Robust mean change point testing in high-dimensional data with heavy tails

Mengchu Li[*,1], Yudong Chen[*,2], Tengyao Wang[2], Yi Yu[1]
[1]Department of Statistics, University of Warwick
[2]Department of Statistics, London School of Economics and Political Science

(May 31, 2023)

Abstract

We study a mean change point testing problem 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. Our results quantify the costs of heavy-tailedness on the fundamental difficulty of change point testing problems for high-dimensional data by comparing to the previous results under Gaussian assumptions.

To be specific, when the error vectors follow sub-Weibull distributions, a CUSUM-type statistic is shown to achieve a minimax testing rate up to $\sqrt{\log\log(8n)}$. When the error distributions have 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. In particular, in the sparse regime, we further propose a computationally-efficient test to achieve the exact optimality. Surprisingly, 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. This phenomenon implies a phase transition of the minimax testing rates at $\alpha = 4$.

1 Introduction

In this paper, we study the single change point detection problem when the observations are corrupted by heavy-tailed errors. To be specific, consider the 'signal plus noise' model

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

where X, θ and E are all $p \times n$ matrices, and the entries of E are independent random variables with zero mean and unit variance. 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. By writing θ_t as the

*Equal contribution

t-th column of θ, our goal can be formalised as testing

$$H_0 : \theta \in \Theta_0(p, n) \quad \text{vs.} \quad H_1 : \theta \in \Theta(p, n, s, \rho) := \bigcup_{t_0=1}^{n-1} \Theta^{(t_0)}(p, n, s, \rho), \tag{2}$$

where

$$\Theta_0(p, n) := \{\theta : \theta_t = \mu \text{ for all } t = 1, 2, \ldots, n, \text{ for some } \mu \in \mathbb{R}^p\}$$

and

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

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) 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 (e.g. Wang and Samworth, 2018; Verzelen et al., 2020; Liu et al., 2021; Wang et al., 2021; Wang and Zhao, 2022; Xu et al., 2022) 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, all these problems 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 rate of testing, $v^*_{\mathcal{Q}}(p, n, s)$, defined below.

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*

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

We say that $v^*_{\mathcal{Q}}(p, n, s)$ is the minimax testing rate if for any $\varepsilon \in (0, 1)$, we have $\mathcal{R}_{\mathcal{Q}}(\rho) \le \epsilon$ when $\rho^2 \ge Cv^*_{\mathcal{Q}}(p, n, s)$, and $\mathcal{R}_{\mathcal{Q}}(\rho) \ge 1/2$ when $\rho^2 \le cv^*_{\mathcal{Q}}(p, n, s)$, where c, C are constants depending on ϵ and $\mathcal{Q}$.

We note that in Definition 1, ε is allowed to be varying with other model parameters and C depends on ε. In the rest of this paper, to highlight the roles and interplay of model parameters, we only consider the case when ε is an absolute constant.

A minimax rate of testing is previously studied in Liu et al. (2021), where the entries of noise matrix E are assumed to be independent standard normal random variables. It is shown that

$$v^*_{N^{\otimes}(0,1)}(p, n, s) = \Big\{ \sqrt{p \log\log(8n)} \wedge \big(s \log\big(ep \log\log(8n)s^{-2}\big)\big) \Big\} \vee \log\log(8n),$$

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 1.1, is to characterise the impact of heavy-tailed distributions on the minimax testing rate. More specifically, we consider two types of heavy-tailedness.

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*

$$\mathbb{E}W = 0, \quad \mathbb{E}W^2 = 1 \quad \textit{and} \quad \mathbb{E}\exp\{|W/K|^\alpha\} \leq 2. \tag{3}$$

The $\mathcal{G}_{\alpha,K}$ class of distributions are also known as sub-Weibull distributions of order α with ψ_α-norm bounded by K – see Definitions 4 and 5 – and with mean zero and variance one. By Proposition 12(a), they possess exponentially-decaying tails, as $\mathbb{P}(|W| \geq x) \leq 2e^{-(x/K)^\alpha}$.

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*

$$\mathbb{E}W = 0, \quad \mathbb{E}W^2 = 1 \quad \textit{and} \quad \mathbb{E}|W/K|^\alpha \leq 1. \tag{4}$$

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 $\mathcal{Q} = \mathcal{G}^{\otimes}_{\alpha,K}$ and $\mathcal{Q} = \mathcal{P}^{\otimes}_{\alpha,K}$, respectively, where $\mathcal{G}^{\otimes}_{\alpha,K}$ and $\mathcal{P}^{\otimes}_{\alpha,K}$ denote the class of joint distributions of all the entries in the error matrix $E \in \mathbb{R}^{p\times n}$ (or simply the class of distributions of E) 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.

1.1 Main results

Our main results are summarised in Figure 1 and Table 1. As shown in Figure 1, when $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$, the minimax testing rate transition boundary between dense and sparse regimes occurs at $s^* = p^{\frac{1}{2}-\frac{1}{\alpha-2}}$ when $\alpha \geq 4$. When $\alpha \in [2, 4)$, there is essentially no sparse regime, i.e. testing sparse change is information-theoretically as hard as testing dense changes. When $P_e \in \mathcal{G}^{\otimes}_{\alpha,K}$, the transition boundary takes a simpler form of $s^* = \sqrt{p}\log^{-2/\alpha}(ep)$ for $\alpha \in (0, 2]$. The corresponding upper bounds and lower bounds on the minimax testing rates $v^*_{\mathcal{P}^{\otimes}_{\alpha,K}}(p, n, s)$ and $v^*_{\mathcal{G}^{\otimes}_{\alpha,K}}(p, n, s)$ are detailed in Table 1. The correspondences between each term in the table and its associated result are detailed in Section 1.3. Note that under both classes of distributions, we achieve matching upper and lower bounds under the aforementioned sparse regimes. For the dense regimes, we characterise the minimax testing rates up to $\sqrt{\log\log(8n)}$ in the case of $\mathcal{G}^{\otimes}_{\alpha,K}$ and up to $\log\log(8n)$ in the case of $\mathcal{P}^{\otimes}_{\alpha,K}$. We provide thorough discussions on these gaps in Section 2.3 and Section 3.3.

Compared to previous works on robust mean change point testing problems (e.g. Yu and Chen, 2022; Jiang et al., 2023), where change point locations are required to be comparable to the length of time series in order to achieve near-optimal guarantees, we consider a more general parameter space, where the change point locations may be arbitrarily close the boundary. Compared to the recent work on optimal mean change point testing problems without robustness (e.g. Liu et al., 2021), our results allow for much more general classes of distributions and quantify the costs of heavy-tailedness. Finally, compared to relevant latest works on robustness (e.g. Comminges et al., 2021), we investigate the more challenging case when $\alpha \in [2, 4]$ under the finite moment noise assumption, and unveil a new phase transition phenomenon that is previously unknown even in sequence models. More in-depth discussions on these works can be found in Section 1.2.

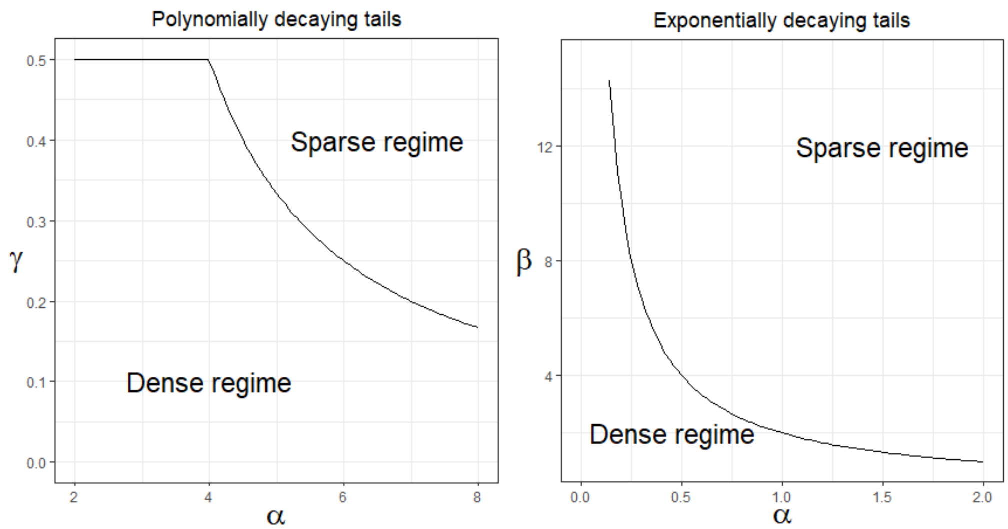

Figure 1: minimax testing rate transition boundaries between dense and sparse regimes when the distribution of the error matrix belongs to $\mathcal{P}^{\otimes}_{\alpha,K}$ (left panel) and $\mathcal{G}^{\otimes}_{\alpha,K}$ (right panel). The testing rates within each regime are detailed in Table 1.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

		Upper bound	Lower bound
$\mathcal{G}^{\otimes}_{\alpha,K}$	Dense	(i) $\sqrt{p \log\log(8n)} + \log\log(8n)$	(ii) $\sqrt{p(\log\log(8n))^{\omega_1}} + \log\log(8n)$
	Sparse	(iii) $s \log^{2/\alpha}(ep/s) + \log\log(8n)$	(iv) $s \log^{2/\alpha}(ep/s) + \log\log(8n)$
$\mathcal{P}^{\otimes}_{\alpha,K}$	Dense	(v) $p^{(2/\alpha)\vee(1/2)} \log\log(n)$	(vi) $p^{(2/\alpha)\vee(1/2)} \sqrt{(\log\log(8n))^{\omega_2}} + \log\log(8n)$
	Sparse	(vii) $s(p/s)^{2/\alpha} + \log\log(8n)$	(viii) $s(p/s)^{2/\alpha} + \log\log(8n)$

Table 1: Minimax testing rates up to $\log\log(8n)$, where $\omega_1 = \mathbb{1}\{s > \sqrt{p\log\log(8n)}\}$ and $\omega_2 = \mathbb{1}\{s > \sqrt{p\log\log(8n)}$ and $\alpha \geq 4\}$. The rows of $\mathcal{G}^{\otimes}_{\alpha,K}$ and $\mathcal{P}^{\otimes}_{\alpha,K}$ correspond to the right and left panel of Figure 1, respectively.

1.2 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 Resnick (2007) for more in-depth discussions. Recent years have witnessed a growing interest among statisticians in quantifying the cost, if there is, of heavy-tailedness on various statistical tasks. Notably, for the seemingly simple task of mean estimation, a range of innovative and sophisticated robust estimators (e.g. Catoni, 2012; Devroye et al., 2016; Lugosi and Mendelson, 2019a,b; Prasad et al., 2019; Depersin, 2020; Depersin and Lecué,

2022) are developed to achieve the same high probability type bounds as in the Gaussian case, even if the data distribution is only assumed to have finite second moment. These results show that, in terms of convergence rates, there is no fundamental cost of allowing for heavy-tailed distributions for the task of mean estimation. With the success of mean estimation, a variety of statistical tasks are considered in the literature with the goal of developing estimators with Gaussian-like performances for heavy-tailed data, including regression (e.g. Fan et al., 2014; Lugosi and Mendelson, 2019a; Sun et al., 2020), empirical risk minimisation (e.g. Lecué and Lerasle, 2020; Prasad et al., 2020), matrix estimation (e.g. Minsker, 2018; Mendelson and Zhivotovskiy, 2020), among others. However, literature regarding change point analysis under heavy-tailed errors has been scarce in general.

One line of recent works (e.g. Cho and Owens, 2022; Wang and Zhao, 2022; Xu et al., 2022) 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 attack develops algorithms with robust components for change point analysis. In particular, in the univariate mean change setting, Fearnhead and Rigaill (2019) propose to modify the commonly used ℓ_2-loss to other loss functions, including the biweight and Huber loss functions to enhance robustness against heavy-tailed errors in localising change points. Li and Yu (2021) 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 showed 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 points inference in stump models (Mukherjee et al., 2022), detecting distributional changes (Chenouri et al., 2020) and covariance changes (Ramsay and Chenouri, 2020), along with a series of work on robust online change point detection (Unnikrishnan et al., 2011; Cao and Xie, 2017; Molloy and Ford, 2017), which is different from the offline version that we study here[1].

Closer to our high-dimensional mean change point setting, Yu and Chen (2022) and Jiang et al. (2023) both consider the testing problem (2) and propose robust methodology targeting sparse change and dense change, respectively. Yu and Chen (2022) 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 (c.f. Theorem 3.3) 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 point. 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 when $t_0 = cn$ for some fixed constant $c \in (0, 1)$. Jiang et al. (2023) consider the same mean change

[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.

point testing problem as ours but without sparsity constraints. They allowed a form of weak spatial dependence across coordinates and we discuss this aspect in Section 5. 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 change point location satisfies $t_0 = cn$ for some fixed constant $c \in (0, 1)$.

In comparison to the results in Yu and Chen (2022) and Jiang et al. (2023), our results as summarised in Section 1.1, 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 the boundary, the fundamental difficulty of the testing problem change dramatically. 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 tighter in the sense that they are optimal up to at most $\log\log(8n)$.

Lastly, we mention two recent works that are technically related to ours. Comminges et al. (2021) study the sparse sequence models where

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

The noise random variables ξ_i 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 s-sparse. They provide minimax rates for estimating $\|\theta\|_2$ among other results (c.f. 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, Comminges et al. (2021) first estimate θ via a penalised least square 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 change. 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 $\alpha \in [2, 4)$ and unveil a previously unknown phase transition behaviour even when n is of constant order.

As mentioned before, Liu et al. (2021) 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 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 1 with the rate $v^*_{N^{\otimes}(0,1)}(p, n, s)$ derived by Liu et al. (2021) under the Gaussianity, we show that the heavy-tailed errors mainly affect the difficulty of testing sparse changes and in the $\mathcal{P}^{\otimes}_{\alpha,K}$ case with $\alpha \in [2, 4)$. 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 heavy-tailed errors in testing change point compared to Gaussian errors.

1.3 Outline

The rest of paper is organised as follows. In Section 2, we study the testing problem (2) under sub-Weibull noise distribution, i.e. when $P_e \in \mathcal{G}^{\otimes}_{\alpha,K}$. We again consider separately the dense regime in Section 2.1 and the sparse regime in Section 2.2. In particular, the rates (i) - (iv) in Table 1 are established in Theorem 1, Proposition 2, Theorem 3 and Proposition 4, respectively. Next, in Section 3, we consider the testing problem under finite moment assumption on the noise distribution, i.e. when $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$. We again consider separately the dense regime in Section 3.1 and the sparse regime in Section 3.2. We present the rates (v) - (viii) in Table 1 in Theorem 5, Proposition 6, Theorem 9 and Proposition 10, respectively. The upper bound rates in Tables 1, particularly those in the sparse regimes, are achieved by procedures that require the knowledge of the sparsity parameter. We propose and study procedures that are adaptive to the unknown sparsity level in Section 4. We end the main part of the paper with some discussions on potential future directions in Section 5. Proofs of our main results are given in Section A of the Appendices, with auxiliary results deferred to Section B.

1.4 Notation

We end this section by introducing some notation used throughout the paper. For $d \in \mathbb{N}$, we write $[d] := \{1, \ldots, d\}$. Given $a, b \in \mathbb{R}$, we denote $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. For a set S, we use $\mathbb{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a vector $v = \big(v(1), \ldots, v(d)\big)^\top \in \mathbb{R}^d$, we define $\|v\|_0 := \sum_{i=1}^d \mathbb{1}_{\{v(i)\neq 0\}}$, $\|v\|_2 := \big\{\sum_{i=1}^d v(i)^2\big\}^{1/2}$ and $\|v\|_\infty := \max_{i\in[d]} |v(i)|$. For a matrix $A = (A_{ij})_{i\in[d_1],j\in[d_2]} = \big(A_j(i)\big)_{i\in[d_1],j\in[d_2]} \in \mathbb{R}^{d_1\times d_2}$, we denote the Frobenius norm $\|A\|_{\mathrm{F}} := \big\{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} A_{ij}^2\big\}^{1/2}$, the operator norm $\|A\|_2 := \max_{v\in\mathbb{R}^{d_2}, v\neq 0} \frac{\|Av\|_2}{\|v\|_2}$, the two-to-infinity norm $\|A\|_{2\to\infty} := \max_{v\in\mathbb{R}^{d_2}, v\neq 0} \frac{\|Av\|_\infty}{\|v\|_2}$ and the max norm $\|A\|_{\max} := \max_{i\in[d_1],j\in[d_2]} |A_{ij}|$. We use $\Gamma(\cdot)$ to denote the gamma function. For two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$, we 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 we define the square of the Hellinger distance between them as $H^2(P, Q) := \int_{\mathcal{X}} \big(\sqrt{p(x)} - \sqrt{q(x)}\big)^2 \lambda(dx)$, where p and q are the Radon–Nikodym derivatives of P and Q with respect to λ respectively. Finally, when the distribution is clear from the context, we use $\mathbb{P}$, $\mathbb{E}$ and Var to denote probability, expectation and variance operators respectively.

2 Testing under sub-Weibull noise distribution

In this section, we consider the entries of the noise matrix E to be independent random variables and each with distribution belonging to the class $\mathcal{G}_{\alpha,K}$, defined in (3). For any measurable test function $\phi : \mathbb{R}^{p\times n} \to \{0, 1\}$ and $\rho > 0$, we write the worst case testing error when $P_e \in \mathcal{G}_{\alpha,K}^{\otimes}$ as

$$\mathcal{R}_{\mathcal{G}}(\rho, \phi) = \sup_{P_e\in\mathcal{G}_{\alpha,K}^{\otimes}} \sup_{\theta\in\Theta_0(p,n)} \mathbb{E}_{\theta,P_e}\phi + \sup_{P_e\in\mathcal{G}_{\alpha,K}^{\otimes}} \sup_{\theta\in\Theta(p,n,s,\rho)} \mathbb{E}_{\theta,P_e}(1-\phi),$$

and recall the minimax testing error as

$$\mathcal{R}_{\mathcal{G}}(\rho) = \inf_{\phi\in\Phi} \mathcal{R}_{\mathcal{G}}(\rho, \phi).$$

In a slight abuse of notation, we use $\mathcal{R}_{\mathcal{G}}$ (resp. $\mathcal{R}_{\mathcal{P}}$) in place of $\mathcal{R}_{\mathcal{G}_{\alpha,K}^{\otimes}}$ (resp. $\mathcal{R}_{\mathcal{G}_{\alpha,K}^{\otimes}}$) in the rest of this paper. We assume the sparsity level s to be known for now and defer the discussion of adaptive procedures to Section 4. Given s, we separately consider the dense and sparse regimes as mentioned in Section 1.1. Recall that the boundary between these two regimes is

$$s^* = \frac{\sqrt{p}}{\log^{2/\alpha}(ep)}. \tag{5}$$

The dense regime refers to the case $s \geq s^*$, while the sparse regime refers to the case $s < s^*$.

2.1 Dense regime

In the dense regime, we consider the testing procedure that is used in Liu et al. (2021). Consider $\mathcal{T} = \{1, 2, 4, \ldots, 2^{\lfloor \log_2(n/2) \rfloor}\}$ and the 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

$$\phi_{\mathcal{G},\mathrm{dense}} := \mathbb{1}_{\{\max_{t\in\mathcal{T}} A_t > r\}}, \tag{6}$$

where

$$A_t := \sum_{j=1}^p \big(Y_t^2(j) - 1\big), \tag{7}$$

and $r > 0$ is the detection threshold specified in (8). The following theorem establishes the theoretical guarantee of the test $\phi_{\mathcal{G},\mathrm{dense}}$.

Theorem 1. *Let $0 < \alpha \le 2$ 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 (6) with*

$$r = C_1\Big(\sqrt{p \log\log(8n)} + \log\log(8n)\Big) \tag{8}$$

satisfies

$$\mathcal{R}_{\mathcal{G}}(\rho, \phi_{\mathcal{G},\mathrm{dense}}) \le \varepsilon,$$

as long as $\rho^2 \ge C_2 v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}$, where

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

Note that this simple test actually achieves the same rate in the dense regime as in $v^*_{N^{\otimes}(0,1)}$, saving for a different separation boundary, even though the noise distributions possess heavier tails than Gaussian distribution. The key argument in showing that there is no cost of allowing sub-Weibull distributions in the dense regime is by a careful analysis of the type-I error instead of using a crude union bound. The following proposition shows that the test $\phi_{\mathcal{G},\mathrm{dense}}$ in (6) is minimax optimal in the dense regime when $s \ge \sqrt{p\log\log(8n)}$ and up to $\sqrt{\log\log(8n)}$ otherwise. We carefully discuss this gap in Section 2.3.

Proposition 2. *Let $0 < \alpha \le 2$, $K \ge K_\alpha$ and $s \ge \sqrt{p}\log^{-2/\alpha}(ep) \vee c$, for some absolute constant $c > 0$ and some constant $K_\alpha > 0$ depending only on α. Then, there exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_{\mathcal{G}}(\rho) \ge 1/2$ whenever $\rho^2 \le c' v^{\mathrm{L}}_{\mathcal{G},\mathrm{dense}}$, where*

$$v^{\mathrm{L}}_{\mathcal{G},\mathrm{dense}} := \sqrt{p(\log\log(8n))^{\omega}} + \log\log(8n),$$

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

2.2 Sparse regime

In the sparse regime, we consider a sampling-splitting testing procedure where, intuitively, we first use half of data to identify coordinates that exhibit strong signal of changing, and then use the second half of the data to aggregate the selected 'signal' coordinates. Such a methodology is applicable for testing potential change locations $t \in \mathcal{T}\backslash\{1\}$ and we deal with testing the special case of $t = 1$ separately.

To be specific, for $t \in \mathcal{T}\backslash\{1\}$, we define the sample-splitting version of (7) as

$$Y_{t,1} := \frac{\sum_{i=1}^{t/2} X_{2i-1} - \sum_{i=1}^{t/2} X_{n+1-2i}}{\sqrt{t}}, \qquad Y_{t,2} := \frac{\sum_{i=1}^{t/2} X_{2i} - \sum_{i=1}^{t/2} X_{n+2-2i}}{\sqrt{t}}, \tag{9}$$

and consider the following test

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

where

$$A_{t,a} := \begin{cases} \sum_{j=1}^{p} \big(Y_{t,1}^2(j) - 1\big) \mathbb{1}_{\{|Y_{t,2}(j)| \geq a\}} & t \geq 2, \\ \sum_{j=1}^{p} \big(Y_{t}^2(j) - 1\big) \mathbb{1}_{\{|Y_{t}(j)| \geq a\}} & t = 1, \end{cases} \tag{11}$$

and a, r, r_1 are parameters specified in (12). The following theorem establishes the theoretical guarantee of the test $\phi_{\mathcal{G},\text{sparse}}$.

Theorem 3. *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 $\phi_{\mathcal{G},\text{dense}}$ defined in (10) with*

$$\begin{aligned} a &= C_1\big(\log^{1/\alpha}(ep/s) + s^{-1/2}\log^{1/2}(\log(8n))\big), \\ r &= C_2\big(\sqrt{s\log\log(8n)} + \log\log(8n)\big), \\ r_1 &= C_3 s \log^{2/\alpha}(ep/s), \end{aligned} \tag{12}$$

satisfies

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

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

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

The idea of selecting coordinates via hard thresholding has been widely used and in particular, in the change point context, considered by Liu et al. (2021) under the Gaussian noise assumption. However, their test does not require sample-splitting due to the tractability of truncated non-central chi-square distribution. Such property disappears even when moving from Gaussian noise assumption to sub-Gaussian noise assumption and results in different rates in the sparse regime (See Section 7.1 in Pilliat et al., 2023). From a technical point of view, our use of sample-splitting allows for the independence between the coordinate selection and ℓ_2 aggregation steps, which simplifies the analysis while achieving the optimal testing rate. A test based on (11) but without using sample splitting can be shown to have a slightly worse testing rate of $s\log^{2/\alpha}(ep/s) + \log(es)\log\log(8n)$ with our current proof technique. The following proposition shows that the test $\phi_{\mathcal{G},\text{sparse}}$ achieves the minimax testing rate in the sparse regime as long as the sparsity level is larger than an absolute constant.

Proposition 4. *Let $0 < \alpha \leq 2$, $K \geq K_\alpha$ and $c \leq s \leq \sqrt{p}\log^{-2/\alpha}(ep)$, for some absolute constant $c > 0$ and some constant $K_\alpha > 0$ depending only on α. Then, there exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_{\mathcal{G}}(\rho) \geq 1/2$ whenever $\rho^2 \leq c' v^{\mathrm{L}}_{\mathcal{G},\mathrm{sparse}}$, where*

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

2.3 Discussion on the minimax gap

We now discuss the gap between the upper bound rates $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}$, $v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}$ and the lower bound rates $v^{\mathrm{L}}_{\mathcal{G},\mathrm{dense}}$, $v^{\mathrm{L}}_{\mathcal{G},\mathrm{sparse}}$. Recall that each rate is a function of p, n and s. A key observation is that Theorems 1 and 3 hold for any sparsity level s. In other words, for any given s, we can simultaneously run the two testing procedures described in Sections 2.1 and 2.2 and take $\phi_{\mathcal{G},\mathrm{dense}} \vee \phi_{\mathcal{G},\mathrm{sparse}}$ as our test. This gives a upper bound

$$v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s) \wedge v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}(p,n,s) \tag{13}$$

for the minimax rate of testing. The separation boundary between dense and sparse regimes based on (13) is $s_{\mathcal{G}}$ which satisfies $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s_{\mathcal{G}}) = v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}(p,n,s_{\mathcal{G}})$, i.e. $\sqrt{p\log\log(8n)} = s_{\mathcal{G}}\log^{2/\alpha}(ep/s_{\mathcal{G}})$. Define $\phi(s) := s\log^{2/\alpha}(e^{2/\alpha}p/s)$, which is an increasing function in s. Then $s_{\mathcal{G}}$ is of the same order as $\phi^{-1}(\sqrt{p\log\log(8n)})$. This is a slightly different boundary from $s^* = \sqrt{p}\log^{-2/\alpha}(ep)$ that we defined in (5), which is the same order $\phi^{-1}(\sqrt{p})$. Note that s^* does not depend on the sample size n while $s_{\mathcal{G}}$ does. Observe also that due to the monotonicity of $s \mapsto \phi(s)$, we have $s^* \leq C_\alpha s_{\mathcal{G}} < \sqrt{p\log\log(8n)}$ for some constant $C_\alpha > 0$ depending only on α.

In our sparse regime $s < s^*$, the rate in (13) simply becomes $v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}(p,n,s)$, and exactly matches the lower bound $v^{\mathrm{L}}_{\mathcal{G},\mathrm{sparse}}(p,n,s)$. In our dense regime, we first focus on the sparsity region $s \geq \sqrt{p\log\log(8n)}$. The rate in (13) is now $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s)$ and again matches the corresponding lower bound $v^{\mathrm{L}}_{\mathcal{G},\mathrm{dense}}(p,n,s)$. We remark that $s \geq \sqrt{p\log\log(8n)}$ is in fact the 'dense regime' under Gaussian noise assumption (Liu et al., 2021, Theorem 1).

The more intriguing region of the sparsity level is $s^* \leq s < \sqrt{p\log\log(8n)}$, where we have used the dense version of the test in Section 3.1 to achieve a rate of $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s)$. Note that a gap of $\sqrt{\log\log(8n)}$ exists between $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s_{\mathcal{G}})$ and $v^{\mathrm{L}}_{\mathcal{G},\mathrm{dense}}(p,n,s_{\mathcal{G}})$. When $s^* \leq s \leq s_{\mathcal{G}}$, we can use the sparse version of test described in Section 2.2 to achieve the rate in (13), which provides a small improvement over $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p,n,s)$. In fact, any sparsity level between s^* and $\sqrt{p\log\log(8n)}$ can be chosen to be the boundary between using the dense test (6) and using the sparse test (10) and leads to an overall minimax gap of order $\sqrt{\log\log(8n)}$. For ease of exposition, we have used s^* as this boundary, as it has a simple closed form expression depending only on p.

Lastly, we note that the minimax gap discussed in the last paragraph even exists when each entry of the noise matrix follows a sub-Gaussian distribution rather than a standard normal; see Pilliat et al. (2023, Section 7.1), where it is suggested that a procedure that explores the exact distribution of the noise may be able to further close this gap.

3 Testing under finite moment noise distribution

In this section, we consider the case when $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$, defined in (4), or equivalently, we assume that the distribution of each entry in the noise matrix E has only finitely α-th moments, for some constant $\alpha \geq 2$. This data-generating mechanism allows a much wider range of noise distributions, including, for example, t distributions and Pareto distributions, compared to the class of sub-Weibull distributions. As a result, it poses a much large statistical challenge, and new approaches to tackle the testing problem are required.

The results developed in the previous section show that, when the noise distributions belong to the sub-Weibull class, standard CUSUM-type testing procedure can already achieve near-optimal minimax testing rate (up to $\sqrt{\log\log(8n)}$) and using more advanced tools from robust statistics is unlikely to result in major improvement on the testing rate. However, when the error distribution has a much heavier tail, i.e. only decays polynomially, which implies the existence of only finitely many moments, the story is different. Using CUSUM-type statistics alone will not be enough to obtain ideal performance due to a lack of sharp concentration under such settings. Instead, we borrow wisdom from robust statistics in pursuit of better concentration around true change point signal.

Similar to the sub-Weibull case, we write the worst case testing error as $\mathcal{R}_{\mathcal{P}}(\rho, \phi)$ and the minimax testing error as $\mathcal{R}_{\mathcal{P}}(\rho)$. Also, we assume the sparsity level to be known and consider the dense and sparse cases separately as in Sections 1.1 and 2. The boundary between these two regimes is now

$$s^* := p^{\frac{1}{2}-(\frac{1}{\alpha-2}\wedge\frac{1}{2})}, \tag{14}$$

and the dense regime refers to the case $s \geq s^*$, while the sparse regime refers to its complement. Notably, when $\alpha \in [2, 4]$, we always have

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

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

3.1 Dense regime

We now describe our testing procedure for dense changes which builds on the median-of-means principle in robust statistics literature. For $i \leq n/2$, we denote $Z_i := (X_i - X_{n-i+1})/\sqrt{2}$. For $t \in \mathcal{T}$, we split $\{Z_1, \dots, Z_t\}$ into G_t groups of equal size

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

where each group contains $t/G_t \geq 1$ elements and the number of groups G_t is specified later in (17). Set $V_{t,g} \in \mathbb{R}^p$ with

$$V_{t,g}(j) := \overline{Z}_{t,g}^2(j) - \frac{G_t}{t},$$

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 (7), 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:

$$A_t^{\mathrm{MoM}} := t \cdot \mathrm{median}\left(\sum_{j=1}^{p} V_{t,1}(j), \sum_{j=1}^{p} V_{t,2}(j), \dots, \sum_{j=1}^{p} V_{t,G_t}(j)\right), \tag{15}$$

and our test is denoted as

$$\phi_{\mathcal{P},\mathrm{dense}} := \mathbb{1}_{\left\{\max_{t\in\mathcal{T}} A_t^{\mathrm{MoM}}/r_t > 1\right\}}, \tag{16}$$

with the detection threshold r_t to be specified in (17). Before presenting the theoretical guarantee of the test $\phi_{\mathcal{P},\mathrm{dense}}$, we first briefly explain the significance of median-of-means type statistics and the novelty of our procedure.

Median-of-means type statistics like (15) have been widely applied in the context of mean estimation (Lugosi and Mendelson, 2019b), empirical risk minimisation (Lerasle and Oliveira, 2011; Lecué and Lerasle, 2020) with applications to regression and density estimation problems (Humbert et al., 2022), estimator selection (Kwon et al., 2021) and classification (Lecué and Lerasle, 2020), among others. Perhaps the most well-known and simplest form is the median-of-means estimator for univariate mean estimation. Suppose we have i.i.d data of 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 samples means. It is shown in Lugosi and Mendelson (2019a, Theorem 2) that when the group size G is chosen to be at least $8\log(1/\delta)$, then with probability at least $1-\delta$, the estimator $\hat{\mu}^{\mathrm{MOM}} = \hat{\mu}^{\mathrm{MOM}}(\delta)$ satisfies

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

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 close to the boundary, as we will not have enough data to ensure good statistical guarantees. Therefore, for $t \in \mathcal{T}$ such that $t \leq \Delta$ with the threshold Δ specified later in (17), we essentially directly take the median of t statistics in (15), i.e. $G_t = 1$. Another challenge in our context lies in analysing the performance of our test when $\alpha \in [2,4]$. Since we compute a second order statistic $V_{t,g}$ within each group g, standard analysis would require a bounded fourth moment condition on the distribution. We are able to extend our result to this more challenging case of $\alpha \in [2,4]$, which is critical in unveiling a new phase transition phenomenon that is previously unknown even when n is of constant order.

Theorem 5. *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 (16) with*

$$r_t = C_1 p^{(1/2)\vee(2/\alpha)} G_t, \quad G_t = t \wedge \Delta, \quad \Delta = 2^{3+\lceil \log_2 \log\log(8n)\rceil}, \tag{17}$$

satisfies

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

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

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

The following proposition shows that the test $\phi_{\mathcal{P},\mathrm{dense}}$ is minimax optimal up to $\log\log(8n)$ in the dense regime. More specifically, the gap is of order $\sqrt{\log\log(8n)}$ when $s > \sqrt{p\log\log(8n)}$ and $\alpha \geq 4$ and of order $\log\log(8n)$ otherwise. We carefully discuss this gap in Section 3.3.

Proposition 6. *Let $\alpha \geq 2$, $K > K_\alpha$ and $s \geq p^{\frac{1}{2}-\frac{1}{\alpha-2}\wedge\frac{1}{2}} \vee c$, for some absolute constant $c > 0$ and some constant $K_\alpha > 0$ depending only on α. Then, there exists some constant $c' > 0$ depending only on α and K, such that $\mathcal{R}_{\mathcal{P}}(\rho) \geq 1/2$ whenever $\rho^2 \leq c' v^{\mathrm{L}}_{\mathcal{P},\mathrm{dense}}$, where*

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

with $\omega = \mathbb{1}_{\left\{s > \sqrt{p\log\log(8n)}\right\}\cap\{\alpha\geq 4\}}$.

Note that when $\alpha \in [2,4]$, the lower bound in Proposition 6 holds for essentially all sparse regimes except when s is less than an absolute constant, which arises as an artefact of our proof. This means that the testing rates in this extreme heavy-tailed setting are, in fact, independent of s, which implies that it is impossible to exploit the sparse structure of the change in pursuit of better results. Therefore, in the subsequent discussion of sparse regime, we only consider the case of $\alpha \geq 4$.

3.2 Sparse regime

In the sparse regime, we take two steps towards constructing a computationally efficient and minimax optimal test denoted as $\phi_{\mathcal{P},\text{sparse}}$, which is presented at the end of this section. This test achieves the testing rate

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

as shown in Theorem 9 and it is the minimax optimal testing rate in the sparse regime as confirmed in Proposition 10.

Our first attempt is, at a high level, combining the median-of-means methodology developed in the dense regime Section 3.1 with the hard-thresholding coordinate selection technique used in Section 2.2. Recall that $Z_i = (X_i - X_{n-i+1})/\sqrt{2}$, for $i \leq n/2$. 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 group of equal size, denoted as

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

with the group number G_t specified later in (19), and use $\overline{Z}_{t,g,1}$ to denote the sample mean of the g-th group. The second set of sample $\{Z_2, Z_4, \ldots, Z_t\}$ is reserved for selecting the signal coordinates as we did in Section 2.2. Now, consider the statistic $V_{t,g,a} \in \mathbb{R}^p$ with

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

where $Y_{t,2}(j)$ is defined in (9) and a is a selection threshold to be specified in (19). Our test statistic takes almost the same form as in the dense case

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

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

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

Finally, our test is

$$\phi^{\mathrm{MoM}}_{\mathcal{P},\text{sparse}} := \mathbb{1}_{\{\max_{t \in \mathcal{T}} A^{\mathrm{MoM}}_{t,a}/r_t > 1\}}. \tag{18}$$

The theoretical guarantee of $\phi^{\mathrm{MoM}}_{\mathcal{P},\text{sparse}}$ is established in Proposition 7 below. It is 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. This is mainly to avoid further complication of the procedure since if, for example, the median-of-means estimator were deployed for each coordinate, then we would need to further split the data into groups and deal with the case when t is close to the boundary.

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 the test $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ defined in (18) with*

$$a = C_1\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log(8n))\big), \quad r_t = C_2\big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + \sqrt{s}G_t\mathbb{1}_{\{t>1\}}\big),$$
$$G_t = (t \wedge \Delta)/2, \quad \Delta = 2^{4+\lceil \log_2 \log\log(8n)\rceil}, \tag{19}$$

satisfies

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

as long as $\rho^2 \geq C_3 v^{\mathrm{U,MoM}}_{\mathcal{P},\mathrm{sparse}}$, where

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

The rate achieved by $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ is slightly worse than the presented rate (vii) in Table 1. Our second attempt, in order to close this gap, is using some robust sparse mean estimator in the literature to directly construct a test. One example of such estimator is given in Prasad et al. (2019):

$$\hat{\mu}^{\mathrm{RSM}}_{n,s}(\{W_i\}_{i=1}^n; \eta) := \inf_{\mu \in \mathcal{L}_s} \sup_{u \in \mathcal{N}^{1/2}_{2s}(\mathcal{S}^{p-1})} |u^\top \mu - \mathrm{1DRobust}(\{u^\top W_i\}_{i=1}^n, \eta/(6ep/s)^s)|, \tag{20}$$

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 $\mathbb{R}^p$, $\mathcal{N}^{1/2}_{2s}(\mathcal{S}^{p-1})$ is a half-cover of the set of $2s$-sparse unit vectors with cardinally $|\mathcal{N}^{1/2}_{2s}(\mathcal{S}^{p-1})| \leq (6ep/s)^s$ (Vershynin, 2009), and 1DRobust is a univariate robust mean estimator defined in Prasad et al. (2019, Algorithm 2) based on the shorth estimator (e.g. Andrews and Hampel, 2015). One may naturally consider other univariate robust mean estimators in place of 1DRobust, including the median-of-means, as discussed briefly in Section 3.1 or trimmed mean variants (Lugosi and Mendelson, 2021). We note that (20) achieves near-optimal result for the task of sparse mean estimation as established in Prasad et al. (2019, Corollary 11), despite its high computation complexity, which scales exponentially with p.

Now, we describe the testing procedure using $\hat{\mu}^{\mathrm{RSM}}_{n,s}(\{W_i\}_{i=1}^n; \eta)$ as an alternative to $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$. For $t \leq \tilde{\Delta}_1$, we simply use the non-robust statistic $A_{t,a}$ as defined in (11). For $t \in \mathcal{T} \cap \{t > \tilde{\Delta}_1\}$, we directly construct the statistic from the ℓ_2 norm of this robust sparse mean estimator:

$$A^{\mathrm{RSM}}_t := t\big\|\hat{\mu}^{\mathrm{RSM}}_{t,s,\eta_t}\big\|_2^2,$$

where we use the notation $\hat{\mu}^{\mathrm{RSM}}_{t,s,\eta_t}$ as a shorthand for $\hat{\mu}^{\mathrm{RSM}}_{t,s}(\{Z_i\}_{i=1}^t; \eta_t)$. With all the parameters $a, \tilde{\Delta}_1, \tilde{\Delta}_2, \tilde{r}_t, r^{\mathrm{RSM}}_t, \eta_t$ specified later in (22), our test is given by

$$\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}} := \mathbb{1}_{\left\{\max_{t\in\mathcal{T}\cap\{t\leq\tilde{\Delta}_1\}} A_{t,a}/\tilde{r}_t > 1\right\}} \vee \mathbb{1}_{\left\{\max_{t\in\mathcal{T}\cap\{t>\tilde{\Delta}_1\}} A^{\mathrm{RSM}}_t/r^{\mathrm{RSM}}_t > 1\right\}}. \tag{21}$$

The theoretical guarantee of $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ is established in the following result.

Proposition 8. *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 $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ defined in (21) with*

$$a = C_1\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log\tilde{\Delta}_1)\big), \quad \tilde{r}_t = C_2\Big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + \sqrt{s\log\tilde{\Delta}_1}\mathbb{1}_{\{t>1\}}\Big),$$
$$\eta_t = \exp\bigg\{s\log(ep/s) - \frac{t \wedge \tilde{\Delta}_2}{C_3}\bigg\}, \quad r^{\mathrm{RSM}}_t = C_4(t \wedge \tilde{\Delta}_2),$$

$$\tilde{\Delta}_1 = C_3\big(s\log(ep/s) + \log(16/\varepsilon)\big), \quad \tilde{\Delta}_2 = C_3\big(s\log(ep/s) + \log(16\log(2n)/\varepsilon)\big), \tag{22}$$

satisfies

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

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

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

Before addressing the issue of computation, several remarks are in order. The main reason we separate $t \in \mathcal{T}$ into different regions is, similar to the issue with median-of-means, that (20) cannot be applied when t is too close to the boundary, and therefore, we need to resort to the non-robust testing statistics $A_{t,a}$. A more subtle issue which is critical in achieving, in fact, the optimal rate $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$ is that we apply the non-robust testing statistics only to a very small number of points in $\mathcal{T}$. Indeed, the boundary $\tilde{\Delta}_1$ is chosen to be independent of n so that the power of (20) can be maximally exploited. Lastly, we mention that our proof of Proposition 8, which establishes the theoretical guarantee of $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$, is actually modular. Any other sparse mean estimator that satisfies a more general condition (Condition 1 in Appendix A.1.5) may be used in place of $\hat{\mu}^{\mathrm{RSM}}_{t,s,\eta_t}$ and the corresponding test achieves the same performance.

As we mentioned before, one caveat of using a robust estimator such as (20), and in fact, a common issue for high dimensional robust statistics in order to achieve good performance, is that the estimator is computationally intractable since it often involves projecting the data onto every $2s$-sparse unit vectors or its covering set. As a result, even though the testing procedure $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ in (21) performs better at detecting change point with weak signal compared to the previous median-of-means type test, the computational cost makes it not implementable in practice, especially when p is large. The main result of this section, as presented in the theorem below, is a testing procedure that achieves the best of both worlds - computational efficiency and minimax optimality.

Theorem 9. *Assume* $\alpha \geq 4$*. Consider the test*

$$\phi_{\mathcal{P},\mathrm{sparse}} := \begin{cases} \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}} & \text{if } p \leq \log^{\alpha-2}(\log(8n)), \\ \phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}} & \text{otherwise}, \end{cases}$$

with the parameters of $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ *and* $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ *chosen according to* (19) *and* (22) *respectively. Then it holds that*

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

as long as $\rho^2 \geq C_1 v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$*, for some constant* C_1*, depending only on* α*,* K *and* ε*. Moreover, the computational complexity of* $\phi_{\mathcal{P},\mathrm{sparse}}$ *is polynomial in* n *and* p*.*

Theorem 9 is based on the following neat observation: by comparing $v^{\mathrm{U,MoM}}_{\mathcal{P},\mathrm{sparse}}$ and $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$, as established in Proposition 7 and Proposition 8 respectively, the improvement offered by $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ over $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ only exists when p is sufficiently small. Therefore, we can bypass the computation barrier by using $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ only when p is sufficiently small, which yields a combined testing procedure that has run time polynomial in n and p while achieving the optimal rate in the sparse regime. The following proposition establish the optimality of the rate $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$ achieved by $\phi_{\mathcal{P},\mathrm{sparse}}$.

Proposition 10. *Let* $\alpha \geq 4$*,* $K \geq K_\alpha$ *and* $c \leq s \leq p^{\frac{1}{2}-\frac{1}{\alpha-2}}$*, for some absolute constant* $c > 0$ *and some constant* $K_\alpha > 0$ *depending only on* α*. Then, there exists some constant* $c' > 0$ *depending only on* α *and* K*, such that* $\mathcal{R}_{\mathcal{P}}(\rho) \geq 1/2$ *whenever* $\rho^2 \leq c' v^{\mathrm{L}}_{\mathcal{P},\mathrm{sparse}}$*, where*

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

3.3 Discussion on the minimax gap

We first focus on the case $\alpha \geq 4$, where there is a non-empty sparse regime $s < s^* = p^{\frac{1}{2}-\frac{1}{\alpha-2}}$. Our first observation is that the upper and lower bounds for the minimax rate of testing again match in the sparse regime. This is the same conclusion as in the sub-Weibull error setting. Second, consider the mean change vector to be fully dense, i.e. $s = p$. Then, we have $v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}}(p, n, p) = \sqrt{p} \log\log(8n)$. Comparing this with the fully dense minimax rate in the sub-Weibull case $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}}(p, n, p)$, we notice an extra factor of $\sqrt{\log\log(8n)}$ when error distributions only have polynomially-decaying tails. We conjecture that our upper bound rate in the dense regime $v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}}(p, n, p)$ may not be tight and discuss a potential way to close this gap. Lugosi and Mendelson (2019b) among others show that there exists a multivariate mean estimator $\hat{\mu}(\delta)$, such that for all distributions with mean μ and covariance I_p, with probability at least $1 - \delta$,

$$\|\hat{\mu}(\delta) - \mu\| \leq C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{23}$$

where n is the sample size and C is a universal constant. Now, consider an estimator $\hat{\mu}'(\delta)$ that satisfies the following stronger condition similar to Laurent and Massart (2000, Lemma 1) for all distributions with finite fourth moment:

$$\|\hat{\mu}'(\delta) - \mu\|_2^2 - \frac{p}{n} \leq C\left(\frac{\sqrt{p\log(1/\delta)}}{n} + \frac{\log(1/\delta)}{n}\right). \tag{24}$$

If we can find an estimator that satisfies (24), then it may help us close this gap from $\sqrt{p}\log\log(8n)$ to $\sqrt{p\log\log(8n)}$, as Laurent and Massart (2000, Lemma 1) lies at the heart of proving the dense rate of $\sqrt{p\log\log(8n)}$ under Gaussian noise assumption in Liu et al. (2021). However, whether such a robust estimator exists remains an open question.

Recall from Section 2.3 that we have a minimax gap of $\sqrt{\log\log(8n)}$ in the dense regime under the sub-Weibull error distribution setting. This is also the case here and, therefore, we omit further discussion on this issue. By combining the two gaps discussed above, we conclude that for any sparsity level in the dense regime, the upper bound and the lower bound of the minimax testing rate are only off by a factor of order at most $\log\log(8n)$.

Lastly, when $2 \leq \alpha < 4$, we first note that for $\alpha = 2$, the minimax optimal rate is $p + \log\log(8n)$. This can be shown by combining the lower bound proof from Section A.3 items (i) and (iii) with a construction of a test to (21) and replacing $\hat{\mu}^{\mathrm{RSM}}$ with a robust mean estimator $\hat{\mu}$ that satisfies (23). In view of the above minimax rate at $\alpha = 2$ and $\alpha \geq 4$, we conjecture that the precise minimax rate of testing for $\alpha \in (2, 4)$ is

$$p^{2/\alpha} \log^{1-2/\alpha} \log(8n) + \log\log(8n).$$

Using Theorem 5 and Proposition 6, the best upper and lower bounds we can achieve in this regime are of the orders $p^{2/\alpha}\log\log(8n)$ and $p^{2/\alpha} + \log\log(8n)$ respectively. The precise exponent on $\log\log(8n)$ may have an interesting dependence on α that will need to be characterised in future research.

4 Adaptation to sparsity

In Section 2 and Section 3, 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, (2) 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 other words, we have assumed the sparsity to be known up until now. In this section, we study the adaptation of these procedures to unknown sparsity level.

We first recall that 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 essentially no sparse regime (see the discussion following Proposition 6). The test $\phi_{\mathcal{P},\mathrm{dense}}$ defined in (16) with its parameters specified in Theorem 5 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 now focus on the case when $P_e \in \mathcal{P}^{\otimes}_{\alpha,K}$ for $\alpha > 4$. Recall from Theorems 5 and 9 that $\phi_{\mathcal{P},\mathrm{dense}}$ 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 is known. In the following, we introduce an adaptive test procedure based on these two tests, while making the dependence on s explicit everywhere in the notation:

$$\begin{aligned}\phi_{\mathcal{P},\mathrm{adaptive}} &:= \phi_{\mathcal{P},\mathrm{dense}} \vee \max_{s\in\mathcal{K}} \phi_{\mathcal{P},\mathrm{sparse},s} \\ &= \begin{cases} \phi_{\mathcal{P},\mathrm{dense}} \vee \max_{s\in\mathcal{K}} \phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse},s} & \text{if } p > \log^{\alpha-2}(\log(8n)), \\ \phi_{\mathcal{P},\mathrm{dense}} \vee \max_{s\in\mathcal{K}} \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse},s} & \text{if } p \leq \log^{\alpha-2}(\log(8n)), \end{cases}\end{aligned} \tag{25}$$

where $\mathcal{K} := \{1, 2, 4, \ldots, 2^{\lceil \log_2(p)\rceil - 1}\}$ is a dyadic grid. Recall that $\phi_{\mathcal{P},\mathrm{dense}}$ does not require the knowledge of s, and we keep its original parameter choices as in (17), with perhaps an enlarged value of the leading constant C_1 in r_t:

$$r_t = C_1 p^{(1/2)\vee(2/\alpha)} G_t, \quad G_t = t \wedge \Delta, \quad \Delta = 2^{3+\lceil \log_2 \log\log(8n)\rceil}. \tag{26}$$

For $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse},s}$, we modify the original parameter choices (19) as follows:

$$\begin{aligned} a_s = C_2\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log(8n))\big), \quad r_{t,s} = C_3\big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + s^{3/4} G_t \mathbb{1}_{\{t>1\}}\big), \\ G_t = (t\wedge\Delta)/2, \quad \Delta = 2^{4+\lceil \log_2 \log\log(8n)\rceil}. \end{aligned} \tag{27}$$

Compared with (19), we use the same a_s (again with perhaps a larger leading constant) and modify $r_{t,s}$. Finally, for $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse},s}$, we modify its original parameter choices (22) to be:

$$\begin{aligned} a_s = C_4\big((p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\log\tilde{\Delta}_{1,s})\big), \quad \tilde{r}_{t,s} = C_5\Big(s(p/s)^{2/\alpha}\mathbb{1}_{\{t=1\}} + s^{3/4}\sqrt{\log\tilde{\Delta}_1}\,\mathbb{1}_{\{t>1\}}\Big), \\ \eta_{t,s} = \exp\bigg\{s\log(ep/s) - \frac{t\wedge\tilde{\Delta}_2}{C_6}\bigg\}, \quad r^{\mathrm{RSM}}_{t,s} = C_7(t\wedge\tilde{\Delta}_{2,s}), \\ \tilde{\Delta}_{1,s} = C_6\big(s\log(ep/s) + \log(80s/\varepsilon)\big), \quad \tilde{\Delta}_{2,s} = C_6\big(s\log(ep/s) + \log(80s\log(2n)/\varepsilon)\big), \end{aligned} \tag{28}$$

Note that $\tilde{\Delta}_{1,s}$ and $\tilde{\Delta}_{2,s}$ now both depend on the sparsity through a logarithmic term.

Theorem 11. *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 $\phi_{\mathcal{P},\mathrm{adaptive}}$ defined in (25) with its parameters specified in (26), (27) and (28) satisfies*

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

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

Note again that $v^{\mathrm{U}}_{\mathcal{P},\mathrm{dense}} \wedge v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$ matches the lower bound in the sparse regime $s < s^*$, where s^* is defined in (14), while a minimax gap of order $\log\log(8n)$ exists in the dense regime; see Sections 2.3 and 3.3 for thorough discussions. When the errors have sub-Weibull tails instead, a similar adaptive testing procedure can be constructed based on $\phi_{\mathcal{G},\mathrm{dense}}$ and $\phi_{\mathcal{G},\mathrm{dense}}$ and can achieve the rate $v^{\mathrm{U}}_{\mathcal{G},\mathrm{dense}} \wedge v^{\mathrm{U}}_{\mathcal{G},\mathrm{sparse}}$. For the sake of brevity, we omit further details here.

5 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 $\sqrt{\log\log(8n)}$ in the case of exponentially-decaying tails, and up to $\log\log(8n)$ in the case of polynomially-decaying tails. Thorough discussions on these gaps are provided in Sections 2.3 and 3.3. In addition, our results quantify the costs of heavy-tailed distributions in this problem by comparing to the previous results under Gaussian error assumption (Liu et al., 2021) 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. It is known that for the mean estimation problem, the fundamental difficulty change dramatically when the distribution has fewer than two finite moments (e.g. Bubeck et al., 2013; Cherapanamjeri et al., 2022). Our results suggest that detecting mean change is rather different from mean estimation and, in fact, has close link to the problem of signal estimation in the sequence model (Comminges et al., 2021). There are several avenues for future research and we briefly discuss them below.

Temporal and spatial dependence. Throughout this paper, we have assumed independence across both coordinates and time. This is possibly the most natural starting point. To relax the independence assumption, one may consider data columns to be stationary with short-range dependence as considered in, for example Wang and Samworth (2018) and Liu et al. (2021). Theory in such settings would require deploying different finite sample analysis tools under weak dependence. As for spatial dependence, similar strategies would work as conducted in e.g. Jiang et al. (2023). 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 decaying 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. Castillo et al., 2014; Bhati, 2020) and on estimating the tail index parameter for sub-Weibull distributions (Vladimirova et al., 2020), which may be combined with our tests to obtain adaptivity. We leave this ambitious task for the future.

Multiple change points setting. A natural extension of this work is to consider the problem of testing and/or estimation multiple change points. A wide range of methodologies in change point analysis literature have been proposed to detect multiple change points via repeatedly testing for a single change point in a collection of sub-intervals of the entire time series data. Many of them share a multiscale nature, including wild binary segmentation (Fryzlewicz, 2014), seeded binary segmentation (Kovács et al., 2020), grid-based approaches (Pilliat et al., 2023), among others. The theoretical performance of these methods are well studied in non-robust change point detection under various models, and we believe they are promising for future research agendas on multiple change point detection within high-dimensional heavy-tailed data streams, where the single change point tests developed in this paper can play a crucial role.

References

Andrews, D. F. and Hampel, F. R. (2015) Robust estimates of location. In *Robust Estimates of Location*. Princeton University Press.

Beiglböck, M. and Siorpaes, P. (2015) Pathwise versions of the Burkholder–Davis–Gundy inequality. *Bernoulli*, **21**, 360–373.

Bhati, D. (2020) A test procedure for distinguishing logarithmically decaying tail from polynomially decaying tail. *J. Korean Stat. Soc.*, **49**, 841–862.

Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013) Bandits with heavy tail. *IEEE Transactions on Information Theory*, **59**, 7711–7717.

Cao, Y. and Xie, Y. (2017) Robust sequential change-point detection by convex optimization. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 1287–1291. IEEE.

Castillo, J. D., Daoudi, J. and Lockhart, R. (2014) Methods to distinguish between polynomial and exponential tails. *Scand. J. Stat.*, **41**, 382–393.

Catoni, O. (2012) Challenging the empirical mean and empirical variance: a deviation study. *Ann. inst. Henri Poincare (B) Probab. Stat.*, **48**, 1148–1185.

Chenouri, S., Mozaffari, A. and Rice, G. (2020) Robust multivariate change point analysis based on data depth. *Can. J. Stat.*, **48**, 417–446.

Cherapanamjeri, Y., Tripuraneni, N., Bartlett, P. and Jordan, M. (2022) Optimal mean estimation without a variance. In *Conference on Learning Theory*, 356–357. PMLR.

Cho, H. and Owens, D. (2022) High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence. *arXiv preprint*, arXiv:2209.08892.

Comminges, L., Collier, O., Ndaoud, M. and Tsybakov, A. B. (2021) Adaptive robust estimation in sparse vector model. *Ann. Statist.*, **49**, 1347–1377.

Depersin, J. (2020) Robust subgaussian estimation with VC-dimension. *arXiv preprint*, arXiv:2004.11734.

Depersin, J. and Lecué, G. (2022) Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.*, **50**, 511–536.

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016) Sub-Gaussian mean estimators. *Ann. Statist.*, **44**, 2695–2725.

Fan, J., Li, Q. and Wang, Y. (2014) Robust estimation of high-dimensional mean regression. *arXiv preprint*, arXiv:1410.2150.

Fearnhead, P. and Rigaill, G. (2019) Changepoint detection in the presence of outliers. *J. Amer. Statist. Assoc.*, **114**, 169–183.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281.

Fuk, D. K. (1973) Some probablistic inequalities for martingales. *Sib. Math. J.*, **14**, 131–137.

Gao, C., Han, F. and Zhang, C.-H. (2020) On estimation of isotonic piecewise constant signals. *Ann. Statist.*, **48**, 629–654.

Götze, F., Sambale, H. and Sinulis, A. (2021) Concentration inequalities for polynomials in α-sub-exponential random variables. *Electron. J. Probab.*, **26**, 1–22.

Hanson, D. L. and Wright, F. T. (1971) A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.*, **42**, 1079–1083.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Humbert, P., Le Bars, B. and Minvielle, L. (2022) Robust kernel density estimation with median-of-means principle. In *International Conference on Machine Learning*, 9444–9465. PMLR.

Ibragimov, I. A. and Has' Minskii, R. Z. (2013) *Statistical Estimation: Asymptotic Theory*, volume 16. Springer Science & Business Media.

Jiang, F., Wang, R. and Shao, X. (2023) Robust inference for change points in high dimension. *J. Multivariate Anal.*, **193**, 105–114.

Kovács, S., Li, H., Bühlmann, P. and Munk, A. (2020) Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633.*

Kuchibhotla, A. K. and Chakrabortty, A. (2022) Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA.*

Kwon, J., Lecué, G. and Lerasle, M. (2021) A MOM-based ensemble method for robustness, subsampling and hyperparameter tuning. *Electron. J. Statist.*, **15**, 1202–1227.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

Lecué, G. and Lerasle, M. (2020) Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, **48**, 906–931.

Lerasle, M. and Oliveira, R. I. (2011) Robust empirical mean estimators. *arXiv preprint*, arXiv:1112.3914.

Li, M. and Yu, Y. (2021) Adversarially robust change point detection. *Advances in Neural Information Processing Systems*, **34**, 22955–22967.

Liu, H., Gao, C. and Samworth, R. J. (2021) Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, **49**, 1081–1112.

Lugosi, G. and Mendelson, S. (2019a) Mean estimation and regression under heavy-tailed distributions: a survey. *Foundations of Computational Mathematics*, **9**, 1145–1190.

Lugosi, G. and Mendelson, S. (2019b) Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, **47**, 783–794.

Lugosi, G. and Mendelson, S. (2021) Robust multivariate mean estimation: the optimality of trimmed mean. *Ann. Statist.*, **49**, 393–410.

Mendelson, S. and Zhivotovskiy, N. (2020) Robust covariance estimation under L_4-L_2 norm equivalence. *Ann. Statist.*, **48**, 1648–1664.

Minsker, S. (2018) Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, **46**, 2871–2903.

Mitzenmacher, M. and Upfal, E. (2017) *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis.* Cambridge University Press.

Molloy, T. L. and Ford, J. J. (2017) Misspecified and asymptotically minimax robust quickest change detection. *IEEE Transactions on Signal Processing*, **65**, 5730–5742.

Mukherjee, D., Banerjee, M. and Ritov, Y. (2022) On robust learning in the canonical change point problem under heavy tailed errors in finite and growing dimensions. *Electron. J. Statist.*, **16**, 1153–1252.

Nagaev, S. (1979) Large deviations of sums of independent random variables. *Ann. Probab.*, **7**, 745–789.

Nesterov, Y. (2003) *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media.

Pilliat, E., Carpentier, A. and Verzelen, N. (2023) Optimal multiple change-point detection for high-dimensional data. *Electron. J. Statist.*, **17**, 1240–1315.

Prasad, A., Balakrishnan, S. and Ravikumar, P. (2019) A unified approach to robust mean estimation. *arXiv preprint*, arXiv:1907.00927.

Prasad, A., Suggala, A. S., Balakrishnan, S. and Ravikumar, P. (2020) Robust estimation via robust gradient estimation. *J. Roy. Statist. Soc., Ser. B*, **82**, 601–627.

Ramsay, K. and Chenouri, S. (2020) Robust multiple change-point detection for multivariate variability using data depth. *arXiv preprint*, arXiv:2011.09558.

Resnick, S. I. (2007) *Heavy-tail Phenomena: Probabilistic and Statistical Modeling.* Springer Science & Business Media.

Rio, E. (2017) About the constants in the Fuk–Nagaev inequalities. *Electron. Commun. Probab.*, **22**, 1–12.

Sun, Q., Zhou, W.-X. and Fan, J. (2020) Adaptive Huber regression. *J. Amer. Statist. Assoc.*, **115**, 254–265.

Unnikrishnan, J., Veeravalli, V. V. and Meyn, S. P. (2011) Minimax robust quickest change detection. *IEEE Transactions on Information Theory*, **57**, 1604–1614.

Vershynin, R. (2009) 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)*, 189–192. IEEE.

Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020) Optimal change-point detection and localization. *arXiv preprint*, arXiv:2010.11470.

Vladimirova, M., Girard, S., Nguyen, H. and Arbel, J. (2020) Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, **9**, e318.

Wang, D., Yu, Y. and Rinaldo, A. (2021) Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.*, **49**, 203–232.

Wang, D. and Zhao, Z. (2022) Optimal change-point testing for high-dimensional linear models with temporal dependence. *arXiv preprint*, arXiv:2205.03880.

Wang, T. and Samworth, R. J. (2018) High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, **80**, 57–83.

Xu, H., Wang, D., Zhao, Z. and Yu, Y. (2022) Change point inference in high-dimensional regression models under temporal dependence. *arXiv preprint*, arXiv:2207.12453.

Yu, M. and Chen, X. (2022) A robust bootstrap change point test for high-dimensional location parameter. *Electron. J. Statist.*, **16**, 1096–1152.

Appendices

The proofs of all theoretical results are presented in the Appendices. Appendix A.1 contains proofs of upper bound results, including Theorem 1, Theorem 3, Theorem 5, Proposition 7, Proposition 8 and Theorem 9. Theorem 11 regarding the adaptive test is proved in Appendix A.2. All lower bound results, including Proposition 2, Proposition 4, Proposition 6 and Proposition 10 are proved in Appendix A.3. Appendix B contain auxiliary results.

A Proofs

A.1 Proofs of upper bound results

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_3 in the proof of Theorem 1 below corresponds to C_1 in the statement of Theorem 1.

A.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:

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

Let $U_i(j) = X_i(j) - \theta_1(j)$, now for any $j \in [p]$, we can write

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

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:

$$\begin{aligned}
\|B\|_{\mathrm{F}} &= \sqrt{2tp\Big(\frac{1}{4t^2}+\frac{2t-1}{t^2}\Big)} \le 2\sqrt{p}, \\
\|B\|_2 &= \frac{1}{2t}+\frac{2t-1}{t} \le 2, \\
\|B\|_{2\to\infty} &= \max_{i\in[2t]} \sqrt{\sum_{j\in[2t]} b_{ij}^2} \le \sqrt{\frac{2}{t}}, \\
\|B\|_{\max} &= 1/t.
\end{aligned}$$

For $\alpha \in [1,2]$, we observe by Proposition 13 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 15, we have

$$\begin{aligned}
\mathbb{P}_\theta(A_t > r) \le \exp\bigg\{1-\Big(\frac{r}{C_1\sqrt{p}}\Big)^2\bigg\} + \exp\bigg\{1-\frac{r}{C_1}\bigg\} + \exp\bigg\{1-\Big(\frac{r\sqrt{t}}{C_1}\Big)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\bigg\} \\
+ \exp\bigg\{1-\Big(\frac{rt}{C_1}\Big)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\bigg\},
\end{aligned}$$

where $C_1 > 0$ is some constant depending only on α and K from Proposition 15. Then, by a union bound and Lemma 18, we obtain that for any $\theta \in \Theta_0(p,n)$ and $r \ge C_1\Big\{\big(2^{\frac{\alpha}{2+\alpha}\wedge\frac{1}{3}}-1\big)^{-\left(\frac{2+\alpha}{\alpha}\vee 3\right)} \vee \big(2^{\frac{\alpha}{2}\wedge\frac{1}{2}}-1\big)^{-\left(\frac{2}{\alpha}\vee 2\right)}\Big\}$

$$\begin{aligned}
\mathbb{E}_\theta\phi_{\mathcal{G},\mathrm{dense}} = \mathbb{P}_\theta\big(\max_{t\in\mathcal{T}} A_{t,0} > r\big) &\le e\log_2(n)\exp\bigg\{-\Big(\frac{r}{C_1\sqrt{p}}\Big)^2\bigg\} + e\log_2(n)\exp\bigg\{-\frac{r}{C_1}\bigg\} \\
&\quad + e\sum_{t\in\mathcal{T}}\exp\bigg\{-\Big(\frac{r\sqrt{t}}{C_1}\Big)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\bigg\} + e\sum_{t\in\mathcal{T}}\exp\bigg\{-\Big(\frac{rt}{C_1}\Big)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\bigg\} \\
&\le e\log_2(n)\exp\bigg\{-\Big(\frac{r}{C_1\sqrt{p}}\Big)^2\bigg\} + e\log_2(n)\exp\bigg\{-\frac{r}{C_1}\bigg\} \\
&\quad + 2e\exp\bigg\{-\Big(\frac{r}{C_1}\Big)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\bigg\} + 2e\exp\bigg\{-\Big(\frac{r}{C_1}\Big)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\bigg\},
\end{aligned} \tag{29}$$

There thus exists a large enough constant $C_2 > 0$ depending only on α and C_1, such that, when

$$r \ge C_2\Big(\sqrt{p\log(8e\varepsilon^{-1}\log_2(n))} + \log(8e\varepsilon^{-1}\log_2(n)) + \log^{\frac{2}{\alpha}\vee 2}(16e\varepsilon^{-1})\Big),$$

or equivalently,

$$r \ge C_3\big(\sqrt{p\log\log(8n)} + \log\log(8n)\big),$$

for some constant $C_3 > 0$, 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

$$Y_{\tilde{t}} = \sqrt{\frac{\tilde{t}}{2}}(\mu_1-\mu_2) + \frac{\sum_{i=1}^{\tilde{t}}(X_i-\mu_1) - \sum_{i=1}^{\tilde{t}}(X_{n+1-i}-\mu_2)}{\sqrt{2\tilde{t}}} =: \delta + Y'_{\tilde{t}}, \tag{30}$$

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 12(b) and Lemma 19(a), we have $\mathbb{E}[(Y'_{\tilde{t}}(j))^4] \leq C_4$ for some constant $C_4 > 0$, depending only on α and K. When $\rho^2 \geq 8r \geq 8C_3\big(\sqrt{p\log\log(8n)} + \log\log(8n)\big)$, we have by Chebyshev's inequality that

$$\begin{aligned}
\mathbb{E}_\theta(1-\phi_{\mathcal{G},\mathrm{dense}}) &\leq \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}}\sum_{j=1}^p Y_t(j)^2 - p \leq \rho^2/8\right) \leq \mathbb{P}_\theta\left(\sum_{j=1}^p\Big(Y_{\tilde{t}}(j)^2 - \delta(j)^2 - 1\Big) \leq -\|\delta\|_2^2/2\right) \\
&\leq \frac{4\sum_{j=1}^p \mathrm{Var}_\theta(Y_{\tilde{t}}(j)^2)}{\|\delta\|_2^4} = \frac{4\sum_{j=1}^p \mathrm{Var}_\theta\big(Y'_{\tilde{t}}(j)^2 + 2\delta(j)Y'_{\tilde{t}}(j)\big)}{\|\delta\|_2^4} \\
&\leq \frac{4\sum_{j=1}^p\big\{2\mathrm{Var}_\theta(Y'_{\tilde{t}}(j)^2) + 8\delta(j)^2\mathrm{Var}_\theta(Y'_{\tilde{t}}(j))\big\}}{\|\delta\|_2^4} \leq \frac{\sum_{j=1}^p\big\{8\mathbb{E}_\theta[Y'_{\tilde{t}}(j)^4] + 32\delta(j)^2\big\}}{\|\delta\|_2^4} \\
&\leq \frac{8C_4p + 32\|\delta\|_2^2}{\|\delta\|_2^4} \leq 128\left(\frac{C_4p}{\rho^4} + \frac{1}{\rho^2}\right) \leq \frac{2C_4}{C_3^2\log\log(8n)} + \frac{16}{C_3\sqrt{p\log\log(8n)}},
\end{aligned} \tag{31}$$

where we have used the fact that $\mathrm{Var}(X+Y) \leq 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$ in the fourth inequality. Therefore, by having $C_3 > \max\{64/\varepsilon, \sqrt{8C_4/\varepsilon}\}$, we are guaranteed that $\mathbb{E}_\theta(1-\phi_{\mathcal{G},\mathrm{dense}}) \leq \varepsilon/2$ and the desired result follows.

A.1.2 Proof of Theorem 3

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

$$\mathbb{E}_\theta\phi_{\mathcal{G},\mathrm{sparse}} \leq \mathbb{P}_\theta(A_{1,a} > r_1) + \sum_{t\in\mathcal{T}\setminus\{1\}}\mathbb{P}_\theta(A_{t,a} > r). \tag{32}$$

We first control the second term in (32). Recall the definition of $Y_{t,1}$ and $Y_{t,2}$ from (9) 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,

$$\begin{aligned}
\sum_{t\in\mathcal{T}\setminus\{1\}}\mathbb{P}_\theta(A_{t,a} > r) &\leq \sum_{t\in\mathcal{T}\setminus\{1\}}\mathbb{P}_\theta\left(\sum_{j\in\mathcal{J}_{t,a}}\big(Y_{t,1}^2(j)-1\big) > r\right) \\
&= \sum_{t\in\mathcal{T}\setminus\{1\}}\mathbb{E}_\theta\left[\mathbb{P}_\theta\left(\sum_{j\in\mathcal{J}_{t,a}}\big(Y_{t,1}^2(j)-1\big) > r \,\middle|\, \mathcal{J}_{t,a}\right)\right] \\
&= \sum_{t\in\mathcal{T}\setminus\{1\}}\sum_{J\subseteq[p]}\left\{\mathbb{P}_\theta\left(\sum_{j\in J}\big(Y_{t,1}^2(j)-1\big) > r\right)\mathbb{P}_\theta(\mathcal{J}_{t,a} = J)\right\} \\
&\leq \sum_{t\in\mathcal{T}\setminus\{1\}}\mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t\in\mathcal{T}\setminus\{1\}}\sup_{J\subseteq[p]:|J|\leq s}\mathbb{P}_\theta\left(\sum_{j\in J}\big(Y_{t,1}^2(j)-1\big) > r\right),
\end{aligned} \tag{33}$$

where the third line follows from the independence of $Y_{t,1}$ and $Y_{t,2}$. We now control the two terms in (33) respectively. Using Proposition 14 with $u_i = t^{-1/2}$ for $i = 1, \dots, t/2$ and $u_i = -t^{-1/2}$ for $i = t/2+1, \dots, t$, we obtain that for any $t \in \mathcal{T}$, $j \in [p]$ and $x \geq 0$

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

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

$$q_{t,a} := \mathbb{P}_\theta(|Y_{t,2}(j)| \geq a) \leq \exp\biggl\{1 - \min\biggl\{\Bigl(\frac{a}{C_1}\Bigr)^2, \Bigl(\frac{a}{C_1 t^{(-\frac{1}{2})\vee(\frac{1}{2}-\frac{1}{\alpha})}}\Bigr)^\alpha\biggr\}\biggr\}. \tag{34}$$

For $0 < \alpha < 2$, by combining (34) and a binomial tail bound (Hoeffding, 1963, eq.(2.1)), we have

$$\begin{aligned}
&\sum_{t \in \mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \sum_{t \in \mathcal{T}\setminus\{1\}} \binom{p}{s} q_{t,a}^s \leq \sum_{t \in \mathcal{T}\setminus\{1\}} \Bigl(\frac{epq_{t,a}}{s}\Bigr)^s \\
&\leq \log_2(n)\Bigl(\frac{2e^2p}{s}\Bigr)^s \exp\Bigl\{-\frac{sa^2}{C_1^2}\Bigr\} + \Bigl(\frac{2e^2p}{s}\Bigr)^s \sum_{t \in \mathcal{T}\setminus\{1\}} \exp\biggl\{-s\biggl(\frac{a^{\frac{2\alpha}{\alpha\wedge(2-\alpha)}}t}{C_1^{\frac{2\alpha}{\alpha\wedge(2-\alpha)}}}\biggr)^{\frac{\alpha\wedge(2-\alpha)}{2}}\biggr\} \\
&\leq \log_2(n)\Bigl(\frac{2e^2p}{s}\Bigr)^s \exp\Bigl\{-\frac{sa^2}{C_1^2}\Bigr\} + 2\Bigl(\frac{2e^2p}{s}\Bigr)^s \exp\Bigl\{-\frac{sa^\alpha}{C_1^\alpha}\Bigr\},
\end{aligned} \tag{35}$$

provided that $a \geq C_1\bigl(2^{\frac{\alpha\wedge(2-\alpha)}{2}} - 1\bigr)^{-1/\alpha}$, where we have used Lemma 18 in the last inequality. In fact, for $\alpha = 2$, by (34), the final bound in (35) remains valid for all $a \geq 0$. Thus, as long as we choose a to satisfy

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

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

$$\sum_{t \in \mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \leq \varepsilon/8. \tag{36}$$

We now bound the second term in (33) by invoking a very similar argument used in (29):

$$\begin{aligned}
&\sum_{t \in \mathcal{T}\setminus\{1\}} \sup_{J \subseteq [p]:|J| \leq s} \mathbb{P}_\theta\biggl(\sum_{j \in J}(Y_{t,1}^2(j) - 1) > r\biggr) \\
&\leq \sum_{t \in \mathcal{T}\setminus\{1\}} \sup_{J \subseteq [p]:|J| \leq s} \biggl\{\exp\biggl\{1 - \Bigl(\frac{r}{C_3\sqrt{|J|}}\Bigr)^2\biggr\} + \exp\Bigl\{1 - \frac{r}{C_3}\Bigr\} + \exp\biggl\{1 - \Bigl(\frac{r\sqrt{t}}{C_3}\Bigr)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\biggr\} \\
&\qquad + \exp\biggl\{1 - \Bigl(\frac{rt}{C_3}\Bigr)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\biggr\}\biggr\} \\
&\leq e\log_2(n)\exp\biggl\{-\Bigl(\frac{r}{C_3\sqrt{s}}\Bigr)^2\biggr\} + e\log_2(n)\exp\Bigl\{-\frac{r}{C_3}\Bigr\} + 2e\exp\biggl\{-\Bigl(\frac{r}{C_3}\Bigr)^{\frac{2\alpha}{2+\alpha}\wedge\frac{2}{3}}\biggr\}
\end{aligned}$$

$$+ 2e\exp\left\{-\left(\frac{r}{C_3}\right)^{\frac{\alpha}{2}\wedge\frac{1}{2}}\right\} \leq \varepsilon/8, \tag{37}$$

whenever

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

where $C_3, C_4 > 0$ are constants, depending only on α and K. Now for the first term in (32), by Proposition 17(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

$$\mathbb{P}_\theta(A_{1,a} > r_1) \leq \varepsilon/4. \tag{38}$$

By combining (32), (33), (36), (37) and (38), 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 (30), 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

$$\begin{aligned}
\mathbb{E}_\theta(1 - \phi_{\mathcal{G},\mathrm{sparse}}) &\leq \mathbb{P}_\theta\Big(\sum_{j\in\mathcal{J}_{\tilde{t},a}}\big(Y_{\tilde{t},1}^2(j) - 1\big) \leq r\Big) \\
&= \mathbb{P}_\theta\Big(\sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}^c}\big(Y_{\tilde{t},1}^2(j) - 1\big) + \sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}}\big(Y_{\tilde{t},1}^2(j) - 1\big) \leq r\Big) \\
&\leq \mathbb{P}_\theta\big(|\mathcal{J}_{\tilde{t},a}| > 2s\big) + \mathbb{P}_\theta\Big(\sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}}\big(Y_{\tilde{t},1}^2(j) - 1\big) \leq r + 2s\Big) \\
&\leq \mathbb{P}_\theta\big(|\mathcal{J}_{\tilde{t},a}| > 2s\big) + \mathbb{P}_\theta\Big(\sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}}\delta(j)^2 < \frac{\|\delta\|_2^2}{12\log(8/\varepsilon)}\Big) \\
&\quad + \mathbb{P}_\theta\Big(\sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}}\big(Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1\big) \leq -\frac{\|\delta\|_2^2}{24\log(8/\varepsilon)}\Big).
\end{aligned} \tag{39}$$

We now control the three terms in (39) respectively. By (36), we have

$$\mathbb{P}_\theta\big(|\mathcal{J}_{\tilde{t},a}| > 2s\big) \leq \mathbb{P}_\theta\big(|\mathcal{J}_{\tilde{t},a} \cap \mathcal{S}_\delta^c| > s\big) \leq \varepsilon/8. \tag{40}$$

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

$$\begin{aligned}
\mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) = \mathbb{P}_\theta(|\delta(j) + Y'_{\tilde{t},2}(j)| < a) &\leq \mathbb{P}_\theta(|Y'_{\tilde{t},2}(j)| > |\delta(j)| - a) \\
&\leq \exp\big\{1 - \big((|\delta(j)| - a)/C_1\big)^\alpha\big\} \leq 1/2,
\end{aligned} \tag{41}$$

where the penultimate inequality follows from (34) and the last two inequalities follow from the choice $a \geq C_1 2^{1/\alpha}$. Consequently,

$$\begin{aligned}\mathrm{Var}_\theta\bigl(\delta(j)^2 \mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}}\bigr) \leq \delta(j)^4 \mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) &\leq e\delta(j)^4 \exp\{-(|\delta(j)|/(2C_1))^\alpha\} \\ &\leq \frac{e\delta(j)^4}{\{(|\delta(j)|/(2C_1))^4\}/\lceil 4/\alpha \rceil!} = 16eC_1^4 \lceil 4/\alpha \rceil! =: C_5, \end{aligned} \tag{42}$$

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

$$\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \geq \|\delta\|_2^2 - s(2a)^2 \geq \|\delta\|_2^2/2. \tag{43}$$

We first consider the case $\|\delta\|_2 \geq \sqrt{12 \log(8/\varepsilon)} \|\delta\|_\infty$. Then, when $\rho^2 \geq 192C_5 s$, by combining (41), (42), (43) and Bernstein's inequality, we have

$$\begin{aligned}&\mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \|\delta\|_2^2/8\biggr) = \mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}} < \|\delta\|_2^2/8\biggr) \\ &\leq \mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{H}_{\delta,a}} \delta(j)^2 \bigl(\mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}} - \mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| \geq a)\bigr) < -\|\delta\|_2^2/8\biggr) \\ &\leq \exp\biggl\{-\frac{\|\delta\|_2^4/64}{2\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\bigl(\delta(j)^2 \mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}}\bigr) + \|\delta\|_\infty^2 \|\delta\|_2^2/12}\biggr\} \\ &\leq \exp\biggl\{-\frac{\|\delta\|_2^4/64}{2C_5 s + \|\delta\|_\infty^2 \|\delta\|_2^2/12}\biggr\} \leq \exp\biggl\{-\frac{\|\delta\|_2^2}{12\|\delta\|_\infty^2}\biggr\} \leq \varepsilon/8. \end{aligned} \tag{44}$$

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 $\rho^2 \geq 384C_1^2 \log^{\frac{\alpha+2}{\alpha}}(8e/\varepsilon)$, we have $|\delta(j^*)| \geq 2C_1 \log^{1/\alpha}(8e/\varepsilon)$ and thus

$$\begin{aligned}&\mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)}\biggr) \leq \mathbb{P}_\theta\biggl(\delta(j^*)^2 \mathbb{1}_{\{|Y_{\tilde{t},2}(j^*)| \geq a\}} < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)}\biggr) \\ &\leq \mathbb{P}_\theta(|Y_{\tilde{t},2}(j^*)| < a) \leq \exp\{1 - (|\delta(j^*)|/(2C_1))^\alpha\} \leq \varepsilon/8. \end{aligned} \tag{45}$$

For the third and final term in (39), we have by Chebyshev's inequality that

$$\begin{aligned}&\mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \bigl(Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1\bigr) \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)}\biggr) \\ &\leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\Bigl(\bigl(Y_{\tilde{t},1}^2(j) - \delta(j)^2 - 1\bigr) \mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}}\Bigr)}{\|\delta\|_2^4/(576 \log^2(8/\varepsilon))} \leq \frac{\sum_{j \in \mathcal{H}_{\delta,a}} \mathrm{Var}_\theta\bigl(Y_{\tilde{t},1}^2(j)\bigr)}{\|\delta\|_2^4/(576 \log^2(8/\varepsilon))} \\ &\leq C_6 \log^2(8/\varepsilon) \biggl(\frac{s}{\rho^4} + \frac{1}{\rho^2}\biggr), \end{aligned} \tag{46}$$

where $C_6 \geq 1$ is a constant depending on α and K and the penultimate inequality follows from a similar argument to (31). Hence, when

$$\rho^2 \geq C_6 \varepsilon^{-1} \max\Bigl\{192(r + 2s) \log(8/\varepsilon), 64a^2 s, 192C_5 s, 384C_1^2 \log^{\frac{\alpha+2}{\alpha}}(8e/\varepsilon)\Bigr\},$$

we have by combining (39), (40), (44), (45) and (46) that

$$\mathbb{E}_\theta(1-\phi_{\mathcal{G},\text{sparse}}) \le \varepsilon/4 + C_6 \log^2(8/\varepsilon)\left(\frac{s}{\rho^4}+\frac{1}{\rho^2}\right) \le \varepsilon/2.$$

Finally, We consider the case that the mean change happens at $t_0 = 1$ instead. Recall that in this case we have $\tilde{t} = 1$. (39) remains true when $\rho^2 \ge 192(r_1+2s)\log(8/\varepsilon)$ if we redefine $\mathcal{J}_{\tilde{t}=1,a} := \{j \in [p] : |Y_1(j)| \ge a\}$. All three terms in (39) can be controlled in the same way as when $t_0 \ge 2$ and this completes the proof.

A.1.3 Proof of Theorem 5

We first prove the result for $\alpha \ge 4$.

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 $\mathbb{E}|E_i(j)|^\alpha \le K^\alpha$, for all $i \in [n]$ and $j \in [p]$ and Jensen's inequality, we deduce $\mathbb{E}E_i(j)^4 \le K^4$. We thus obtain, for all $i \le n/2$ and $j \in [p]$

$$\mathbb{E}_\theta Z_i^4(j) = \mathbb{E}_\theta\left[\frac{X_i(j)-X_{n-i}(j)}{\sqrt{2}}\right]^4 = \frac{\mathbb{E}_\theta\big[E_i(j)-E_{n-i}(j)\big]^4}{4} \le \frac{K^4+3}{2} =: C_1. \tag{47}$$

Then, by Chebyshev's inequality (or, alternatively, Lemma 20) and Lemma 19(a), with $r_t = C_2\sqrt{p}G_t$, we have for all $t \in \mathcal{T}$ and $g \in [G_t]$ that

$$\begin{aligned}\mathbb{P}_\theta\left(t\sum_{j=1}^p V_{t,g}(j) > r_t\right) &= \mathbb{P}_\theta\left(\sum_{j=1}^p\left(\overline{Z}^2_{t,g}(j) - \frac{G_t}{t}\right) > \frac{C_2\sqrt{p}G_t}{t}\right) \le \frac{t^2\sum_{j=1}^p \mathbb{E}_\theta \overline{Z}^4_{t,g}(j)}{C_2^2 p G_t^2} \\ &\le \frac{3pt^2 C_1 (G_t/t)^2}{C_2^2 p G_t^2} \le \frac{3C_1}{C_2^2} \le \varepsilon/36,\end{aligned} \tag{48}$$

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

$$\mathcal{B}_t := \left\{g \in [G_t] : \frac{t}{2}\sum_{j=1}^p V_{t,g}(j) > r_t\right\}.$$

By (48) and the multiplicative Chernoff bound (e.g. Mitzenmacher and Upfal, 2017, Corollary 4.9), we have for $t \in \mathcal{T}$

$$\begin{aligned}\mathbb{P}_\theta(A_t^{\mathrm{MoM}} > r_t) &\le \mathbb{P}_\theta(|\mathcal{B}_t| \ge G_t/2) = \mathbb{P}_\theta\left(|\mathcal{B}_t| \ge \frac{\varepsilon G_t}{36}\left(1+\left(\frac{18}{\varepsilon}-1\right)\right)\right) \\ &\le \exp\left\{-\frac{\varepsilon G_t}{36}\left(\frac{18}{\varepsilon}\log\left(\frac{18}{\varepsilon}\right) - \frac{18}{\varepsilon}+1\right)\right\} \le \exp\left\{-\frac{G_t}{2}\log(6/\varepsilon)\right\}.\end{aligned} \tag{49}$$

Thus, by (48), (49), the choices of G_t and Δ in (17) and a union bound, we conclude that

$$\begin{aligned}\mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} &\le \mathbb{P}_\theta\left(\sum_{j=1}^p V_{t=1,1}(j) > r_{t=1}\right) + \sum_{t\in\mathcal{T}:\,2\le t\le\Delta} \mathbb{P}_\theta\big(A_t^{\mathrm{MoM}} > r_t\big) + \sum_{t\in\mathcal{T}:\,t>\Delta} \mathbb{P}_\theta\big(A_t^{\mathrm{MoM}} > r_t\big) \\ &\le \varepsilon/36 + \sum_{t\in\mathcal{T}:\,2\le t\le\Delta} (6/\varepsilon)^{-t/2} + \sum_{t\in\mathcal{T}:\,t>\Delta} (6/\varepsilon)^{-\Delta/2}\end{aligned}$$

$$\leq \varepsilon/36 + \frac{(6/\varepsilon)^{-1}}{1-(6/\varepsilon)^{-1}} + \log_2(n/2)(6/\varepsilon)^{-\Delta/2} \leq \varepsilon/36 + \varepsilon/5 + \varepsilon/5 < \varepsilon/2, \tag{50}$$

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 (47). When $\rho^2 \geq 16C_2\sqrt{p}\Delta$, we have

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

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

$$\begin{aligned}
\mathbb{P}_\theta\left(\tilde{t}\sum_{j=1}^p V_{\tilde{t},g}(j) \leq r_{\tilde{t}}\right) &= \mathbb{P}_\theta\left(\sum_{j=1}^p \left(\left(\overline{Z}'_{\tilde{t},g}(j) + \frac{\mu_1(j) - \mu_2(j)}{\sqrt{2}}\right)^2 - \frac{G_{\tilde{t}}}{\tilde{t}}\right) \leq \frac{r_{\tilde{t}}}{\tilde{t}}\right) \\
&= \mathbb{P}_\theta\left(\sum_{j=1}^p \left(\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 - \frac{G_{\tilde{t}}}{\tilde{t}} + \sqrt{2}\big(\mu_1(j) - \mu_2(j)\big)\overline{Z}'_{\tilde{t},g}(j)\right) \leq \frac{r_{\tilde{t}}}{\tilde{t}} - \frac{\|\mu_1 - \mu_2\|_2^2}{2}\right) \\
&\leq \mathbb{P}_\theta\left(\sum_{j=1}^p \left(\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 - \frac{G_{\tilde{t}}}{\tilde{t}} + \sqrt{2}\big(\mu_1(j) - \mu_2(j)\big)\overline{Z}'_{\tilde{t},g}(j)\right) \leq -\frac{\rho^2}{16\tilde{t}} - \frac{\|\mu_1 - \mu_2\|_2^2}{4}\right).
\end{aligned} \tag{51}$$

By Chebyshev's inequality and Lemma 19(a), we obtain

$$\mathbb{P}_\theta\left(\sum_{j=1}^p \left(\left(\overline{Z}'_{\tilde{t},g}(j)\right)^2 - G_{\tilde{t}}/\tilde{t}\right) \leq -\frac{\rho^2}{16\tilde{t}}\right) \leq \frac{256(\tilde{t})^2\sum_{j=1}^p \mathbb{E}_\theta\left(\overline{Z}'_{\tilde{t},g}(j)\right)^4}{\rho^4} \leq \frac{768C_1pG_{\tilde{t}}^2}{\rho^4}, \tag{52}$$

and

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

Combining these with (51), as long as

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

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

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

If $\tilde{t} \ge 2$, then $G_{\tilde{t}} \ge 2$ and we use the same binomial tail bound argument as in (49) to conclude that

$$\mathbb{E}_\theta(1 - \phi_{\mathcal{P},\text{dense}}) \le \mathbb{P}_\theta\big(A^{\text{MoM}}_{\tilde{t}} \le r_{\tilde{t}}\big) \le \exp\bigg\{-\frac{\varepsilon G_{\tilde{t}}}{12}\bigg(\frac{6}{\varepsilon}\log\Big(\frac{6}{\varepsilon}\Big) - \frac{6}{\varepsilon} + 1\bigg)\bigg\} \le \Big(\frac{2}{\varepsilon}\Big)^{-1}.$$

This completes the proof for $\alpha \ge 4$. We now consider the case $\alpha < 4$. The proof is similar to above and we essentially replace Chebyshev's inequality wherever used by Lemma 20. We only highlight the difference for brevity.

Null term. Note that for all $t \in \mathcal{T}$ and $g \in [G_t]$, using Lemma 20 with $k = \alpha/2 < 2$ and $L = t/G_t$, we have with $r_t = C_2 p^{2/\alpha} G_t$ that

$$\mathbb{P}_\theta\bigg(t\sum_{j=1}^p V_{t,g}(j) > r_t\bigg) = \mathbb{P}_\theta\bigg(\frac{t}{G_t}\sum_{j=1}^p\Big(\overline{Z}^2_{t,g}(j) - \frac{G_t}{t}\Big) > C_2 p^{2/\alpha}\bigg) \le \frac{\varepsilon}{36}, \tag{53}$$

for $C_2 \ge C_{\alpha/2}(36/\varepsilon)^{2/\alpha}$, where $C_{\alpha/2} > 0$ is the constant depending on α and K from Lemma 20. By substituting (48) with (53) and following the rest of the argument in the above proof, we prove that $\mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} \le \epsilon/2$ for all $\theta \in \Theta_0(p, n)$.

Alternative term. For all $g \in [G_{\tilde{t}}]$, again using Lemma 20 with $k = \alpha/2 < 2$ and $L = \tilde{t}/G_{\tilde{t}}$, we have

$$\mathbb{P}_\theta\bigg(\sum_{j=1}^p\Big(\big(\overline{Z}'_{\tilde{t},g}(j)\big)^2 - G_{\tilde{t}}/\tilde{t}\Big) \le -\frac{\rho^2}{16\tilde{t}}\bigg) = \mathbb{P}_\theta\bigg(\frac{\tilde{t}}{G_{\tilde{t}}}\sum_{j=1}^p\Big(\big(\overline{Z}'_{\tilde{t},g}(j)\big)^2 - G_{\tilde{t}}/\tilde{t}\Big) \le -\frac{\rho^2}{16G_{\tilde{t}}}\bigg) \le \frac{\varepsilon}{24}, \tag{54}$$

for $\rho^2 \ge 24^{(2+\alpha)/\alpha} C_{\alpha/2}\varepsilon^{-2/\alpha}p^{2/\alpha}\Delta$, where $C_{\alpha/2}$ is, as above, a constant depending only on α and K. By substituting (52) with (54) and following the rest of argument in the above proof, we prove that as long as

$$\rho^2 \ge \max\bigg\{16C_2p^{2/\alpha}\Delta, 24^{(2+\alpha)/\alpha}C_{\alpha/2}\varepsilon^{-2/\alpha}p^{2/\alpha}\Delta, \frac{1536\Delta}{\varepsilon}\bigg\},$$

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

A.1.4 Proof of Proposition 7

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

$$\mathbb{E}_\theta \phi^{\text{MoM}}_{\mathcal{P},\text{sparse}} \le \mathbb{P}_\theta(A_{1,a} > r_1) + \sum_{t \in \mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\big(A^{\text{MoM}}_{t,a} > r_t\big). \tag{55}$$

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

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

Note that $A^{\mathrm{MoM}}_{t,a} = A^{\mathrm{MoM}}_{t,*,\mathcal{J}_{t,a}}$. Using the same technique as (33) in the proof of Theorem 3, we have

$$\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(A^{\mathrm{MoM}}_{t,a} > r_t) \le \sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t\in\mathcal{T}\setminus\{1\}} \sup_{J\subseteq[p]:|J|\le s} \mathbb{P}_\theta(A^{\mathrm{MoM}}_{t,*,J} > r_t), \tag{56}$$

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

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

Then, by Fuk–Nagaev inequality (Proposition 16), we have

$$\begin{aligned} q_{t,a} = \mathbb{P}_\theta(|Y_{t,2}(j)| \ge a) &\le 2\left(\frac{(\alpha+2)(K^\alpha 2^{\alpha/2} t/2)^{1/\alpha}}{\alpha a\sqrt{t/2}}\right)^\alpha + 2\exp\left\{-\frac{2a^2}{(\alpha+2)^2 e^\alpha}\right\}. \\ &\le \frac{K^\alpha}{(a/3)^\alpha t^{\alpha/2-1}} + \exp\left\{1 - \frac{a^2}{2\alpha^2 e^\alpha}\right\}, \end{aligned} \tag{57}$$

where we have used $\alpha \ge 4$ in the last inequality. Similar to (35), by a binomial tail bound, we have

$$\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) \le \sum_{t\in\mathcal{T}\setminus\{1\}} \left(\frac{epq_{t,a}}{s}\right)^s \le \left(\frac{2epK^\alpha}{s(a/3)^\alpha}\right)^s + \log_2(n)\left(\frac{2e^2p}{s}\right)^s \exp\left\{-\frac{sa^2}{2\alpha^2 e^\alpha}\right\}. \tag{58}$$

Thus, as long as we choose a to satisfy

$$a \ge C_1\big(\varepsilon^{-1}(p/s)^{1/\alpha} + s^{-1/2}\log^{1/2}(\varepsilon^{-1}\log(8n))\big) \tag{59}$$

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

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

Furthermore, By setting $r_t = C_2\sqrt{s}G_t$ with a suffcently large $C_2 > 0$ and $\Delta = 2^{4+\lceil \log_2\log\log(8n)\rceil}$ and by following the argument from (48) to (50), we can upper bound the second term in (56) at $\varepsilon/8$ as well. Finally, to control the first term in (55), by Proposition 17(b), whenever $r_1 \ge C_1' s(p/s)^{2/\alpha}$ for sufficiently large $C_1' > 0$, depending on α, K and ε, we have $\mathbb{P}_\theta(A_{1,a} > r_1) \le \varepsilon/4$. Hence, we conclude that $\mathbb{E}_\theta\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}} \le \varepsilon/2$ for all $\theta \in \Theta_0(p,n)$.

Alternative term. Recall the definitions of δ, $\mathcal{S}_\delta$ and $\mathcal{H}_{\delta,a}$ from the alternative term part of the proof of Theorem 3:

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

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 5. We first consider the case $t_0 \ge 2$, which implies $\tilde{t} \ge 2$. For $J \subseteq [p]$, we further denote

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

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

$$\begin{aligned}
&\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \biggl(\overline{Z}^2_{\tilde{t},g,1}(j) - \frac{(\mu_1(j) - \mu_2(j))^2}{2} - \frac{2G_{\tilde{t}}}{\tilde{t}} \biggr) - \sum_{j \in \mathcal{J}_{\tilde{t},a}} V_{\tilde{t},g,a}(j) \\
&= -\frac{\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} (\mu_1(j) - \mu_2(j))^2}{2} + \sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} V_{\tilde{t},g,a}(j) - \sum_{j \in \mathcal{J}_{\tilde{t},a}} V_{\tilde{t},g,a}(j) \\
&\leq -\frac{\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} (\mu_1(j) - \mu_2(j))^2}{2} + \frac{2G_{\tilde{t}} |\mathcal{J}_{\tilde{t},a}|}{\tilde{t}}.
\end{aligned}$$

Then, on the event $\{|\mathcal{J}_{\tilde{t},a}| \leq 2s\} \cap \Bigl\{ \sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 \geq \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)} \Bigr\}$, by Lemma 21, we deduce

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

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

$$\begin{aligned}
&\mathbb{E}_\theta(1 - \phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}) \leq \mathbb{P}_\theta\bigl(A^{\mathrm{MoM}}_{\tilde{t},a} \leq r_{\tilde{t}}\bigr) = \mathbb{P}_\theta\bigl(A^{\mathrm{MoM}}_{\tilde{t},*,\mathcal{J}_{\tilde{t},a}} \leq r_{\tilde{t}}\bigr) \\
&\leq \mathbb{P}_\theta\bigl(|\mathcal{J}_{\tilde{t},a}| > 2s\bigr) + \mathbb{P}_\theta\biggl(\sum_{j \in \mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \delta(j)^2 < \frac{\|\delta\|_2^2}{12 \log(8/\varepsilon)} \biggr) + \mathbb{P}_\theta\biggl(A^{\mathrm{MoM}'}_{\tilde{t},*,\mathcal{J}_{\tilde{t},a} \cap \mathcal{H}_{\delta,a}} \leq -\frac{\|\delta\|_2^2}{24 \log(8/\varepsilon)} \biggr).
\end{aligned} \tag{60}$$

We control the three terms respectively. The arguments below mirror those made in the proof of Theorem 3 between (40) and (46) and we will omit details in places where the same reasoning is used in the last proof. First, it remains true that

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

For all $j \in \mathcal{H}_{\delta,a}$, We have by (57) and the choice $a \geq \{4^{1+1/\alpha} K\} \vee \{3\alpha e^{\alpha/2}\}$ that

$$\mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) \leq \frac{K^\alpha}{((|\delta(j)| - a)/3)^\alpha} + \exp\biggl\{ 1 - \frac{(|\delta(j)| - a)^2}{2\alpha^2 e^\alpha} \biggr\} \leq 1/2,$$

and thus

$$\mathrm{Var}_\theta\bigl(\delta(j)^2 \mathbb{1}_{\{|Y_{\tilde{t},2}(j)| \geq a\}}\bigr) \leq \delta(j)^4 \mathbb{P}_\theta(|Y_{\tilde{t},2}(j)| < a) \leq (6K)^\alpha + 128\alpha^4 e^{2\alpha} =: C_3,$$

At this point, we consider

$$\rho^2 \geq C_4 \max\Bigl\{ 192 C_2 s \Delta \log^2(8/\varepsilon), 64a^2 s, 192 C_3 s, 3456 K^2 (16/\varepsilon)^{2/\alpha} \log(8/\varepsilon), 768 \alpha^2 e^\alpha \log^2(16e/\varepsilon) \Bigr\},$$

with some $C_4 > 0$. Then, by repeating the argument in (43), (44) and (45), as long as $C_4 \geq 1$, we obtain

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

We now bound the third and final term in (60). By Chebyshev's inequality, we deduce that for $g \in [G_t]$

$$\begin{aligned}
&\mathbb{P}\biggl(\frac{\tilde{t}}{2}\sum_{j\in\mathcal{J}_{\tilde{t},a}\cap\mathcal{H}_{\delta,a}}\Bigl(\overline{Z}_{\tilde{t},g,1}^2(j) - \frac{(\mu_1(j)-\mu_2(j))^2}{2} - \frac{2G_{\tilde{t}}}{\tilde{t}}\Bigr) \leq -\frac{\|\delta\|_2^2}{24\log(8/\varepsilon)}\biggr)\\
&\leq \frac{\sum_{j\in\mathcal{H}_{\delta,a}}\mathrm{Var}_\theta\Bigl(\overline{Z}_{\tilde{t},g,1}^2(j)\Bigr)}{\|\delta\|_2^4/(144\tilde{t}^2\log^2(8/\varepsilon))} \leq \frac{\sum_{j\in\mathcal{H}_{\delta,a}} 2\tilde{t}^2\mathrm{Var}_\theta\Bigl(\bigl(\overline{Z}'_{\tilde{t},g,1}(j)\bigr)^2\Bigr) + \sum_{j\in\mathcal{H}_{\delta,a}} 16\tilde{t}\delta(j)^2\mathrm{Var}_\theta\Bigl(\overline{Z}'_{\tilde{t},g,1}(j)\Bigr)}{\|\delta\|_2^4/(144\log^2(8/\varepsilon))}\\
&\leq C_5\log^2(8/\varepsilon)\Bigl(\frac{s\Delta^2}{\rho^4} + \frac{\Delta}{\rho^2}\Bigr) \leq \varepsilon/48,
\end{aligned}$$

when $C_4 \geq 1$ is sufficiently large. The third inequality above follows from (52) and C_5 is a constant depending only on α and K. If $\tilde{t} = 2$, then $G_{\tilde{t}} = 1$ and we immediately have

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

If $\tilde{t} > 2$, then $G_{\tilde{t}} \geq 2$ and we again use the binomial tail bound argument as in (49) to obtain

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

By (60), we conclude $\mathbb{E}_\theta(1 - \phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}) \leq \varepsilon/2$. Finally, for the case that the mean change happens at $t_0 = 1$ instead, similar to the last paragraph of the proof of Theorem 3, we can still control the three terms in (60) in the same way respectively when we redefine $\mathcal{J}_{\tilde{t}=1,a} := \{j \in [p] : |Y_1(j)| \geq a\}$ instead.

A.1.5 Proof of Proposition 8

We actually prove a more general result. Any mean estimator that satisfies the following condition can be used in place of $\hat{\mu}^{\mathrm{RSM}}_{n,s,\eta}(\cdot) = \hat{\mu}^{\mathrm{RSM}}_{n,s}(\cdot;\eta)$ introduced in Section 3.2 while Proposition 8 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 $n \geq C_1\bigl(s\log(ep/s) + \log(1/\eta)\bigr)$, then with probability at least $1 - \eta$, we have*

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

In particular, the robust sparse mean estimator that we use from Prasad et al. (2019) satisfies the condition above as shown in Corollary 11[2] therein.

[2]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 $C = (\sqrt{2}K)^\alpha$.

In the result 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 $t \in \mathcal{T}\backslash\{1\}$.

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

$$\begin{aligned}\mathbb{E}_\theta \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}} = \mathbb{P}_\theta(A_{1,a} > \tilde{r}_1) &+ \sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t) \\ &+ \sum_{t \in \tilde{\mathcal{T}}_2} \mathbb{P}_\theta(A^{\mathrm{RSM}}_t > r^{\mathrm{RSM}}_t) + \sum_{t \in \tilde{\mathcal{T}}_3} \mathbb{P}_\theta(A^{\mathrm{RSM}}_t > r^{\mathrm{RSM}}_t). \end{aligned} \tag{61}$$

For the first term, similar to the proof of Theorem 3, by Proposition 17(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 (61), we closely follow the arguments in the null term part of the proof of Theorem 3 and Proposition 7. By (33), we have

$$\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t) \leq \sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(|\mathcal{J}_{t,a}| > s) + \sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \sup_{J \subseteq [p] : |J| \leq s} \mathbb{P}_\theta\biggl(\sum_{j \in J} \bigl(Y^2_{t,1}(j) - 1\bigr) > \tilde{r}_t\biggr). \tag{62}$$

For the first term on the right hand side, by (58), we obtain

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

The choice of a in (22) 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 (48), we obtain

$$\mathbb{P}_\theta\biggl(\sum_{j \in J} \bigl(Y^2_{t,1}(j) - 1\bigr) > \tilde{r}_t\biggr) \leq \frac{\varepsilon}{16 \log_2(\tilde{\Delta}_1)},$$

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. We thus deduce that $\sum_{t \in \tilde{\mathcal{T}}_1 \backslash \{1\}} \mathbb{P}_\theta(A_{t,a} > \tilde{r}_t) \leq \varepsilon/8$.

Now, we control the third and fourth terms in (61). For $t \in \tilde{\mathcal{T}}_2 \cup \tilde{\mathcal{T}}_3$, we observe that

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

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

$$\mathbb{P}_\theta(A^{\mathrm{RSM}}_t > r^{\mathrm{RSM}}_t) = \mathbb{P}_\theta\Bigl(t \bigl\|\hat{\mu}^{\mathrm{RSM}}_{t,s,\eta_t}\bigr\|_2^2 > 2C_2\bigl(s \log(ep/s) + \log(1/\eta_t)\bigr)\Bigr) \leq \eta_t,$$

and therefore,

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

$$\leq 2\exp\Big\{s\log(ep/s) - \frac{\tilde{\Delta}_1}{C_1}\Big\} + \frac{\varepsilon \log_2(n/2)}{16\log 2n} < \varepsilon/4, \tag{63}$$

where we use Lemma 18 in the second inequality. Hence, we conclude that $\mathbb{E}_\theta \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}} \leq \varepsilon/2$ 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 $\mathbb{E}_\theta(1 - \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}) \leq \varepsilon/2$. Note that if $\rho^2 \geq C_5 v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$, 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

$$\begin{aligned}
&\mathbb{E}_\theta(1 - \phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}) = \mathbb{P}_\theta(A^{\mathrm{RSM}}_{\tilde{t},a} \leq r^{\mathrm{RSM}}_{\tilde{t}}) = \mathbb{P}_\theta\Big(\tilde{t}\big\|\hat{\mu}^{\mathrm{RSM}}_{\tilde{t},s,\eta_{\tilde{t}}}\big\|_2^2 \leq 2C_2\big(s\log(ep/s) + \log(1/\eta_{\tilde{t}})\big)\Big) \\
&\leq \mathbb{P}_\theta\bigg(\sqrt{\tilde{t}}\bigg|\Big\|\frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2 - \Big\|\hat{\mu}^{\mathrm{RSM}}_{\tilde{t},s,\eta_{\tilde{t}}} - \frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2\bigg| \leq \sqrt{2C_2}\big(\sqrt{s\log(ep/s)} + \sqrt{\log(1/\eta_{\tilde{t}})}\big)\bigg) \\
&\leq \mathbb{P}_\theta\bigg(\sqrt{\tilde{t}}\Big\|\hat{\mu}^{\mathrm{RSM}}_{\tilde{t},s,\eta_{\tilde{t}}} - \frac{\mu_1 - \mu_2}{\sqrt{2}}\Big\|_2 > \sqrt{C_2}\big(\sqrt{s\log(ep/s)} + \sqrt{\log(1/\eta_{\tilde{t}})}\big)\bigg) \leq \eta_{\tilde{t}} \\
&\leq \exp\Big\{s\log(ep/s) - \frac{\min(t, \tilde{\Delta}_2)}{C_1}\Big\} \leq \frac{\varepsilon}{16},
\end{aligned}$$

as desired.

A.1.6 Proof of Theorem 9

We first consider the statistical property of $\phi_{\mathcal{P},\mathrm{sparse}}$. By comparing the two rates $v^{\mathrm{U,MoM}}_{\mathcal{P},\mathrm{sparse}}$ and $v^{\mathrm{U}}_{\mathcal{P},\mathrm{sparse}}$, we note that the improvement offered by $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ over $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ only exists when

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

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

Now onto the computational complexity claim. For each $t \in \mathcal{T}$, computing the statistics $A^{\mathrm{MoM}}_{t,a}$ and $A_{t,a}$ in $\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse}}$ and $\phi^{\mathrm{RSM}}_{\mathcal{P},\mathrm{sparse}}$ 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 $\hat{\mu}_{t,s,\eta_t}^{\mathrm{RSM}}$. Note that we are using this only when $p \le \log^{\alpha-2}(\log(8n))$. For each fixed t, we claim that the computation/approximation of $\hat{\mu}_{t,s,\eta_t}^{\mathrm{RSM}}$ 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 $\hat{\mu}^{\mathrm{RSM}}$ for clarity.

1. Each evaluation of the function $\mathrm{1DRobust}(\cdot)$ (c.f. Prasad et al., 2019, Algorithm 2) of t data point requires time of order $t \log t \le n \log n$ (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 $|\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})| \le (6ep/s)^s \le (6ep)^p \le \exp(6ep^2) \le \exp(C_\alpha \log(n)) = n^{C_\alpha}$ for some constant $C_\alpha > 0$, depending only on α. Denote
$$g(\mu) := \max_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^\top \mu - \mathrm{1DRobust}(\{u^\top Z_i\}_{i=1}^t, \eta_t/(6ep/s)^s)|.$$
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 $\hat{\mu}^{\mathrm{RSM}}$ 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 $\binom{p}{s} \le n^{C_\alpha}$ minima.

4. Fix $\mathcal{U} \subseteq \mathbb{R}^p$ with $|\mathcal{U}| = s$. We solve the optimisation problem
$$\min_{\mu \in \mathbb{R}: \mu(j)=0\, \forall j \in \mathcal{U}^c} g(\mu)$$
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. Nesterov, 2003, Theorem 3.2.2) shows that $\big(\min_{k\in[K]} g\big(\mu_{\mathcal{U}}^{(k)}\big)\big) - g_{*,\mathcal{U}} \le \epsilon$ in $K \asymp 1/\epsilon^2$ steps, where we choose $\epsilon = \sqrt{s \log(ep/s) t^{-1}}$. The computational complexity is again at most polynomial in n. Denote $\tilde{\mu}_{\mathcal{U}}^{\mathrm{RSM}}$ to be the update that attains the best objective value in K iterations.

Write
$$\tilde{\mu}^{\mathrm{RSM}} := \operatorname*{argmin}_{\mu \in \{\tilde{\mu}_{\mathcal{U}}^{\mathrm{RSM}} :\, |\mathcal{U}|=s\}} g(\mu),$$
as our final estimator (an approximation of $\hat{\mu}^{\mathrm{RSM}}$). We have now shown that $\tilde{\mu}^{\mathrm{RSM}}$ can be obtained in time that is polynomial in n. Finally, we prove that $\tilde{\mu}^{\mathrm{RSM}}$ still satisfies Condition 1. Indeed, following the proof of Lemma 4 and Corollary 12 in Prasad et al. (2019), we have
$$\begin{aligned}
\|\tilde{\mu}^{\mathrm{RSM}} - \mu_Z\|_2 &\le \|\tilde{\mu}^{\mathrm{RSM}} - \hat{\mu}^{\mathrm{RSM}}\|_2 + \|\hat{\mu}^{\mathrm{RSM}} - \mu_Z\|_2 \le g\big(\tilde{\mu}^{\mathrm{RSM}}\big) + g\big(\hat{\mu}^{\mathrm{RSM}}\big) + g\big(\hat{\mu}^{\mathrm{RSM}}\big) + g(\mu_Z) \\
&\le g\big(\hat{\mu}^{\mathrm{RSM}}\big) + \epsilon + 2g\big(\hat{\mu}^{\mathrm{RSM}}\big) + g(\mu_Z) \le \varepsilon + 4g(\mu_Z) \\
&\le \sqrt{C}\bigg(\sqrt{\frac{s\log(ep/s)}{t}} + \sqrt{\frac{\log(1/\eta_t)}{t}}\bigg),
\end{aligned}$$
for some $C \ge 1$, where $\mu_Z = \mathbb{E} Z_1$.

A.2 Proof of the adaptation result in Section 4

Proof of Theorem 11. This proof is based on the proofs of Theorem 5, Propositions 7, 8. For brevity, we only highlight the main steps and differences.

Null term. By a union bound, (55) and (61), we have

$$
\begin{aligned}
&\mathbb{E}_\theta \phi_{\mathcal{P},\text{adaptive}} \\
&\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \mathbb{E}_\theta\left[\max_{s\in\mathcal{K}} \phi^{\text{MoM}}_{\mathcal{P},\text{sparse},s}\right]\mathbb{1}_{\{p>\log^{\alpha-2}(\log(8n))\}} + \mathbb{E}_\theta\left[\max_{s\in\mathcal{K}} \phi^{\text{RSM}}_{\mathcal{P},\text{sparse},s}\right]\mathbb{1}_{\{p\leq\log^{\alpha-2}(\log(8n))\}} \\
&\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \left(\mathbb{P}_\theta\left(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{r_{1,s}} > 1\right) + \sum_{s\in\mathcal{K}}\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\left(A^{\text{MoM}}_{t,a_s,s} > r_{t,s}\right)\right)\mathbb{1}_{\{p>\log^{\alpha-2}(\log(8n))\}} \\
&\quad + \left(\mathbb{P}_\theta\left(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{\tilde{r}_{1,s}} > 1\right) + \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})\right. \\
&\qquad\quad \left. + \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \mathbb{P}_\theta(A^{\text{RSM}}_{t,s} > r^{\text{RSM}}_{t,s}) + \sum_{s\in\mathcal{K}}\sum_{t\in\tilde{\mathcal{T}}_{3,s}} \mathbb{P}_\theta(A^{\text{RSM}}_{t,s} > r^{\text{RSM}}_{t,s})\right)\mathbb{1}_{\{p\leq\log^{\alpha-2}(\log(8n))\}} \\
&\leq \mathbb{E}_\theta \phi_{\mathcal{P},\text{dense}} + \mathbb{P}_\theta\left(\max_{s\in\mathcal{K}} \frac{A_{1,a_s,s}}{r_{1,s}\wedge\tilde{r}_{1,s}} > 1\right) + \sum_{s\in\mathcal{K}}\sum_{t\in\mathcal{T}\setminus\{1\}} \mathbb{P}_\theta\left(A^{\text{MoM}}_{t,a_s,s} > r_{t,s}\right) \\
&\quad + \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}}\left(\sum_{t\in\tilde{\mathcal{T}}_{2,s}} \mathbb{P}_\theta(A^{\text{RSM}}_{t,s} > r^{\text{RSM}}_{t,s}) + \sum_{t\in\tilde{\mathcal{T}}_{3,s}} \mathbb{P}_\theta(A^{\text{RSM}}_{t,s} > r^{\text{RSM}}_{t,s})\right),
\end{aligned}
\tag{64}
$$

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 (64) by $\varepsilon/10$.

Term 1. By closely following the null term part of the proof of Theorem 5, with a sufficiently large constant C_1, we deduce, similar to (50), 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 17(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 (56). The first term can be controlled via (58). A careful inspection reveals that the condition on a (same as a_s here) given in (59) with a possibly larger value of the leading constant can guarantee the control of both terms in (58) at $\varepsilon/(160s^{1/2})$. Bounding the second term in (56) required the argument from (48) to (50), 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 (48) (dimension being s). Correspondingly, the RHS of (49) now becomes

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

Thus, The second term in (56) can now be bounded instead by

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

Putting everything together, we conclude that

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

Term 4. We follow the null term part of the proof of Proposition 8. More specifically, this term can be split into two terms according to (62). Similar to the argument made for the second term above, with C_4 sufficiently large, the first term in (62) 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 Proposition 8. By Condition 1 and similar to (63), we can now bound

$$\begin{aligned}
&\sum_{s\in\mathcal{K}}\Bigg(\sum_{t\in\tilde{\mathcal{T}}_{2,s}}\mathbb{P}_\theta(A^{\mathrm{RSM}}_{t,s}>r^{\mathrm{RSM}}_{t,s})+\sum_{t\in\tilde{\mathcal{T}}_{3,s}}\mathbb{P}_\theta(A^{\mathrm{RSM}}_{t,s}>r^{\mathrm{RSM}}_{t,s})\Bigg)\\
&\le\sum_{s\in\mathcal{K}}\Bigg(\sum_{t\in\tilde{\mathcal{T}}_{2,s}}\exp\Big\{s\log(ep/s)-\frac{t}{C_6}\Big\}+\sum_{t\in\tilde{\mathcal{T}}_{3,s}}\exp\Big\{s\log(ep/s)-\frac{\tilde{\Delta}_{2,s}}{C_6}\Big\}\Bigg)\\
&\le\sum_{s\in\mathcal{K}}\Bigg(2\exp\Big\{s\log(ep/s)-\frac{\tilde{\Delta}_{1,s}}{C_6}\Big\}+\log_2(n/2)\frac{\varepsilon}{80s\log n}\Bigg)<\sum_{s\in\mathcal{K}}\frac{\varepsilon}{20s}\le\frac{\varepsilon}{10},
\end{aligned}$$

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) $p>\log^{\alpha-2}(\log(8n))$ 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 5, 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) $p>\log^{\alpha-2}(\log(8n))$ 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 $\mathbb{E}_\theta(1-\phi_{\mathcal{P},\mathrm{adaptive}})\le\mathbb{E}_\theta(1-\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse},\tilde{s}})$. Now, by carefully inspecting the alternative term part of the proof of Proposition 7, we can still deduce $\mathbb{E}_\theta(1-\phi^{\mathrm{MoM}}_{\mathcal{P},\mathrm{sparse},\tilde{s}})\le\varepsilon/2$ as long as ρ satisfies

$$\begin{aligned}
\rho^2&\ge C'\Big(s\big((p/s)^{2/\alpha}+\log\log(8n)\big)\Big)\ge\frac{C'}{2}\Big(\tilde{s}\big((p/\tilde{s})^{2/\alpha}+\log\log(8n)\big)\Big)\\
&\ge C''\max\Big\{\log^2(8/\varepsilon)\max_{t\in\mathcal{T}\setminus\{1\}}(r_{t,\tilde{s}}+2\tilde{s}G_t),(r_{1,\tilde{s}}+2\tilde{s})\log^2(8/\varepsilon),a^2_{\tilde{s}}\tilde{s}\Big\},
\end{aligned}\tag{65}$$

for sufficiently large C' and C'', where the final inequality in (65) remains true with our modified choice of $r_{t,s}$.

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

$$\begin{aligned}\rho^2 &\ge C'\big(s(p/s)^{2/\alpha} + \log\log(8n)\big) \ge \frac{C'}{2}\big(\tilde{s}(p/\tilde{s})^{2/\alpha} + \log\log(8n)\big)\\ &\ge C''\max\big\{(\tilde{r}_{t\neq 1,\tilde{s}} + 2\tilde{s})\log^2(8/\varepsilon), (\tilde{r}_{1,\tilde{s}} + 2\tilde{s})\log^2(8/\varepsilon), a_{\tilde{s}}^2\tilde{s}, \tilde{s}\log(ep/\tilde{s}) + \log\log(8n)\big\}, \end{aligned} \tag{66}$$

for sufficiently large C' and C'', where the final inequality in (66) remains true with our new choices of a_s and $\tilde{r}_{t,s}$.

The desired result then follows from Theorem 9 and the first part of its proof. $\square$

A.3 Proofs of lower bound results

In this section, we prove all lower bound results presented in the paper, namely Propositions 2, 4, 6 and 10. 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 claims below, as they immediately imply all the lower bound results in the paper.

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

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

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

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

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

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

$$\begin{aligned}\mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|E_i(j)|}{x}\bigg)^\alpha\bigg\}\bigg] &= \mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|E_i(j)|}{x}\bigg)^\alpha\bigg\}\mathbb{1}_{\{|E_i(j)|\ge 2\}}\bigg] + \mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|E_i(j)|}{x}\bigg)^\alpha\bigg\}\mathbb{1}_{\{|E_i(j)|<2\}}\bigg]\\ &\le \mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|E_i(j)|}{2}\bigg)^2\bigg\}\mathbb{1}_{\{|E_i(j)|\ge 2\}}\bigg] + \exp\big\{(2/x)^\alpha\big\}\\ &= \sqrt{2} - \mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|E_i(j)|}{2}\bigg)^2\bigg\}\mathbb{1}_{\{|E_i(j)|<2\}}\bigg] + \exp\big\{(2/x)^\alpha\big\}\\ &\le \sqrt{2} - \big(1 - \exp(-2)\big) + \exp\big\{(2/x)^\alpha\big\} < 2,\end{aligned}$$

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 $\alpha > 2$

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

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 Gao et al. (2020, Proposition 4.2). Note that this lower bound holds for all $1 \leq s \leq p$.

(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 Liu et al. (2021, Proposition 3).

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

$$\tilde{\omega} = \begin{cases} +1 & \text{w.p. } \frac{s}{4p}\big(1+\frac{\gamma^2 s}{2p}\big)^{-1} \\ -1 & \text{w.p. } \frac{s}{4p}\big(1+\frac{\gamma^2 s}{2p}\big)^{-1} \\ 0 & \text{otherwise,} \end{cases} \tag{67}$$

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

$$\operatorname{supp}(\bar{\nu}) \subseteq \big\{v \in \mathbb{R}^p : \|v\|_0 \leq s, \|v\|_2^2 \geq s\gamma^2/6\big\}. \tag{68}$$

We also have

$$-\nu(\mathcal{V}_s^c) = -\Big(\frac{1}{\nu(\mathcal{V}_s)} - 1\Big)\nu(\mathcal{V}_s) \leq \nu(A) - \bar{\nu}(A) = \nu(A\cap\mathcal{V}_s^c) - \Big(\frac{1}{\nu(\mathcal{V}_s)} - 1\Big)\nu(A\cap\mathcal{V}_s) \leq \nu(\mathcal{V}_s^c). \tag{69}$$

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 (69), we have

$$\mathrm{TV}(\mathbf{P}^*, \bar{\mathbf{P}}^*) \leq \mathrm{TV}(\nu,\bar{\nu}) = \sup_A|\nu(A) - \bar{\nu}(A)| \leq \nu(\mathcal{V}_s^c) = \mathbb{P}\big(\|\underline{\tilde{\omega}}\|_0 > s\big) + \mathbb{P}\big(\|\underline{\tilde{\omega}}\|_0 < s/6\big). \tag{70}$$

Suppose γ is chosen to satisfy $\gamma \le \sqrt{p/s}$. Then from (67), we deduce $\frac{s}{3p} \le \mathbb{P}(\tilde{\omega}(1) \neq 0) < \frac{s}{2p}$. By Chernoff bounds, we have

$$\begin{aligned} \mathbb{P}\big(\|\underline{\tilde{\omega}}\|_0 > s\big) &\le \frac{\mathbb{E}\big[e^{\|\underline{\tilde{\omega}}\|_0 \log 2}\big]}{e^{s\log 2}} \le \frac{\big(1+s/(2p)\big)^p}{e^{s\log 2}} \le e^{-s/6}, \\ \mathbb{P}\big(\|\underline{\tilde{\omega}}\|_0 < s/6\big) &\le \frac{\mathbb{E}\big[e^{-\|\underline{\tilde{\omega}}\|_0 \log 2}\big]}{e^{-(s\log 2)/6}} \le \frac{\big(1-s/(6p)\big)^p}{e^{-(s\log 2)/6}} \le e^{-s/20}. \end{aligned} \tag{71}$$

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:

$$\Xi^* \in \mathcal{G}_{\alpha,K} \text{ (resp. } \mathcal{P}_{\alpha,K}), \tag{72}$$

$$\tilde{\Pi} \in \mathcal{G}_{\alpha,K} \text{ (resp. } \mathcal{P}_{\alpha,K}), \tag{73}$$

$$H^2(P_\xi, P_{\tilde{\xi}}) \le \frac{1}{16p}, \tag{74}$$

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

$$\begin{aligned} \mathrm{TV}(\tilde{\mathbf{P}}^*, \mathbf{P}^*) &\le \mathrm{TV}(P_{0,\tilde{\Xi}^*}, P_{0,\Xi^*}) \le \mathrm{TV}\big(P_{\underline{\tilde{\xi}}}, P_{\underline{\xi}}\big) \le H\big(P_{\underline{\tilde{\xi}}}, P_{\underline{\xi}}\big) = \sqrt{2\big(1-(1-H^2(P_{\tilde{\xi}}, P_\xi)/2)^p\big)} \\ &\le \sqrt{pH^2(P_\xi, P_{\tilde{\xi}})} \le 1/4, \end{aligned} \tag{75}$$

where the penultimate inequality follows from the fact that $(1-x)^p \ge 1 - px$ for all $0 \le x \le 1$ and $p \ge 1$. Combining (68), (70), (71), and (75), when $s \ge 30$, for all $\rho^2 \le s\gamma^2/12$, we have

$$\begin{aligned} \mathcal{R}_{\mathcal{Q}}(\rho) &= \inf_{\phi\in\Phi}\left\{ \sup_{P_e\in\mathcal{Q}} \sup_{\theta\in\Theta_0(p,n)} \mathbb{E}_{\theta,P_e}\phi + \sup_{P_e\in\mathcal{Q}} \sup_{\theta\in\Theta(p,n,s,\rho)} \mathbb{E}_{\theta,P_e}(1-\phi)\right\} \\ &\ge 1 - \mathrm{TV}(P_{0,\tilde{\Pi}}, \bar{\mathbf{P}}^*) = 1 - \mathrm{TV}(\tilde{\mathbf{P}}^*, \bar{\mathbf{P}}^*) \ge 1 - \mathrm{TV}(\tilde{\mathbf{P}}^*, \mathbf{P}^*) - \mathrm{TV}(\mathbf{P}^*, \bar{\mathbf{P}}^*) \\ &\ge 3/4 - e^{-s/6} - e^{-s/20} \ge 1/2, \end{aligned}$$

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 (72), (73) and (74), and specify the corresponding choices of γ. Each construction corresponds to a rate given at the beginning of the proof.

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

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

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

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

Note that we have $\gamma \leq \sqrt{p/s}$. Now, to check (72) and (73), it suffices to only verify that $\mathbb{E}|\xi|^\alpha \leq K^\alpha$ and that $\mathbb{E}\bigl|\tilde{\xi} + \gamma\tilde{\omega}\bigr|^\alpha \leq K^\alpha$ respectively. Indeed, as $\alpha > 2$ and $K \geq \sqrt{2}$, we have

$$\begin{aligned}
\mathbb{E}|\xi|^\alpha &\leq 1 + t_0^\alpha \frac{1 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}}{t_0^2 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}} \leq 1 + t_0^\alpha \frac{\gamma^2 s/(2p)}{t_0^2 - 1} = 1 + \frac{\gamma^2 s t_0^{\alpha-2}}{p} \leq 1 + \frac{32^{\alpha-2}\gamma^\alpha s}{p} \\
&\leq 1 + \max\biggl\{K^\alpha - 1, 2^{\alpha/2 - 10}\biggr\} \leq K^\alpha,
\end{aligned}$$

and

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

We also verify (74):

$$\begin{aligned}
H^2(P_\xi, P_{\tilde{\xi}}) &= \left(1 - \sqrt{\frac{t_0^2 - 1}{t_0^2 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}}}\right)^2 + \frac{1 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}}{t_0^2 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}} \leq \frac{2\Bigl(1 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}\Bigr)}{t_0^2 - \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1}} \\
&\leq \frac{2\gamma^2 s}{p t_0^2} = \frac{60\gamma^2}{(32\gamma)^2 p} \leq \frac{1}{16p}.
\end{aligned}$$

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

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

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.

(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:

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

Observe that $\mathbb{E}\xi_{\text{aux}} = 0$ and $\sigma_{\text{aux}}^2 := \mathbb{E}\xi_{\text{aux}}^2 \in (1, 1.01)$. Now let ξ and $\tilde{\xi}$ be independent random variables such that $\xi \overset{d}{=} \sigma_{\text{aux}}^{-1}\xi_{\text{aux}}$ and $\tilde{\xi} \overset{d}{=} \bigl(1 + \frac{\gamma^2 s}{2p}\bigr)^{-1/2}\sigma_{\text{aux}}^{-1}\xi_{\text{aux}}$. Again, direct calculations show that both ξ and $\tilde{\xi} + \gamma\tilde{\omega}$ have mean 0 and variance 1. For condition (72), 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 $\gamma \le \sqrt{p/s}$. We verify (73) as follows:

$$\begin{aligned}\mathbb{E}\big|\tilde{\xi}+\gamma\tilde{\omega}\big|^{\alpha} &\le 1.1^{\alpha}+(1.1+\gamma)^{\alpha}\cdot\mathbb{P}(\tilde{\omega}\neq 0)\le 1.1^{\alpha}+\frac{(1.1+\gamma)^{\alpha}s}{2p}\le 1.1^{\alpha}+\frac{\max\{1.2,12\gamma\}^{\alpha}s}{2p}\\ &\le \max\Big\{1.1^{\alpha}+\frac{1.2^{\alpha}}{2},1.1^{\alpha}+\frac{1}{2}\Big\}\le 2^{\alpha/2}\le K^{\alpha}. \end{aligned}\tag{76}$$

Finally, by Ibragimov and Has' Minskii (2013, Theorem 7.6), we have when $s \le p^{\frac{\alpha-4}{2\alpha-4}}$

$$\begin{aligned}H^2(P_{\xi},P_{\tilde{\xi}}) &\le \frac{\Big(\sigma_{\mathrm{aux}}^{-1}-\big(1+\frac{\gamma^2 s}{2p}\big)^{-1/2}\sigma_{\mathrm{aux}}^{-1}\Big)^2}{4}\sup_{u\in\big[\big(1+\frac{\gamma^2 s}{2p}\big)^{-1/2}\sigma_{\mathrm{aux}}^{-1},\sigma_{\mathrm{aux}}^{-1}\big]}\frac{\int_{\mathrm{supp}(\xi_{\mathrm{aux}})}\big(f'_{\xi_{\mathrm{aux}}}(x)\big)^2/f_{\xi_{\mathrm{aux}}}(x)\,dx}{u^2}\\ &\le \frac{\Big(\big(1+\frac{\gamma^2 s}{2p}\big)^{1/2}-1\Big)^2}{4}\int_{\mathrm{supp}(\xi_{\mathrm{aux}})}\big(f'_{\xi_{\mathrm{aux}}}(x)\big)^2/f_{\xi_{\mathrm{aux}}}(x)\,dx\\ &\le \frac{\gamma^4 s^2}{64p^2}\cdot 4\Big(\int_0^{0.05}\frac{(-2000x)^2}{5-1000x^2}\,dx+\int_{0.05}^{0.1}\frac{(2000(x-0.1))^2}{1000(x-0.1)^2}\,dx\Big)\le\frac{25\gamma^4 s^2}{p^2}\le\frac{1}{16p},\end{aligned}\tag{77}$$

and this verifies (74). Therefore, under the noise distribution class $\mathcal{P}_{\alpha,K}^{\otimes}$ with $\alpha\ge 4$ and $K\ge\sqrt{2}$, whenever $30\le s\le p^{\frac{\alpha-4}{2\alpha-4}}$ and $\rho^2\le s(p/s)^{2/\alpha}/1728$, we have $\mathcal{R}_{\mathcal{P}}(\rho)\ge 1/2$.

(v). We work within the noise distribution class $\mathcal{G}_{\alpha,K}^{\otimes}$ with $0<\alpha\le 2$ and $K\ge 2^{1+2/\alpha}$. 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 $\log x \le \frac{2}{e\alpha}x^{\alpha/2}$ for all $x\ge e$, we can verify that $\gamma\le\sqrt{p/s}$. Again, for condition (72), since $|\xi|<1.1$ holds with probability one, we have $\Xi^*\in\mathcal{G}_{\alpha,K}$ for all $\alpha\ge 4$ and $K\ge 2^{1+2/\alpha}$, as $\exp\{(1.1/K)^{\alpha}\}\le e^{1/4}<2$. We now verify (73) using the technique in (76):

$$\begin{aligned}\mathbb{E}\bigg[\exp\bigg\{\bigg(\frac{|\tilde{\xi}+\gamma\tilde{\omega}|}{K}\bigg)^{\alpha}\bigg\}\bigg] &\le \exp\bigg\{\bigg(\frac{1.1}{K}\bigg)^{\alpha}\bigg\}+\frac{s}{2p}\exp\bigg\{\bigg(\frac{\max\{2,3\gamma\}}{K}\bigg)^{\alpha}\bigg\}\\ &\le e^{1/4}+\max\bigg\{e^{1/4}\frac{s}{2p},\Big(\frac{s}{2p}\Big)^{1-\frac{2\alpha}{8K^{\alpha}}}\bigg\}\\ &\le e^{1/4}+\max\big\{e^{1/4}/2,\sqrt{1/2}\big\}<2.\end{aligned}$$

We then follow (77) to verify (74) as well:

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

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

B 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 Vladimirova et al. (2020) and Kuchibhotla and Chakrabortty (2022, Section 2).

Definition 4 (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 5 (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 12 (Vladimirova et al., 2020, 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

$$\mathbb{E}\exp\big\{(|X|/K'')^\alpha\big\} \leq 2.$$

In other words, X is a sub-Weibull(α) random variable with $\|X\|_{\psi_\alpha} \leq K'' < \infty$.

Proposition 13 (Vladimirova et al., 2020, 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 15 below can be viewed as an extension of the Hanson–Wright inequality (Hanson and Wright, 1971).

Proposition 14 (Kuchibhotla and Chakrabortty, 2022, Theorem 3.1). *Let $\alpha > 0$ and $n \in \mathbb{N}$. 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 $i \in \mathbb{N}$ 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 15 (Götze et al., 2021, 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 $\|X_i\|_{\psi_\alpha} \le K$ for all $i \in \mathbb{N}$ and for some $K > 0$. Then, there exists a constant $C > 0$, depending only on α and K, such that for any $x \ge 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{1\le i,j\le n} a_{ij}X_iX_j - \sum_{i=1}^n a_{ii}\sigma_i^2\Big| \ge x\Big) \le \exp\big(1 - \eta_\alpha(x/C; A)\big),$$

where

$$\eta_\alpha(x; A) := \min\bigg\{\Big(\frac{x}{\|A\|_{\mathrm{F}}}\Big)^2, \frac{x}{\|A\|_2}, \Big(\frac{x}{\|A\|_{2\to\infty}}\Big)^{\frac{2\alpha}{2+\alpha}}, \Big(\frac{x}{\|A\|_{\max}}\Big)^{\frac{\alpha}{2}}\bigg\}.$$

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 Rio (2017).

Proposition 16 (Fuk, 1973; Nagaev, 1979)**.** *Let $X_1, \ldots, X_n$ be independent random variables, each having mean 0 and variance σ^2. Assume further that for some $q \ge 2$ and $C_q > 0$, we have for all $i \in [n]$*

$$\mathbb{E}[\{\max(X_i, 0)\}^q] \le C_q.$$

Then for any $x > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \ge x\Big) \le \bigg(\frac{(q+2)(nC_q)^{1/q}}{qx}\bigg)^q + \exp\bigg\{-\frac{2x^2}{n(q+2)^2e^q\sigma^2}\bigg\}.$$

Proposition 17. *Let $X_1, \ldots, X_p$ be independent random variables, each with mean zero and unit variance. Let $a \ge 0$ and $Z := \sum_{i=1}^p (X_i^2 - 1)\mathbb{1}_{\{|X_i|\ge a\}}$.*
(a) Let $\alpha > 0$, $K > 0$, $0 < \varepsilon \le 1$ and $1 \le s \le \sqrt{p}$. Assume that $X_1, \ldots, X_p$ are independent sub-Weibull random variables of order α, with $\|X_i\|_{\psi_\alpha} \le K$ for all $i \in [p]$. By setting

$$a \ge K \log^{1/\alpha}\Big(\frac{4ep}{\varepsilon s}\Big) \qquad \text{and} \qquad r = 2^{2/\alpha}K^2 s \log^{2/\alpha}\Big(\frac{4p}{\sqrt{\varepsilon} s}\Big),$$

we have $\mathbb{P}(Z > r) \le \varepsilon$.
(b) Let $\alpha \ge 2$, $K > 0$, $0 < \varepsilon \le 1$ and $1 \le s \le p$. Assume that $\mathbb{E}|X_i/K|^\alpha \le 1$ for all $i \in [p]$. By setting

$$a \ge K\Big(\frac{2ep}{\varepsilon s}\Big)^{1/\alpha} \qquad \text{and} \qquad r = \frac{\alpha - 2}{\alpha}K^2 s\Big(\frac{2ep}{\varepsilon s}\Big)^{2/\alpha},$$

we have $\mathbb{P}(Z > r) \le \varepsilon$.
(c) Assume the same conditions as in (b). Write $Z_s := \sum_{i=1}^p (X_i^2 - 1)\mathbb{1}_{\{|X_i|\ge a_s\}}$ to make the dependence on s explicit. By choosing the same a_s and r_s as in (b), we have

$$\mathbb{P}\big(\max_{s\in[p]} Z_s/r_s > 1\big) \le 2\varepsilon.$$

Proof. We denote the order statistics of $|X_1|, \ldots, |X_p|$ as $|X|_{(1)} \le \ldots \le |X|_{(p)}$. For $x \ge 0$, we write $q_x := \min_{i\in[p]} \mathbb{P}(|X_i| \ge x)$ and $\mathcal{J}_x := \{i \in [p] : |X_i| \ge x\}$.
(a) Note that $q_x \le 2\exp\{-(x/K)^\alpha\}$ by Proposition 12(a). Since $s \le \sqrt{p}$, we observe that

$$\sum_{j=1}^s K^2 \log^{2/\alpha}\Big(\frac{4ep}{\varepsilon j}\Big) \le \sum_{j=1}^s K^2 \log^{2/\alpha}\Big(\frac{4ep^2}{\varepsilon s^2}\Big) \le 2^{2/\alpha}K^2 s \log^{2/\alpha}\Big(\frac{4p}{\sqrt{\varepsilon} s}\Big) = r.$$

Then, by a union bound and a binomial tail bound, we have

$$\begin{aligned}
\mathbb{P}(Z > r) &\le \mathbb{P}(|\mathcal{J}_a| > s) + \mathbb{P}\Biggl(\sum_{j=1}^{s}\bigl(|X|^2_{(p-j+1)} - 1\bigr) > r\Biggr) \\
&\le \mathbb{P}(|\mathcal{J}_a| > s) + \sum_{j=1}^{s}\mathbb{P}\biggl(|X|_{(p-j+1)} > K\log^{1/\alpha}\Bigl(\frac{4ep}{\varepsilon j}\Bigr)\biggr) \\
&\le \Bigl(\frac{ep}{s}\Bigr)^{s}\biggl\{2\exp\Bigl\{-\Bigl(\frac{a}{K}\Bigr)^{\alpha}\Bigr\}\biggr\}^{s} + \sum_{j=1}^{s}\Bigl(\frac{ep}{j}\Bigr)^{j}\biggl\{2\exp\biggl\{-\biggl(\frac{K\log^{1/\alpha}\bigl(\frac{4ep}{\varepsilon j}\bigr)}{K}\biggr)^{\alpha}\biggr\}\biggr\}^{j} \\
&\le (\varepsilon/2)^{s} + \sum_{j=1}^{s}(\varepsilon/2)^{j} \le \varepsilon.
\end{aligned}$$

(b) Note that $q_x \le (K/a)^{\alpha}$ by Chebyshev's inequality. We now observe that

$$\sum_{j=1}^{s} K^2\Bigl(\frac{2ep}{\varepsilon j}\Bigr)^{2/\alpha} \le K^2\Bigl(\frac{2ep}{\varepsilon}\Bigr)^{2/\alpha}\biggl\{1 + \int_1^s x^{-2/\alpha}\,dx\biggr\} \le \frac{\alpha-2}{\alpha}K^2 s\Bigl(\frac{2ep}{\varepsilon s}\Bigr)^{2/\alpha} = r$$

The rest then follows from the proof for part (a).
(c) By a union bound and the proof for the previous parts, we have

$$\begin{aligned}
\mathbb{P}\bigl(\max_{s\in[p]} Z_s/r_s > 1\bigr) &\le \Biggl(\sum_{s=1}^{p}\mathbb{P}(|\mathcal{J}_{a(s)}| > s)\Biggr) + \mathbb{P}\Biggl(\max_{s\in[p]}\frac{\sum_{j=1}^{s}\bigl(|X|^2_{(p-j+1)} - 1\bigr)}{r_s} > 1\Biggr) \\
&\le \sum_{s=1}^{p}(\varepsilon/2)^{s} + \sum_{j=1}^{p}\mathbb{P}\biggl(|X|_{(p-j+1)} > K\Bigl(\frac{2ep}{\varepsilon j}\Bigr)^{2/\alpha}\biggr) \\
&\le \sum_{s=1}^{p}(\varepsilon/2)^{s} + \sum_{j=1}^{p}(\varepsilon/2)^{j} \le 2\varepsilon.
\end{aligned}$$

□

Lemma 18. *Let $\gamma > 0$. Then, for all $x \ge (2^{\gamma} - 1)^{-1/\gamma}$ we have*

$$\sum_{i=0}^{\infty}\exp\bigl\{-(x2^{i})^{\gamma}\bigr\} \le 2\exp(-x^{\gamma}).$$

Proof. By the convexity of $y \mapsto 2^{\gamma y}$, we have that $2^{(i+1)\gamma} - 2^{i\gamma} \ge 2^{i\gamma} - 2^{(i-1)\gamma}$ and thus

$$2^{i\gamma} = 1 + \sum_{j=1}^{i}(2^{j\gamma} - 2^{(j-1)\gamma}) \ge 1 + i(2^{\gamma} - 1).$$

for all $i \in \mathbb{N}$. Denote $\tilde{x} := \exp(x^{\gamma})$. We hence deduce that when $\tilde{x} > 2^{\frac{1}{2^{\gamma}-1}}$,

$$\sum_{i=0}^{\infty}\exp\bigl\{-(x2^{i})^{\gamma}\bigr\} = \sum_{i=0}^{\infty}\tilde{x}^{-2^{i\gamma}} \le \sum_{i=0}^{\infty}\tilde{x}^{-1-i(2^{\gamma}-1)} = \frac{1}{\tilde{x}\bigl(1 - \tilde{x}^{-(2^{\gamma}-1)}\bigr)} \le 2\tilde{x}^{-1}.$$

□

Lemma 19. *Let $Z_1, \ldots, Z_n$ be independent mean zero random variables.*
(a) Assume that there exists $C > 0$ such that $\mathbb{E}Z_i^4 \le C$ for all $i \in [n]$. Then for any $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, we have

$$\mathbb{E}\Big[\Big(\sum_{i=1}^n v_i Z_i\Big)^4\Big] \le 3C\|v\|_2^4.$$

(b) Assume that there exists $C > 0$ such that $\mathbb{E}\big(|Z_i|^{2k}\big) \le C$ for some $k \ge 1$. Then, there exists a constant $C_k > 0$, depending only on k and C such that

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n Z_i/n\Big|^{2k}\Big] \le C_k n^{-k}.$$

Proof. (a) Since $\mathbb{E}Z_i^0 = 1$ and $\mathbb{E}Z_i^1 = 0$ for all $i \in [n]$, we have

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{i=1}^n v_i Z_i\Big)^4\Big] &= \sum_{i=1}^n v_i^4 \mathbb{E}Z_i^4 + \sum_{1\le i<j\le n} 6v_i^2 v_j^2 \mathbb{E}(Z_i^2 Z_j^2) \le C\Big(\sum_{i=1}^n v_i^4 + \sum_{1\le i<j\le n} 6v_i^2 v_j^2\Big) \\ &\le 3C\Big(\sum_{i=1}^n v_i^2\Big)^2,\end{aligned}$$

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. Beiglböck and Siorpaes, 2015, Theorem 1.1), we have for any $k \ge 1$,

$$\mathbb{E}\big[|S_n|^{2k}\big] \le \mathbb{E}\Big[\Big(\max_{j\le n}|S_j|\Big)^{2k}\Big] \le C_{k,1}\mathbb{E}\Big[\big([S]_n\big)^k\Big],$$

for some constant $C_{k,1} > 0$, depending only on k. Thus, we have

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n Z_i/n\Big|^{2k}\Big] \le \frac{C_{k,1}}{n^{2k}}\mathbb{E}\Big[\Big(\sum_{i=1}^n Z_i^2\Big)^k\Big] \le \frac{C_{k,1}}{n^k}\mathbb{E}\Big[\frac{\sum_{i=1}^n |Z_i|^{2k}}{n}\Big] \le \frac{CC_{k,1}}{n^k},$$

where we have used Jensen's inequality in the second inequality. □

Lemma 20. *Let $k \ge 1$ 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 $\mathbb{E}\big[|V_i(j)|^{2k}\big] \le C$ 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*

$$\mathbb{P}\left(\Big|\sum_{j=1}^p L\big(\overline{V}^2(j) - 1/L\big)\Big| > C_k \frac{p^{\frac{1}{2}\vee\frac{1}{k}}}{\delta^{1/k}}\right) \le \delta$$

for some constant $C_k > 0$, depending only on C and k.

Proof. We first prove the result for $1 \le k \le 2$. Note that for any $\eta > 0$

$$\mathbb{P}\Big(\Big|\sum_{j=1}^p L\big(\overline{V}^2(j) - 1/L\big)\Big| > \eta\Big) \le p\mathbb{P}\Big(\big|L\big(\overline{V}^2(1) - 1/L\big)\big| > \eta\Big)$$

$$+\mathbb{P}\Bigg(\Bigg|\sum_{j=1}^{p} L\big(\overline{V}^2(j)-1/L\big)\mathbb{1}\Big\{\big|L\big(\overline{V}^2(j)-1/L\big)\big|\le\eta\Big\}\Bigg|>\eta\Bigg). \tag{78}$$

We control the two terms separately. For the first term, we have

$$\mathbb{P}\Big(\big|L\big(\overline{V}^2(1)-1/L\big)\big|>\eta\Big)\le\frac{\mathbb{E}\Big[\big|L\overline{V}^2(1)-1\big|^k\Big]}{\eta^k}\le\frac{2^{k-1}\Big(\mathbb{E}\big|L\overline{V}^2(1)\big|^k+1\Big)}{\eta^k}\le\frac{2^{k-1}(C_{0,k}+1)}{\eta^k}, \tag{79}$$

where the three inequalities follow, respectively, from Markov's inequality, Jensen's inequality and Lemma 19(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 (78), we have

$$\begin{aligned}
&\mathbb{P}\Bigg(\Bigg|\sum_{j=1}^{p} L\big(\overline{V}^2(j)-1/L\big)\mathbb{1}\Big\{\big|L\big(\overline{V}^2(j)-1/L\big)\big|\le\eta\Big\}\Bigg|>\eta\Bigg)\\
&\le\frac{1}{\eta^2}\Bigg\{p\mathbb{E}\Big[\big(L\overline{V}^2(1)-1\big)^2\mathbb{1}\Big\{\big|L\big(\overline{V}^2(1)-1/L\big)\big|\le\eta\Big\}\Big]\\
&\qquad+p^2\bigg(\mathbb{E}\Big[L\big(\overline{V}^2(1)-1/L\big)\mathbb{1}\Big\{\big|L\big(\overline{V}^2(1)-1/L\big)\big|>\eta\Big\}\Big]\bigg)^2\Bigg\}\\
&\le\frac{p}{\eta^2}\mathbb{E}\Big[\big|L\overline{V}^2(1)-1\big|^k\eta^{2-k}\Big]+\frac{p^2}{\eta^2}\Big\{\mathbb{E}\Big[\big|L\overline{V}^2(1)-1\big|^k\Big]\Big\}^{2/k}\Big\{\mathbb{P}\Big(\big|L\big(\overline{V}^2(1)-1/L\big)\big|>\eta\Big)\Big\}^{2(k-1)/k}\\
&\le\frac{C_{1,k}p}{\eta^k}+\frac{C_{1,k}^{2/k}p^2}{\eta^2}\left(\frac{C_{1,k}}{\eta^k}\right)^{2(k-1)/k},
\end{aligned}\tag{80}$$

where we have used Markov's inequality for the first inequality, Hölder's inequality for the second one and (79) for the last one. Combining (78), (79) and (80), we have

$$\mathbb{P}\Bigg(\Bigg|\sum_{j=1}^{p} L\big(\overline{V}^2(j)-1/L\big)\Bigg|>\eta\Bigg)\le\frac{2C_{1,k}p}{\eta^k}+\frac{C_{1,k}^2p^2}{\eta^{2k}}.$$

Note that if $C_{1,k}p/\eta^k>1$, the bound above holds trivially. Therefore we obtain

$$\mathbb{P}\Bigg(\sum_{j=1}^{p} L(\overline{V}^2(j)-1/L)>\eta\Bigg)\le\frac{3C_{1,k}p}{\eta^k},$$

for any $\eta>0$, which is equivalent to the claimed bound.

For $k>2$, by Markov's inequality, (79) and Lemma 19(b), there exists a constant $C_{2,k}>0$, depending only on k and C such that

$$\mathbb{P}\Big(\big|L\big(\overline{V}^2(1)-1/L\big)\big|>\eta\Big)\le\frac{\mathbb{E}\Big[\Big|\sum_{j=1}^{p}L\big(\overline{V}^2(j)-1/L\big)\Big|^k\Big]}{\eta^k}\le\frac{C_{2,k}p^{k/2}}{\eta^k},$$

which proves the desired result.

□

Lemma 21. *Let $n \in \mathbb{N}$ and $c \in \mathbb{R}$. Let $a_1, \dots, a_n, b_1, \dots, b_n$ be $2n$ real numbers. Suppose that $a_i - b_i \le c$ for all $i \in [n]$. Then*

$$\operatorname{median}(a_1, \dots, a_n) - \operatorname{median}(b_1, \dots, b_n) \le c.$$

Proof. We sort the two arrays respectively and obtain $a_{(1)} \le \dots \le a_{(n)}$ and $b_{(1)} \le \dots \le b_{(n)}$. We show that $a_{(i)} - b_{(i)} \le c$ for all $i \in [n]$. Indeed, there exists a set $\mathcal{I}_i \subseteq [n]$ with $|\mathcal{I}_i| \ge i$ such that

$$b_{(i)} = \max\{b_j : j \in \mathcal{I}_i\} \ge \max\{a_j - c : j \in \mathcal{I}_i\} \ge a_{(i)} - c.$$

The desired results follows by observing that the median is a convex combination of the order statistics. $\square$