilon/48.$$

If $\tilde{t} > 2$, then $G_{\tilde{t}} \geq 2$ and we again use the binomial tail bound argument as in (64) to obtain

$$\mathbb{P}_\theta\left(A_{\tilde{t},*,\mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}}^{\mathrm{MoM}'} \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)}\right) \leq \exp\left\{-\frac{\varepsilon G_{\tilde{t}}}{48}\left(\frac{24}{\varepsilon} \log\left(\frac{24}{\varepsilon}\right) - \frac{24}{\varepsilon} + 1\right)\right\} \leq \frac{\varepsilon}{8}.$$

By (75), we conclude $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM}}) \leq \varepsilon/2$. Finally, for the case that the mean change happens at $t_0 = 1$ instead, similar to the last paragraph of the proof of Theorem 2, we can still control the three terms in (75) in the same way respectively when we redefine $\mathcal{J}_{\tilde{t}=1,a} := \{j \in [p] : |Y_1(j)| \geq a\}$ instead.

5) Proof of Theorem 5: We actually prove a more general result. Any mean estimator that satisfies the following condition can be used in place of $\hat{\mu}_{n,s,\eta}^{\mathrm{RSM}}(\cdot) = \hat{\mu}_{n,s}^{\mathrm{RSM}}(\cdot; \eta)$ introduced in Section III-B while Theorem 5 still holds.

Condition 1. *Assume $\alpha \geq 4$. Let $W_1, \ldots, W_n$ be independent random vectors in $\mathbb{R}^p$, each with mean μ_W and covariance matrix I_p. Assume $\|\mu_W\|_0 \leq s$ and $\mathbb{E}|W_i(j) - \mu_W(j)|^\alpha \leq (\sqrt{2}K)^\alpha$ for $i \in [n]$ and $j \in [p]$. Then there exist constants $C_1, C_2 \geq 1$, depending only on α and K such that for any given $0 < \eta < 1$, when $n \geq C_1\big(s \log(ep/s) + \log(1/\eta)\big)$, then with probability at least $1 - \eta$, we have*

$$\big\|\hat{\mu}_{n,s}^{\mathrm{RSM}}(W_1, \ldots, W_n; \eta) - \mu_W\big\|_2 \leq \sqrt{C_2}\left(\sqrt{\frac{s \log(ep/s)}{n}} + \sqrt{\frac{\log(1/\eta)}{n}}\right).$$

In particular, the robust sparse mean estimator that we use from [36] satisfies the condition above as shown in Corollary 11[3] therein.

In the rest of the proof, we denote $\tilde{\mathcal{T}}_1 := \{t \in \mathcal{T} : t \leq \tilde{\Delta}_1\}$, $\tilde{\mathcal{T}}_2 := \{t \in \mathcal{T} : \tilde{\Delta}_1 < t \leq \tilde{\Delta}_2\}$ and $\tilde{\mathcal{T}}_3 := \{t \in \mathcal{T} : t > \tilde{\Delta}_2\}$ and recall that $\mathcal{J}_{t,a} = \{j \in [p] : |Y_{t,2}(j)| \geq a\}$ for $t \in \mathcal{T} \setminus \{1\}$.

Null term. For $\theta \in \Theta_0(p, n)$, we have

$$\mathbb{E}_\theta \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{RSM}} = \mathbb{P}_\theta(A_{1,a} > \tilde{r}_1) + \sum_{t \in \tilde{\mathcal{T}}_1 \setminus \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t)$$

$$+ \sum_{t \in \tilde{\mathcal{T}}_2} \mathbb{P}_\theta(A_t^{\mathrm{RSM}} > r_t^{\mathrm{RSM}}) + \sum_{t \in \tilde{\mathcal{T}}_3} \mathbb{P}_\theta(A_t^{\mathrm{RSM}} > r_t^{\mathrm{RSM}}). \tag{76}$$

---

[3]Note that their result is under the assumption that for each vector v with $\|v\|_2 = 1$, $\mathbb{E}(v^\top(W - \mu_W))^\alpha \leq C[\mathbb{E}(v^\top(W - \mu_W))^2]^{\alpha/2}$ for some absolute constant C, which is certainly satisfied by our assumption $\mathbb{E}|W(j) - \mu_W(j)|^\alpha \leq (\sqrt{2}K)^\alpha$ for $j \in [p]$ in Condition 1 with $C = (\sqrt{2}K)^\alpha$.(Rio09-(1.2))

For the first term, similar to the proof of Theorem 2, by Proposition 34(b), when $\tilde{r}_1 \geq C_4' s(p/s)^{2/\alpha}$, for some large enough $C_4' > 0$, depending only on α, K and ε, we have $\mathbb{P}_\theta(A_{1,a} > \tilde{r}_1) \leq \varepsilon/8$. To control the second term in (76), we closely follow the arguments in the null term part of the proof of Theorem 2 and Proposition 7. By (46), we have

$$\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t) \leq \sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \sup_{J \subseteq [p]:|J| \leq s} \mathbb{P}_\theta\left(\sum_{j \in J}(Y_{t,1}^2(j) - 1) > \tilde{r}_t\right). \tag{77}$$

For the first term on the right hand side, by (73), we obtain

$$\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \left(\frac{2epK}{s(a/3)^\alpha}\right)^s + \log_2(\tilde{\Delta}_1)\left(\frac{2e^2p}{s}\right)^s \exp\left\{-\frac{sa^2}{2\alpha^2 e^\alpha}\right\}.$$

The choice of a in (21) with a large enough constant $C_3 > 0$ guarantees that $\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \varepsilon/16$. For the second term, we fix $J \subseteq [p]$ with $|J| \leq s$. By the same technique as in (63), we obtain

$$\mathbb{P}_\theta\left(\sum_{j \in J}(Y_{t,1}^2(j) - 1) > \tilde{r}_t\right) \leq \frac{\varepsilon}{16 \log_2(\tilde{\Delta}_1)},$$

when $\tilde{r}_t = C_4\sqrt{s \log \tilde{\Delta}_1}$, for some large enough $C_4 > 0$, depending only on α, K and ε. Thus, we deduce $\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t) \leq \varepsilon/8$.

Now, we control the third and fourth terms in (76). For $t \in \tilde{\mathcal{T}}_2 \cup \tilde{\mathcal{T}}_3$, we observe that

$$C_1\big(s \log(ep/s) + \log(1/\eta_t)\big) = \min(t, \tilde{\Delta}_2) \leq t.$$

Since $Z_1, \ldots, Z_t$ are independent and identically distributed random vectors with mean 0 and covariance matrix I_p and satisfy $\mathbb{E}|Z_i(j)|^\alpha \leq 2^{\alpha/2}K^\alpha$ for $i \in [t], j \in [p]$ under the null, by Condition 1, we obtain

$$\mathbb{P}_\theta(A_t^{\text{RSM}} > r_t^{\text{RSM}}) = \mathbb{P}_\theta\Big(t\big\|\hat{\mu}_{t,s,\eta_t}^{\text{RSM}}\big\|_2^2 > 2C_2\big(s \log(ep/s) + \log(1/\eta_t)\big)\Big) \leq \eta_t,$$

and therefore,

$$\sum_{t \in \tilde{\mathcal{T}}_2} \mathbb{P}_\theta(A_t^{\text{RSM}} > r_t^{\text{RSM}}) + \sum_{t \in \tilde{\mathcal{T}}_3} \mathbb{P}_\theta(A_t^{\text{RSM}} > r_t^{\text{RSM}}) \leq \sum_{t \in \tilde{\mathcal{T}}_2} \exp\left\{s \log(ep/s) - \frac{t}{C_1}\right\} + \sum_{t \in \tilde{\mathcal{T}}_3} \frac{\varepsilon}{16 \log 2n}$$

$$\leq 2\exp\left\{s \log(ep/s) - \frac{\tilde{\Delta}_1}{C_1}\right\} + \frac{\varepsilon \log_2(n/2)}{16 \log 2n} < \varepsilon/4, \tag{78}$$

where we use Lemma 35 in the second inequality. Hence, we conclude that $\mathbb{E}_\theta \phi_{\mathcal{P},\text{sparse}}^{\text{RSM}} \leq \varepsilon/2$ for all $\theta \in \Theta_0(p, n)$.

Alternative term. As in all previous proofs of alternative term, we consider the unique $\tilde{t} \in \mathcal{T}$, such that $t_0/2 \leq \tilde{t} \leq t_0$, where $t_0(\leq n/2)$ is the true change point location. When $t_0 = 1$, we simply use the final paragraph of the proof of Proposition 7. When $t_0 \geq 2$, we consider separately the two cases $\tilde{t} \in \tilde{\mathcal{T}}_1 \backslash \{1\}$ and $\tilde{t} \in \tilde{\mathcal{T}}_2 \cup \tilde{\mathcal{T}}_3$. When $\tilde{t} \in \tilde{\mathcal{T}}_1 \backslash \{1\}$, the arguments are again almost the same as those used in the alternative term part of the proof of Proposition 7. We thus omit the details and directly state the conclusion: as long as

$$\rho^2 \geq C_6 \max\big\{(\tilde{r}\mathbb{1}_{\{t \neq 1\}} + 2s) \log^2(8/\varepsilon), a^2 s, (1/\varepsilon)^{2/\alpha} \log(8/\varepsilon)\big\},$$

for some large enough $C_6 > 0$, depending only on α and K, we have $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}) \leq \varepsilon/2$. Note that if $\rho^2 \geq C_5 v_{\mathcal{P},\text{sparse}}^{\text{U}}$, for some large enough $C_5 > 0$, depending only on α, K and ε, then the above condition is satisfied.

If $\tilde{t} \in \tilde{\mathcal{T}}_2 \cup \tilde{\mathcal{T}}_3$ instead, then $Z_1, \ldots, Z_t$ are independent and identically distributed random vectors with mean $(\mu_1 - \mu_2)/\sqrt{2}$ and covariance matrix I_p and satisfy $\mathbb{E}\big|Z_i(j) - \frac{\mu_1(j) - \mu_2(j)}{\sqrt{2}}\big|^\alpha \leq 2^{\alpha/2}K$ for $i \in [t], j \in [p]$. Recall that $\tilde{t}\|\mu_1 - \mu_2\|_2^2 \geq \rho^2/2$. Hence, when $\rho^2 \geq 24C_2\big(s \log(ep/s) + \log(16 \log(2n)/\varepsilon)\big)$, we have by Condition 1 that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}) = \mathbb{P}_\theta(A_{\tilde{t},a}^{\text{RSM}} \leq r_{\tilde{t}}^{\text{RSM}}) = \mathbb{P}_\theta\Big(\tilde{t}\big\|\hat{\mu}_{\tilde{t},s,\eta_{\tilde{t}}}^{\text{RSM}}\big\|_2^2 \leq 2C_2\big(s \log(ep/s) + \log(1/\eta_{\tilde{t}})\big)\Big)$$

$$\leq \mathbb{P}_\theta\left(\sqrt{\tilde{t}}\Big\|\frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2 - \Big\|\hat{\mu}_{\tilde{t},s,\eta_{\tilde{t}}}^{\text{RSM}} - \frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2 \leq \sqrt{2C_2}\big(\sqrt{s \log(ep/s)} + \sqrt{\log(1/\eta_{\tilde{t}})}\big)\right)$$

$$\leq \mathbb{P}_\theta\left(\sqrt{\tilde{t}}\Big\|\hat{\mu}_{\tilde{t},s,\eta_{\tilde{t}}}^{\text{RSM}} - \frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2 > \sqrt{C_2}\big(\sqrt{s \log(ep/s)} + \sqrt{\log(1/\eta_{\tilde{t}})}\big)\right) \leq \eta_{\tilde{t}}$$

$$\leq \exp\left\{s \log(ep/s) - \frac{\min(t, \tilde{\Delta}_2)}{C_1}\right\} \leq \frac{\varepsilon}{16},$$

as desired.

6) Proof of Corollary 8: We first consider the statistical property of $\phi_{\mathcal{P},\text{sparse}}$. By comparing the two rates $v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}}$ and $v_{\mathcal{P},\text{sparse}}^{\text{U}}$, we note that the improvement offered by $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ over $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ only exists when

$$(p/s)^{2/\alpha} < \log\log(8n),$$

since otherwise $v_{\mathcal{P},\text{sparse}}^{\text{U,MoM}} = v_{\mathcal{P},\text{sparse}}^{\text{U}}$. Combining this with the fact that we are in the sparse regime $s < p^{(\alpha-4)/(2\alpha-4)}$, we deduce that $p < \log^{\alpha-2}(\log(8n))$. The desired result is then an immediate consequence of Theorem 5 and Proposition 7.

Now onto the computational complexity claim. For each $t \in \mathcal{T}$, computing the statistics $A_{t,a}^{\text{MoM}}$ and $A_{t,a}$ in $\phi_{\mathcal{P},\text{sparse}}^{\text{MoM}}$ and $\phi_{\mathcal{P},\text{sparse}}^{\text{RSM}}$ take time polynomial in n and p since they only involve performing basic operations and finding the median of $G_t \leq 8\log\log(8n)$ quantities. The computationally demanding part lies in computing A_t^{RSM}, or equivalently the robust sparse mean estimator $\hat{\mu}_{t,s,\eta_t}^{\text{RSM}}$. Note that we are using this only when $p < \log^{\alpha-2}(\log(8n))$. For each fixed t, we claim that the computation/approximation of $\hat{\mu}_{t,s,\eta_t}^{\text{RSM}}$ can be performed in time that is polynomial in n. We now show this by arguing that each component below has time complexity that is polynomial in n. In the rest of the proof, we omit the subscripts and adopt the notation $\hat{\mu}^{\text{RSM}}$ for clarity.

1) Each evaluation of the function $1\text{DRobust}(\cdot)$ in Algorithm 2 of [36] of t data point requires time of order $t\log t \leq n\log n$ (in order to find the shortest interval).
2) The total number of projection $|\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})|$ can be bounded by $|\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})| \leq (6ep/s)^s \leq (6ep)^p \leq \exp(6ep^2) \leq \exp(C_\alpha \log(n)) = n^{C_\alpha}$ for some constant $C_\alpha > 0$, depending only on α. Denote

$$g(\mu) := \max_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^\top \mu - 1\text{DRobust}(\{u^\top Z_i\}_{i=1}^t, \eta_t/(6ep/s)^s)|.$$

Thus for a fixed $\mu \in \mathbb{R}^p$, the computational complexity of evaluating $g(\mu)$ is polynomial in n.
3) The optimisation problem defining $\hat{\mu}^{\text{RSM}}$ can be written as

$$\min_{\mu \in \mathcal{L}_s} g(\mu).$$

We solve this by first considering each possible s-sparsity coordinate pattern individually before working out the minimum among these $\binom{p}{s} \leq n^{C_\alpha}$ minima.
4) Fix $\mathcal{U} \subseteq \mathbb{R}^p$ with $|\mathcal{U}| = s$. We solve the optimisation problem

$$\min_{\mu \in \mathbb{R}^p : \mu(j)=0 \, \forall j \in \mathcal{U}^c} g(\mu)$$

by subgradient descent. Denote the optimal value to be $g_{*,\mathcal{U}}$ and the k-th iterate to be $\mu_{\mathcal{U}}^{(k)}$. Note that $g(\mu)$ is 1-Lipschitz and $\partial g(\mu) \subseteq \{\pm u : u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})\}$. Standard result on the convergence of subgradient descent, e.g. Theorem 3.2.2 of [48], shows that $\left(\min_{k \in [K]} g(\mu_{\mathcal{U}}^{(k)})\right) - g_{*,\mathcal{U}} \leq \upsilon$ in $K \asymp 1/\upsilon^2$ steps, where we choose $\upsilon = \sqrt{s\log(ep/s)t^{-1}}$. The computational complexity is again at most polynomial in n. Denote $\tilde{\mu}_{\mathcal{U}}^{\text{RSM}}$ to be the update that attains the best objective value in K iterations.

Write

$$\tilde{\mu}^{\text{RSM}} := \operatorname*{argmin}_{\mu \in \left\{\tilde{\mu}_{\mathcal{U}}^{\text{RSM}} : |\mathcal{U}|=s\right\}} g(\mu),$$

as our final estimator (an approximation of $\hat{\mu}^{\text{RSM}}$). We have now shown that $\tilde{\mu}^{\text{RSM}}$ can be obtained in time that is polynomial in n. Finally, we prove that $\tilde{\mu}^{\text{RSM}}$ still satisfies Condition 1. Indeed, following the proof of Lemma 4 and Corollary 12 in [36], we have

$$\|\tilde{\mu}^{\text{RSM}} - \mu_Z\|_2 \leq \|\tilde{\mu}^{\text{RSM}} - \hat{\mu}^{\text{RSM}}\|_2 + \|\hat{\mu}^{\text{RSM}} - \mu_Z\|_2 \leq g(\tilde{\mu}^{\text{RSM}}) + g(\hat{\mu}^{\text{RSM}}) + g(\hat{\mu}^{\text{RSM}}) + g(\mu_Z)$$

$$\leq g(\hat{\mu}^{\text{RSM}}) + \upsilon + 2g(\hat{\mu}^{\text{RSM}}) + g(\mu_Z) \leq \upsilon + 4g(\mu_Z)$$

$$\leq \sqrt{C}\left(\sqrt{\frac{s\log(ep/s)}{t}} + \sqrt{\frac{\log(1/\eta_t)}{t}}\right),$$

for some $C \geq 1$, where $\mu_Z = \mathbb{E}Z_1$.

B. *Proof of the adaptation result in Section IV*

1) Adaptive testing procedure: We focus on the case when $P_e \in \mathcal{P}_{\alpha,K}^{\otimes}$ for $\alpha > 4$. We introduce an adaptive testing procedure based on these two tests:

$$\phi_{\mathcal{P},\text{adaptive}} := \phi_{\mathcal{P},\text{dense}} \vee \max_{s \in \mathcal{K}} \phi_{\mathcal{P},\text{sparse},s}$$

$$= \begin{cases} \phi_{\mathcal{P},\text{dense}} \vee \max_{s \in \mathcal{K}} \phi_{\mathcal{P},\text{sparse},s}^{\text{MoM}}, & \text{if } p > \log^{\alpha-2}(\log(8n)), \\ \phi_{\mathcal{P},\text{dense}} \vee \max_{s \in \mathcal{K}} \phi_{\mathcal{P},\text{sparse},s}^{\text{RSM}}, & \text{if } p \leq \log^{\alpha-2}(\log(8n)), \end{cases} \tag{79}$$

where the dependence of $\phi_{\mathcal{P},\text{sparse}}$ on s is made explicit by writing it as $\phi_{\mathcal{P},\text{sparse},s}$, and the set $\mathcal{K} := \{1, 2, 4, \ldots, 2^{\lceil \log_2(p)\rceil - 1}\}$ is a dyadic grid. Recall that $\phi_{\mathcal{P},\text{dense}}$ does not require the knowledge of s, and we keep its original parameter choices as in (18), with a potentially larger constant C_1 in r_t:

$$r_t = C_1 p^{(1/2)\vee(2/\alpha)} G_t, \quad G_t = t \wedge \Delta \quad \text{and} \quad \Delta = 2^{3 + \lceil \log_2 \log\log(8n)\rceil}. \tag{80}$$

For $\phi_{\mathcal{P},\text{sparse},s}^{\text{MoM}}$, we modify the original parameter choices (29) as follows:

$$a_s = C_2\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log(8n))\big), \quad r_{t,s} = C_3\big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + s^{3/4}G_t\mathbb{1}_{\{t>1\}}\big),$$
$$G_t = (t \wedge \Delta)/2 \quad \text{and} \quad \Delta = 2^{4 + \lceil \log_2 \log\log(8n)\rceil}. \tag{81}$$

Comparing with (29), we use the same a_s (but potentially larger constants) and modify $r_{t,s}$. Finally, for $\phi_{\mathcal{P},\text{sparse},s}^{\text{RSM}}$, we modify its original parameter choices (21) to be:

$$a_s = C_4\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log \tilde{\Delta}_{1,s})\big), \quad \tilde{r}_{t,s} = C_5\big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + s^{3/4}\sqrt{\log \tilde{\Delta}_1}\mathbb{1}_{\{t>1\}}\big),$$
$$\eta_{t,s} = \exp\bigg\{s\log(ep/s) - \frac{t \wedge \tilde{\Delta}_2}{C_6}\bigg\}, \quad r_{t,s}^{\text{RSM}} = C_7(t \wedge \tilde{\Delta}_{2,s}), \tag{82}$$
$$\tilde{\Delta}_{1,s} = C_6\big(s\log(ep/s) + \log(80s/\varepsilon)\big) \quad \text{and} \quad \tilde{\Delta}_{2,s} = C_6\big(s\log(ep/s) + \log(80s\log(2n)/\varepsilon)\big).$$

2) Proof of Theorem 9: We prove the following theorem on the theoretical guarantee of the test $\phi_{\mathcal{P},\text{adaptive}}$, fully constructed and specified above in Appendix B1. Theorem 9 then follows as an immediate consequence.

Theorem 16. *Assume $\alpha \geq 4$. For any $\varepsilon \in (0,1)$, there exist $C_1, \ldots, C_8 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{P},\text{adaptive}}$ defined in (79) with its parameters specified in (80), (81) and (82) satisfies*

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\text{adaptive}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_8\big(v_{\mathcal{P},\text{dense}}^{\text{U}} \wedge v_{\mathcal{P},\text{sparse}}^{\text{U}}\big)$.

Proof. This proof is based on the proofs of Theorem 4, Theorem 5 and Proposition 7. For brevity, we only highlight the main steps and differences.

Null term. By a union bound, (70) and (76), we have

$$\mathbb{E}_\theta \phi_{\mathcal{P},\text{adaptive}}$$
$$\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \mathbb{E}_\theta\bigg[\max_{s\in\mathcal{K}} \phi_{\mathcal{P},\text{sparse},s}^{\text{MoM}}\bigg]\mathbb{1}_{\{p>\log^{\alpha-2}(\log(8n))\}} + \mathbb{E}_\theta\bigg[\max_{s\in\mathcal{K}} \phi_{\mathcal{P},\text{sparse},s}^{\text{RSM}}\bigg]\mathbb{1}_{\{p\leq\log^{\alpha-2}(\log(8n))\}}$$
$$\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \bigg(\mathbb{P}_\theta\bigg(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{r_{1,s}} > 1\bigg) + \sum_{s\in\mathcal{K}}\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\big(A_{t,a_s,s}^{\text{MoM}} > r_{t,s}\big)\bigg)\mathbb{1}_{\{p>\log^{\alpha-2}(\log(8n))\}}$$
$$+ \bigg(\mathbb{P}_\theta\bigg(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{\tilde{r}_{1,s}} > 1\bigg) + \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{1,s}\setminus\{1\}} \mathbb{P}_\theta(A_{t,a_s,s} > \tilde{r}_{t,s})$$
$$+ \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \mathbb{P}_\theta(A_{t,s}^{\text{RSM}} > r_{t,s}^{\text{RSM}}) + \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{3,s}} \mathbb{P}_\theta(A_{t,s}^{\text{RSM}} > r_{t,s}^{\text{RSM}})\bigg)\mathbb{1}_{\{p\leq\log^{\alpha-2}(\log(8n))\}}$$
$$\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \mathbb{P}_\theta\bigg(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{r_{1,s} \wedge \tilde{r}_{1,s}} > 1\bigg) + \sum_{s\in\mathcal{K}}\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\big(A_{t,a_s,s}^{\text{MoM}} > r_{t,s}\big)$$
$$+ \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{1,s}\setminus\{1\}} \mathbb{P}_\theta(A_{t,a_s,s} > \tilde{r}_{t,s}) + \sum_{s\in\mathcal{K}}\bigg(\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \mathbb{P}_\theta(A_{t,s}^{\text{RSM}} > r_{t,s}^{\text{RSM}}) + \sum_{t\in\tilde{\mathcal{T}}_{3,s}} \mathbb{P}_\theta(A_{t,s}^{\text{RSM}} > r_{t,s}^{\text{RSM}})\bigg), \tag{83}$$

where we denote $\tilde{\mathcal{T}}_{1,s} := \{t \in \mathcal{T} : t \leq \tilde{\Delta}_{1,s}\}$, $\tilde{\mathcal{T}}_{2,s} := \{t \in \mathcal{T} : \tilde{\Delta}_{1,s} < t \leq \tilde{\Delta}_{2,s}\}$ and $\tilde{\mathcal{T}}_{3,s} := \{t \in \mathcal{T} : t > \tilde{\Delta}_{2,s}\}$. In the following, we bound each of the five terms in (83) by $\varepsilon/10$.

Term 1. By closely following the null term part of the proof of Theorem 4, with a sufficiently large constant C_1, we deduce, similar to (65), that

$$\mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} \leq \varepsilon/180 + \frac{(32/\varepsilon)^{-1}}{1-(32/\varepsilon)^{-1}} + \log_2(n/2)(32/\varepsilon)^{-\Delta/2} \leq \varepsilon/180 + \varepsilon/31 + \varepsilon/31 < \varepsilon/10.$$

Term 2. By having C_3 and C_5 sufficiently large, by Proposition 34(c), we can control this term at level $\varepsilon/10$.

Term 3. For this, we follow the null term part of the proof of Proposition 7. The key step in that proof was to bound both terms

in (71). The first term can be controlled via (73). A careful inspection reveals that the condition on a (same as a_s here) given in (74) with a possibly larger value of the leading constant can guarantee the control of both terms in (73) at $\varepsilon/(160s^{1/2})$. Bounding the second term in (71) required the argument from (63) to (65), within which the dimension p was replaced by s. Our new choice of $r_{t,s} = C_3 s^{3/4} G_t$ for $t > 1$ with a sufficiently large C_3 allows us to have $\varepsilon/(1100s^{1/2})$ as the RHS bound in (63) (dimension being s). Correspondingly, the RHS of (64) now becomes

$$\exp\left\{-\frac{\varepsilon G_t}{1100\sqrt{s}}\left(\frac{550\sqrt{s}}{\varepsilon}\log\left(\frac{550\sqrt{s}}{\varepsilon}\right) - \frac{550\sqrt{s}}{\varepsilon} + 1\right)\right\} \le \exp\left\{-\frac{G_t}{2}\log\left(200\sqrt{s}/\varepsilon\right)\right\}.$$

Thus, the second term in (71) can now be bounded instead by

$$\frac{\varepsilon}{1100\sqrt{s}} + \frac{(200\sqrt{s}/\varepsilon)^{-1}}{1 - (200\sqrt{s}/\varepsilon)^{-1}} + \log_2(n/2)(200\sqrt{s}/\varepsilon)^{-\Delta/4} \le \frac{\varepsilon}{1100\sqrt{s}} + \frac{\varepsilon}{199\sqrt{s}} + \frac{\varepsilon}{199\sqrt{s}} < \frac{\varepsilon}{80\sqrt{s}}.$$

Putting everything together, we conclude that

$$\sum_{s\in\mathcal{K}}\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\left(A_{t,a_s,s}^{\mathrm{MoM}} > r_{t,s}\right) \le \sum_{s\in\mathcal{K}} \frac{\varepsilon}{40\sqrt{s}} < \frac{\varepsilon}{10}.$$

Term 4. We follow the null term part of the proof of Theorem 5. More specifically, this term can be split into two terms according to (77). Similar to the argument made for the second term above, with C_4 sufficiently large, the first term in (77) can be guaranteed to be at most $\varepsilon/(80s^{1/2})$. The second term, with the new choice of $\tilde{r}_{t,s}$ and its leading constant C_5 being sufficiently large, can also be bounded above by

$$\frac{\varepsilon|\tilde{\mathcal{T}}_{1,s}\setminus\{1\}|}{80\sqrt{s}\log_2(\tilde{\Delta}_{1,s})} \le \varepsilon/(80s^{1/2}).$$

Therefore, we can again control the fourth term at level $\varepsilon/10$.

Term 5. We again follow the null term part of the proof of Theorem 5. By Condition 1 and similar to (78), we can now bound

$$\sum_{s\in\mathcal{K}}\left(\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \mathbb{P}_\theta(A_{t,s}^{\mathrm{RSM}} > r_{t,s}^{\mathrm{RSM}}) + \sum_{t\in\tilde{\mathcal{T}}_{3,s}} \mathbb{P}_\theta(A_{t,s}^{\mathrm{RSM}} > r_{t,s}^{\mathrm{RSM}})\right)$$

$$\le \sum_{s\in\mathcal{K}}\left(\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \exp\left\{s\log(ep/s) - \frac{t}{C_6}\right\} + \sum_{t\in\tilde{\mathcal{T}}_{3,s}} \exp\left\{s\log(ep/s) - \frac{\tilde{\Delta}_{2,s}}{C_6}\right\}\right)$$

$$\le \sum_{s\in\mathcal{K}}\left(2\exp\left\{s\log(ep/s) - \frac{\tilde{\Delta}_{1,s}}{C_6}\right\} + \log_2(n/2)\frac{\varepsilon}{80s\log n}\right) < \sum_{s\in\mathcal{K}} \frac{\varepsilon}{20s} \le \frac{\varepsilon}{10},$$

as desired.

Alternative term. First, let s_1 satisfy $s_1(p/s_1)^{2/\alpha} + \log\log(8n) = \sqrt{p}\log\log(8n)$ and s_2 satisfy $s_2\big((p/s_2)^{2/\alpha} + \log\log(8n)\big) = \sqrt{p}\log\log(8n)$. Note that $s_1 \ge s_2$. For $\theta\in\Theta(p,n,s,\rho)$, we consider all four possible (p,s) regimes below.

(1) $p \ge \log^{\alpha-2}(\log(8n))$ and $s \ge s_2/2$. We have $\mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{adaptive}}) \le \mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{dense}})$. By the alternative term part of the proof of Theorem 4, we can bound the above quantity by $\varepsilon/2$ as long as $\rho^2 \ge C'\sqrt{p}\log\log(8n)$ with a sufficiently large C'. We also note that when $s_2/2 \le s < s_2$, we have

$$\frac{1}{2}\sqrt{p}\log\log(8n) \le s\big((p/s)^{2/\alpha} + \log\log(8n)\big) \le \sqrt{p}\log\log(8n).$$

(2) $p \ge \log^{\alpha-2}(\log(8n))$ and $s < s_2/2$. By the definition of $\mathcal{K}$, there exists an $\tilde{s}\in\mathcal{K}$ such that $s \le \tilde{s} < 2s$. We have $\mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{adaptive}}) \le \mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{sparse},\tilde{s}}^{\mathrm{MoM}})$. Now, by carefully inspecting the alternative term part of the proof of Proposition 7, we can still deduce $\mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{sparse},\tilde{s}}^{\mathrm{MoM}}) \le \varepsilon/2$ as long as ρ satisfies

$$\rho^2 \ge C'\Big(s\big((p/s)^{2/\alpha} + \log\log(8n)\big)\Big) \ge \frac{C'}{2}\Big(\tilde{s}\big((p/\tilde{s})^{2/\alpha} + \log\log(8n)\big)\Big)$$

$$\ge C''\max\Big\{\log^2(8/\varepsilon)\max_{t\in\mathcal{T}\setminus\{1\}}(r_{t,\tilde{s}} + 2\tilde{s}G_t), (r_{1,\tilde{s}} + 2\tilde{s})\log^2(8/\varepsilon), a_{\tilde{s}}^2\tilde{s}\Big\}, \tag{84}$$

for sufficiently large C' and C'', where the final inequality in (84) remains true with our modified choice of $r_{t,s}$.

(3) $p < \log^{\alpha-2}(\log(8n))$ and $s \ge s_1/2$. We use the same argument as in (1) to obtain the same condition $\rho^2 \ge C'\sqrt{p}\log\log(8n)$. Similarly, we also note that when $s_1/2 \le s < s_1$, we have

$$\frac{1}{2}\sqrt{p}\log\log(8n) \le s(p/s)^{2/\alpha} + \log\log(8n) \le \sqrt{p}\log\log(8n).$$

(4) $p < \log^{\alpha-2}(\log(8n))$ and $s < s_1/2$. Similar to (2), we have $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{adaptive}}) \le \mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{sparse},\tilde{s}}^{\text{RSM}})$. By carefully examining the alternative term part of the proof of Theorem 5, we can obtain $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{sparse},\tilde{s}}^{\text{RSM}}) \le \varepsilon/2$ as long as

$$\rho^2 \ge C'\big(s(p/s)^{2/\alpha} + \log\log(8n)\big) \ge \frac{C'}{2}\big(\tilde{s}(p/\tilde{s})^{2/\alpha} + \log\log(8n)\big)$$
$$\ge C'' \max\big\{(\tilde{r}_{t\neq1,\tilde{s}} + 2\tilde{s})\log^2(8/\varepsilon), (\tilde{r}_{1,\tilde{s}} + 2\tilde{s})\log^2(8/\varepsilon), a_s^2\tilde{s}, \tilde{s}\log(ep/\tilde{s}) + \log\log(8n)\big\}, \tag{85}$$

for sufficiently large C' and C'', where the final inequality in (85) remains true with our new choices of a_s and $\tilde{r}_{t,s}$.

The desired result then follows from Corollary 8 and the first part of its proof. $\qquad\square$

C. Proofs of lower bound results in Sections II and III

For $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$, to prove Theorem 3, we establish the lower bound separately for the sparse and dense regimes. In the sparse regime $s < \sqrt{p}\log^{-2/\alpha}(ep)$, we have

$$s\log^{2/\alpha}(ep/s) \le \sqrt{p} = \sqrt{p(\log\log(8n))^{\omega_1}},$$

and we shall prove the lower bound $s\log^{2/\alpha}(ep/s) + \log\log(8n)$, as stated in Proposition 17 below. In the dense regime, we first consider when $\sqrt{p}\log^{-2/\alpha}(ep) \le s \le \sqrt{p\log\log(8n)}$, we have

$$s\log^{2/\alpha}(ep/s) \gtrsim \sqrt{p} = \sqrt{p(\log\log(8n))^{\omega_1}}.$$

When $s > \sqrt{p\log\log(8n)}$, we still have

$$s\log^{2/\alpha}(ep/s) \ge s \ge \sqrt{p\log\log(8n)}.$$

Thus, in the dense regime, it suffices to prove the lower bound $\sqrt{p(\log\log(8n))^{\omega_1}} + \log\log(8n)$, as stated in Proposition 18 below.

Similarly for $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$, to prove Theorem 6, it suffices to establish Proposition 19 for the sparse regime and Proposition 20 for the dense regime.

Proposition 17. *For $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$ with $0 < \alpha \le 2$ and $K \ge K_\alpha$, for some constant $K_\alpha > 0$ depending only on α. Assume $c \le s < \sqrt{p}\log^{-2/\alpha}(ep)$, for some absolute constant $c \ge 1$. There exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_\mathcal{G}(\rho) \ge 1/2$ whenever $\rho^2 \le c'v_{\mathcal{G},\text{sparse}}^{\mathrm{L}}$, where*

$$v_{\mathcal{G},\text{sparse}}^{\mathrm{L}} := s\log^{2/\alpha}(ep/s) + \log\log(8n).$$

Proposition 18. *For $\mathcal{Q} = \mathcal{G}_{\alpha,K}^\otimes$ with $0 < \alpha \le 2$ and $K \ge K_\alpha$, for some constant $K_\alpha > 0$ depending only on α. Assume $s \ge \sqrt{p}\log^{-2/\alpha}(ep) \vee c$, for some absolute constant $c \ge 1$. There exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_\mathcal{G}(\rho) \ge 1/2$ whenever $\rho^2 \le c'v_{\mathcal{G},\text{dense}}^{\mathrm{L}}$, where*

$$v_{\mathcal{G},\text{dense}}^{\mathrm{L}} := \sqrt{p(\log\log(8n))^{\omega_1}} + \log\log(8n)$$

and $\omega_1 = \mathbb{1}_{\left\{s \ge \sqrt{p\log\log(8n)}\right\}}$.

Proposition 19. *For $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \ge 2$ and $K \ge K_\alpha$, for some constant $K_\alpha > 0$ depending only on α. Assume $c \le s < p^{\frac{1}{2}-(\frac{1}{\alpha-2}\wedge\frac{1}{2})}$, for some absolute constant $c \ge 1$. There exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_\mathcal{P}(\rho) \ge 1/2$ whenever $\rho^2 \le c'v_{\mathcal{P},\text{sparse}}^{\mathrm{L}}$, where*

$$v_{\mathcal{P},\text{sparse}}^{\mathrm{L}} := s(p/s)^{2/\alpha} + \log\log(8n).$$

Proposition 20. *For $\mathcal{Q} = \mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \ge 2$ and $K \ge K_\alpha$, for some constant $K_\alpha > 0$ depending only on α. Assume $s \ge p^{\frac{1}{2}-(\frac{1}{\alpha-2}\wedge\frac{1}{2})} \vee c$, for some absolute constant $c \ge 1$. There exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_\mathcal{P}(\rho) \ge 1/2$ whenever $\rho^2 \le c'v_{\mathcal{P},\text{dense}}^{\mathrm{L}}$, where*

$$v_{\mathcal{P},\text{dense}}^{\mathrm{L}} := p^{(2/\alpha)\vee(1/2)}(\log\log(8n))^{\omega_2} + \log\log(8n)$$

and $\omega_2 = (1/2)\mathbb{1}_{\left\{s > \sqrt{p\log\log(8n)}\right\}\cap\{\alpha \ge 4\}}$.

We now prove all these four lower bound results. Throughout the proof, we use $P_{\theta,\Xi}$ to denote the probability distribution of $X \in \mathbb{R}^{p\times n}$ that satisfies $X - \theta \sim \Xi$, and $E_{\theta,\Xi}$ the corresponding expectation under this distribution. It suffices to prove the five lower bound rate claims below, as they then immediately imply Propositions 17, 18, 19 and 20.

(i). $\log\log(8n)$, for $\mathcal{G}_{\alpha,K}^\otimes$ with $0 < \alpha \le 2$ and $K \ge 2^{1+2/\alpha}$ and for $\mathcal{P}_{\alpha,K}^\otimes$ with $\alpha > 2$ and $K \ge \sqrt{\alpha+1}$ or $\alpha = 2$ and $K \ge 1$;

(ii). $\sqrt{p \log\log(8n)}$ when $s \geq \sqrt{p\log\log(8n)}$, for $\mathcal{G}^{\otimes}_{\alpha,K}$ with $0 < \alpha \leq 2$ and $K \geq 2^{1+2/\alpha}$ and for $\mathcal{P}^{\otimes}_{\alpha,K}$ with $\alpha > 2$ and $K \geq \sqrt{\alpha+1}$ or $\alpha = 2$ and $K \geq 1$;

(iii). $p^{2/\alpha}$ when $s \geq 30$, for $\mathcal{P}^{\otimes}_{\alpha,K}$ with $\alpha > 2$ and $K \geq \sqrt{2}$ or $\alpha = 2$ and $K \geq 1$;

(iv). $s(p/s)^{2/\alpha}$ when $30 \leq s \leq p^{\frac{\alpha-4}{2\alpha-4}}$, for $\mathcal{P}^{\otimes}_{\alpha,K}$ with $\alpha \geq 4$ and $K \geq \sqrt{2}$;

(v). $s\log^{2/\alpha}(ep/s)$ when $30 \leq s \leq \sqrt{p}\log^{-2/\alpha}(ep)$, for $\mathcal{G}^{\otimes}_{\alpha,K}$ with $0 < \alpha \leq 2$ and $K \geq 2^{1+2/\alpha}$.

Proof of claim (i). We first consider that each entry of the noise matrix E follows an independent standard normal distribution. Then for $0 < \alpha \leq 2$, $i \in [n]$, $j \in [p]$ and $x \geq 2^{1+2/\alpha}$, we have

$$
\mathbb{E}\left[\exp\left\{\left(\frac{|E_i(j)|}{x}\right)^{\alpha}\right\}\right] = \mathbb{E}\left[\exp\left\{\left(\frac{|E_i(j)|}{x}\right)^{\alpha}\right\}\mathbb{1}_{\{|E_i(j)|\geq 2\}}\right] + \mathbb{E}\left[\exp\left\{\left(\frac{|E_i(j)|}{x}\right)^{\alpha}\right\}\mathbb{1}_{\{|E_i(j)|<2\}}\right]
$$

$$
\leq \mathbb{E}\left[\exp\left\{\left(\frac{|E_i(j)|}{2}\right)^{2}\right\}\mathbb{1}_{\{|E_i(j)|\geq 2\}}\right] + \exp\{(2/x)^{\alpha}\}
$$

$$
= \sqrt{2} - \mathbb{E}\left[\exp\left\{\left(\frac{|E_i(j)|}{2}\right)^{2}\right\}\mathbb{1}_{\{|E_i(j)|<2\}}\right] + \exp\{(2/x)^{\alpha}\}
$$

$$
\leq \sqrt{2} - \left(1 - \exp(-2)\right) + \exp\{(2/x)^{\alpha}\} < 2,
$$

where the penultimate inequality follows from the standard Gaussian tail bound. Thus, for any $K \geq 2^{1+2/\alpha}$, we have $\|E_i(j)\|_{\psi_\alpha} \leq K$. Furthermore, by Jensen's inequality, we obtain for $\alpha > 2$

$$
\mathbb{E}|E_i(j)|^{\alpha} \leq \left\{\mathbb{E}|E_i(j)|^{2\lceil\alpha/2\rceil}\right\}^{\frac{\alpha/2}{\lceil\alpha/2\rceil}} = \left\{\prod_{i=1}^{\lceil\alpha/2\rceil}(2i-1)\right\}^{\frac{\alpha/2}{\lceil\alpha/2\rceil}} \leq \left(2\lceil\alpha/2\rceil-1\right)^{\alpha/2} \leq (\alpha+1)^{\alpha/2}.
$$

Therefore $P_e \in \mathcal{G}^{\otimes}_{\alpha,K}$ for all $0 < \alpha \leq 2$ and $K \geq 2^{1+\alpha/2}$ and $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$ for all $\alpha \geq 2$ and $K \geq \sqrt{\alpha+1}$ or $\alpha = 2$ and $K \geq 1$. For the mean vectors μ_1 and μ_2 in the definition of $\Theta^{(t_0)}(p,n,s,\rho)$, we restrict them to be equal in all coordinates except perhaps the first. Then under this setting, the lower bound $\log\log(8n)$ of the detection rate is established in Proposition 4.2 of [49]. Note that this lower bound holds for all $1 \leq s \leq p$.

Proof of claim (ii). When $s \geq \sqrt{p\log\log(8n)}$, we again consider the independent standard normal noise structure. The lower bound $\sqrt{p\log\log(8n)}$ is shown in Proposition 3 of [2].

We now use a unified approach to establish the three remaining rates. Let ξ and $\tilde{\xi}$ be two independent random variables on $\mathbb{R}$, whose distributions are to be specified later; let $\tilde{\omega}$ be an discrete random variable (independent of $\xi, \tilde{\xi}$), taking values

$$
\tilde{\omega} = \begin{cases} +1 & \text{w.p. } \frac{s}{4p}\left(1+\frac{\gamma^2 s}{2p}\right)^{-1} \\ -1 & \text{w.p. } \frac{s}{4p}\left(1+\frac{\gamma^2 s}{2p}\right)^{-1} \\ 0 & \text{otherwise,} \end{cases} \tag{86}
$$

where $\gamma > 0$ is also to be specified later; let $\tilde{\pi} := \tilde{\xi} + \gamma\tilde{\omega}$. We remark that $\tilde{\omega}$ can be viewed as a Rademacher random variable being multiplied by a Bernoulli random variable. Denote $\underline{\xi} := (\xi(1),\ldots,\xi(p))^{\top} \in \mathbb{R}^p$, where the coordinates are i.i.d. copies of ξ and we use similar notations $\underline{\tilde{\xi}}, \underline{\tilde{\omega}}, \underline{\tilde{\pi}}$. Let ν denote the distribution of $\gamma\underline{\tilde{\omega}} \in \mathbb{R}^p$, and $\bar{\nu}$ the distribution restricted to $\mathcal{V}_s := \{v \in \mathbb{R}^p : s/6 \leq \|v\|_0 \leq s\}$, i.e. $\bar{\nu}(A) = \frac{\nu(A\cap\mathcal{V}_s)}{\nu(\mathcal{V}_s)}$ for any Borel set $A \subseteq \mathbb{R}^p$. Consequently, the support of this restricted measure satisfies

$$
\text{supp}(\bar{\nu}) \subseteq \{v \in \mathbb{R}^p : \|v\|_0 \leq s, \|v\|_2^2 \geq s\gamma^2/6\}. \tag{87}
$$

We also have

$$
-\nu(\mathcal{V}_s^c) = -\left(\frac{1}{\nu(\mathcal{V}_s)}-1\right)\nu(\mathcal{V}_s) \leq \nu(A) - \bar{\nu}(A) = \nu(A\cap\mathcal{V}_s^c) - \left(\frac{1}{\nu(\mathcal{V}_s)}-1\right)\nu(A\cap\mathcal{V}_s) \leq \nu(\mathcal{V}_s^c). \tag{88}
$$

for any Borel set A. Denote Ξ^* to be the distribution of $(\underline{\xi}, R_2,\ldots,R_n) \in \mathbb{R}^{p\times n}$, $\tilde{\Xi}^*$ the distribution of $(\underline{\tilde{\xi}}, R_2,\ldots,R_n)$, and $\tilde{\Pi}$ the distribution of $(\underline{\tilde{\pi}}, R_2,\ldots,R_n)$, where $(R_i(j))_{i\in\{2,\ldots,n\},j\in[p]}$ are i.i.d. Rademacher random variables, independent of $\underline{\xi}, \underline{\tilde{\xi}}, \underline{\tilde{\pi}}$. Now we consider the following mixture measures:

$$
\bar{\mathbf{P}}^* := \int P_{\theta^{(1)},\Xi^*}\,\bar{\nu}(d\theta_1), \quad \mathbf{P}^* := \int P_{\theta^{(1)},\Xi^*}\,\nu(d\theta_1), \quad \text{and} \quad \tilde{\mathbf{P}}^* := \int P_{\theta^{(1)},\tilde{\Xi}^*}\,\nu(d\theta_1),
$$

where $\theta^{(1)} := (\theta_1, 0,\ldots,0) \in \mathbb{R}^{p\times n}$. Observe that $\tilde{\mathbf{P}}^* = P_{0,\tilde{\Pi}}$, as both sides represent the distribution of $(\underline{\tilde{\pi}}, R_2,\ldots,R_n)$. We first provide an upper bound on the total variation distance between $\mathbf{P}^*$ and $\bar{\mathbf{P}}^*$. By (88), we have

$$
\text{TV}(\mathbf{P}^*, \bar{\mathbf{P}}^*) \leq \text{TV}(\nu,\bar{\nu}) = \sup_A |\nu(A) - \bar{\nu}(A)| \leq \nu(\mathcal{V}_s^c) = \mathbb{P}(\|\underline{\tilde{\omega}}\|_0 > s) + \mathbb{P}(\|\underline{\tilde{\omega}}\|_0 < s/6). \tag{89}
$$

Suppose γ is chosen to satisfy $\gamma \leq \sqrt{p/s}$. Then from (86), we deduce $\frac{s}{3p} \leq \mathbb{P}(\tilde{\omega}(1) \neq 0) < \frac{s}{2p}$. By Chernoff bounds, we have

$$\mathbb{P}\big(\|\tilde{\omega}\|_0 > s\big) \leq \frac{\mathbb{E}\big[e^{\|\tilde{\omega}\|_0 \log 2}\big]}{e^{s \log 2}} \leq \frac{\big(1 + s/(2p)\big)^p}{e^{s \log 2}} \leq e^{-s/6},$$

$$\mathbb{P}\big(\|\tilde{\omega}\|_0 < s/6\big) \leq \frac{\mathbb{E}\big[e^{-\|\tilde{\omega}\|_0 \log 2}\big]}{e^{-(s \log 2)/6}} \leq \frac{\big(1 - s/(6p)\big)^p}{e^{-(s \log 2)/6}} \leq e^{-s/20}. \tag{90}$$

The key step of the proof is to carefully construct two random variables ξ and $\tilde{\xi}$ such that the following three conditions are satisfied:

$$\Xi^* \in \mathcal{G}_{\alpha,K} \ (\text{resp. } \mathcal{P}_{\alpha,K}), \tag{91}$$

$$\tilde{\Pi} \in \mathcal{G}_{\alpha,K} \ (\text{resp. } \mathcal{P}_{\alpha,K}), \tag{92}$$

$$\mathrm{H}^2(P_\xi, P_{\tilde{\xi}}) \leq \frac{1}{16p}, \tag{93}$$

where, in a slight abuse of notation, we denote P_ξ and $P_{\tilde{\xi}}$ to be the distribution of ξ and $\tilde{\xi}$ respectively. Then, by data processing inequality as well as some basic properties of the total variation distance and the Hellinger distance, we obtain

$$\mathrm{TV}(\tilde{\mathbf{P}}^*, \mathbf{P}^*) \leq \mathrm{TV}(P_{0,\tilde{\Xi}^*}, P_{0,\Xi^*}) \leq \mathrm{TV}\big(P_{\tilde{\xi}}, P_{\underline{\xi}}\big) \leq \mathrm{H}\big(P_{\tilde{\xi}}, P_{\underline{\xi}}\big) = \sqrt{2\big(1 - (1 - \mathrm{H}^2(P_{\tilde{\xi}}, P_{\xi})/2)^p\big)}$$

$$\leq \sqrt{p\mathrm{H}^2(P_\xi, P_{\tilde{\xi}})} \leq 1/4, \tag{94}$$

where the penultimate inequality follows from the fact that $(1 - x)^p \geq 1 - px$ for all $0 \leq x \leq 1$ and $p \geq 1$. Combining (87), (89), (90), and (94), when $s \geq 30$, for all $\rho^2 \leq s\gamma^2/12$, we have

$$\mathcal{R}_{\mathcal{Q}}(\rho) = \inf_{\phi \in \Phi}\left\{\sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_{\theta,P_e}\phi + \sup_{P_e \in \mathcal{Q}} \sup_{\theta \in \Theta(p,n,s,\rho)} \mathbb{E}_{\theta,P_e}(1 - \phi)\right\}$$

$$\geq 1 - \mathrm{TV}(P_{0,\tilde{\Pi}}, \bar{\mathbf{P}}^*) = 1 - \mathrm{TV}(\tilde{\mathbf{P}}^*, \bar{\mathbf{P}}^*) \geq 1 - \mathrm{TV}(\tilde{\mathbf{P}}^*, \mathbf{P}^*) - \mathrm{TV}(\mathbf{P}^*, \bar{\mathbf{P}}^*)$$

$$\geq 3/4 - e^{-s/6} - e^{-s/20} \geq 1/2,$$

where the class $\mathcal{Q}$ is either $\mathcal{G}_{\alpha,K}^{\otimes}$ or $\mathcal{P}_{\alpha,K}^{\otimes}$. Below, we give three constructions of ξ and $\tilde{\xi}$ that satisfy conditions (91), (92) and (93), and specify the corresponding choices of γ. Each construction corresponds to a rate given at the beginning of the proof.

Proof of claim (iii). We work within the noise distribution class $\mathcal{P}_{\alpha,K}^{\otimes}$ with $\alpha > 2$ and $K \geq \sqrt{2}$ or $\alpha = 2$ and $K \geq 1$ and we only consider $s = 30$ (a constant) in this construction. Let ξ and $\tilde{\xi}$ be two independent discrete random variables such that

$$\tilde{\xi} = \begin{cases} \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2} & \text{w.p. } 1/2 \\ -\big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2} & \text{w.p. } 1/2 \end{cases} \quad \text{and} \quad \xi = \begin{cases} \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2} & \text{w.p. } \frac{t_0^2 - 1}{2\big(t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}\big)} \\ -\big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2} & \text{w.p. } \frac{t_0^2 - 1}{2\big(t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}\big)} \\ t_0 & \text{w.p. } \frac{1 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}{2\big(t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}\big)} \\ -t_0 & \text{w.p. } \frac{1 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}{2\big(t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}\big)}. \end{cases}$$

Direct calculations show that both ξ and $\tilde{\xi} + \gamma\tilde{\omega}$ have mean 0 and variance 1. Choose

$$\gamma = \max\left\{-1 + \frac{\big\{(K^\alpha - 1)p/s\big\}^{1/\alpha}}{\max\{32, K\}}, \frac{\sqrt{2}}{32}\right\} \quad \text{and} \quad t_0 = 32\gamma \geq \sqrt{2}.$$

Note that we have $\gamma \leq \sqrt{p/s}$. Now, to check (91) and (92), it suffices to only verify that $\mathbb{E}|\xi|^\alpha \leq K^\alpha$ and that $\mathbb{E}\big|\tilde{\xi} + \gamma\tilde{\omega}\big|^\alpha \leq K^\alpha$ respectively. Indeed, as $\alpha > 2$ and $K \geq \sqrt{2}$, we have

$$\mathbb{E}|\xi|^\alpha \leq 1 + t_0^\alpha \frac{1 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}{t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}} \leq 1 + t_0^\alpha \frac{\gamma^2 s/(2p)}{t_0^2 - 1} = 1 + \frac{\gamma^2 s t_0^{\alpha-2}}{p} \leq 1 + \frac{32^{\alpha-2} \gamma^\alpha s}{p}$$

$$\leq 1 + \max\left\{K^\alpha - 1, 2^{\alpha/2-10}\right\} \leq K^\alpha,$$

and

$$\mathbb{E}\big|\tilde{\xi} + \gamma\tilde{\omega}\big|^\alpha \leq 1 + (1+\gamma)^\alpha \cdot \mathbb{P}(\tilde{\omega} \neq 0) \leq 1 + \frac{(1+\gamma)^\alpha s}{2p} \leq 1 + \max\left\{K^\alpha - 1, \frac{(17/16)^\alpha}{2}\right\} \leq K^\alpha.$$

We also verify (93):

$$\mathrm{H}^2(P_\xi, P_{\tilde{\xi}}) = \left(1 - \sqrt{\frac{t_0^2 - 1}{t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}}\right)^2 + \frac{1 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}{t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}} \leq \frac{2\big(1 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}\big)}{t_0^2 - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1}}$$

$$\leq \frac{2\gamma^2 s}{p t_0^2} = \frac{60\gamma^2}{(32\gamma)^2 p} \leq \frac{1}{16p}.$$

We thus conclude that under the noise distribution class $\mathcal{P}_{\alpha,K}^\otimes$ with $\alpha > 2$ and $K \geq \sqrt{2}$, whenever $s \geq 30$ and

$$\rho^2 \leq \frac{30}{12}\left(\max\left\{-1 + \frac{\{(K^\alpha - 1)p/30\}^{1/\alpha}}{\max\{32, K\}}, \frac{\sqrt{2}}{32}\right\}\right)^2 \leq c \cdot p^{2/\alpha},$$

for some $c > 0$ depending only on α and K, we have $\mathcal{R}_{\mathcal{P}}(\rho) \geq 1/2$. When $\alpha = 2$, we can simply set $\gamma = \sqrt{p/s}$ and $t_0 = 32\gamma$ and reach the same conclusion.

Proof of claim (iv). We work within the noise distribution class $\mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \geq 4$ and $K \geq \sqrt{2}$. We first define an auxiliary random variable ξ_{aux} and with the following density elsewhere:

$$f_{\xi_{\mathrm{aux}}}(x) = \begin{cases} 1000(x - \mathrm{sgn}(x) \cdot 0.9)^2 & 0.9 \leq |x| < 0.95 \\ 5 - 1000(x - \mathrm{sgn}(x))^2 & 0.95 \leq |x| \leq 1.05 \\ 1000(x - \mathrm{sgn}(x) \cdot 1.1)^2 & 1.05 < |x| \leq 1.1 \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\mathbb{E}\xi_{\mathrm{aux}} = 0$ and $\sigma_{\mathrm{aux}}^2 := \mathbb{E}\xi_{\mathrm{aux}}^2 \in (1, 1.01)$. Now let ξ and $\tilde{\xi}$ be independent random variables such that $\xi \overset{d}{=} \sigma_{\mathrm{aux}}^{-1}\xi_{\mathrm{aux}}$ and $\tilde{\xi} \overset{d}{=} \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2}\sigma_{\mathrm{aux}}^{-1}\xi_{\mathrm{aux}}$. Again, direct calculations show that both ξ and $\tilde{\xi} + \gamma\tilde{\omega}$ have mean 0 and variance 1. For condition (91), since $|\xi| < 1.1$ holds with probability one, we have $\Xi^* \in \mathcal{P}_{\alpha,K}$ for all $\alpha \geq 4$ and $K \geq \sqrt{2}$. We choose

$$\gamma = \frac{1}{12}(p/s)^{1/\alpha}.$$

Note that $\gamma \leq \sqrt{p/s}$. We verify (92) as follows:

$$\mathbb{E}\big|\tilde{\xi} + \gamma\tilde{\omega}\big|^\alpha \leq 1.1^\alpha + (1.1 + \gamma)^\alpha \cdot \mathbb{P}(\tilde{\omega} \neq 0) \leq 1.1^\alpha + \frac{(1.1 + \gamma)^\alpha s}{2p} \leq 1.1^\alpha + \frac{\max\{1.2, 12\gamma\}^\alpha s}{2p}$$

$$\leq \max\left\{1.1^\alpha + \frac{1.2^\alpha}{2}, 1.1^\alpha + \frac{1}{2}\right\} \leq 2^{\alpha/2} \leq K^\alpha. \tag{95}$$

Finally, by Theorem 7.6 of [50], we have when $s \leq p^{\frac{\alpha-4}{2\alpha-4}}$

$$\mathrm{H}^2(P_\xi, P_{\tilde{\xi}}) \leq \frac{\left(\sigma_{\mathrm{aux}}^{-1} - \big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2}\sigma_{\mathrm{aux}}^{-1}\right)^2}{4} \sup_{u \in \left[\big(1 + \frac{\gamma^2 s}{2p}\big)^{-1/2}\sigma_{\mathrm{aux}}^{-1}, \sigma_{\mathrm{aux}}^{-1}\right]} \frac{\int_{\mathrm{supp}(\xi_{\mathrm{aux}})}\big(f_{\xi_{\mathrm{aux}}}'(x)\big)^2/f_{\xi_{\mathrm{aux}}}(x)\,dx}{u^2}$$

$$\leq \frac{\left(\big(1 + \frac{\gamma^2 s}{2p}\big)^{1/2} - 1\right)^2}{4} \int_{\mathrm{supp}(\xi_{\mathrm{aux}})}\big(f_{\xi_{\mathrm{aux}}}'(x)\big)^2/f_{\xi_{\mathrm{aux}}}(x)\,dx$$

$$\leq \frac{\gamma^4 s^2}{64p^2} \cdot 4\left(\int_0^{0.05}\frac{(-2000x)^2}{5 - 1000x^2}\,dx + \int_{0.05}^{0.1}\frac{(2000(x - 0.1))^2}{1000(x - 0.1)^2}\,dx\right) \leq \frac{25\gamma^4 s^2}{p^2} \leq \frac{1}{16p}, \tag{96}$$

and this verifies (93). Therefore, under the noise distribution class $\mathcal{P}_{\alpha,K}^\otimes$ with $\alpha \geq 4$ and $K \geq \sqrt{2}$, whenever $30 \leq s \leq p^{\frac{\alpha-4}{2\alpha-4}}$ and $\rho^2 \leq s(p/s)^{2/\alpha}/1728$, we have $\mathcal{R}_{\mathcal{P}}(\rho) \geq 1/2$.

Proof of claim (v). We work within the noise distribution class $\mathcal{G}_{\alpha,K}^\otimes$ with $0 < \alpha \leq 2$ and $K \geq 2^{1+2/\alpha}$. We use the same construction as in (iv), but now choose instead

$$\gamma = \frac{1}{3 \cdot (8/\alpha)^{1/\alpha}}\log^{1/\alpha}(ep/s).$$

Since $\log x \leq \frac{2}{e\alpha} x^{\alpha/2}$ for all $x \geq e$, we can verify that $\gamma \leq \sqrt{p/s}$. Again, for condition (91), since $|\xi| < 1.1$ holds with probability one, we have $\Xi^* \in \mathcal{G}_{\alpha, K}$ for all $\alpha \geq 4$ and $K \geq 2^{1+2/\alpha}$, as $\exp\{(1.1/K)^\alpha\} \leq e^{1/4} < 2$. We now verify (92) using the technique in (95):

$$\mathbb{E}\left[\exp\left\{\left(\frac{|\tilde{\xi} + \gamma\tilde{\omega}|}{K}\right)^\alpha\right\}\right] \leq \exp\left\{\left(\frac{1.1}{K}\right)^\alpha\right\} + \frac{s}{2p}\exp\left\{\left(\frac{\max\{2, 3\gamma\}}{K}\right)^\alpha\right\}$$

$$\leq e^{1/4} + \max\left\{e^{1/4}\frac{s}{2p}, \left(\frac{s}{2p}\right)^{1-\frac{2\alpha}{8K^\alpha}}\right\}$$

$$\leq e^{1/4} + \max\left\{e^{1/4}/2, \sqrt{1/2}\right\} < 2.$$

We then follow (96) to verify (93) as well:

$$\mathrm{H}^2(P_\xi, P_{\tilde{\xi}}) \leq \frac{25\gamma^4 s^2}{p^2} \leq \frac{1}{16p},$$

when $s \leq \sqrt{p}\log^{-2/\alpha}(ep)$. Therefore, under the noise distribution class $\mathcal{G}_{\alpha, K}^{\otimes}$ with $\alpha \leq 2$ and $K \geq 2^{1+2/\alpha}$, whenever $30 \leq s \leq \sqrt{p}\log^{-2/\alpha}(ep)$ and $\rho^2 \leq \frac{s\log^{2/\alpha}(ep/s)}{36 \cdot (8/\alpha)^{1/\alpha}}$, we have $\mathcal{R}_\mathcal{G}(\rho) \geq 1/2$.

D. Technical details of Section VI

1) Multiple change points: To describe our testing procedure for multiple change points, we first denote

$$J := \left\{(\ell, t) : t = 2^{1+\lceil\log_2(\log(n))\rceil}, 2^{2+\lceil\log_2(\log(n))\rceil}, \ldots, 2^{\lfloor\log_2(n)\rfloor - 1}, \ell = t, \ldots, n - t\right\}.$$

Recall that in Section III-A, we use a median-of-means-type statistic A_t^{MoM} in (16) to determine whether there is a single change at or near each $t \in \mathcal{T}$ based on data $X_1, \ldots, X_t, X_{n-t+1}, \ldots X_n$. Here, we compute the same statistic using $X_{\ell-t+1}, \ldots, X_\ell, X_{\ell+1}, \ldots, X_{\ell+t}$ for each pair $(\ell, t) \in J$, i.e. $Z_{i,\ell,t} := (X_{\ell-t+i} - X_{\ell+t+1-i})/\sqrt{2}$ for $i \in [t]$. We then split $\{Z_{1,\ell,t}, Z_{2,\ell,t}, \ldots, Z_{t,\ell,t}\}$ into G groups of equal size, with G specified later in Theorem 21. Let $V_{g,\ell,t} \in \mathbb{R}^p$ be with j-th coordinate $V_{g,\ell,t}(j) := \overline{Z}_{g,\ell,t}^2(j) - G/t$, $j \in [p]$ and $g \in [G]$, where $\overline{Z}_{g,\ell,t}$ is the sample mean of the g-th group. We then have that

$$T_{\ell,t} := \mathbb{1}_{\{A_{\ell,t}^{\mathrm{MoM}} > r\}},$$

where

$$A_{\ell,t}^{\mathrm{MoM}} := t \cdot \mathrm{median}\left(\sum_{j=1}^p V_{1,\ell,t}(j), \sum_{j=1}^p V_{2,\ell,t}(j), \ldots, \sum_{j=1}^p V_{G_t,\ell,t}(j)\right),$$

and the threshold r is specified in Theorem 21. Our test for the multiple change points case is

$$\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}} := \max_{(\ell,t)\in J} T_{\ell,t}. \tag{97}$$

We prove the following theorem on the theoretical guarantee of the test $\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}$, constructed above. Theorem 10 then follows as an immediate consequence.

Theorem 21. *Assume $n \geq 50$ and $\alpha \geq 4$. For any $\varepsilon \in (0, 1)$, there exist $C_1, C_2 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}$ defined in (97) with*

$$r = C_1\sqrt{p}G \quad \text{and} \quad G = 2^{1+\lceil\log_2(\log(n))\rceil}$$

satisfies that

$$\mathcal{R}_{\mathcal{P},\mathrm{multi}}(\rho, \phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_2 v_{\mathcal{P}}^{\mathrm{U}}$, where

$$v_{\mathcal{P}}^{\mathrm{U}} := \sqrt{p}\log(n).$$

Proof. **Null term.** For any $\theta \in \Theta_0(p, n)$, we have, by a union bound

$$\mathbb{E}_\theta\phi_{\mathcal{P},\mathrm{multi}}^{\mathrm{MoM}} = \mathbb{P}_\theta\left(\max_{(\ell,t)\in J} T_{\ell,t} = 1\right) \leq \sum_{(\ell,t)\in J} \mathbb{P}_\theta(T_{\ell,t} = 1).$$

Note that, for a fixed pair (ℓ, t), the test variable $T_{\ell,t}$ is constructed in the exact same way as the MoM test $\phi_{\mathcal{P},\mathrm{dense}}$ we studied in Section III-A. Therefore, by following the proof of Theorem 4 and, in particular, (64), we have

$$\mathbb{P}_\theta(T_{\ell,t} = 1) = \mathbb{P}_\theta(A_{\ell,t}^{\mathrm{MoM}} > r) \leq \exp\left\{-\frac{G}{2}\log(6/\varepsilon)\right\},$$

when we choose $r = C_1 \sqrt{p} G$ for some sufficiently large constant C_1 that depends on ε. Now, with $G = 2^{1+\lceil \log_2(\log(n)) \rceil} \geq 2 \log(n)$ and $n \geq 50$, we have for all $(\ell, t) \in J$ that

$$\mathbb{P}_\theta(T_{\ell,t} = 1) \leq \exp\left\{ -\frac{G}{2} \log(6/\varepsilon) \right\} \leq \exp\left\{ -\frac{4}{3} \log(n) - \log(6/\varepsilon) \right\} = \frac{\varepsilon}{6n^{4/3}},$$

where the second inequality is derived from the calculation $\big(\log(n) - 1\big)\big(\log(6/\varepsilon) - 4/3\big) \geq 4/3$. Therefore, as long as $n \geq 50$, we conclude

$$\sum_{(\ell,t) \in J} \mathbb{P}_\theta(T_{\ell,t} = 1) \leq \frac{\varepsilon}{6n^{4/3}} |J| \leq \frac{\varepsilon}{6n^{4/3}} n \log_2(n/2) \leq \frac{\varepsilon}{4}.$$

Alternative term. For any $\theta \in \Theta_{\text{multi}}(p, n, \rho)$, by definition, there exists an $i^* \in \mathbb{Z}^+$ such that $\Delta_{i^*} \geq 4 \log(n)$ and $\kappa_{i^*}^2 \Delta_{i^*} \geq \rho^2$. Consequently, there exists a corresponding $(\tau_{i^*}, t^*) \in J$ such that $\Delta_{i^*}/2 \leq t^* \leq \Delta_{i^*}$. Then, we have

$$\mathbb{E}_\theta\big(1 - \phi_{\mathcal{P},\text{multi}}^{\text{MoM}}\big) = \mathbb{P}_\theta\Big(\max_{(\ell,t) \in J} T_{\ell,t} = 0 \Big) \leq \mathbb{P}_\theta\Big(T_{\tau_{i^*},t^*} = 0 \Big).$$

Since we use the same type of test statistic as $\phi_{\mathcal{P},\text{dense}}$, and that $\{Z_{1,\tau_{i^*},t^*}, \ldots, Z_{t^*,\tau_{i^*},t^*}\}$ are independent random vectors each with mean $(\mu_1 - \mu_2)/\sqrt{2}$, we can again follow the proof of Theorem 4 to obtain

$$\mathbb{P}_\theta\Big(T_{\tau_{i^*},t^*} = 0 \Big) \leq \exp\left\{ -\frac{\varepsilon G}{12}\left(\frac{6}{\varepsilon} \log\left(\frac{6}{\varepsilon}\right) - \frac{6}{\varepsilon} + 1 \right) \right\} \leq \frac{\varepsilon}{4},$$

provided that $t^* \|\mu_1 - \mu_2\|_2^2/2 \geq \rho^2/4 \geq \frac{C_2}{4} \sqrt{p} \log(n)$ for some sufficiently large constant C_2 that depends on ε and this completes the proof. $\square$

Proof of Proposition 11. Throughout this proof, we take $P_e = N^\otimes(0,1)$. Recall that $N^\otimes(0,1)$ denotes the joint distribution of all pn independent $N(0,1)$ entries in E. Following the calculation in the lower bound proof of claim (i) in Section C, we confirm that $P_e \in \mathcal{P}_{\alpha,K}^\otimes$ for all $\alpha \geq 2$ and $K \geq \sqrt{\alpha+1}$. We write $P_{\theta,E}$ to denote the distribution of $\theta + E$ where $\theta \in \mathbb{R}^{p \times n}$, and $E \in \mathbb{R}^{p \times n}$ is a matrix with entries being i.i.d standard normal random variables. We also write $P_{\theta \sim \pi, E}$ to denote the distribution of $\theta + E$ when $\theta \sim \pi$.

This result essentially follows from Theorem 2 in [29]. To lower bound the minimax testing error, we have

$$\mathcal{R}_{\mathcal{P},\text{multi}}(\rho) = \inf_{\phi \in \Phi} \left\{ \sup_{P_e \in \mathcal{P}_{\alpha,K}^\otimes} \sup_{\theta \in \Theta_0(p,n)} \mathbb{E}_{\theta,P_e}(\phi) + \sup_{P_e \in \mathcal{P}_{\alpha,K}^\otimes} \sup_{\theta \in \Theta_{\text{multi}}(p,n,\rho)} \mathbb{E}_{\theta,P_e}(1 - \phi) \right\}$$
$$\geq 1 - \text{TV}(P_{0,E}, P_{\theta \sim \pi, E}),$$

for any π that has support only on θ with two change points such that $\min\{\Delta_1, \Delta_2\} \geq 4 \log(n)$. The constructions of $\tilde{\Theta}_{(2)}$ and $\tilde{\Theta}_{(3)}$ in Case 2 and Case 3 in the proof of Theorem 2 in [29] with $r = 4 \log(n)$ both have support on θ with two change points such that $\min\{\Delta_1, \Delta_2\} \geq 4 \log(n)$. The remaining calculation therein shows that when $n/4 \geq \lceil 4 \log(n) \rceil$ and

$$\rho^2 \leq c\big\{ \sqrt{p \log(n)} + \log(n) \big\},$$

for some sufficiently small constant $c > 0$, we have $\text{TV}(P_{0,E}, P_{\theta \sim \pi, E}) \leq 1/2$. $\square$

A heuristic example for requiring minimum spacing: We provide a heuristic example illustrating why the heavy-tailed nature of the data necessitates a minimum spacing condition to achieve a rate with logarithmic dependence on n.

Consider independent univariate random variables $U_1, \ldots, U_n \in \mathbb{R}$ with $\mathbb{E}U_i = \mu_i$ for $i \in [n]$. For $\alpha \geq 2$ and $K < \infty$, we assume that $\mathbb{E}\big[\big(|U_i - \mu_i|/K\big)^\alpha\big] \leq 1$ for all i. Now, consider the testing problem

$$\text{H}_0 : \mu_1 = \cdots = \mu_n = 0 \quad \text{vs.} \quad \text{H}_1 : \exists j \in [n] \text{ s.t. } \mu_i = 0 \text{ for all } i \neq j \text{ and } |\mu_j| \geq \rho.$$

This can be viewed as testing for a single outlier caused by a mean shift at an unknown time, which corresponds to two change points with *no minimum spacing requirement*.

A natural test to consider is

$$\max_{i \in [n]} \mathbb{1}_{\{|U_i| \geq r\}},$$

for some threshold r. Suppose we want to control the Type I error probability at some $\varepsilon > 0$. Under the null, applying a union bound and Chebyshev's inequality, we obtain the upper bound

$$\text{Type I error prob.} \leq \sum_{i=1}^n \mathbb{P}(|U_i| \geq r) \leq \frac{n}{(r/K)^\alpha}.$$

This is at most ε when $r \geq K(n/\varepsilon)^{1/\alpha}$. To distinguish the null from the alternative, the signal strength ρ must exceed this threshold, implying that ρ must be at least of order $n^{1/\alpha}$, i.e. polynomial in n.

2) Temporal dependence: Denote $Z_i := (X_i - X_{n+1-i})/\sqrt{2}$ for $i \in [n/2]$. For $t \in \mathcal{T}$, we split $\{Z_1, \ldots, Z_t\}$ into G_t groups of equal size that

$$\mathcal{Z}_{t,1} := \left\{Z_1, \ldots, Z_{\frac{t}{G_t}}\right\}, \quad \mathcal{Z}_{t,2} := \left\{Z_{\frac{t}{G_t}+1}, \ldots, Z_{\frac{2t}{G_t}}\right\}, \quad \ldots, \quad \mathcal{Z}_{t,G_t} := \left\{Z_{\frac{(G_t-1)t}{G_t}+1}, \ldots, Z_t\right\}.$$

The procedure to form the test remains the same as in (15) and (16) from Section III-A, except that we replace G_t/t with $\mathbb{E}\overline{Z}_{t,g}^2(j)$ in $V_{t,g}(j)$. Our test is

$$\phi_{\mathcal{P}_{\mathrm{Temp}}} := \mathbb{1}_{\left\{\max_{t \in \mathcal{T}} A_t^{\mathrm{MoM}}/r_t^{\mathrm{Temp}} > 1\right\}}, \tag{98}$$

with the detection threshold r_t^{Temp} specified in Theorem 22. We assume $\mathbb{E}\overline{Z}_{t,g}^2(j)$ to be known, though this can be relaxed if this quantity can be estimated reasonably well. We discuss this aspect in detail and provide a specific example where the noise is generated by a moving average process at the end of this subsection.

We prove the following theorem on the theoretical guarantee of the test $\phi_{\mathcal{P}_{\mathrm{Temp}}}$, constructed above. Theorem 12 then follows as an immediate consequence.

Theorem 22. *Assume $\alpha > 4$. For any $\varepsilon \in (0,1)$, there exist $C_1, C_2, C_3 > 0$ depending only on α, K, c_1, c_2 and ε, such that the test $\phi_{\mathcal{P}_{\mathrm{Temp}}}$ defined in (98) with*

$$r_t^{\mathrm{Temp}} = C_1 p^{1/2} G_t, \quad G_t = t \wedge \Delta \quad and \quad \Delta = 2^{\lceil \log_2(C_2(\log\log(8n))(\log\log\log(16n))^2)\rceil},$$

satisfies that

$$\mathcal{R}_{\mathcal{P}_{\mathrm{Temp}}}(\rho, \phi_{\mathcal{P}_{\mathrm{Temp}}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_3 v_{\mathcal{P}_{\mathrm{Temp}}}^{\mathrm{U}}$, where

$$v_{\mathcal{P}_{\mathrm{Temp}}}^{\mathrm{U}} := p^{1/2}(\log\log(8n))(\log\log\log(64n))^2.$$

Proof. If we define $\{\tilde{E}_i\}_{i \in [n]}$ as the reordered sequence of $\{E_i\}_{i \in [n]}$ with

$$\tilde{E}_i = \begin{cases} E_{(i+1)/2} & \text{for odd } i, \\ E_{n+1-i/2} & \text{for even } i, \end{cases} \tag{99}$$

then we verify that the (usual) α-mixing coefficient of this reordered process $\{\tilde{E}_i\}_{i \in [n]}$ satisfies

$$\alpha_{\mathrm{pa}}(i) := \sup_{\ell \in [n-i]} \sup_{A \in \sigma(\tilde{E}_j : 1 \leq j \leq \ell), B \in \sigma(\tilde{E}_j : \ell+i \leq j \leq n)} \left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right|$$
$$\leq \alpha^*(\lfloor i/2 \rfloor) \leq c_1 e^{-c_2 i/3}, \tag{100}$$

for all $i \in [n-1]$.

Null term. Similar to the proof of Theorem 4, we still have for all $t \in \mathcal{T}$ and $g \in [G_t]$ that

$$\mathbb{P}_\theta\left(t\sum_{j=1}^p V_{t,g}(j) > r_t^{\mathrm{Temp}}\right) = \mathbb{P}_\theta\left(\sum_{j=1}^p \left(\overline{Z}_{t,g}^2(j) - \mathbb{E}\overline{Z}_{t,g}^2(j)\right) > \frac{C_1\sqrt{p}G_t}{t}\right) \leq \frac{t^2 \sum_{j=1}^p \mathbb{E}_\theta \overline{Z}_{t,g}^4(j)}{C_1^2 p G_t^2}$$
$$\leq \frac{pt^2 C_4 (G_t/t)^2}{C_1^2 p G_t^2} = \frac{C_4}{C_1^2}. \tag{101}$$

for some $C_4 > 0$, depending on α, K and c. The second inequality above now follows from Theorem 4.1 of [51]. According to the reordered sequence $\{\tilde{E}_i\}_{i \in [n]}$ defined in (99), we have, for $g \in [G_t]$

$$\mathbb{1}_{\left\{t\sum_{j=1}^p V_{t,g}(j) > r_t^{\mathrm{Temp}}\right\}} \in \sigma\left(\tilde{E}_i : \frac{2(g-1)t}{G_t} + 1 \leq i \leq \frac{2gt}{G_t}\right).$$

Thus, as long as $C_4/C_1^2 \leq 1/4$ in (101), by (100) and Theorem 1 of [52][4], we have for all $t \in \mathcal{T}$ with $G_t \geq 4$

$$\mathbb{P}_\theta\left(A_t^{\mathrm{MoM}} > r_t^{\mathrm{Temp}}\right) \leq \mathbb{P}_\theta\left(\sum_{g=1}^{G_t} \mathbb{1}_{\left\{t\sum_{j=1}^p V_{t,g}(j) > r_t^{\mathrm{Temp}}\right\}} \geq G_t/2\right) \leq \exp\left\{-\frac{C_5 G_t}{\log G_t \log\log G_t}\right\}, \tag{102}$$

for some $C_5 > 0$, depending only on c_1 and c_2, the constants in the interlaced α-mixing condition (34). To this end, we observe that there exists sufficiently large $C_2 > 0$, depending only on C_5 and ε, such that for all $t \geq C_2/12$ and for our choice $\Delta \geq C_2(\log\log(8n))(\log\log\log(64n))^2 > C_2/12$, we have

$$\exp\left\{-\frac{C_5 t}{\log t \log\log t}\right\} \leq \frac{\varepsilon}{10} e^{-\sqrt{t}}, \tag{103}$$

---

[4]Theorem 1 in the cited work assumes that the α-mixing coefficient is bounded by $\alpha(i) \leq e^{-ci}$ for some $c > 0$. A closer examination of the proof reveals that this result remains valid even if the bound is relaxed to $c' e^{-ci}$ for some constants $c, c' > 0$.

and

$$\exp\left\{-\frac{C_5\Delta}{\log\Delta\log\log\Delta}\right\} \leq \frac{\varepsilon}{10\log_2(n/2)}, \tag{104}$$

Thus, by choosing $C_1 > \sqrt{C_2^2 C_4/\varepsilon}$ and combining (101), (102), (103), (104) and a union bound, we conclude, when C_2 is sufficiently large, that

$$\mathbb{E}_\theta \phi_{\mathcal{P}_{\text{Temp}}} \leq \sum_{t\in\mathcal{T}:\,t<C_2/12}\sum_{g\in[t]}\mathbb{P}_\theta\left(t\sum_{j=1}^p V_{t,g}(j) > r_t^{\text{Temp}}\right) + \sum_{t\in\mathcal{T}:\,C_2/12\leq t\leq\Delta}\mathbb{P}_\theta\left(A_t^{\text{MoM}} > r_t^{\text{Temp}}\right)$$
$$+ \sum_{t\in\mathcal{T}:\,t>\Delta}\mathbb{P}_\theta\left(A_t^{\text{MoM}} > r_t^{\text{Temp}}\right)$$
$$\leq \frac{C_2^2 C_4}{12 C_1^2} + \sum_{t\in\mathcal{T}:\,C_2/12\leq t\leq\Delta}\frac{\varepsilon}{10}e^{-\sqrt{t}} + \sum_{t\in\mathcal{T}:\,t>\Delta}\frac{\varepsilon}{10\log_2(n/2)} \leq \varepsilon/12 + \varepsilon/10 + \varepsilon/10 < \varepsilon/2.$$

for all $\theta \in \Theta_0(p,n)$.

Alternative term. We again follow the proof of Theorem 4 and reach

$$\mathbb{P}_\theta\left(\tilde{t}\sum_{j=1}^p V_{\tilde{t},g}(j) \leq r_{\tilde{t}}^{\text{Temp}}\right)$$
$$\leq \mathbb{P}_\theta\left(\sum_{j=1}^p\left(\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 - \mathbb{E}\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 + \sqrt{2}\left(\mu_1(j) - \mu_2(j)\right)\overline{Z}'_{\tilde{t},g}(j)\right) \leq -\frac{\rho^2}{16\tilde{t}} - \frac{\|\mu_1 - \mu_2\|_2^2}{4}\right). \tag{105}$$

Again by Theorem 4.1 of [51], we have

$$\mathbb{P}_\theta\left(\sum_{j=1}^p\left(\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 - \mathbb{E}\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2\right) \leq -\frac{\rho^2}{16\tilde{t}}\right) \leq \frac{256(\tilde{t})^2\sum_{j=1}^p\mathbb{E}_\theta\left(\overline{Z}'_{\tilde{t},g}(j)\right)^4}{\rho^4} \leq \frac{C_6 p G_{\tilde{t}}^2}{\rho^4},$$

and

$$\mathbb{P}_\theta\left(\sum_{j=1}^p\sqrt{2}\left(\mu_1(j) - \mu_2(j)\right)\overline{Z}'_{\tilde{t},g}(j) \leq -\frac{\|\mu_1 - \mu_2\|_2^2}{4}\right) \leq \frac{32\sum_{j=1}^p(\mu_1(j) - \mu_2(j))^2\mathbb{E}_\theta\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2}{\|\mu_1 - \mu_2\|_2^4}$$
$$\leq \frac{C_6 G_{\tilde{t}}}{\rho^2},$$

for some $C_6 > 0$, depending only on α, K and c. Combining these with (105), as long as

$$\rho^2 \geq C_3 p^{1/2}(\log\log(8n))(\log\log\log(64n))^2,$$

for sufficiently large C_3, we are guaranteed that

$$\mathbb{P}_\theta\left(\tilde{t}\sum_{j=1}^p V_{\tilde{t},g}(j) \leq r_{\tilde{t}}^{\text{Temp}}\right) \leq \frac{C_4}{C_1^2} < \frac{\varepsilon}{12}.$$

Note that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P}_{\text{Temp}}}) \leq \mathbb{P}_\theta\left(A_{\tilde{t}}^{\text{MoM}} \leq r_{\tilde{t}}^{\text{Temp}}\right) \leq \mathbb{P}_\theta\left(\sum_{g=1}^{G_{\tilde{t}}}\mathbb{1}_{\left\{\tilde{t}\sum_{j=1}^p V_{\tilde{t},g}(j)\leq r_{\tilde{t}}^{\text{Temp}}\right\}} \geq G_{\tilde{t}}/2\right).$$

We consider three cases separately: (i) $\tilde{t} < C_\varepsilon$, (ii) $C_\varepsilon \leq \tilde{t} < \Delta$, and (iii) $\tilde{t} \geq \Delta$. Using the same argument as in controlling the null term, we can show that $\mathbb{E}_\theta(1 - \phi_{\mathcal{P}_{\text{Temp}}}) \leq \varepsilon/2$ and this completes the proof for $\alpha > 4$. □

Remark 1. *A similar result can be obtained for $2 < \alpha \leq 4$. In this case, Theorem 22 holds with $r_t^{\text{Temp}} = C_1 p^{2/\alpha'} G_t$ and $v_{\mathcal{P}_{\text{Temp}}}^{\text{U}} = p^{2/\alpha'}(\log\log(8n))(\log\log\log(64n))^2$ for any $\alpha' < \alpha$, with G_t and Δ remaining unchanged.*

An MA(1) example: When forming the test $\phi_{\mathcal{P}_{\text{Temp}}}$, we assume $\mathbb{E}\overline{Z}_{t,g}^2(j)$ is known for all $t \in \mathcal{T}$, $g \in [G_t]$, and $j \in [p]$. This term also appears in the setting with independent observations, where it simplifies to G_t/t, given the assumption that the variance of each error term is 1; see Definition 3. A close examination of the proof of Theorem 22 reveals that if, with high probability, $\sum_{j=1}^p \mathbb{E}\overline{Z}_{t,g}^2(j)$ can be accurately estimated with an error of $O(\sqrt{p}G_t/t)$ for all $t \in \mathcal{T}$, $g \in [G_t]$, and $j \in [p]$, then the conclusion of Theorem 22 still holds for the modified test $\phi_{\mathcal{P}_{\text{Temp}}}$, where the exact expectations $\mathbb{E}\overline{Z}_{t,g}^2(j)$ are replaced by their estimators.

We consider an example of a specific temporal dependence model. Assume that E has independent component series, and for each $j \in [p]$, the j-th component series $\{E_i(j)\}_{i \in [n]}$ follows a moving average process of order 1 (MA(1))

$$E_i(j) = \omega_i(j) + \pi_{\text{ma}}\omega_{i-1}(j), \tag{106}$$

where $\{\omega_i(j)\}_{i=0,1,\dots}$ is an independent white noise sequence satisfying $\mathbb{E}[\omega_i(j)^2] = (1 + \pi_{\text{ma}}^2)^{-1}$ and $\mathbb{E}\big|\omega_i(j) + \pi_{\text{ma}}\omega_{i-1}(j)\big|^\alpha \le K^\alpha$ for all i. The lag-1 autocorrelation is given by $r_1 := \pi_{\text{ma}}/(1 + \pi_{\text{ma}}^2)$. As discussed in Section VI-B, the interlaced α-mixing coefficient of our noise sequence $\{E_i\}_{i \in [n]}$ satisfies $\alpha^*(i) \le e^{1-i}$ for all $i \in [n-1]$. Now, if we can estimate r_1 well, then a plug-in estimator for $\mathbb{E}\overline{Z}_{t,g}^2(j)$ can be used. We formalise this in the following corollary.

Corollary 23. *Assume $\alpha > 4$ and let $\varepsilon \in (0,1)$. Consider the MA(1) data-generating mechanism for the noise sequence described by (106) and in the last paragraph. Let $\hat{r}_1$ be any estimator for r_1 satisfying*

$$\mathbb{P}\big(|\hat{r}_1 - r_1| > cp^{-1/2}\big) \le \varepsilon/2, \tag{107}$$

for some $c > 0$, depending on ε. Then, after modifying $V_{t,g}$ to become

$$V_{t,g}(j) := \overline{Z}_{t,g}^2(j) - \frac{(t/G_t) + 2\{(t/G_t) - 1\}\hat{r}_1}{(t/G_t)^2},$$

the theoretical guarantee on $\phi_{\mathcal{P}_{\text{Temp}}}$ in Theorem 22 remains valid, with possibly increased values of C_1, C_2 and C_3.

Suppose we have a (historical) dataset $X_1^{(h)}, \dots, X_m^{(h)}$ of size $m \in \mathbb{Z}^+$, within which no change point is present. Each observation can be written as $X_i^{(h)} = \mu_0 + E_i^{(h)}$, where the noise follows the data-generating mechanism described in the previous paragraph. The lag-1 autocorrelation can be estimated by

$$\hat{r}_1 := \frac{1}{(m-1)p}\sum_{j=1}^p \sum_{i=1}^{m-1}\left(X_i^{(h)}(j) - \frac{1}{m}\sum_{k=1}^m X_k^{(h)}(j)\right)\left(X_{i+1}^{(h)}(j) - \frac{1}{m}\sum_{k=1}^m X_k^{(h)}(j)\right).$$

This estimator can be shown to satisfy condition (107) when, informally speaking, the sample size m is significantly larger than $\sqrt{p}$.

Note that in (106), all components share the same lag-1 coefficient. If this assumption does not hold, then the requirement on the estimators becomes $\mathbb{P}\big(\sum_{j=1}^p |\hat{r}_1(j) - r_1(j)| > cp^{-1/2}\big) \le \varepsilon/2$ in place of (107) in Corollary 23, where $r_1(j)$ denotes the lag-1 autocorrelation for the j-th component series and $\hat{r}_1(j)$ its estimator.

3) Fewer than two finite moments: As mentioned in Section VI-C of the main text, our test procedure has two components. One component utilises a robust mean estimator $\hat{\mu}^{\text{RM}}$ from Algorithm 1-7 of [42], developed specifically for distributions satisfying (36).

Proposition 24 ([42], Theorem 1.2). *Let $1 \le \alpha \le 2$. For $t \in \mathbb{Z}^+$, let $X_1, \dots, X_t$ be independent random vectors in $\mathbb{R}^p$ with mean μ. Assume that the distribution of $W_i := X_i - \mu$ belongs to $\mathcal{W}_\alpha$ for each $i \in [t]$. Then, there exists a polynomial-time algorithm that, given inputs $X_1, \dots, X_t$ and $\eta > 0$, outputs $\hat{\mu}_t^{\text{RM}}(\{X_i\}_{i=1}^t; \eta)$. There exist absolute constants $C_{01}, C_{02} > 0$ such that, for any $0 < \eta < 1$, when $t \ge C_{01}\log(1/\eta)$, with probability at least $1 - \eta$, it holds that*

$$\|\hat{\mu}_t^{\text{RM}}(\{X_i\}_{i=1}^t; \eta) - \mu\|_2 \le C_{02}\left\{\sqrt{\frac{p}{t}} + \left(\frac{p}{t}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log(1/\eta)}{t}\right)^{\frac{\alpha-1}{\alpha}}\right\}.$$

Similar to the limitations of using $\hat{\mu}^{\text{RSM}}$ as discussed in Section III-B, test statistics based on $\hat{\mu}^{\text{RM}}$ only have theoretical guarantee when the change is sufficiently away from the boundary, due to the condition $t \ge C_{01}\log(1/\eta)$ in Proposition 24. To cover the case when the potential change occurs near the boundary, we need to adopt a different strategy that does not have this limitation on the minimal sample size.

Recall that $Z_i = (X_i - X_{n-i+1})/2$, for $i \le n/2$. We let

$$\begin{cases} \tilde{A}_t = \|\sum_{i=1}^t Z_i/t\|_2, & t \in \mathcal{T} \cap \{t \le \tilde{\Delta}_1\}, \\ \tilde{A}_t^{\text{RM}} = \|\hat{\mu}_t^{\text{RM}}(\{Z_i\}_{i=1}^t; \eta_t)\|_2, & t \in \mathcal{T} \cap \{t > \tilde{\Delta}_1\}, \end{cases}$$

and define the test

$$\phi_{\mathcal{W}_\alpha}^{\text{RM}} := \mathbb{1}_{\{\max_{t \in \mathcal{T} \cap \{t \le \tilde{\Delta}_1\}} \tilde{A}_t/\tilde{r}_t > 1\}} \vee \mathbb{1}_{\{\max_{t \in \mathcal{T} \cap \{t > \tilde{\Delta}_1\}} \tilde{A}_t^{\text{RM}}/\tilde{r}_t^{\text{RM}} > 1\}}, \tag{108}$$

where $\tilde{\Delta}_1, \tilde{r}_t, \tilde{r}_t^{\mathrm{RM}}$ and η_t are specified later in Theorem 25.

We prove the following theorem on the theoretical guarantee of the test $\phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}$, constructed above. Theorem 13 then follows as an immediate consequence.

Theorem 25. *Assume $1 \leq \alpha \leq 2$. For any $\varepsilon \in (0,1)$, there exist $C_1, C_2, C_3, C_4 > 0$ depending only on α and ε, such that the test defined in* (108) *with*

$$\tilde{r}_t = C_1 \tilde{\Delta}_1^{\frac{2-\alpha}{2\alpha}} \log^{\frac{1}{\alpha}}(\tilde{\Delta}_1) \sqrt{\frac{p}{t}},$$

$$\eta_t = \exp\left\{-\frac{t \wedge \tilde{\Delta}_2}{C_2}\right\}, \quad \tilde{r}_t^{\mathrm{RM}} = C_3\left(\sqrt{\frac{p}{t}} + \left(\frac{p}{t}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log(1/\eta_t)}{t}\right)^{\frac{\alpha-1}{\alpha}}\right),$$

$$\tilde{\Delta}_1 = C_2 \log(16/\varepsilon) \quad \text{and} \quad \tilde{\Delta}_2 = C_2 \log(16\log(2n)/\varepsilon),$$

satisfies that

$$\mathcal{R}_{\mathcal{W}_\alpha^\otimes}(\{\rho_{t_0}\}_{t_0\in[n-1]}, \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}) \leq \varepsilon,$$

as long as

$$\rho_{t_0} \geq C_4\left\{\sqrt{\frac{p}{t_0 \wedge (n-t_0)}} + \left(\frac{p}{t_0 \wedge (n-t_0)}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log\log(n)}{t_0 \wedge (n-t_0)}\right)^{\frac{\alpha-1}{\alpha}}\right\}, \quad (109)$$

for all $t_0 \in [n-1]$.

Proof. We denote $\tilde{\mathcal{T}}_1 := \{t \in \mathcal{T} : t \leq \tilde{\Delta}_1\}$, $\tilde{\mathcal{T}}_2 := \{t \in \mathcal{T} : \tilde{\Delta}_1 < t \leq \tilde{\Delta}_2\}$ and $\tilde{\mathcal{T}}_3 := \{t \in \mathcal{T} : t > \tilde{\Delta}_2\}$. **Null term.** Under the null hypothesis, we need to control

$$\mathbb{E}_\theta \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}} = \sum_{t\in\tilde{\mathcal{T}}_1} \mathbb{P}_\theta(\tilde{A}_t > \tilde{r}_t) + \sum_{t\in\tilde{\mathcal{T}}_2} \mathbb{P}_\theta(\tilde{A}_t^{\mathrm{RM}} > \tilde{r}_t^{\mathrm{RM}}) + \sum_{t\in\tilde{\mathcal{T}}_3} \mathbb{P}_\theta(\tilde{A}_t^{\mathrm{RM}} > \tilde{r}_t^{\mathrm{RM}}).$$

We first control the first term. Notice that Z_i are independent random vectors satisfying (36), since $\mathbb{E}_\theta Z_i = 0$ and

$$\mathbb{E}_\theta|\langle Z_i, v\rangle|^\alpha = \frac{1}{2^\alpha}\mathbb{E}_\theta|\langle X_i - X_{n-i+1}, v\rangle|^\alpha \leq \mathbb{E}_\theta|\langle X_i - \mu, v\rangle|^\alpha \leq 1,$$

for any unit vector v. Thus, we have by Lemma 39

$$\mathbb{P}_\theta(\tilde{A}_t > \tilde{r}_t) \leq \frac{\pi}{C_1^\alpha (\tilde{\Delta}_1/t)^{\frac{2-\alpha}{2}} \log(\tilde{\Delta}_1)}.$$

Now, by setting $C_1 \geq (16\pi/\varepsilon)^{1/\alpha}$, we are guaranteed to have

$$\sum_{t\in\tilde{\mathcal{T}}_1} \mathbb{P}_\theta(\tilde{A}_t > \tilde{r}_t) \leq \varepsilon/4.$$

Now, for the second and third terms, we note that $t \geq C_2 \log(1/\eta_t)$. Thus, by using Proposition 24 with $\eta = \eta_t$, we have $\mathbb{P}(\tilde{A}_t^{\mathrm{RM}} > \tilde{r}_t^{\mathrm{RM}}) \leq \eta_t$ and thus

$$\sum_{t\in\tilde{\mathcal{T}}_2} \mathbb{P}_\theta(\tilde{A}_t^{\mathrm{RM}} > \tilde{r}_t^{\mathrm{RM}}) + \sum_{t\in\tilde{\mathcal{T}}_3} \mathbb{P}_\theta(\tilde{A}_t^{\mathrm{RM}} > \tilde{r}_t^{\mathrm{RM}}) \leq \sum_{t\in\tilde{\mathcal{T}}_2} \exp\left\{-\frac{t}{C_2}\right\} + \sum_{t\in\tilde{\mathcal{T}}_3} \frac{\varepsilon}{16\log(2n)}$$

$$\leq 2\exp\left\{-\frac{\tilde{\Delta}_1}{C_2}\right\} + \frac{\varepsilon\log_2(n/2)}{16\log(2n)} < \varepsilon/4.$$

We therefore conclude $\mathbb{E}_\theta \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}} < \varepsilon/2$.

Alternative term. Consider, as before, the point $\tilde{t} \in \mathcal{T}$ such that $t_0/2 < \tilde{t} \leq t_0 \leq n/2$. We shall deal with the case $t_0 > n/2$ later. Note that $Z_1, \ldots, Z_{\tilde{t}}$ are independent with mean $(\mu_1 - \mu_2)/2$ and that $Z_i - (\mu_1 - \mu_2)/2$ satisfies (36) for each i. We first consider the case $\tilde{t} \leq \tilde{\Delta}_1$. In this case, we have

$$\mathbb{P}_\theta(\tilde{A}_{\tilde{t}} < \tilde{r}_{\tilde{t}}) = \mathbb{P}_\theta\left(\left\|\sum_{i=1}^{\tilde{t}} Z_i/\tilde{t}\right\|_2 < \tilde{r}_{\tilde{t}}\right) \leq \mathbb{P}_\theta\left(\left|\left\|\frac{\mu_1-\mu_2}{2}\right\|_2 - \left\|\sum_{i=1}^{\tilde{t}} Z_i/\tilde{t} - \frac{\mu_1-\mu_2}{2}\right\|_2\right| < \tilde{r}_{\tilde{t}}\right). \quad (110)$$

By Lemma 39 and the choice of $\tilde{r}_{\tilde{t}}$ with $C_1 \geq (16\pi/\varepsilon)^{1/\alpha}$, we have

$$\mathbb{P}_\theta\left(\left\|\sum_{i=1}^{\tilde{t}} Z_i/\tilde{t} - \frac{\mu_1-\mu_2}{2}\right\|_2 > \tilde{r}_{\tilde{t}}\right) \leq \frac{\varepsilon}{16},$$

Therefore, as long as

$$\left\|\frac{\mu_1 - \mu_2}{2}\right\|_2 \geq 4C_1 \tilde{\Delta}_1^{\frac{2-\alpha}{2\alpha}} \log^{\frac{1}{\alpha}}(\tilde{\Delta}_1)\sqrt{\frac{p}{t_0}},$$

we obtain $\left\|\frac{\mu_1-\mu_2}{2}\right\|_2 \geq 2\tilde{r}_{\tilde{t}}$ and thus $\mathbb{E}_\theta(1 - \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}) \leq \mathbb{P}_\theta(\tilde{A}_{\tilde{t}} < r_{\tilde{t}}) \leq \varepsilon/16$. In the case of $t_0 > n/2$, we should consider instead the point $\tilde{t} \in \mathcal{T}$ such that $(n-t_0)/2 < \tilde{t} \leq n - t_0 \leq n/2$, and the same arguments as above show that $\mathbb{E}_\theta(1 - \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}) \leq \mathbb{P}_\theta(\tilde{A}_{\tilde{t}} < r_{\tilde{t}}) \leq \varepsilon/16$ as long as

$$\left\|\frac{\mu_1 - \mu_2}{2}\right\|_2 \geq 4C_1 \tilde{\Delta}_1^{\frac{2-\alpha}{2\alpha}} \log^{\frac{1}{\alpha}}(\tilde{\Delta}_1)\sqrt{\frac{p}{n-t_0}}.$$

Now, consider the case $\tilde{t} > \tilde{\Delta}_1$, when $t_0 \leq n/2$. Similar to (110), we now have

$$\mathbb{P}_\theta(\tilde{A}_{\tilde{t}}^{\mathrm{RM}} < \tilde{r}_{\tilde{t}}^{\mathrm{RM}}) = \mathbb{P}_\theta\left(\left\|\hat{\mu}_{\tilde{t}}^{\mathrm{RM}}(\{Z_i\}_{i=1}^{\tilde{t}}; \eta_{\tilde{t}})\right\|_2 < \tilde{r}_{\tilde{t}}^{\mathrm{RM}}\right)$$

$$\leq \mathbb{P}_\theta\left(\left|\left\|\frac{\mu_1 - \mu_2}{2}\right\|_2 - \left\|\hat{\mu}_{\tilde{t}}^{\mathrm{RM}}(\{Z_i\}_{i=1}^{\tilde{t}}; \eta_{\tilde{t}}) - \frac{\mu_1 - \mu_2}{2}\right\|_2\right| < \tilde{r}_{\tilde{t}}^{\mathrm{RM}}\right).$$

Thus, as long as

$$\left\|\frac{\mu_1 - \mu_2}{2}\right\|_2 \geq 4C_3\left(\sqrt{\frac{p}{t_0}} + \left(\frac{p}{t_0}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log\log(n)}{t_0}\right)^{\frac{\alpha-1}{\alpha}}\right),$$

we have $\mathbb{P}_\theta(\tilde{A}_{\tilde{t}}^{\mathrm{RM}} < \tilde{r}_{\tilde{t}}^{\mathrm{RM}}) \leq \varepsilon/16$. In the case of $t_0 > n/2$, we should consider instead the point $\tilde{t} \in \mathcal{T}$ such that $(n-t_0)/2 < \tilde{t} \leq n - t_0 \leq n/2$, and the same arguments as above show that $\mathbb{P}_\theta(\tilde{A}_{\tilde{t}}^{\mathrm{RM}} < \tilde{r}_{\tilde{t}}^{\mathrm{RM}}) \leq \varepsilon/16$ as long as

$$\left\|\frac{\mu_1 - \mu_2}{2}\right\|_2 \geq 4C_3\left(\sqrt{\frac{p}{n-t_0}} + \left(\frac{p}{n-t_0}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log\log(n)}{n-t_0}\right)^{\frac{\alpha-1}{\alpha}}\right).$$

Together, we obtain $\mathbb{E}_\theta(1 - \phi_{\mathcal{W}_\alpha}^{\mathrm{RM}}) \leq \varepsilon/16$, as long as

$$\rho_{t_0} \geq C_4\left(\sqrt{\frac{p}{t_0 \wedge (n-t_0)}} + \left(\frac{p}{t_0 \wedge (n-t_0)}\right)^{\frac{\alpha-1}{\alpha}} + \left(\frac{\log\log(n)}{t_0 \wedge (n-t_0)}\right)^{\frac{\alpha-1}{\alpha}}\right),$$

for all $t_0 \in [n-1]$ and some $C_4 > 0$. $\qquad\square$

Proposition 26. *For the testing problem*

$$\mathrm{H}_0 : \theta \in \Theta_0(p,n) \quad \text{vs.} \quad \mathrm{H}_1 : \theta \in \bigcup_{t_0=1}^{n-1} \Theta(p,n,\rho_{t_0}),$$

when $p = 1$, it holds that

$$\inf_{\phi \in \Phi} \mathcal{R}_{\mathcal{W}_\alpha^\otimes}(\{\rho_{t_0}\}_{t_0 \in [n-1]}, \phi) \geq 1/2,$$

if

$$\rho_{t_0} \leq \left(t_0 \wedge (n-t_0)\right)^{-(1-\frac{1}{\alpha})}, \tag{111}$$

for some $t_0 \in [n-1]$.

Proof. Consider two distributions P_+ and P_- on $\mathbb{R}$ such that

$$P_+(\{0\}) = P_-(\{0\}) = 1 - u, \quad P_+(\{c\}) = P_-(\{-c\}) = u,$$

where $u \in [0,1]$ and $c > 0$ are to be specified. Let $\mu_+ = cu$ and $\mu_- = -cu$ denote the mean of P_+ and P_-, respectively. Note that we have $|\mu_+ - \mu_-| = 2cu$. Also, we have the α-th central moment of both distributions satisfies

$$\mathbb{E}_{P_+}(|X - \mu_+|^\alpha) = \mathbb{E}_{P_-}(|X - \mu_-|^\alpha) = c^\alpha u(1-u)(u^{\alpha-1} + (1-u)^{\alpha-1}) \leq 2c^\alpha u,$$

if $u \leq 1/2$.

We first focus on the case where $t_0 \leq n/2$, i.e. when $\left(\frac{1}{t_0}\right)^{\frac{\alpha-1}{\alpha}}$ dominates. Consider the following two sequences of random variables

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} P_+, \quad Y_1, \ldots, Y_{t_0} \overset{i.i.d}{\sim} P_-, \; Y_{t_0+1}, \ldots, Y_n \overset{i.i.d}{\sim} P_+.$$

Let $P_+^{\otimes n}$ and $P_-^{\otimes n}$ denote the n-fold product distribution of P_+ and P_-, respectively. With the choice of $u = 1/(2t_0)$ and $c = (2u)^{-\frac{1}{\alpha}}$, we have $2c^\alpha u \leq 1$ and

$$1 - \mathrm{TV}(P_+^{\otimes t_0}, P_-^{\otimes t_0}) \geq 1 - t_0\mathrm{TV}(P_+, P_-) = 1 - t_0 u \geq 1/2,$$

with $|\mu_+ - \mu_-| = t_0^{-\frac{\alpha-1}{\alpha}}$, which proves the claim.

In the case of $t_0 > n/2$, one simply chooses

$$Y_1, \ldots, Y_{t_0} \overset{i.i.d}{\sim} P_+, \ Y_{t_0+1}, \ldots, Y_n \overset{i.i.d}{\sim} P_-,$$

and the same arguments lead to the corresponding result. $\qquad\square$

4) Change away from boundary: For $\mathcal{Q} = \mathcal{P}_{\alpha,K}^{\otimes}$ with $\alpha \geq 4$, we modify the median-of-means-type test proposed in Section III-B. Recall that $\{Z_2, Z_4, \ldots, Z_t\}$ is used for coordinate selection in (25). We now split this set into G^{res} groups of equal size, and use $\overline{Z}_{t,g,2}$ to denote the sample mean of the g-th group. Our new test relies on a more robust coordinate selection step compared to (25) by considering the following statistic:

$$V_{t,g,a^{\mathrm{res}}}^{\mathrm{res}}(j) := \left(\overline{Z}_{t,g,1}^2(j) - \frac{2G_t}{t} \right) \mathbb{1}_{\left\{ \sqrt{t/(2G^{\mathrm{res}})} \left| \mathrm{median}\left(\overline{Z}_{t,1,2}(j), \ldots, \overline{Z}_{t,G^{\mathrm{res}},2}(j) \right) \right| \geq a^{\mathrm{res}} \right\}}, \quad j \in [p],$$

where both the number of groups G^{res} and the threshold a^{res} are to be specified in Theorem 27. With $\mathcal{T}^{\mathrm{res}} := \mathcal{T} \cap \left\{ \lceil (t^{\mathrm{res}} + 1)/2 \rceil, \ldots, n+1 - \lceil (t^{\mathrm{res}} + 1)/2 \rceil \right\}$, our test is

$$\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}^{\mathrm{res}}} A_{t,a^{\mathrm{res}}}^{\mathrm{MoM}}/r_t > 1\}}, \tag{112}$$

where $A_{t,a^{\mathrm{res}}}^{\mathrm{MoM}}$ is the same as (26) but with each $V_{t,g,a}(j)$ replaced by $V_{t,g,a^{\mathrm{res}}}^{\mathrm{res}}(j)$, for $g \in [G_t]$.

For $\mathcal{Q} = \mathcal{G}_{\alpha,K}^{\otimes}$ with $0 < \alpha < 2$, we adopt a similar robust strategy for selecting signal coordinates in the sparse regime by replacing (12) with the following statistic:

$$A_{t,a} := \sum_{j=1}^p \{Y_{t,1}^2(j) - 1\} \mathbb{1}_{\left\{ \sqrt{t/(2G^{\mathrm{res}})} \left| \mathrm{median}\left(\overline{Z}_{t,1,2}(j), \ldots, \overline{Z}_{t,G^{\mathrm{res}},2}(j) \right) \right| \geq a^{\mathrm{res}} \right\}},$$

and the test takes the form of

$$\phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}^{\mathrm{res}}} A_{t,a} > r\}}. \tag{113}$$

We prove the following theorem on the theoretical guarantee of the test $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$, constructed above. Theorem 14 then follows as an immediate consequence.

Theorem 27. *Let* $t^{\mathrm{res}} = 32\{\log(e^2 p/s) + s^{-1} \log\log(8n)\}$ *and* $\mathcal{Q} = \mathcal{P}_{\alpha,K}^{\otimes}$ *with* $\alpha \geq 4$. *For any* $\varepsilon \in (0,1)$, *there exist* $C_1, C_2, C_3 > 0$ *depending only on* α, K *and* ε, *such that the test* $\phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$ *defined in* (112) *with*

$$a^{\mathrm{res}} = C_1, \ G^{\mathrm{res}} = 2^{\lfloor \log_2(t^{\mathrm{res}}/2) \rfloor}, \ r_t = C_2 \sqrt{s} G_t, \ G_t = (t \wedge \Delta)/2 \text{ and } \Delta = 2^{4 + \lceil \log_2 \log\log(8n) \rceil},$$

satisfies that

$$\mathcal{R}_{\mathcal{P}}(\rho, \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_3 v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}$, *where*

$$v_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}} := s\{\log(ep/s) + \log\log(8n)\}.$$

Proof. The proof is analogous to that of Proposition 7; we thus omit many details and highlight those places where the arguments differ.

Null term. For any $\theta \in \Theta_0(p,n)$, we have by a union bound that

$$\mathbb{E}_\theta \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}} \leq \sum_{t \in \mathcal{T}^{\mathrm{res}}} \mathbb{P}_\theta\left(A_{t,a}^{\mathrm{MoM}} > r_t \right) \leq \sum_{t \in \mathcal{T}^{\mathrm{res}}} \mathbb{P}_\theta\left(|\mathcal{J}_{t,a^{\mathrm{res}}}| > s \right) + \sum_{t \in \mathcal{T}^{\mathrm{res}}} \sup_{J \subseteq [p]:|J| \leq s} \mathbb{P}_\theta\left(A_{t,*,J}^{\mathrm{MoM}} > r_t \right), \tag{114}$$

with $A_{t,*,J}^{\mathrm{MoM}}$ unchanged from the proof of Proposition 7 and $\mathcal{J}_{t,a^{\mathrm{res}}}$ modified to be

$$\mathcal{J}_{t,a^{\mathrm{res}}} := \left\{ j \in [p] : \sqrt{t/2G^{\mathrm{res}}} \left| \mathrm{median}\left(\overline{Z}_{t,1,2}(j), \ldots, \overline{Z}_{t,G^{\mathrm{res}},2}(j) \right) \right| \geq a^{\mathrm{res}} \right\},$$

for $t \in \mathcal{T}^{\mathrm{res}}$. The second term in (114) can still be bounded by $\varepsilon/8$ with r_t, G_t and Δ unchanged from Proposition 7. We now bound the first term. By Fuk–Nagaev inequality (Proposition 33), we have for any $j \in [p]$, $g \in [G^{\mathrm{res}}]$ and $t \in \mathcal{T}^{\mathrm{res}}$

$$\mathbb{P}_\theta\left(\sqrt{\frac{t}{2G^{\mathrm{res}}}} |\overline{Z}_{t,g,2}(j)| \geq a^{\mathrm{res}} \right) \leq 2 \left(\frac{(\alpha+2)(K^\alpha 2^{\alpha/2} t/(2G^{\mathrm{res}}))^{1/\alpha}}{\alpha a^{\mathrm{res}} \sqrt{t/(2G^{\mathrm{res}})}} \right)^\alpha + 2\exp\left\{ -\frac{2(a^{\mathrm{res}})^2}{(\alpha+2)^2 e^\alpha} \right\}.$$

$$\leq \frac{K^\alpha}{(a^{\mathrm{res}}/3)^\alpha (t/G^{\mathrm{res}})^{\alpha/2-1}} + \exp\left\{ 1 - \frac{(a^{\mathrm{res}})^2}{2\alpha^2 e^\alpha} \right\}$$

$$\leq (3K/a^{\mathrm{res}})^\alpha + \exp\left\{ 1 - \frac{(a^{\mathrm{res}})^2}{2\alpha^2 e^\alpha} \right\} \leq \varepsilon/36,$$

when $a^{\mathrm{res}} = C_1$ is sufficiently large, depending on K, α and ε. Consequently, by the multiplicative Chernoff bound, we have for $j \in [p]$

$$\mathbb{P}_\theta(j \in \mathcal{J}_{t,a^{\mathrm{res}}}) \leq \mathbb{P}_\theta\left(\left|\left\{g \in [G^{\mathrm{res}}] : \sqrt{\frac{t}{2G^{\mathrm{res}}}}|\overline{Z}_{t,g,2}(j)| \geq a^{\mathrm{res}}\right\}\right| \geq G^{\mathrm{res}}/2\right)$$

$$\leq \mathbb{P}_\theta\left(\left|\left\{g \in [G^{\mathrm{res}}] : \sqrt{\frac{t}{2G^{\mathrm{res}}}}|\overline{Z}_{t,g,2}(j)| \geq a^{\mathrm{res}}\right\}\right| \geq \frac{\varepsilon G^{\mathrm{res}}}{36}\left(1 + \left(\frac{18}{\varepsilon} - 1\right)\right)\right)$$

$$\leq \exp\left\{-\frac{\varepsilon G^{\mathrm{res}}}{36}\left(\frac{18}{\varepsilon}\log\left(\frac{18}{\varepsilon}\right) - \frac{18}{\varepsilon} + 1\right)\right\} \leq \exp\left\{-\frac{G^{\mathrm{res}}}{2}\log(6/\varepsilon)\right\} \leq (\varepsilon/6)^{G^{\mathrm{res}}/2}. \quad (115)$$

Then, again by a standard binomial tail bound and a union bound

$$\sum_{t \in \mathcal{T}^{\mathrm{res}}} \mathbb{P}_\theta(|\mathcal{J}_{t,a^{\mathrm{res}}}| > s) \leq \log_2(n)\left(\frac{ep(\varepsilon/6)^{G^{\mathrm{res}}/2}}{s}\right)^s \leq \varepsilon/4,$$

when $G^{\mathrm{res}} \geq 2\big(\log(e^2 p/s) + s^{-1}\log\log(8n)\big)$. Putting things together, we reach $\mathbb{E}_\theta \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}} \leq \varepsilon/2$.

Alternative term. First, according to our definition of $\mathcal{T}^{\mathrm{res}}$, there still exists a unique $\tilde{t} \in \mathcal{T}^{\mathrm{res}}$ such that $t_0/2 < \tilde{t} \leq t_0$, where $t_0 \geq t^{\mathrm{res}} + 1$ (and without loss of generality $t_0 \leq n/2$) is the true change point location. We retain most notation used in the proof of Proposition 7, with one modification $\mathcal{H}_{\delta,a^{\mathrm{res}}} := \big\{j \in [p] : |\delta(j)| \geq 2a^{\mathrm{res}}\sqrt{G^{\mathrm{res}}}\big\}$. The reasoning behind this change is

$$\sqrt{\frac{\tilde{t}}{2G^{\mathrm{res}}}}\,\mathrm{median}\big(\overline{Z}_{\tilde{t},1,2}(j), \ldots, \overline{Z}_{\tilde{t},G^{\mathrm{res}},2}(j)\big) = \sqrt{\frac{\tilde{t}}{2G^{\mathrm{res}}}}\,\mathrm{median}\big(\overline{Z}'_{\tilde{t},1,2}(j), \ldots, \overline{Z}'_{\tilde{t},G^{\mathrm{res}},2}(j)\big) + \frac{\delta(j)}{\sqrt{G^{\mathrm{res}}}}. \quad (116)$$

The only major difference in the proof lies in the argument for establishing

$$\mathbb{P}_\theta\left(\sum_{j \in \mathcal{J}_{\tilde{t},a^{\mathrm{res}}} \cap \mathcal{H}_{\delta,a^{\mathrm{res}}}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12\log(8/\varepsilon)}\right) \leq \varepsilon/8. \quad (117)$$

For $j \in \mathcal{H}_{\delta,a^{\mathrm{res}}}$, we have by (116) and (115)

$$\mathbb{P}_\theta(j \notin \mathcal{J}_{\tilde{t},a^{\mathrm{res}}}) = \mathbb{P}_\theta\left(\sqrt{\tilde{t}/2G^{\mathrm{res}}}\big|\mathrm{median}\big(\overline{Z}_{\tilde{t},1,2}(j), \ldots, \overline{Z}_{\tilde{t},G^{\mathrm{res}},2}(j)\big)\big| < a^{\mathrm{res}}\right)$$

$$\leq \mathbb{P}_\theta\left(\sqrt{\tilde{t}/2G^{\mathrm{res}}}\big(\mathrm{median}\big(\overline{Z}'_{\tilde{t},1,2}(j), \ldots, \overline{Z}'_{\tilde{t},G^{\mathrm{res}},2}(j)\big)\big) < -a^{\mathrm{res}}\right)$$

$$\leq (\varepsilon/6)^{G^{\mathrm{res}}/2} \leq \frac{\varepsilon}{2048} \leq \frac{1}{256\log(8/\varepsilon)},$$

whenever $G^{\mathrm{res}} \geq 10$. Thus, we still have

$$\sum_{j \in \mathcal{H}_{\delta,a^{\mathrm{res}}}} \mathrm{Var}_\theta\big(\delta(j)^2 \mathbb{1}_{\{j \in \mathcal{J}_{\tilde{t},a^{\mathrm{res}}}\}}\big) \leq \frac{\|\delta\|_2^4}{256\log(8/\varepsilon)}.$$

The condition $\rho^2 \geq 64a^2 s$ in the previous proof is replaced by $\rho^2 \geq 64(a^{\mathrm{res}})^2 sG^{\mathrm{res}}$ in order to obtain

$$\sum_{j \in \mathcal{H}_{\delta,a^{\mathrm{res}}}} \delta(j)^2 \geq \|\delta\|_2^2 - s\big(2a^{\mathrm{res}}\sqrt{G^{\mathrm{res}}}\big)^2 \geq \|\delta\|_2^2/2.$$

With these in place, (117) can be established and we can deduce $\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\mathrm{sparse}}^{\mathrm{MoM+res}}) \leq \varepsilon/2$ whenever

$$\rho^2 \geq C\big(s\Delta + (a^{\mathrm{res}})^2 sG^{\mathrm{res}}\big)$$

with a sufficiently large C. $\qquad\square$

Theorem 15 is again a direct consequence of the following result, which provides a theoretical guarantee for the test $\phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{res}}$ constructed earlier in (113).

Theorem 28. *Let $t^{\mathrm{res}} = 32\big\{\log(e^2 p/s) + s^{-1}\log\log(8n)\big\}$ and $\mathcal{Q} = \mathcal{G}_{\alpha,K}^{\otimes}$ with $0 < \alpha < 2$. For any $\varepsilon \in (0,1)$, there exist $C_1, C_2, C_3 > 0$ depending only on α, K and ε, such that the test $\phi_{\mathcal{G},\mathrm{sparse}}^{\mathrm{MoM+res}}$ defined in (113) with*

$$a^{\mathrm{res}} = C_1, \quad G^{\mathrm{res}} = 2^{\lfloor \log_2(t^{\mathrm{res}}/2)\rfloor} \quad \text{and} \quad r = C_2\big\{\sqrt{s\log\log(8n)} + \log\log(8n)\big\},$$

satisfies that

$$\mathcal{R}_\mathcal{G}(\rho, \phi^{\mathrm{res}}_{\mathcal{G},\mathrm{sparse}}) \leq \varepsilon,$$

as long as $\rho^2 \geq C_3 v^{\mathrm{res}}_{\mathcal{G},\mathrm{sparse}}$, where

$$v^{\mathrm{res}}_{\mathcal{G},\mathrm{sparse}} := s \log(ep/s) + \log\log(8n).$$

The proof of Theorem 28 is omitted, as it is very similar to the last proof.

Comparison with [7]: We compare Theorem 15 with the minimum signal strength requirement for a bootstrapped U-statistics-based test in Theorem 3.3 of [7]. Their result indicates that, for the sub-exponential noise class $\mathcal{G}^{\otimes}_{1,K}$, if $\log(p) = o(n^{1/7})$ (a mild dimension condition required for size control in Theorem 3.1 therein) and the change is away from the boundary by at least $O(\log^{5/2}(np))$, then the sum of Type I error and Type II error probabilities can be controlled provided that

$$t_0(n - t_0)\|\mu_1 - \mu_2\|_\infty \gtrsim n^{3/2} \log^{1/2}(np). \tag{118}$$

One immediate observation is that our requirement on the boundary removal, $O\big(\log(ep/s) + s^{-1}\log\log(8n)\big)$, is much smaller than their required $O(\log^{5/2}(np))$.

Now, suppose the s-sparse mean shift $\mu_1 - \mu_2$ takes the form

$$\mu_1 - \mu_2 = (\underbrace{a, \ldots, a}_{s}, \underbrace{0, \ldots, 0}_{p-s})^\top.$$

Without loss of generality, assume $t_0 \leq n/2$. Under our framework, the signal strength condition $\rho^2 \geq C_3 v^{\mathrm{res}}_{\mathcal{G},\mathrm{sparse}}$ implies

$$a \gtrsim t_0^{-1/2}\big(\log^{1/2}(ep/s) + s^{-1/2}\log^{1/2}\log(8n)\big),$$

while (118) requires

$$a \gtrsim (n/t_0)^{1/2} t_0^{-1/2}\big(\log^{1/2}(p) + \log^{1/2}(n)\big).$$

Our rate is clearly smaller across all parameter settings, with the advantage becoming particularly pronounced when the change location is not near the middle of the sequence, i.e. when t_0 is not of the same order as n.

E. Auxiliary results

We first present the definition and some basic properties of sub-Weibull random variables. For a more in-depth introduction and discussion, we refer to [45] and Section 2 of [53].

Definition 6 (Orlicz norms). *Let $f : [0, \infty) \to [0, \infty)$ be a non-decreasing function with $f(0) = 0$. The f-Orlicz norm of a real-valued random variable X is*

$$\|X\|_f := \inf\{t > 0 : \mathbb{E}f(|X|/t) \leq 1\}.$$

Definition 7 (sub-Weibull random variables). *A random variable X is sub-Weibull of order $\alpha > 0$, denoted sub-Weibull(α), if it has mean zero and*

$$\|X\|_{\psi_\alpha} < \infty,$$

with the function ψ_α defined by $\psi_\alpha(x) := \exp(x^\alpha) - 1$ for $x \geq 0$.

Proposition 29 ([45], Theorem 2.1). *Let X be a sub-Weibull(α) random variable with $0 < \alpha \leq 2$ and $\|X\|_{\psi_\alpha} = K < \infty$. Then, we have the following properties:*
(a) the tails of X satisfy

$$\mathbb{P}(|X| \geq x) \leq 2\exp\{-(x/K)^\alpha\} \qquad \text{for all } x \geq 0;$$

(b) Let $\|X\|_k := \mathbb{E}(|X|^k)^{1/k}, k \geq 1$, then

$$\|X\|_k \leq K'k^{1/\alpha}$$

for some absolute constant $K' > 0$.
(c) Conversely, if a random variable X has mean zero and satisfies $\mathbb{P}(|X| \geq x) \leq 2\exp\{-(x/K)^\alpha\}$ for all $x \geq 0$, then there exists $K'' > 0$, depending only on α and K, such that

$$\mathbb{E}\exp\{(|X|/K'')^\alpha\} \leq 2.$$

In other words, X is a sub-Weibull(α) random variable with $\|X\|_{\psi_\alpha} \leq K'' < \infty$.

Proposition 30 ([45], Proposition 2.1). *Let $\alpha > \alpha' > 0$ and X be a sub-Weibull(α) random variable with $\|X\|_{\psi_\alpha} = K < \infty$. Then there exists $K' > 0$, depending only on α' and K, such that X is a sub-Weibull(α') random variable with $\|X\|_{\psi_{\alpha'}} \leq K' < \infty$.*

We now provide two tail bound results from literature for sums and quadratic forms of independent sub-Weibull random variables respectively. Proposition 32 below can be viewed as an extension of the Hanson–Wright inequality [54].

Proposition 31 ([53], Theorem 3.1). *Let $\alpha > 0$ and $n \in \mathbb{Z}^+$. Let $X_1, \ldots, X_n$ be independent mean zero sub-Weibull random variables of order α, with $\|X_i\|_{\psi_\alpha} \leq K$ for all $i \in \mathbb{Z}^+$ and for some $K > 0$. Then, there exists a constant $C > 0$, depending only on α and K, such that for any vector $u = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$ and $x \geq 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n u_i X_i\Big| \geq x\Big) \leq \exp\Big\{1 - \min\Big\{\Big(\frac{x}{C\|u\|_2}\Big)^2, \Big(\frac{x}{C\|u\|_{\beta(\alpha)}}\Big)^\alpha\Big\}\Big\},$$

where $\beta(\alpha) = \infty$ when $\alpha \leq 1$ and $\beta(\alpha) = \alpha/(\alpha - 1)$ when $\alpha > 1$.

Proposition 32 ([55], Proposition 1.5). *Let $\alpha \in (0, 1] \cup \{2\}$, $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $X_1, \ldots, X_n$ be independent mean zero sub-Weibull random variables of order α, with $\mathbb{E}X_i^2 = \sigma_i^2$ and $\|X_i\|_{\psi_\alpha} \leq K$ for all $i \in \mathbb{Z}^+$ and for some $K > 0$. Then, there exists a constant $C > 0$, depending only on α and K, such that for any $x \geq 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{1 \leq i,j \leq n} a_{ij} X_i X_j - \sum_{i=1}^n a_{ii}\sigma_i^2\Big| \geq x\Big) \leq \exp\big(1 - \eta_\alpha(x/C; A)\big),$$

where

$$\eta_\alpha(x; A) := \min\Big\{\Big(\frac{x}{\|A\|_{\mathrm{F}}}\Big)^2, \frac{x}{\|A\|_2}, \Big(\frac{x}{\|A\|_{2\to\infty}}\Big)^{\frac{2\alpha}{2+\alpha}}, \Big(\frac{x}{\|A\|_{\max}}\Big)^{\frac{\alpha}{2}}\Big\}.$$

The following proposition presents a concentration inequality for sums of independent random variables with only finite certain number of moments. We use the form of the Fuk–Nagaev type inequalities appeared in [56].

Proposition 33 ([57], [58]). *Let $X_1, \ldots, X_n$ be independent random variables, each having mean 0 and variance σ^2. Assume further that for some $q \geq 2$ and $C_q > 0$, we have for all $i \in [n]$*

$$\mathbb{E}[\{\max(X_i, 0)\}^q] \leq C_q.$$

Then for any $x > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \geq x\Big) \leq \Big(\frac{(q+2)(nC_q)^{1/q}}{qx}\Big)^q + \exp\Big\{-\frac{2x^2}{n(q+2)^2 e^q \sigma^2}\Big\}.$$

Proposition 34. *Let $X_1, \ldots, X_p$ be independent random variables, each with mean zero and unit variance. Let $a \geq 0$ and $Z := \sum_{i=1}^p (X_i^2 - 1)\mathbb{1}_{\{|X_i| \geq a\}}$.*
(a) Let $\alpha > 0$, $K > 0$, $0 < \varepsilon \leq 1$ and $1 \leq s \leq \sqrt{p}$. Assume that $X_1, \ldots, X_p$ are independent sub-Weibull random variables of order α, with $\|X_i\|_{\psi_\alpha} \leq K$ for all $i \in [p]$. By setting

$$a \geq K \log^{1/\alpha}\Big(\frac{4ep}{\varepsilon s}\Big) \qquad and \qquad r = 2^{2/\alpha} K^2 s \log^{2/\alpha}\Big(\frac{4p}{\sqrt{\varepsilon}s}\Big),$$

we have $\mathbb{P}(Z > r) \leq \varepsilon$.
(b) Let $\alpha \geq 2$, $K > 0$, $0 < \varepsilon \leq 1$ and $1 \leq s \leq p$. Assume that $\mathbb{E}|X_i/K|^\alpha \leq 1$ for all $i \in [p]$. By setting

$$a \geq K\Big(\frac{2ep}{\varepsilon s}\Big)^{1/\alpha} \qquad and \qquad r = \frac{\alpha - 2}{\alpha} K^2 s \Big(\frac{2ep}{\varepsilon s}\Big)^{2/\alpha},$$

we have $\mathbb{P}(Z > r) \leq \varepsilon$.
(c) Assume the same conditions as in (b). Write $Z_s := \sum_{i=1}^p (X_i^2 - 1)\mathbb{1}_{\{|X_i| \geq a_s\}}$ to make the dependence on s explicit. By choosing the same a_s and r_s as in (b), we have

$$\mathbb{P}\big(\max_{s \in [p]} Z_s/r_s > 1\big) \leq 2\varepsilon.$$

Proof. We denote the order statistics of $|X_1|, \ldots, |X_p|$ as $|X|_{(1)} \leq \ldots \leq |X|_{(p)}$. For $x \geq 0$, we write $q_x := \min_{i \in [p]} \mathbb{P}(|X_i| \geq x)$ and $\mathcal{J}_x := \{i \in [p] : |X_i| \geq x\}$.
(a) Note that $q_x \leq 2\exp\{-(x/K)^\alpha\}$ by Proposition 29(a). Since $s \leq \sqrt{p}$, we observe that

$$\sum_{j=1}^s K^2 \log^{2/\alpha}\Big(\frac{4ep}{\varepsilon j}\Big) \leq \sum_{j=1}^s K^2 \log^{2/\alpha}\Big(\frac{4ep^2}{\varepsilon s^2}\Big) \leq 2^{2/\alpha} K^2 s \log^{2/\alpha}\Big(\frac{4p}{\sqrt{\varepsilon}s}\Big) = r.$$

Then, by a union bound and a binomial tail bound, we have

$$\mathbb{P}(Z > r) \leq \mathbb{P}(|\mathcal{J}_a| > s) + \mathbb{P}\Big(\sum_{j=1}^s \big(|X|_{(p-j+1)}^2 - 1\big) > r\Big)$$

$$\leq \mathbb{P}(|\mathcal{J}_a| > s) + \sum_{j=1}^{s} \mathbb{P}\left(|X|_{(p-j+1)} > K \log^{1/\alpha}\left(\frac{4ep}{\varepsilon j}\right)\right)$$

$$\leq \left(\frac{ep}{s}\right)^s \left\{2\exp\left\{-\left(\frac{a}{K}\right)^\alpha\right\}\right\}^s + \sum_{j=1}^{s} \left(\frac{ep}{j}\right)^j \left\{2\exp\left\{-\left(\frac{K\log^{1/\alpha}\left(\frac{4ep}{\varepsilon j}\right)}{K}\right)^\alpha\right\}\right\}^j$$

$$\leq (\varepsilon/2)^s + \sum_{j=1}^{s} (\varepsilon/2)^j \leq \varepsilon.$$

(b) Note that $q_x \leq (K/a)^\alpha$ by Chebyshev's inequality. We now observe that

$$\sum_{j=1}^{s} K^2 \left(\frac{2ep}{\varepsilon j}\right)^{2/\alpha} \leq K^2 \left(\frac{2ep}{\varepsilon}\right)^{2/\alpha} \left\{1 + \int_1^s x^{-2/\alpha}\, dx\right\} \leq \frac{\alpha-2}{\alpha} K^2 s \left(\frac{2ep}{\varepsilon s}\right)^{2/\alpha} = r$$

The rest then follows from the proof for part (a).

(c) By a union bound and the proof for the previous parts, we have

$$\mathbb{P}(\max_{s\in[p]} Z_s/r_s > 1) \leq \left(\sum_{s=1}^{p} \mathbb{P}(|\mathcal{J}_{a(s)}| > s)\right) + \mathbb{P}\left(\max_{s\in[p]} \frac{\sum_{j=1}^{s}\left(|X|_{(p-j+1)}^2 - 1\right)}{r_s} > 1\right)$$

$$\leq \sum_{s=1}^{p} (\varepsilon/2)^s + \sum_{j=1}^{p} \mathbb{P}\left(|X|_{(p-j+1)} > K\left(\frac{2ep}{\varepsilon j}\right)^{2/\alpha}\right)$$

$$\leq \sum_{s=1}^{p} (\varepsilon/2)^s + \sum_{j=1}^{p} (\varepsilon/2)^j \leq 2\varepsilon.$$

$\square$

Lemma 35. *Let* $\gamma > 0$. *Then, for all* $x \geq (2^\gamma - 1)^{-1/\gamma}$ *we have*

$$\sum_{i=0}^{\infty} \exp\left\{-(x2^i)^\gamma\right\} \leq 2\exp(-x^\gamma).$$

Proof. By the convexity of $y \mapsto 2^{\gamma y}$, we have that $2^{(i+1)\gamma} - 2^{i\gamma} \geq 2^{i\gamma} - 2^{(i-1)\gamma}$ and thus

$$2^{i\gamma} = 1 + \sum_{j=1}^{i} (2^{j\gamma} - 2^{(j-1)\gamma}) \geq 1 + i(2^\gamma - 1).$$

for all $i \in \mathbb{Z}^+$. Denote $\tilde{x} := \exp(x^\gamma)$. We hence deduce that when $\tilde{x} > 2^{\frac{1}{2^\gamma - 1}}$,

$$\sum_{i=0}^{\infty} \exp\left\{-(x2^i)^\gamma\right\} = \sum_{i=0}^{\infty} \tilde{x}^{-2^{i\gamma}} \leq \sum_{i=0}^{\infty} \tilde{x}^{-1-i(2^\gamma-1)} = \frac{1}{\tilde{x}\left(1 - \tilde{x}^{-(2^\gamma-1)}\right)} \leq 2\tilde{x}^{-1}.$$

$\square$

Lemma 36. *Let* $Z_1, \ldots, Z_n$ *be independent mean zero random variables.*
(a) Assume that there exists $C > 0$ *such that* $\mathbb{E}Z_i^4 \leq C$ *for all* $i \in [n]$. *Then for any* $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, *we have*

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} v_i Z_i\right)^4\right] \leq 3C\|v\|_2^4.$$

(b) Assume that there exists $C > 0$ *such that* $\mathbb{E}\left(|Z_i|^{2k}\right) \leq C$ *for some* $k \geq 1$. *Then, there exists a constant* $C_k > 0$, *depending only on* k *and* C *such that*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} Z_i/n\right|^{2k}\right] \leq C_k n^{-k}.$$

Proof. (a) Since $\mathbb{E}Z_i^0 = 1$ and $\mathbb{E}Z_i^1 = 0$ for all $i \in [n]$, we have

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} v_i Z_i\right)^4\right] = \sum_{i=1}^{n} v_i^4 \mathbb{E}Z_i^4 + \sum_{1\leq i<j\leq n} 6v_i^2 v_j^2 \mathbb{E}(Z_i^2 Z_j^2) \leq C\left(\sum_{i=1}^{n} v_i^4 + \sum_{1\leq i<j\leq n} 6v_i^2 v_j^2\right)$$

$$\leq 3C\left(\sum_{i=1}^{n} v_i^2\right)^2,$$

where the first inequality follows from Jensen's inequality.

(b) Note that $S_j := \sum_{i=1}^{j} Z_i$ is a martingale (adapted to the natural filtration) and $[S]_j := \sum_{i=1}^{j} Z_i^2$ can be viewed as the quadratic variation of this martingale. By Burkholder–Davis–Gundy inequality, e.g. Theorem 1.1 of [59], we have for any $k \geq 1$,

$$\mathbb{E}\big[|S_n|^{2k}\big] \leq \mathbb{E}\bigg[\Big(\max_{j \leq n} |S_j|\Big)^{2k}\bigg] \leq C_{k,1}\mathbb{E}\big[\big([S]_n\big)^k\big],$$

for some constant $C_{k,1} > 0$, depending only on k. Thus, we have

$$\mathbb{E}\bigg[\Big|\sum_{i=1}^{n} Z_i/n\Big|^{2k}\bigg] \leq \frac{C_{k,1}}{n^{2k}}\mathbb{E}\bigg[\Big(\sum_{i=1}^{n} Z_i^2\Big)^k\bigg] \leq \frac{C_{k,1}}{n^k}\mathbb{E}\bigg[\frac{\sum_{i=1}^{n} |Z_i|^{2k}}{n}\bigg] \leq \frac{CC_{k,1}}{n^k},$$

where we have used Jensen's inequality in the second inequality. $\qquad\square$

Lemma 37. *Let $k \geq 1$ and $V_1, \ldots, V_L$ be independent random vectors in $\mathbb{R}^p$, each having zero mean and independent coordinates. Assume that there exists $C > 0$ such that $\mathbb{E}\big[|V_i(j)|^{2k}\big] \leq C$ for all $i \in [L]$ and $j \in [p]$. Denote $\overline{V} := \sum_{i=1}^{L} V_i/L$. Then for any $\delta \in (0,1)$, we have*

$$\mathbb{P}\bigg(\Big|\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big)\Big| > C_k \frac{p^{\frac{1}{2} \vee \frac{1}{k}}}{\delta^{1/k}}\bigg) \leq \delta$$

for some constant $C_k > 0$, depending only on C and k.

Proof. We first prove the result for $1 \leq k \leq 2$. Note that for any $\eta > 0$

$$\mathbb{P}\bigg(\Big|\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big)\Big| > \eta\bigg) \leq p\mathbb{P}\Big(\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\Big)$$

$$+ \mathbb{P}\bigg(\Big|\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big)\mathbb{1}\big\{\big|L\big(\overline{V}^2(j) - 1/L\big)\big| \leq \eta\big\}\Big| > \eta\bigg). \quad (119)$$

We control the two terms separately. For the first term, we have

$$\mathbb{P}\Big(\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\Big) \leq \frac{\mathbb{E}\big[\big|L\overline{V}^2(1) - 1\big|^k\big]}{\eta^k} \leq \frac{2^{k-1}\big(\mathbb{E}\big|L\overline{V}^2(1)\big|^k + 1\big)}{\eta^k} \leq \frac{2^{k-1}(C_{0,k} + 1)}{\eta^k}, \quad (120)$$

where the three inequalities follow, respectively, from Markov's inequality, Jensen's inequality and Lemma 36(b), with $C_{0,k}$ being the constant that depends only on C and k in that lemma. For convenience, we denote $C_{1,k} := 2^{k-1}(C_{0,k} + 1)$ hereafter. For the second term in (119), we have

$$\mathbb{P}\bigg(\Big|\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big)\mathbb{1}\big\{\big|L\big(\overline{V}^2(j) - 1/L\big)\big| \leq \eta\big\}\Big| > \eta\bigg)$$

$$\leq \frac{1}{\eta^2}\bigg\{p\mathbb{E}\Big[\big(L\overline{V}^2(1) - 1\big)^2\mathbb{1}\big\{\big|L\big(\overline{V}^2(1) - 1/L\big)\big| \leq \eta\big\}\Big]$$

$$+ p^2\Big(\mathbb{E}\Big[L\big(\overline{V}^2(1) - 1/L\big)\mathbb{1}\big\{\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\big\}\Big]\Big)^2\bigg\}$$

$$\leq \frac{p}{\eta^2}\mathbb{E}\Big[\big|L\overline{V}^2(1) - 1\big|^k \eta^{2-k}\Big] + \frac{p^2}{\eta^2}\Big\{\mathbb{E}\Big[\big|L\overline{V}^2(1) - 1\big|^k\Big]\Big\}^{2/k}\Big\{\mathbb{P}\Big(\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\Big)\Big\}^{2(k-1)/k}$$

$$\leq \frac{C_{1,k}p}{\eta^k} + \frac{C_{1,k}^{2/k}p^2}{\eta^2}\Big(\frac{C_{1,k}}{\eta^k}\Big)^{2(k-1)/k}, \quad (121)$$

where we have used Markov's inequality for the first inequality, Hölder's inequality for the second one and (120) for the last one. Combining (119), (120) and (121), we have

$$\mathbb{P}\bigg(\Big|\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big)\Big| > \eta\bigg) \leq \frac{2C_{1,k}p}{\eta^k} + \frac{C_{1,k}^2 p^2}{\eta^{2k}}.$$

Note that if $C_{1,k}p/\eta^k > 1$, the bound above holds trivially. Therefore we obtain

$$\mathbb{P}\bigg(\sum_{j=1}^{p} L\big(\overline{V}^2(j) - 1/L\big) > \eta\bigg) \leq \frac{3C_{1,k}p}{\eta^k},$$

for any $\eta > 0$, which is equivalent to the claimed bound.

For $k > 2$, by Markov's inequality, (120) and Lemma 36(b), there exists a constant $C_{2,k} > 0$, depending only on k and C such that

$$\mathbb{P}\Big(\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\Big) \leq \frac{\mathbb{E}\left[\Big|\sum_{j=1}^p L\big(\overline{V}^2(j) - 1/L\big)\Big|^k\right]}{\eta^k} \leq \frac{C_{2,k} p^{k/2}}{\eta^k},$$

which proves the desired result.

$\square$

Lemma 38. *Let $n \in \mathbb{Z}^+$ and $c \in \mathbb{R}$. Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be $2n$ real numbers. Suppose that $a_i - b_i \leq c$ for all $i \in [n]$. Then*

$$\mathrm{median}(a_1, \ldots, a_n) - \mathrm{median}(b_1, \ldots, b_n) \leq c.$$

Proof. We sort the two arrays respectively and obtain $a_{(1)} \leq \ldots \leq a_{(n)}$ and $b_{(1)} \leq \ldots \leq b_{(n)}$. We show that $a_{(i)} - b_{(i)} \leq c$ for all $i \in [n]$. Indeed, there exists a set $\mathcal{I}_i \subseteq [n]$ with $|\mathcal{I}_i| \geq i$ such that

$$b_{(i)} = \max\{b_j : j \in \mathcal{I}_i\} \geq \max\{a_j - c : j \in \mathcal{I}_i\} \geq a_{(i)} - c.$$

The desired results follows by observing that the median is a convex combination of the order statistics. $\square$

Lemma 39. *Let $1 \leq \alpha \leq 2$ and $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^p$ with mean μ. Assume that the distribution of $X_i - \mu$ belongs to $\mathcal{W}_\alpha$ for each $i \in [n]$. Then*

$$\mathbb{P}\left(\big|\|\overline{X}\|_2 - \|\mu\|_2\big| \geq u\right) \leq \frac{\pi p^{\frac{\alpha}{2}}}{n^{\alpha-1} u^\alpha},$$

for any $u > 0$.

Proof of Lemma 39. By Lemma 4.2 in [42], we have

$$\mathbb{E}|\langle \overline{X} - \mu, v\rangle|^\alpha \leq \frac{2}{n^{\alpha-1}},$$

for any unit vector $v \in \mathbb{R}^p$. This means that $\overline{X} - \mu$ also satisfy (36) but with a different constant instead of 1. Then, we use Lemma 4.1 in [42] to deduce

$$\mathbb{E}\|\overline{X} - \mu\|_2^\alpha \leq \frac{\pi}{n^{\alpha-1}} p^{\frac{\alpha}{2}}.$$

By triangle inequality and Markov's inequality, we have

$$\mathbb{P}\left(\big|\|\overline{X}\|_2 - \|\mu\|_2\big| \geq u\right) \leq \mathbb{P}(\|\overline{X} - \mu\|_2 \geq u) \leq \frac{\pi p^{\frac{\alpha}{2}}}{n^{\alpha-1} u^\alpha},$$

as desired. $\square$

REFERENCES

[1] T. Wang and R. J. Samworth, "High dimensional change point estimation via sparse projection," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 80, pp. 57–83, 2018.

[2] H. Liu, C. Gao, and R. J. Samworth, "Minimax rates in sparse, high-dimensional change point detection," *Annals of Statistics*, vol. 49, no. 2, pp. 1081–1112, 2021.

[3] D. Wang, Y. Yu, and A. Rinaldo, "Optimal change point detection and localization in sparse dynamic networks," *Annals of Statistics*, vol. 49, no. 1, pp. 203–232, 2021.

[4] D. Wang and Z. Zhao, "Optimal change-point testing for high-dimensional linear models with temporal dependence," *arXiv preprint*, p. arXiv:2205.03880, 2022.

[5] H. Xu, D. Wang, Z. Zhao, and Y. Yu, "Change-point inference in high-dimensional regression models under temporal dependence," *Annals of Statistics*, vol. 52, no. 3, pp. 999–1026, 2024.

[6] N. Verzelen, M. Fromont, M. Lerasle, and P. Reynaud-Bouret, "Optimal change-point detection and localization," *Annals of Statistics*, vol. 51, pp. 1586–1610, 2023.

[7] M. Yu and X. Chen, "A robust bootstrap change point test for high-dimensional location parameter," *Electronic Journal of Statistics*, vol. 16, no. 1, pp. 1096–1152, 2022.

[8] F. Jiang, R. Wang, and X. Shao, "Robust inference for change points in high dimension," *Journal of Multivariate Analysis*, vol. 193, pp. 105–114, 2023.

[9] L. Comminges, O. Collier, M. Ndaoud, and A. B. Tsybakov, "Adaptive robust estimation in sparse vector model," *Annals of Statistics*, vol. 49, no. 3, pp. 1347–1377, 2021.

[10] S. I. Resnick, *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media, 2007.

[11] H. Cho and D. Owens, "High-dimensional data segmentation in regression settings permitting temporal dependence and non-gaussianity," *Electronic Journal of Statistics*, vol. 18, no. 1, pp. 2620–2664, 2024.

[12] P. Fearnhead and G. Rigaill, "Changepoint detection in the presence of outliers," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 169–183, 2019.

[13] M. Li and Y. Yu, "Adversarially robust change point detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22955–22967, 2021.

[14] D. Mukherjee, M. Banerjee, and Y. Ritov, "On robust learning in the canonical change point problem under heavy tailed errors in finite and growing dimensions," *Electronic Journal of Statistics*, vol. 16, no. 1, pp. 1153–1252, 2022.

[15] B. Liu, Z. Qi, X. Zhang, and Y. Liu, "Change point detection for high-dimensional linear models: A general tail-adaptive approach," *arXiv preprint*, p. arXiv:2207.11532, 2022.

[16] L. Wegner and M. Wendler, "Robust change-point detection for functional time series based on u-statistics and dependent wild bootstrap," *arXiv preprint*, p. arXiv:2206.01458, 2022.

[17] K. Ramsay and S. Chenouri, "Robust multiple change-point detection for multivariate variability using data depth," *arXiv preprint*, p. arXiv:2011.09558, 2020.

[18] S. Chenouri, A. Mozaffari, and G. Rice, "Robust multivariate change point analysis based on data depth," *Canadian Journal of Statistics*, vol. 48, no. 3, pp. 417–446, 2020.

[19] E. Gombay and M. Hušková, "Rank based estimators of the change-point," *Journal of Statistical Planning and Inference*, vol. 67, no. 1, pp. 137–154, 1998.

[20] H. Dehling, A. Rooch, and M. S. Taqqu, "Non-parametric change-point tests for long-range dependent data," *Scandinavian Journal of Statistics*, vol. 40, no. 1, pp. 153–173, 2013.

[21] A. Betken, "Testing for change-points in long-range dependent time series by means of a self-normalized wilcoxon test," *Journal of Time Series Analysis*, vol. 37, no. 6, pp. 785–809, 2016.

[22] C. Gerstenberger, "Robust wilcoxon-type estimation of change-point location under short-range dependence," *Journal of Time Series Analysis*, vol. 39, no. 1, pp. 90–104, 2018.

[23] A. Betken and M. Wendler, "Rank-based change-point analysis for long-range dependent time series," *Bernoulli*, vol. 28, no. 4, pp. 2209–2233, 2022.

[24] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1604–1614, 2011.

[25] Y. Cao and Y. Xie, "Robust sequential change-point detection by convex optimization," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1287–1291.

[26] T. L. Molloy and J. J. Ford, "Misspecified and asymptotically minimax robust quickest change detection," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5730–5742, 2017.

[27] E. S. Page, "A test for a change in a parameter occurring at an unknown point," *Biometrika*, vol. 42, pp. 523–527, 1955.

[28] S. Kovács, H. Li, L. Haubner, and P. Munk, Axel andBühlmann, "Optimistic search: change point estimation for large-scale data via adaptive logarithmic queries," *Journal of Machine Learning Research*, vol. 25, pp. 1–64, 2024.

[29] E. Pilliat, A. Carpentier, and N. Verzelen, "Optimal multiple change-point detection for high-dimensional data," *Electronic Journal of Statistics*, vol. 17, pp. 1240–1315, 2023.

[30] G. Lugosi and S. Mendelson, "Sub-gaussian estimators of the mean of a random vector," *Annals of Statistics*, vol. 47, no. 2, pp. 783–794, 2019.

[31] M. Lerasle and R. I. Oliveira, "Robust empirical mean estimators," *arXiv preprint*, p. arXiv:1112.3914, 2011.

[32] G. Lecué and M. Lerasle, "Robust machine learning by median-of-means: theory and practice," *Annals of Statistics*, vol. 48, no. 2, pp. 906–931, 2020.

[33] P. Humbert, B. Le Bars, and L. Minvielle, "Robust kernel density estimation with median-of-means principle," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9444–9465.

[34] J. Kwon, G. Lecué, and M. Lerasle, "A mom-based ensemble method for robustness, subsampling and hyperparameter tuning," *Electronic Journal of Statistics*, vol. 15, pp. 1202–1227, 2021.

[35] G. Lugosi and S. Mendelson, "Mean estimation and regression under heavy-tailed distributions: a survey," *Foundations of Computational Mathematics*, vol. 9, pp. 1145–1190, 2019.

[36] A. Prasad, S. Balakrishnan, and P. Ravikumar, "A unified approach to robust mean estimation," *arXiv preprint*, p. arXiv:1907.00927, 2019.

[37] R. Vershynin, "On the role of sparsity in compressed sensing and random matrix theory," in *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2009, pp. 189–192.

[38] G. Lugosi and S. Mendelson, "Robust multivariate mean estimation: the optimality of trimmed mean," *Annals of Statistics*, vol. 49, no. 1, pp. 393–410, 2021.

[39] K. Frick, A. Munk, and H. Sieling, "Multiscale change point inference," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 76, no. 3, pp. 495–580, 2014.

[40] R. C. Bradley, "Basic properties of strong mixing conditions. a survey and some open questions." *Probability Surveys*, vol. 2, pp. 107–144, 2005.

[41] ——, "Every "lower psi-mixing" markov chain is "interlaced rho-mixing"," *Stochastic Processes and Their Applications*, vol. 72, pp. 221–239, 1997.

[42] Y. Cherapanamjeri, N. Tripuraneni, P. Bartlett, and M. Jordan, "Optimal mean estimation without a variance," in *Conference on Learning Theory*. PMLR, 2022, pp. 356–357.

[43] J. D. Castillo, J. Daoudi, and R. Lockhart, "Methods to distinguish between polynomial and exponential tails," *Scandinavian Journal of Statistics*, vol. 41, no. 2, pp. 382–393, 2014.

[44] D. Bhati, "A test procedure for distinguishing logarithmically decaying tail from polynomially decaying tail," *Journal of the Korean Statistical Society*, vol. 49, pp. 841–862, 2020.

[45] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel, "Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions," *Stat*, vol. 9, no. 1, p. e318, 2020.

[46] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[47] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2017.

[48] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003, vol. 87.

[49] C. Gao, F. Han, and C.-H. Zhang, "On estimation of isotonic piecewise constant signals," *Annals of Statistics*, vol. 48, no. 2, pp. 629–654, 2020.

[50] I. A. Ibragimov and R. Z. Has' Minskii, *Statistical Estimation: Asymptotic Theory*. Springer Science & Business Media, 2013, vol. 16.

[51] Q.-M. Shao and H. Yu, "Weak convergence for weighted empirical processes of dependent sequences," *Annals of Probability*, vol. 24, pp. 2098–2127, 1996.

[52] F. Merlevède, M. Peligrad, and E. Rio, "Bernstein inequality and moderate deviations under strong mixing conditions," *Inst. Math. Stat. (IMS) Collect.*, vol. 5, pp. 273–292, 2009.

[53] A. K. Kuchibhotla and A. Chakrabortty, "Moving beyond sub-gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression," *Information and Inference: A Journal of the IMA*, 2022.

[54] D. L. Hanson and F. T. Wright, "A bound on tail probabilities for quadratic forms in independent random variables." *Annals of Mathematical Statistics*, vol. 42, pp. 1079–1083, 1971.

[55] F. Götze, H. Sambale, and A. Sinulis, "Concentration inequalities for polynomials in α-sub-exponential random variables," *Electronic Journal of Probability*, vol. 26, pp. 1–22, 2021.

[56] E. Rio, "About the constants in the fuk–nagaev inequalities," *Electronic Communications in Probability*, vol. 22, no. 28, pp. 1–12, 2017.

[57] D. K. Fuk, "Some probablistic inequalities for martingales," *Siberian Mathematical Journal*, vol. 14, pp. 131–137, 1973.

[58] S. V. Nagaev, "Large deviations of sums of independent random variables," *Annals of Probability*, vol. 7, pp. 745–789, 1979.

[59] M. Beiglböck and P. Siorpaes, "Pathwise versions of the burkholder–davis–gundy inequality," *Bernoulli*, vol. 21, no. 1, pp. 360–373, 2015.