Bernoulli **26**(1), 2020, 694–727
https://doi.org/10.3150/19-BEJ1149

Robust modifications of U-statistics and applications to covariance estimation problems

STANISLAV MINSKER[1] and XIAOHAN WEI[2]

[1]*Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA. E-mail: minsker@usc.edu*
[2]*Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA. E-mail: xiaohanw@usc.edu*

Let Y be a d-dimensional random vector with unknown mean μ and covariance matrix Σ. This paper is motivated by the problem of designing an estimator of Σ that admits exponential deviation bounds in the operator norm under minimal assumptions on the underlying distribution, such as existence of only 4th moments of the coordinates of Y. To address this problem, we propose robust modifications of the operator-valued U-statistics, obtain non-asymptotic guarantees for their performance, and demonstrate the implications of these results to the covariance estimation problem under various structural assumptions.

Keywords: covariance estimation; heavy tails; robust estimators; U-statistics

1. Introduction

In mathematical statistics, it is common to assume that data satisfy an underlying model along with a set of assumptions on this model – for example, that the sequence of vector-valued observations is i.i.d. and has multivariate normal distribution. Since real-world data typically do not fit the model or satisfy the assumptions exactly (e.g., due to outliers and noise), reducing the number and strictness of the assumptions helps to reduce the gap between the "mathematical" world and the "real" world. The concept of robustness occupies one the central roles in understanding this gap. One of the viable ways to model noisy data and outliers is to assume that the observations are generated by a heavy-tailed distribution, and this is precisely the approach that we follow in this work.

Robust M-estimators introduced by P. Huber [22] constitute a powerful method in the toolbox for the analysis of heavy-tailed data. Huber noted that "it is an empirical fact that the best [outlier] rejection procedures do not quite reach the performance of the best robust procedures." His conclusion remains valid in today's age of high-dimensional data that pose new challenging questions and demand novel methods.

The goal of this work is to introduce robust modifications for the class of operator-valued U-statistics, which naturally appear in the problems related to the estimation of covariance matrices. Statistical estimation in the presence of outliers and heavy-tailed data has recently attracted attention of the research community, and the literature on the topic covers the wide range of topics. A comprehensive review is beyond the scope of this section, so we mention only few notable

1350-7265

contributions. Several popular approaches to robust covariance estimation and robust principal component analysis are discussed in [7,24,37], including the Minimum Covariance Determinant (MCD) estimator and the Minimum Volume Ellipsoid estimator (MVE). Maronna's [32] and Tyler's [39,42] M-estimators are other well-known alternatives. Rigorous results for these estimators are available only for special families of distributions, such as elliptically symmetric distributions. Robust estimators based on Kendall's tau have been recently studied in [19,41], again for the class of elliptically symmetric distributions and their generalizations.

The papers [10,11,18] discuss robust covariance estimation for heavy-tailed distributions and are all based on the ideas originating in the work [9] that provided detailed non-asymptotic analysis of robust M-estimators of the univariate mean. The present paper can be seen as a direct extension of these ideas to the case of matrix-valued U-statistics, and continues the line of work initiated in [16] and [34]; the main advantage of the techniques proposed here is that they result in estimators that can be computed efficiently, and cover scenarios beyond covariance estimation problem. Recent advances in this direction include the works [15] and [35] that present new results on robust covariance estimation; see Remark 4.1 for more details.

Finally, let us mention the paper [25] that investigates robust analogues of U-statistics obtained via the median-of-means technique [2,13,29,36]. We include a more detailed discussion and comparison with the methods of this work in Section 3 below.

The rest of the paper is organizes as follows. Section 2 explains the main notation and background material. Section 3 introduces the main results. Implications for covariance estimation problem and its versions are outlined in Section 4. Finally, the proofs of the main results are contained in Section 5.

2. Preliminaries

In this section, we introduce the main notation and recall useful facts that we rely on in the subsequent exposition.

2.1. Definitions and notation

Given $A \in \mathbb{C}^{d_1\times d_2}$, let $A^* \in \mathbb{C}^{d_2\times d_1}$ be the Hermitian adjoint of A. The set of all $d \times d$ self-adjoint matrices will be denoted by $\mathbb{H}^d$. For a self-adjoint matrix A, we will write $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ for the largest and smallest eigenvalues of A. Hadamard (entry-wise) product of matrices $A, B \in \mathbb{C}^{d_1\times d_2}$ will be denoted by $A_1 \odot A_2$. Next, we will introduce the matrix norms used in the paper.

Everywhere below, $\|\cdot\|$ stands for the operator norm $\|A\| := \sqrt{\lambda_{\max}(A^*A)}$. If $d_1 = d_2 = d$, we denote by $\operatorname{tr} A$ the trace of A. Next, for $A \in \mathbb{C}^{d_1\times d_2}$, the nuclear norm $\|\cdot\|_1$ is defined as $\|A\|_1 = \operatorname{tr}(\sqrt{A^*A})$, where $\sqrt{A^*A}$ is a nonnegative definite matrix such that $(\sqrt{A^*A})^2 = A^*A$. The Frobenius (or Hilbert–Schmidt) norm is $\|A\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(A^*A)}$, and the associated inner product is $\langle A_1, A_2\rangle = \operatorname{tr}(A_1^* A_2)$. Finally, define $\|A\|_{\max} := \sup_{i,j} |A_{i,j}|$. For a vector $Y \in \mathbb{R}^d$, $\|Y\|_2$ stands for the usual Euclidean norm of Y.

Given two self-adjoint matrices A and B, we will write $A \succeq B$ (or $A \succ B$) iff $A - B$ is nonnegative (or positive) definite.

For a random matrix $Y \in \mathbb{C}^{d_1 \times d_2}$ with $\mathbb{E}\|Y\| < \infty$, the expectation $\mathbb{E}Y$ denotes a $d_1 \times d_2$ matrix such that $(\mathbb{E}Y)_{i,j} = \mathbb{E}Y_{i,j}$.

For $a, b \in \mathbb{R}$, set $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. Finally, recall the definition of the function of a matrix-valued argument.

Definition 2.1. Given a real-valued function f defined on an interval $\mathbb{T} \subseteq \mathbb{R}$ and a self-adjoint $A \in \mathbb{H}^d$ with the eigenvalue decomposition $A = U\Lambda U^*$ such that $\lambda_j(A) \in \mathbb{T}$, $j = 1, \dots, d$, define $f(A)$ as $f(A) = Uf(\Lambda)U^*$, where

$$f(\Lambda) = f\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}\right) = \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{pmatrix}.$$

Finally, we introduce the Hermitian dilation which allows to reduce the problems involving general rectangular matrices to the case of Hermitian matrices.

Definition 2.2. Given the rectangular matrix $A \in \mathbb{C}^{d_1 \times d_2}$, the Hermitian dilation $\mathcal{D} : \mathbb{C}^{d_1 \times d_2} \mapsto \mathbb{C}^{(d_1+d_2)\times(d_1+d_2)}$ is defined as

$$\mathcal{D}(A) = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}. \tag{2.1}$$

Since $\mathcal{D}(A)^2 = \begin{pmatrix} AA^* & 0 \\ 0 & A^*A \end{pmatrix}$, it is easy to see that $\|\mathcal{D}(A)\| = \|A\|$.

2.2. U-statistics

Consider a sequence of i.i.d. random variables $X_1, \dots, X_n$ $(n \geq 2)$ taking values in a measurable space $(\mathcal{S}, \mathcal{B})$, and let P be the distribution of X_1. Assume that $H : \mathcal{S}^m \to \mathbb{H}^d$ $(2 \leq m \leq n)$ is a $\mathcal{S}^m$-measurable permutation-symmetric kernel, meaning that $H(x_1, \dots, x_m) = H(x_{\pi_1}, \dots, x_{\pi_m})$ for any $(x_1, \dots, x_m) \in \mathcal{S}^m$ and any permutation π. The U-statistic with kernel H is defined as [20]

$$U_n := \frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} H(X_{i_1}, \dots, X_{i_m}), \tag{2.2}$$

where $I_n^m := \{(i_1, \dots, i_m) : 1 \leq i_j \leq n, , i_j \neq i_k \text{ if } j \neq k\}$; clearly, it is an unbiased estimator of $\mathbb{E}H(X_1, \dots, X_m)$. Throughout this paper, we will impose a mild assumption stating that $\mathbb{E}\|H(X_1, \dots, X_m)^2\| < \infty$.

One of the key questions in statistical applications is to understand the concentration of a given estimator around the unknown parameter of interest. Majority of existing results for U-statistics assume that the kernel H is bounded [4], or that $\|H(X_1, \dots, X_m)\|$ has sub-Gaussian tails [17]. However, in the case when only the moments of low orders of $\|H(X_1, \dots, X_m)\|$ are finite, deviations of the random variable

$$\left\| H(X_1, \dots, X_m) - \mathbb{E} H(X_1, \dots, X_m) \right\|$$

do not satisfy exponential concentration inequalities. At the same time, as we show in this paper, it is possible to construct "robust modifications" of U_n for which sub-Gaussian type deviation results hold.

In the remainder of this section, we recall several useful facts about U-statistics. The projection operator $\pi_{m,k} (k \le m)$ is defined as

$$\pi_{m,k} H(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}) := (\delta_{\mathbf{x}_{i_1}} - P) \dots (\delta_{\mathbf{x}_{i_k}} - P) P^{m-k} H,$$

where

$$Q^m H := \int \dots \int H(\mathbf{y}_1, \dots, \mathbf{y}_m) \, dQ(\mathbf{y}_1) \dots dQ(\mathbf{y}_m),$$

for any probability measure Q in $(\mathcal{S}, \mathcal{B})$, and δ_x is a Dirac measure concentrated at $x \in \mathcal{S}$. For example, $\pi_{m,1} H(x) = \mathbb{E}[H(X_1, \dots, X_m) | X_1 = x] - \mathbb{E} H(X_1, \dots, X_m)$.

Definition 2.3. An $\mathcal{S}^m$-measurable function $F : \mathcal{S}^m \to \mathbb{H}^d$ is P-degenerate of order r $(1 \le r < m)$, if

$$\mathbb{E} F(\mathbf{x}_1, \dots, \mathbf{x}_r, X_{r+1}, \dots, X_m) = 0, \quad \forall \mathbf{x}_1, \dots, \mathbf{x}_r \in \mathcal{S},$$

and $\mathbb{E} F(\mathbf{x}_1, \dots, \mathbf{x}_r, \mathbf{x}_{r+1}, X_{r+2}, \dots, X_m)$ is not a constant function. Otherwise, F is non-degenerate.

The following result is commonly referred to as Hoeffding's decomposition; see [14] for the details.

Proposition 2.1. *The following equality holds almost surely*:

$$U_n = \sum_{k=0}^{m} \binom{m}{k} V_n(\pi_{m,k} H),$$

where

$$V_n(\pi_{m,k} H) = \frac{(n-k)!}{n!} \sum_{(i_1, \dots, i_k) \in I_n^k} \pi_{m,k} H(X_{i_1}, \dots, X_{i_k}).$$

For instance, the first order term $(k = 1)$ in the decomposition is

$$m V_n(\pi_{m,1} H) = \frac{m}{n} \sum_{j=1}^{n} \pi_{m,1} H(X_j).$$

It is well known that

$$\mathbb{E}\big(U_n - \mathbb{E}H(X_1,\dots,X_m)\big)^2 = \binom{n}{m}^{-1}\sum_{k=1}^{m}\binom{m}{k}\binom{n-m}{m-k}\Sigma_k^2,$$

where $\Sigma_k^2 = \mathbb{E}(\pi_{m,k}H(X_1,X_2,\dots,X_k))^2$, $k=1,\dots,m$. As n gets large, the first term in the sum above dominates the rest that are of smaller order, so that

$$\big\|\mathbb{E}\big[\big(U_n - \mathcal{P}^m H\big)^2\big]\big\| = \left\|\binom{n}{m}^{-1} m\binom{n-m}{m-1}\Sigma_1^2\right\| + o\big(n^{-1}\big) = \left\|\frac{m^2}{n}\Sigma_1^2\right\| + o\big(n^{-1}\big)$$

as $n \to \infty$. In this paper, we consider non-degenerate U-statistics which commonly appear in applications such as estimation of covariance matrices and that serve as a main motivation for this work.

3. Robust modifications of U-statistics

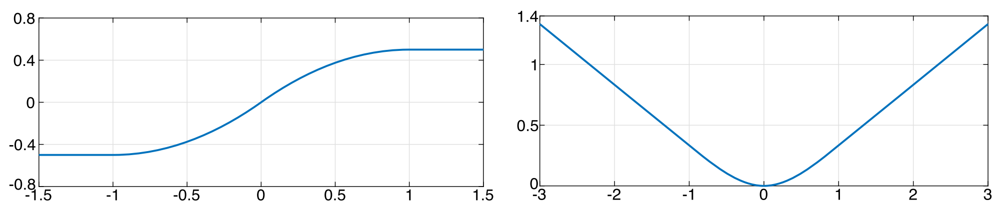
The goal of this section is to introduce the robust versions of U-statistics, and state the main results about their performance. Define

$$\psi(x) = \begin{cases} 1/2, & x > 1, \\ x - \operatorname{sign}(x)\cdot x^2/2, & |x| \le 1, \\ -1/2, & x < -1 \end{cases} \tag{3.1}$$

and its antiderivative

$$\Psi(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{|x|^3}{6}, & |x| \le 1, \\ \dfrac{1}{3} + \dfrac{1}{2}\big(|x|-1\big), & |x| > 1, \end{cases} \tag{3.2}$$

both of which are depicted in Figure 1. The function $\Psi(x)$ is closely related to Huber's loss [23]; concrete choice of $\Psi(x)$ is motivated by its properties, namely convexity and the fact that its



(a) $\psi(x)$ (b) $\Psi(x)$

Figure 1. Graphs of the functions $\psi(x)$ and $\Psi(x)$.

derivative $\psi(x)$ is operator Lipschitz and bounded (see Lemma 3.1 below). Let U_n be $\mathbb{H}^d$-valued U-statistic,

$$U_n := \frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} H(X_{i_1},\dots,X_{i_m}).$$

Since U_n is the average of matrices of the form $H(X_{i_1},\dots,X_{i_m}), (i_1,\dots,i_m) \in I_n^m$, it can be equivalently written as

$$\begin{aligned} U_n &= \operatorname*{argmin}_{U\in\mathbb{H}^d} \sum_{(i_1,\dots,i_m)\in I_n^m} \left\| H(X_{i_1},\dots,X_{i_m}) - U \right\|_{\mathrm{F}}^2 \\ &= \operatorname*{argmin}_{U\in\mathbb{H}^d} \operatorname{tr}\left[\sum_{(i_1,\dots,i_m)\in I_n^m} \left(H(X_{i_1},\dots,X_{i_m}) - U\right)^2\right]. \end{aligned}$$

A robust version of U_n is then defined by replacing the quadratic loss by the (rescaled) loss $\Psi(x)$. Namely, let $\theta > 0$ be a scaling parameter, and define

$$\widehat{U}_n^\star = \operatorname*{argmin}_{U\in\mathbb{H}^d} \operatorname{tr}\left[\sum_{(i_1,\dots,i_m)\in I_n^m} \Psi\left(\theta\left(H(X_{i_1},\dots,X_{i_m}) - U\right)\right)\right]. \tag{3.3}$$

For brevity, we will set

$$H_{i_1\dots i_m} := H(X_{i_1},\dots,X_{i_m}) \quad \text{and} \quad \mathbb{E}H := \mathbb{E}H_{i_1\dots i_m}$$

in what follows. Define

$$F_\theta(U) := \frac{1}{\theta^2}\frac{(n-m)!}{n!} \operatorname{tr}\left[\sum_{(i_1,\dots,i_m)\in I_n^m} \Psi\left(\theta(H_{i_1\dots i_m} - U)\right)\right]. \tag{3.4}$$

Clearly, $\widehat{U}_n^\star$ can be equivalently written as

$$\widehat{U}_n^\star = \operatorname*{argmin}_{U\in\mathbb{H}^d} \operatorname{tr}\left[F_\theta(U)\right].$$

The following result describes the basic properties of this optimization problem.

Lemma 3.1. *The following statements hold*:

1. *Problem* (3.3) *is a convex optimization problem.*
2. *The gradient $\nabla F_\theta(U)$ can be represented as*

$$\nabla F_\theta(U) = -\frac{1}{\theta}\frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} \psi\left(\theta(H_{i_1\dots i_m} - U)\right).$$

Moreover, $\nabla F_\theta(\cdot) : \mathbb{H}^d \mapsto \mathbb{H}^d$ is Lipschitz continuous in Frobenius and operator norms with Lipschitz constant equal to 1.

3. *Problem* (3.3) *is equivalent to*

$$\sum_{(i_1,\dots,i_m)\in I_n^m} \psi\big(\theta\big(H_{i_1\dots i_m} - \widehat{U}_n^\star\big)\big) = 0_{d\times d}. \tag{3.5}$$

Proofs of these facts are given in Section 5.2. Next, we present our main result regarding performance of the estimator $\widehat{U}_n^\star$. Define the *effective rank* [40] of a nonnegative definite matrix $A \in \mathbb{H}^d$ as

$$\mathrm{r}(A) := \frac{\operatorname{tr} A}{\|A\|}.$$

It is easy to see that for any matrix $A \in \mathbb{H}^d$, $\mathrm{r}(A) \le d$. We will be interested in the effective rank of the matrix $\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2$, and will denote

$$\mathrm{r}_H := \mathrm{r}\big(\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2\big).$$

Theorem 3.1. *Let $k = \lfloor n/m \rfloor$, and assume that $t \ge 1$ is such that*

$$\mathrm{r}_H \frac{t}{k} \le \frac{1}{104}.$$

Then for any $\sigma \ge \|\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2\|^{1/2}$ and $\theta := \theta_\sigma = \frac{1}{\sigma}\sqrt{\frac{2t}{k}}$,

$$\big\|\widehat{U}_n^\star - \mathbb{E}H\big\| \le 23\sigma\sqrt{\frac{t}{k}}$$

with probability at least $1 - \min(4d+1, 18\mathrm{r}_H)e^{-t}$.

The proof is presented in Section 5.3.

Remark 3.1. Condition $\mathrm{r}_H \frac{t}{k} \le \frac{1}{104}$ in Theorem 3.1 can be weakened to

$$\frac{\operatorname{tr}(\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2)}{\sigma^2}\frac{t}{k} \le \frac{1}{104},$$

where $\sigma^2 \ge \|\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2\|$. This fact follows from the straightforward modification of the proof of Theorem 3.1 and can be useful in applications.

Remark 3.2. Paper [25] investigates robust analogues of univariate U-statistics based on the median-of-means (MOM) technique. This approach can be extended to higher dimensions via replacing the univariate median by an appropriate multivariate generalization (e.g., the spatial median). When applied to covariance estimation problem, it yields estimates for the error measured in Frobenius norm; however, is not clear whether it can be used to obtain the error bounds in the operator norm. More specifically, to obtain such a bound via the MOM method, one would need to estimate $\mathbb{E}\|\frac{1}{n}\sum_{j=1}^n (Y_j - \mathbb{E}Y)(Y_j - \mathbb{E}Y)^T - \Sigma\|^2$, where $Y_1,\dots,Y_j$ are i.i.d. copies of a random vector $Y \in \mathbb{R}^d$ such that $\mathbb{E}(Y-\mathbb{E}Y)(Y-\mathbb{E}Y)^T = \Sigma$ and $\mathbb{E}\|Y\|_2^4 < \infty$. We are not aware of any existing (non-trivial) upper bounds for the aforementioned expectation that require only 4 finite moments of $\|Y\|_2$. On the other hand, it is straightforward to obtain an upper bound in the Frobenius norm since $\mathbb{E}\|\frac{1}{n}\sum_{j=1}^n (Y_j-\mathbb{E}Y)(Y_j-\mathbb{E}Y)^T - \Sigma\|_{\mathrm{F}}^2 = \frac{1}{n}(\mathbb{E}\|Y-\mathbb{E}Y\|_2^4 - \|\Sigma\|_{\mathrm{F}}^2)$.

Remark 3.3. We note that, due to the condition $\mathrm{r}_H\frac{t}{k} \le \frac{1}{104}$ that can be viewed as an upper bound on t, the inequality of Theorem 3.1 can not be integrated to obtain a bound in expectation. To establish such a bound, new techniques, beyond the ones developed in this paper, might be required.

3.1. Construction of the adaptive estimator

The downside of the estimator $\widehat{U}_n^\star$ defined in (3.3) is the fact that it is not completely data-dependent as the choice of θ requires the knowledge of an upper bound on

$$\sigma_*^2 := \left\|\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^2\right\|.$$

To alleviate this difficulty, we propose an adaptive construction based on a variant of Lepski's method [28].

Assume that $\sigma_{\min}$ is a known (possible crude) lower bound on σ_*. Choose $\gamma > 1$, let $\sigma_j := \sigma_{\min}\gamma^j$, and for each integer $j \ge 0$, set $t_j := t + \log[j(j+1)]$ and

$$\theta_j = \theta(j,t) = \sqrt{\frac{2t_j}{k}}\frac{1}{\sigma_j},$$

where $k = \lfloor n/m \rfloor$ as before. Let

$$\widehat{U}_{n,j} = \operatorname*{argmin}_{U\in\mathbb{H}^d} F_{\theta_j}(U),$$

where F_θ was defined in (3.4). Finally, set

$$\mathcal{L} := \mathcal{L}(t) = \left\{ l \in \mathbb{N} : \mathrm{r}_H \frac{t_l}{k} \le \frac{1}{104} \right\}$$

and

$$j_* := \min\left\{ j \in \mathcal{L} : \forall l \in \mathcal{L}, l > j, \|\widehat{U}_{n,l} - \widehat{U}_{n,j}\| \le 46\sigma_l\sqrt{\frac{t_l}{k}} \right\} \tag{3.6}$$

and $\widetilde{U}_n^\star := \widehat{U}_{n,j_*}$; if condition (3.6) is not satisfied by any $j \in \mathcal{L}$, we set $j_* = +\infty$ and $\widetilde{U}_n^\star = 0_{d\times d}$. Let

$$\Xi = \log\left[\left(\left\lfloor \frac{\log(\sigma_*/\sigma_{\min})}{\log\gamma}\right\rfloor + 1\right)\left(\left\lfloor \frac{\log(\sigma_*/\sigma_{\min})}{\log\gamma}\right\rfloor + 2\right)\right]. \tag{3.7}$$

Theorem 3.2. *Assume that* $t \geq 1$ *is such that*

$$\mathrm{r}_H \frac{(t+\Xi)}{k} \leq \frac{1}{104}.$$

Then with probability at least $1 - \min(4d+1, 18\mathrm{r}_H)e^{-t}$,

$$\left\| \widetilde{U}_n^{\star} - \mathbb{E}H \right\| \leq 69\gamma \cdot \sigma_* \sqrt{\frac{t+\Xi}{k}},$$

In other words, adaptive estimator can be obtained at the cost of an additional multiplicative factor 3γ in the error bound.

Proof. Let $\bar{j} = \min\{j \geq 1 : \sigma_j \geq \sigma_*\}$, and note that $\bar{j} \leq \lfloor \frac{\log(\sigma_*/\sigma_{\min})}{\log \gamma} \rfloor + 1$ and $\sigma_{\bar{j}} \leq \gamma \sigma_*$. Observe that condition of Theorem 3.2 guarantees that $\bar{j} \in \mathcal{L}$. We will show that $j_* \leq \bar{j}$ with high probability. Indeed,

$$\begin{aligned}
\Pr(j_* > \bar{j}) &\leq \Pr\Bigg(\bigcup_{l \in \mathcal{L}: l > \bar{j}} \Bigg\{ \|\widehat{U}_{n,l} - \widehat{U}_{n,\bar{j}}\| > 46\sigma_l \sqrt{\frac{t_l}{k}} \Bigg\} \Bigg) \\
&\leq \Pr\Bigg(\|\widehat{U}_{n,\bar{j}} - \mathbb{E}H\| > 23\sigma_{\bar{j}} \sqrt{\frac{t_{\bar{j}}}{k}} \Bigg) + \sum_{l \in \mathcal{L}: l > \bar{j}} \Pr\Bigg(\|\widehat{U}_{n,l} - \mathbb{E}H\| > 23\sigma_l \sqrt{\frac{t_l}{k}} \Bigg) \\
&\leq \tilde{d} e^{-t} \frac{1}{\bar{j}(\bar{j}+1)} + \tilde{d} e^{-t} \sum_{l > \bar{j}} \frac{1}{l(l+1)} \leq \tilde{d} e^{-t},
\end{aligned}$$

where we used Theorem 3.1 to bound each of the probabilities in the sum, and $\tilde{d} = \min(4d + 1, 18\mathrm{r}_H)$. The display above implies that the event

$$\mathcal{B} = \bigcap_{l \in \mathcal{L}: l \geq \bar{j}} \Bigg\{ \|\widehat{U}_{n,l} - \mathbb{E}H\| \leq 23\sigma_l \sqrt{\frac{t_l}{k}} \Bigg\}$$

of probability at least $1 - \min(4d+1, 18\mathrm{r}_H)e^{-t}$ is contained in $\mathcal{E} = \{j_* \leq \bar{j}\}$. Hence, on $\mathcal{B}$ we have

$$\begin{aligned}
\left\| \widetilde{U}_n^{\star} - \mathbb{E}H \right\| &\leq \left\| \widetilde{U}_n^{\star} - \widehat{U}_{n,\bar{j}} \right\| + \|\widehat{U}_{n,\bar{j}} - \mathbb{E}H\| \leq 46\sigma_{\bar{j}} \sqrt{\frac{t_{\bar{j}}}{k}} + 23\sigma_{\bar{j}} \sqrt{\frac{t_{\bar{j}}}{k}} \\
&\leq \gamma \cdot 69\sigma_* \sqrt{\frac{t+\Xi}{k}},
\end{aligned}$$

where $\Xi = \log[(\lfloor \frac{\log(\sigma_*/\sigma_{\min})}{\log \gamma} \rfloor + 1)(\lfloor \frac{\log(\sigma_*/\sigma_{\min})}{\log \gamma} \rfloor + 2)]$. □

3.2. Extension to rectangular matrices

In this section, we assume a more general setting where $H : \mathcal{S}^m \mapsto \mathbb{C}^{d_1\times d_2}$ is a $\mathbb{C}^{d_1\times d_2}$-valued permutation-symmetric function. As before, our goal is to construct an estimator of $\mathbb{E}H$. We reduce this general problem to the case of $\mathbb{H}^{d_1+d_2}$-valued functions via the self-adjoint dilation defined in (2.1). Let

$$\mathcal{D}(H_{i_1\dots i_m}) = \begin{pmatrix} 0 & H(X_{i_1},\dots,X_{i_m}) \\ \big[H(X_{i_1},\dots,X_{i_m})\big]^* & 0 \end{pmatrix},$$

and

$$\bar{U}_n^\star = \underset{U\in\mathbb{H}^{d_1+d_2}}{\operatorname{argmin}}\ \operatorname{tr}\Bigg[\sum_{(i_1,\dots,i_m)\in I_n^m} \Psi\big(\theta\big(\mathcal{D}(H_{i_1\dots i_m}) - U\big)\big)\Bigg].$$

Let $\hat{U}_{11}^\star \in \mathbb{C}^{d_1\times d_1}$, $\hat{U}_{22}^\star \in \mathbb{C}^{d_2\times d_2}$, $\hat{U}_{12}^\star \in \mathbb{C}^{d_1\times d_2}$ be such that $\bar{U}_n^\star$ can be written in the block form as $\bar{U}_n^\star = \left(\begin{smallmatrix} \hat{U}_{11}^\star & \hat{U}_{12}^\star \\ (\hat{U}_{12}^\star)^* & \hat{U}_{22}^\star \end{smallmatrix}\right)$. Moreover, define

$$\sigma_\star^2 := \max\Big(\big\|\mathbb{E}(H_{1\dots m} - \mathbb{E}H)(H_{1\dots m} - \mathbb{E}H)^*\big\|, \big\|\mathbb{E}(H_{1\dots m} - \mathbb{E}H)^*(H_{1\dots m} - \mathbb{E}H)\big\|\Big)$$

and

$$\tilde{\mathrm{r}}_H := 2\cdot\frac{\operatorname{tr}[\mathbb{E}(H_{1\dots m} - \mathbb{E}H)(H_{1\dots m} - \mathbb{E}H)^*]}{\sigma_\star^2}.$$

Corollary 3.1. *Let $k = \lfloor n/m\rfloor$, and assume that $t \ge 1$ is such that*

$$\tilde{\mathrm{r}}_H \frac{t}{k} \le \frac{1}{104}.$$

Then for any $\sigma \ge \sigma_\star$ and $\theta := \theta_\sigma = \frac{1}{\sigma}\sqrt{\frac{2t}{k}}$,

$$\big\|\hat{U}_{12}^\star - \mathbb{E}H\big\| \le 23\sigma\sqrt{\frac{t}{k}}$$

with probability at least $1 - \min(4(d_1 + d_2) + 1, 18\tilde{\mathrm{r}}_H)e^{-t}$.

The proof is outlined in Section 5.7.

3.3. Computational considerations

Since the estimator $\widehat{U}_n^\star$ is defined as the solution of the convex optimization problem (3.3), it can be approximated via the gradient descent. We consider the simplest gradient descent scheme with constant step size equal 1. Note that the Lipschitz constant of $F_\theta(U)$ is $L_F = 1$ by Lemma 3.1, hence this step choice is exactly equal to $\frac{1}{L_F}$. Given a starting point $U_0 \in \mathbb{H}^d$, the gradient descent iteration for minimization of $\operatorname{tr} F_\theta(U)$ is

$$
\begin{aligned}
U_n^{(0)} &:= U_0, \\
U_n^{(j)} &:= U_n^{(j-1)} - \nabla\big(\operatorname{tr} F_\theta\big(U_n^{(j-1)}\big)\big) \\
&= U_n^{(j-1)} + \frac{1}{\theta}\frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} \psi\big(\theta\big(H_{i_1\dots i_m} - U_n^{(j-1)}\big)\big), \quad j \geq 1.
\end{aligned}
$$

Lemma 3.2. *The following inequalities hold for all* $j \geq 1$:

$$
(a) \quad \operatorname{tr}\big[F_\theta\big(U_n^{(j)}\big) - F_\theta\big(\widehat{U}_n^\star\big)\big] \leq \frac{\|U_0 - \widehat{U}_n^\star\|_F^2}{2j};
$$

Moreover, under the assumptions of Theorem 3.1,

$$
(b) \quad \big\|U_n^{(j)} - \mathbb{E}H\big\| \leq \left(\frac{3}{4}\right)^j \|U_0 - \mathbb{E}H\| + 23\sigma\sqrt{\frac{t}{k}}
$$

with probability at least $1 - \min(4d+1, 18\mathrm{r}_H)e^{-t}$.

The proof is given is Section 5.6. Note that part (b) implies that a small number of iterations suffice to get an estimator of $\mathbb{E}H$ that achieves performance bound similar to $\widehat{U}_n^\star$.

4. Estimation of covariance matrices

In this section, we consider applications of the previously discussed results to covariance estimation problems. Let $Y \in \mathbb{R}^d$ be a random vector with mean $\mathbb{E}Y = \mu$, covariance matrix $\Sigma = \mathbb{E}[(Y-\mu)(Y-\mu)^T]$, and such that $\mathbb{E}\|Y-\mu\|_2^4 < \infty$. Assume that $Y_1, \dots, Y_n$ be i.i.d. copies of Y. Our goal is to estimate Σ; note that when the observations are the heavy-tailed, mean estimation problem becomes non-trivial, so the assumption $\mu = 0$ is not plausible.

U-statistics offer a convenient way to avoid explicit mean estimation. Indeed, observe that $\Sigma = \frac{1}{2}\mathbb{E}[(Y_1 - Y_2)(Y_1 - Y_2)^T]$, hence the natural estimator of Σ is the U-statistic

$$
\widetilde{\Sigma}_n = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}. \tag{4.1}
$$

It is easy to check that $\widetilde{\Sigma}$ coincides with the usual sample covariance estimator

$$
\widetilde{\Sigma}_n = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T.
$$

The robust version is defined according to (3.3) as

$$\widehat{\Sigma}_\star = \underset{S \in \mathbb{R}^{d\times d}, S=S^T}{\operatorname{argmin}} \left[\operatorname{tr} \sum_{i\neq j} \Psi\left(\theta\left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S\right)\right)\right], \tag{4.2}$$

which, by Lemma 3.1, is equivalent to

$$\sum_{i\neq j} \psi\left(\theta\left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma}_\star\right)\right) = 0_{d\times d}.$$

Remark 4.1. Assume that $\Sigma_n^{(0)} = 0_{d\times d}$, then the first iteration of the gradient descent for the problem (4.2) is

$$\Sigma_n^{(1)} = \frac{1}{\theta}\frac{1}{n(n-1)} \sum_{i\neq j} \psi\left(\theta \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}\right).$$

$\Sigma_n^{(1)}$ can itself be viewed as an estimator of the covariance matrix. It has been proposed in [34] (see Remark 7 in that paper), and its performance has been later analyzed in [15] (see Theorem 3.2). These results support the claim that a small number of gradient descent steps for problem (3.3) suffice in applications.

To assess performance of $\widehat{\Sigma}_\star$, we will apply Theorem 3.1. First, let us discuss the "matrix variance" parameter σ^2 appearing in the statement. Direct computation shows that for $H(Y_1, Y_2) = \frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2}$,

$$\mathbb{E}\big(H(Y_1, Y_2) - \mathbb{E}H\big)^2 = \frac{1}{2}\Big(\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2 + \operatorname{tr}(\Sigma)\Sigma\Big).$$

The following result (which is an extension of Lemma 2.3 in [35]) connects $\|\mathbb{E}(H - \mathbb{E}H)^2\|$ with $r(\Sigma)$, the effective rank of the covariance matrix Σ.

Lemma 4.1.

(a) *Assume that kurtoses of the linear forms $\langle Y, v\rangle$ are uniformly bounded by K, meaning that* $\sup_{v:\|v\|_2=1} \frac{\mathbb{E}\langle Y-\mathbb{E}Y, v\rangle^4}{[\mathbb{E}\langle Y-\mathbb{E}Y, v\rangle^2]^2} \le K$. *Then*

$$\Big\|\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\Big\| \le K \operatorname{tr}(\Sigma)\|\Sigma\|.$$

(b) *Assume that the kurtoses of coordinates $Y^{(j)} := \langle Y, e_j\rangle$ of Y are uniformly bounded by $K' < \infty$, meaning that* $\max_{j=1,\dots,d} \frac{\mathbb{E}(Y^{(j)}-\mathbb{E}Y^{(j)})^4}{[\mathbb{E}(Y^{(j)}-\mathbb{E}Y^{(j)})^2]^2} \le K'$. *Then*

$$\operatorname{tr}\Big[\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\Big] \le K'\big(\operatorname{tr}(\Sigma)\big)^2.$$

(c) *The following inequality holds*:

$$\left\|\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\right\| \geq \mathrm{tr}(\Sigma)\|\Sigma\|.$$

Lemma 4.1 immediately implies that under the bounded kurtosis assumption,

$$\left\|\mathbb{E}(H-\mathbb{E}H)^2\right\| \leq K\mathrm{r}(\Sigma)\|\Sigma\|^2.$$

The following corollary of Theorem 3.1 (together with Remark 3.1) is immediate.

Corollary 4.1. *Assume that the kurtosis of linear forms $\langle Y, v\rangle$, $v \in \mathbb{R}^d$, is uniformly bounded by K. Moreover, let $t > 0$ be such that*

$$r(\Sigma)\frac{t}{\lfloor n/2\rfloor} \leq \frac{1}{104}.$$

Then for any $\sigma \geq \sqrt{K\mathrm{r}(\Sigma)}\|\Sigma\|$ and $\theta := \theta_\sigma = \frac{1}{\sigma}\sqrt{\frac{2t}{\lfloor n/2\rfloor}}$,

$$\|\widehat{\Sigma}_\star - \Sigma\| \leq 23\sigma\sqrt{\frac{t}{\lfloor n/2\rfloor}}$$

with probability at least $1-\min(4d+1, 18K'\mathrm{r}(\Sigma))e^{-t}$.

An adaptive version of the estimator $\widetilde{\Sigma}_\star$ can be constructed as in (3.6), and its performance follows similarly from Theorem 3.2.

Remark 4.2. It is known [27] that the quantity $\sqrt{\mathrm{r}(\Sigma)}\|\Sigma\|$ controls the expected error of the sample covariance estimator in the Gaussian setting. On the other hand, fluctuations of the error around its expected value in the Gaussian case [27] are controlled by the "weak variance" $\sup_{v\in\mathbb{R}^d:\|v\|_2=1}\mathbb{E}^{1/2}\langle Z, v\rangle^4 \leq \sqrt{K}\|\Sigma\|$, while in our bounds fluctuations are controlled by the larger quantity σ^2; this fact leaves room for improvement in our results.

4.0.1. *Numerical simulation*

We provide a short numerical illustration of the performance of the proposed covariance estimator. Data was generated as follows: let $U = (U^{(1)}, \ldots, U^{(100)})^T \in \mathbb{R}^{100}$ be a vector with i.i.d. coordinates such that $U^{(j)}$, $j = 1, \ldots, 100$ are independent random variables with Student's t-distribution with 4.01 degrees of freedom scaled so that $\mathrm{Var}(U^{(j)}) = 1$.

Next, let $Y = \sqrt{\Sigma}U$, where Σ is a diagonal matrix with $\Sigma_{11} = 10$, $\Sigma_{22} = 5$, $\Sigma_{33} = 1$, and $\Sigma_{jj} = \frac{1}{97}$, $j \geq 4$. In particular, $\mathbb{E}Y = 0$ and $\mathbb{E}YY^T = \Sigma$. The goal of numerical experiment was to evaluate performance of the adaptive estimator $\widetilde{\Sigma}_\star$ of the covariance matrix Σ, constructed using a version of Lepski's method as described in Section 3.1. The (relative) error was measured in the operator norm, defined via the ratio $\frac{\|\widetilde{\Sigma}_\star-\Sigma\|}{\|\Sigma\|}$. We compared $\widetilde{\Sigma}_\star$ with a standard sample covariance estimator $\widetilde{\Sigma}_n$, introduced in (4.1). We performed 400 repetitions

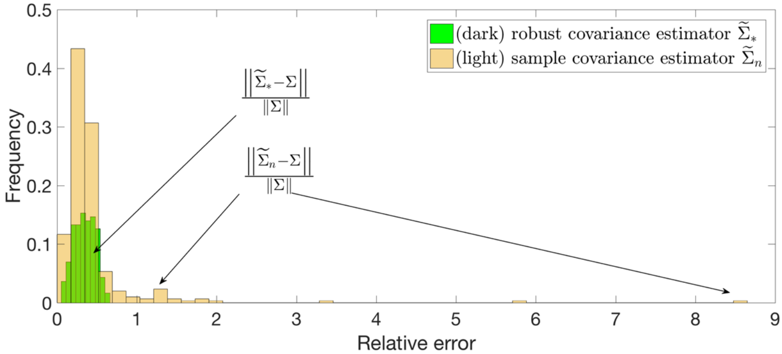


Figure 2. Histogram over 400 repetitions of the experiment; sample size $n = 100$, dimension $d = 100$.

of the experiment with the sample size set to $n = 100$, and recorded the estimation error of $\widetilde{\Sigma}_\star$ and $\widetilde{\Sigma}_n$ for each run. For each value of σ, estimator $\widetilde{\Sigma}_\star$ has been approximated by performing 2 gradient descent iterations for the problem (4.2) (additional iterations, while being computationally intensive, did not provide noticeable improvement in our experiments). The average and maximum relative errors (over 400 repetitions) for the robust estimator $\widetilde{\Sigma}_\star$ were 0.34 and 0.65 respectively, with the standard deviation of 0.13. The corresponding values for the sample covariance estimator were 0.44, 8.62 and 0.66. Histograms illustrating performance of both estimators are presented in Figure 2. It is clear from the graphs that in the considered scenario, estimator $\widetilde{\Sigma}_\star$ performed noticeably better than the sample covariance $\widetilde{\Sigma}_n$.

4.1. Estimation in the Frobenius norm

Next, we show that thresholding of the singular values of the adaptive estimator $\widetilde{\Sigma}_\star$ (defined as in (3.6) for some $\gamma > 1$) yields the estimator that achieves optimal performance in the Frobenius norm. Given $\tau > 0$, define

$$\widetilde{\Sigma}_\star^\tau = \sum_{j=1}^{d} \max\big(\lambda_j(\widetilde{\Sigma}_\star) - \tau/2, 0\big) v_j(\widetilde{\Sigma}_\star) v_j(\widetilde{\Sigma}_\star)^T, \tag{4.3}$$

where $\lambda_j(\widetilde{\Sigma}_\star)$ and $v_j(\widetilde{\Sigma}_\star)$ are the eigenvalues and the corresponding eigenvectors of $\widetilde{\Sigma}_\star$.

Corollary 4.2. *Assume that kurtoses of linear forms $\langle Y, v\rangle$, $v \in \mathbb{R}^d$, are uniformly bounded by K. Moreover, let $t > 0$ be such that*

$$\mathrm{r}(\Sigma)\frac{t + \Xi}{\lfloor n/2 \rfloor} \le \frac{1}{104},$$

where Ξ *was defined in* (3.7) *with* $\sigma_* := \sqrt{K\,\mathrm{r}(\Sigma)}\|\Sigma\|$. *Then for any*

$$\tau \geq \gamma \cdot 138\sqrt{K}\|\Sigma\|\sqrt{\frac{\mathrm{r}(\Sigma)(t+\Xi)}{\lfloor n/2\rfloor}},$$
$$\big\|\widetilde{\Sigma}_\star^\tau - \Sigma\big\|_{\mathrm{F}}^2 \leq \inf_{S\in\mathbb{R}^{d\times d}, S=S^T}\left[\|S-\Sigma\|_{\mathrm{F}}^2 + \frac{(1+\sqrt{2})^2}{8}\tau^2\,\mathrm{rank}(S)\right]. \tag{4.4}$$

with probability at least $1-\min(4d+1, 18K'\mathrm{r}(\Sigma))e^{-t}$.

The proof of this corollary is given in Section 5.9.

4.2. Masked covariance estimation

Masked covariance estimation framework is based on the assumption that some entries of the covariance matrix Σ are "more important." This is quantified by a symmetric mask matrix $M \in \mathbb{R}^{d\times d}$, whence the goal is to estimate the matrix $M\odot\Sigma$ that "downweights" the entries of Σ that are deemed less important, or incorporates the prior information on Σ. This problem formulation has been introduced in [30], and later studied in a number of papers including [12] and [26].

We will be interested in finding an estimator $\widehat{\Sigma}_\star^M$ such that $\|\widehat{\Sigma}_*^M - M\odot\Sigma\|$ is small with high probability, and specifically in dependence of the estimation error on the mask matrix M. Consider the following estimator:

$$\widehat{\Sigma}_\star^M = \operatorname*{argmin}_{S\in\mathbb{R}^{d\times d}, S=S^T}\left[\mathrm{tr}\sum_{i\neq j}\Psi\left(\theta\left(\frac{M\odot(Y_i-Y_j)(Y_i-Y_j)^T}{2} - S\right)\right)\right], \tag{4.5}$$

which is a "robust" version of the estimator $M\odot\widetilde{\Sigma}_n$, where $\widetilde{\Sigma}_n$ is the sample covariance matrix defined in (4.1). Next, following [12] we introduce additional parameters that appear in the performance bound for $\widehat{\Sigma}_\star^M$. Let

$$\|M\|_{1\to 2} := \max_{j=1,\ldots,d}\sqrt{\sum_{i=1}^d M_{ij}^2}$$

be the maximum $\|\cdot\|_2$ norm of the columns of M. We also define

$$\nu_4(Y) := \sup_{\|\mathbf{v}\|_2\leq 1}\mathbb{E}^{1/4}\langle\mathbf{v}, Y-\mathbb{E}Y\rangle^4$$

and

$$\mu_4(Y) = \max_{j=1\ldots d}\mathbb{E}^{1/4}\big(Y^{(j)}-\mathbb{E}Y^{(j)}\big)^4.$$

The following result describes the finite-sample performance guarantees for $\widehat{\Sigma}_\star^M$.

Corollary 4.3. *Assume that kurtoses of the coordinates $Y^{(j)} = \langle Y, e_j\rangle$ of Y are uniformly bounded by K'. Moreover, let $t > 0$ be such that*

$$\sqrt{K'}\frac{\operatorname{tr}(\Sigma)}{\nu_4^2(Y)}\frac{t}{\lfloor n/2\rfloor} \le \frac{1}{104}.$$

Then for any $\Delta \ge \sqrt{2}\|M\|_{1\to 2}\nu_4(Y)\mu_4(Y)$ and $\theta = \frac{1}{\Delta}\sqrt{\frac{2t}{\lfloor n/2\rfloor}}$,

$$\big\|\widehat{\Sigma}_\star^M - M\odot\Sigma\big\| \le 23\Delta\sqrt{\frac{t}{\lfloor n/2\rfloor}}$$

with probability at least $1-\min(4d+1, 18K'\mathrm{r}(\Sigma))e^{-t}$.

Proof. Let X and X' be independent and identically distributed random variables. Then it is easy to check that

$$\mathbb{E}\big(X-X'\big)^4 \le 8\mathbb{E}(X-\mathbb{E}X)^4. \tag{4.6}$$

It implies that $\nu_4^2(Y_1-Y_2) \le 2\sqrt{2}\nu_4^2(Y)$ and $\mu_4^2(Y_1-Y_2)\le 2\sqrt{2}\mu_4^2(Y)$.

Next, Lemma 4.1 in [12] yields that

$$\left\|\mathbb{E}\left(\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2}\odot M\right)^2\right\| \le 2\|M\|_{1\to 2}^2\mu_4^2(Y)\nu_4^2(Y). \tag{4.7}$$

Now we will find an upper bound for the trace of $\mathbb{E}(\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2}\odot M)^2$. It is easy to see that (e.g., see equation (4.1) in [12])

$$\begin{aligned}
&\mathbb{E}\left(\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2}\odot M\right)^2 \\
&\quad=\sum_{j=1}^d M^{(j)}\big(M^{(j)}\big)^T\odot\mathbb{E}\left(\frac{Y_1^{(j)}-Y_2^{(j)}}{\sqrt{2}}\right)^2\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2},
\end{aligned}$$

where $M^{(j)}$ denotes the j-th column of the matrix M. It follows from (4.6), Hölder's inequality and the bounded kurtosis assumption that

$$\begin{aligned}
\operatorname{tr}\left[\mathbb{E}\left(\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2}\odot M\right)^2\right] &= \sum_{i,j=1}^d M_{i,j}^2\mathbb{E}\left[\left(\frac{Y_1^{(i)}-Y_2^{(i)}}{\sqrt{2}}\right)^2\left(\frac{Y_1^{(j)}-Y_2^{(j)}}{\sqrt{2}}\right)^2\right] \\
&\le 2\sum_{i,j=1}^d M_{i,j}^2\mathbb{E}^{1/2}\big(Y^{(i)}-\mathbb{E}Y^{(i)}\big)^4\mathbb{E}^{1/2}\big(Y^{(j)}-\mathbb{E}Y^{(j)}\big)^4 \\
&\le 2\sqrt{K'}\mu_4^2(Y)\|M\|_{1\to 2}^2\operatorname{tr}(\Sigma).
\end{aligned}$$

Next, we deduce that for $\Delta^2 \geq 2\|M\|_{1\to 2}^2 \mu_4^2(Y)\nu_4^2(Y)$,

$$\frac{\mathrm{tr}[\mathbb{E}(\frac{(Y_1-Y_2)(Y_1-Y_2)^T}{2} \odot M)^2]}{\Delta^2} \leq \sqrt{K'} \frac{\mathrm{tr}(\Sigma)}{\nu_4^2(Y)}.$$

Result now follows from Theorem 3.1 and Remark 3.1. □

Remark 4.3. Let

$$K := \sup_{v:\|v\|_2=1} \frac{\mathbb{E}\langle Y - \mathbb{E}Y, v\rangle^4}{[\mathbb{E}\langle Y - \mathbb{E}Y, v\rangle^2]^2}.$$

Since $\nu_4^2(Y) \leq \sqrt{K}\|\Sigma\|$ and $\mu_4^2 \leq \sqrt{K'}\|\Sigma\|_{\max}$, we can state a slightly modified version of Corollary 4.3. Namely, let $t > 0$ be such that

$$\sqrt{\frac{K'}{K}} \mathrm{r}(\Sigma) \frac{t}{\lfloor n/2 \rfloor} \leq \frac{1}{104}.$$

Then for any $\Delta \geq \sqrt{2K}\|M\|_{1\to 2}\sqrt{\|\Sigma\|_{\max}\|\Sigma\|}$ and $\theta = \frac{1}{\Delta}\sqrt{\frac{2t}{\lfloor n/2 \rfloor}}$,

$$\left\| \widehat{\Sigma}_\star^M - M \odot \Sigma \right\| \leq 23\Delta\sqrt{\frac{t}{\lfloor n/2 \rfloor}}$$

with probability at least $1 - \min(4d+1, 18K'\mathrm{r}(\Sigma))e^{-t}$. In particular, if $\|M\|_{1\to 2}^2 \ll \mathrm{r}(\Sigma)\frac{\|\Sigma\|}{\|\Sigma\|_{\max}}$, then our bounds show that $M \odot \Sigma$ can be estimated at a faster rate than Σ itself. This conclusion is consistent with results in [12] for Gaussian random vectors (e.g., see Theorem 1.1 in that paper); however, we should note that our bounds were obtained under much weaker assumptions.

5. Proofs of the mains results

In this section, we present the proofs that were omitted from the main exposition.

5.1. Technical tools

We recall several useful facts from probability theory and matrix analysis that our arguments rely on.

Fact 1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. Then $A \mapsto \mathrm{tr}\, f(A)$ is convex on the set of self-adjoint matrices. In particular, for any self-adjoint matrices A, B,*

$$\mathrm{tr}\, f\left(\frac{A+B}{2}\right) \leq \frac{1}{2}\mathrm{tr}\, f(A) + \frac{1}{2}\mathrm{tr}\, f(B).$$

Proof. This is a consequence of Peierls inequality, see Theorem 2.9 in [8] and the comments following it. □

Fact 2. *Let $F : \mathbb{R} \mapsto \mathbb{R}$ be a continuously differentiable function, and $S \in \mathbb{C}^{d\times d}$ be a self-adjoint matrix. Then the gradient of $G(S) := \operatorname{tr} F(S)$ is*

$$\nabla G(S) = F'(S),$$

where F' is the derivative of F and $F'(S) : \mathbb{C}^{d\times d} \mapsto \mathbb{C}^{d\times d}$ is the matrix function in the sense of the definition 2.1.

Proof. See Lemma A.1 in [34]. □

Fact 3. *Function $\psi(x)$ defined in (3.1) satisfies*

$$-\log\bigl(1 - x + x^2\bigr) \le \psi(x) \le \log\bigl(1 + x + x^2\bigr) \tag{5.1}$$

for all $x \in \mathbb{R}$. Moreover, as a function of $\mathbb{H}^d$-valued argument (see definition 2.1), $\psi(\cdot)$ is Lipschitz continuous in the Frobenius and operator norms with Lipschitz constant 1, meaning that for all $A_1, A_2 \in \mathbb{H}^d$,

$$\begin{aligned}\bigl\|\psi(A_1) - \psi(A_2)\bigr\|_F &\le \|A_1 - A_2\|_F,\\ \bigl\|\psi(A_1) - \psi(A_2)\bigr\| &\le \|A_1 - A_2\|.\end{aligned}$$

Proof. To show (5.1), it is enough to check that $x - x^2/2 \ge -\log(1 - x + x^2)$ for $x \in [0, 1]$ and that $x - x^2/2 \le \log(1 + x + x^2)$, $x \in [0, 1]$. Other inequalities follow after the change of variable $y = -x$. To check that $f(x) := x - x^2/2 \ge -\log(1 - x + x^2) := g(x)$ for $x \in [0, 1]$, note that $f(0) = g(0) = 0$ and that $f'(x) = 1 - x \ge 1 - \frac{x(1+x)}{1-x+x^2} = g'(x)$ for $x \in [0, 1]$. Inequality $x - x^2/2 \le \log(1 + x + x^2)$, $x \in [0, 1]$ can be established similarly.

Note that the function $\psi : \mathbb{R} \mapsto \mathbb{R}$ is Lipshitz continuous with Lipschitz constant 1 as a function of real variable. Lemma 5.5 (Chapter 7) in [6] immediately implies that it is also Lipshitz continuous in the Frobenius norm, still with Lipschitz constant 1.

Lipshitz property of ψ in the operator norm follows from Corollary 1.1.2 in [1] which states that if $g \in C^1(\mathbb{R})$ and g' is positive definite, then the Lipschitz constant of g (as a function on $\mathbb{H}^d$) is equal to $g'(0)$. It is easy to check that

$$\psi'(x) = \begin{cases} 1 - |x|, & |x| \le 1,\\ 0, & \text{otherwise,}\end{cases}$$

which is the Fourier transform of the positive integrable function $\operatorname{sinc}(y) = (\frac{\sin(\pi y)}{\pi y})^2$, hence ψ' is positive definite and the (operator) Lipschitz constant of ψ is equal to 1. □

Fact 4. *Let $T_1, \dots, T_L$ be arbitrary $\mathbb{H}^d$-valued random variables, and $p_1, \dots, p_L$ be non-negative weights such that $\sum_{j=1}^L p_j = 1$. Moreover, let $T = \sum_{j=1}^L p_j T_j$ be convex combination of $T_1, \dots, T_L$. Then*

$$\Pr\big(\lambda_{\max}(T) \ge t\big) \le \max_{j=1,\dots,L} \Big[\inf_{\theta>0} e^{-\theta t} \mathbb{E}\,\mathrm{tr}\, e^{\theta T_j}\Big].$$

Proof. This fact is a corollary of the well-known Hoeffding's inequality (see Section 5 in [21]). Indeed, for any $\theta > 0$,

$$\begin{aligned}\Pr\left(\lambda_{\max}\left(\sum_{j=1}^L p_j T_j\right) \ge t\right) &\le \Pr\left(\exp\left(\theta\lambda_{\max}\left(\sum_{j=1}^L p_j T_j\right)\right) \ge e^{\theta t}\right) \\ &\le e^{-\theta t}\mathbb{E}\,\mathrm{tr}\exp\left(\theta\sum_{j=1}^L p_j T_j\right) \le e^{-\theta t}\sum_{j=1}^L p_j \mathbb{E}\,\mathrm{tr}\exp(\theta T_j),\end{aligned}$$

where the last inequality follows from Fact 1. □

Fact 5 (Chernoff bound). *Let $\xi_1, \dots, \xi_n$ be a sequence of i.i.d. copies of ξ such that* $\Pr(\xi = 1) = 1 - \Pr(\xi = 0) = p \in (0, 1)$, *and define $S_n := \sum_{j=1}^n \xi_j$. Then*

$$\Pr\big(S_n/n \ge (1+\tau)p\big) \le \inf_{\theta>0}\big[e^{-\theta np(1+\tau)}\mathbb{E}e^{\theta S_n}\big] \le \begin{cases} e^{-\frac{\tau^2 np}{2+\tau}}, & \tau > 1, \\ e^{-\frac{\tau^2 np}{3}}, & 0 < \tau \le 1. \end{cases}$$

Proof. See Proposition 2.4 in [3]. □

Let π_n be the collection of all permutations $i : \{1, \dots, n\} \mapsto \{1, \dots, n\}$. For integers $m \le \lfloor n/2 \rfloor$, let $k = \lfloor n/m \rfloor$. Given a permutation $(i_1, \dots, i_n) \in \pi_n$ and a U-statistic U_n defined in (2.2), let

$$\begin{aligned} W_{i_1,\dots,i_n} := \frac{1}{k}\big(&H(X_{i_1}, \dots, X_{i_m}) + H(X_{i_{m+1}}, \dots, X_{i_{2m}}) + \cdots \\ &+ H(X_{i_{(k-1)m+1}}, \dots, X_{i_{km}})\big). \end{aligned} \tag{5.2}$$

Fact 6. *The following equality holds*:

$$U_n = \frac{1}{n!}\sum_{(i_1,\dots,i_n)\in\pi_n} W_{i_1,\dots,i_n}.$$

Proof. See Section 5 in [21]. □

Let $Z_1, \dots, Z_n$ be a sequence of independent copies of $Z \in \mathbb{H}^d$ such that $\|\mathbb{E}Z^2\| < \infty$.

Fact 7 (Matrix Bernstein Inequality). *Assume that* $\|Z - \mathbb{E}Z\| \le M$ *almost surely. Then for any* $\sigma \ge \|\mathbb{E}(Z - \mathbb{E}Z)^2\|$ *and any* $t > 0$,

$$\left\| \frac{\sum_{j=1}^n Z_j}{n} - \mathbb{E}Z \right\| \le 2\sigma\sqrt{\frac{t}{n}} \vee \frac{4Mt}{3n}$$

with probability at least $1 - 2de^{-t}$. *Moreover, for all* $t \ge 1$,

$$\left\| \frac{\sum_{j=1}^n Z_j}{n} - \mathbb{E}Z \right\| \le 2\sigma\sqrt{\frac{t}{n}} \vee \frac{4Mt}{3n}$$

with probability at least $1 - 14\mathrm{r}(\mathbb{E}(Z - \mathbb{E}Z)^2)e^{-t}$, *where* $\mathrm{r}(A)$ *stands for the effective rank of* A.

Proof. See Theorem 1.4 in [38] for the first statement and Theorem 3.1 in [33] for the second bound. □

Assume that $\|H(X_{i_1}, \dots, X_{i_m})\| \le M$ almost surely. Together with Facts 6 and 4, Bernstein's inequality can be used to show that

$$\|U_n - \mathbb{E}H\| \le 2\left\|\mathbb{E}(H - \mathbb{E}H)^2\right\|^{1/2}\sqrt{\frac{t}{k}} \vee \frac{4Mt}{3k} \tag{5.3}$$

with probability $\ge 1 - 2de^{-t}$, where as before $k = \lfloor n/m \rfloor$. This corollary will be useful in the sequel.

Fact 8. *Let* $\psi(\cdot)$ *be defined by* (3.1). *Then the following inequalities hold for all* $\theta > 0$:

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum_{j=1}^n \big(\psi(\theta Z_j) - \theta\mathbb{E}Z\big)\right) \le \mathrm{tr}\exp\big(n\theta^2\mathbb{E}Z^2\big),$$

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum_{j=1}^n \big(\theta\mathbb{E}Z - \psi(\theta Z_j)\big)\right) \le \mathrm{tr}\exp\big(n\theta^2\mathbb{E}Z^2\big).$$

Proof. These inequalities follow from (5.1) and Lemma 3.1 in [34]. Note that we did not assume boundedness of $\|Z - \mathbb{E}Z\|$. □

Finally, we will need the following statement related to the self-adjoint dilation (2.1).

Fact 9. *Let* $S \in \mathbb{C}^{d_1 \times d_1}$, $T \in \mathbb{C}^{d_2 \times d_2}$ *be self-adjoint matrices, and* $A \in \mathbb{C}^{d_1 \times d_2}$. *Then*

$$\left\| \begin{pmatrix} S & A \\ A^* & T \end{pmatrix} \right\| \ge \left\| \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \right\|.$$

Proof. See Lemma 2.1 in [34]. □

5.2. Proof of Lemma 3.1

(1) Convexity follows from Fact 1 since the sum of convex functions is a convex function.

(2) The expression for the gradient follows from Fact 2. To show that $\nabla F_\theta(U)$ is Lipschitz continuous, note that

$$\begin{aligned}
&\left\| \frac{1}{\theta}\psi\big(\theta(H_{i_1,\dots,i_m} - U_1)\big) - \frac{1}{\theta}\psi\big(\theta(H_{i_1,\dots,i_m} - U_2)\big) \right\| \\
&\quad \le \left\| \frac{1}{\theta}\big(\theta(H_{i_1,\dots,i_m} - U_1) - \theta(H_{i_1,\dots,i_m} - U_2)\big) \right\| = \|U_1 - U_2\|, \\
&\left\| \frac{1}{\theta}\psi\big(\theta(H_{i_1,\dots,i_m} - U_1)\big) - \frac{1}{\theta}\psi\big(\theta(H_{i_1,\dots,i_m} - U_2)\big) \right\|_{\mathrm{F}} \\
&\quad \le \left\| \frac{1}{\theta}\big(\theta(H_{i_1,\dots,i_m} - U_1) - \theta(H_{i_1,\dots,i_m} - U_2)\big) \right\|_{\mathrm{F}} = \|U_1 - U_2\|_{\mathrm{F}}
\end{aligned}$$

by Fact 3. Since the convex combination of Lipschitz continuous functions is still Lipschitz continuous, the claim follows.

(3) Since $\widehat{U}_n^\star$ is the solution of the problem (3.3), the directional derivative

$$dF_\theta\big(\widehat{U}_n^\star; B\big) := \lim_{t\to 0} \frac{F_\theta(\widehat{U}_n^\star + tB) - F_\theta(\widehat{U}_n^\star)}{t} = \mathrm{tr}\big(\nabla F_\theta\big(\widehat{U}_n^\star\big) B\big)$$

is equal to 0 for any $B \in \mathbb{H}^d$. Result follows by taking consecutively $B_{i,j} = e_i e_j^T + e_j e_i^T, i \ne j$ and $B_{i,i} = e_i e_i^T, i = 1, \dots, d$, where $\{e_1, \dots, e_d\}$ is the standard Euclidean basis.

5.3. Proof of Theorem 3.1

The proof is based on the analysis of the gradient descent iteration for the problem (3.3). Let

$$G(U) := \mathrm{tr}\, F_\theta(U) = \mathrm{tr}\left[\frac{1}{\theta^2} \frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} \Psi\big(\theta\big(H(X_{i_1}, \dots, X_{i_m}) - U\big)\big) \right],$$

and define

$$\begin{aligned}
U_n^{(0)} &:= \mathbb{E}H = \mathbb{E}H(X_1, \dots, X_m), \\
U_n^{(j)} &:= U_n^{(j-1)} - \nabla G\big(U_n^{(j-1)}\big) \\
&= U_n^{(j-1)} + \frac{1}{\theta} \frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} \psi\big(\theta\big(H_{i_1\dots i_m} - U_n^{(j-1)}\big)\big), \quad j \ge 1,
\end{aligned}$$

which is the gradient descent for (3.3) with the step size equal to 1. We will show that with high probability (and for an appropriate choice of θ), $U_n^{(j)}$ does not escapes a small neighborhood of $\mathbb{E}H(X_1, \dots, X_m)$. The claim of the theorem then easily follows from this fact.

Give a permutation $(i_1, \dots, i_n) \in \pi_n$ and $U \in \mathbb{H}^d$, let $k = \lfloor n/m \rfloor$ and

$$Y_{i_1 \dots i_m}(U;\theta) := \psi\big(\theta(H_{i_1 \dots i_m} - U)\big),$$
$$W_{i_1 \dots i_n}(U;\theta) := \frac{1}{k}\big(Y_{i_1 \dots i_m}(U;\theta) + Y_{i_{m+1} \dots i_{2m}}(U;\theta) + \dots + Y_{i_{(k-1)m+1} \dots i_{km}}(U;\theta)\big).$$

Fact 6 implies that

$$\nabla G(U) = \frac{(n-m)!}{n!} \sum_{(i_1 \dots i_m) \in I_n^m} \frac{1}{\theta} \psi\big(\theta(H_{i_1 \dots i_m} - U)\big) = \frac{1}{n!} \sum_{(i_1 \dots i_n) \in \pi_n} \frac{1}{\theta} W_{i_1 \dots i_n}(U;\theta), \tag{5.4}$$

where π_n ranges over all permutations of $(1, \dots, n)$. Next, for $j \geq 1$ we have

$$\begin{aligned}
\big\| U_n^{(j)} - \mathbb{E}H \big\| &= \left\| \frac{1}{\theta} \frac{(n-m)!}{n!} \sum_{(i_1, \dots, i_m) \in I_n^m} \psi\big(\theta\big(H_{i_1 \dots i_m} - U_n^{(j-1)}\big)\big) - \big(\mathbb{E}H - U_n^{(j-1)}\big) \right\| \\
&= \left\| \frac{1}{\theta n!} \sum_{(i_1 \dots i_n) \in \pi_n} W_{i_1 \dots i_n}\big(U_n^{(j-1)};\theta\big) - \big(\mathbb{E}H - U_n^{(j-1)}\big) \right\| \\
&\leq \left\| \frac{1}{\theta n!} \sum_{(i_1 \dots i_n) \in \pi_n} \big(W_{i_1 \dots i_n}\big(U_n^{(j-1)};\theta\big) - W_{i_1, \dots, i_n}(\mathbb{E}H;\theta_\sigma)\big) - \big(\mathbb{E}H - U_n^{(j-1)}\big) \right\| \\
&\quad + \left\| \frac{1}{\theta_\sigma} \frac{1}{n!} \sum_{\pi_n} W_{i_1, \dots, i_n}(\mathbb{E}H;\theta_\sigma) \right\|.
\end{aligned} \tag{5.5}$$

The following two lemmas provide the bounds that allows to control the size of $\|U_n^{(j)} - \mathbb{E}H\|$. For a given $\sigma^2 \geq \|\mathbb{E}(H - \mathbb{E}H)^2\|$ and $\theta_\sigma = \frac{1}{\sigma}\sqrt{\frac{2t}{k}}$, consider the random variable

$$L_n(\delta) = \sup_{\|U - \mathbb{E}H\| \leq \delta} \left\| \frac{1}{\theta_\sigma} \frac{1}{n!} \sum_{\pi_n} \big(W_{i_1, \dots, i_n}(U;\theta_\sigma) - W_{i_1, \dots, i_n}(\mathbb{E}H;\theta_\sigma)\big) - (\mathbb{E}H - U) \right\|.$$

Lemma 5.1. *With probability at least* $1 - (2d+1)e^{-t}$, *for all* $\delta \leq \frac{1}{2}\frac{1}{\theta_\sigma}$ *simultaneously,*

$$L_n(\delta) \leq \left(\mathrm{r}_H \frac{26t}{k} + \frac{1}{2} \right)\delta + \frac{3(1+\sqrt{2})}{2} \sigma \sqrt{\frac{t}{k}}.$$

Moreover, the same bound holds with probability at least $1 - 14\mathrm{r}(\mathbb{E}(H - \mathbb{E}H)^2)e^{-t}$ *given that* $t \geq 1$.

The proof of this lemma is given in Section 5.4.

Lemma 5.2. *With probability at least* $1 - 2d \cdot e^{-t}$,

$$\left\| \frac{1}{\theta_\sigma} \frac{1}{n!} \sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H; \theta_\sigma) \right\| \le \frac{3}{\sqrt{2}} \sigma \sqrt{\frac{t}{k}}.$$

Moreover, the same bounds holds with probability at least $1 - \frac{8}{3}\mathrm{r}(\mathbb{E}(H - \mathbb{E}H)^2)e^{-t}$ *given that* $t \ge 1$.

The proof is given in Section 5.5. Next, define the sequence

$$\delta_0 = 0,$$
$$\delta_j = \left(\mathrm{r}_H \frac{26t}{k} + \frac{1}{2} \right) \delta_{j-1} + 5.75\sigma \sqrt{\frac{t}{k}}.$$

If $\mathrm{r}_H \frac{26t}{k} \le \frac{1}{4}$, then $t \le \frac{k}{104}$, hence $5.75\sigma\sqrt{\frac{t}{k}} \le \frac{1}{8}\frac{1}{\theta_\sigma}$ and

$$\delta_j \le \frac{3}{4}\delta_{j-1} + \frac{1}{8}\frac{1}{\theta_\sigma} \le \frac{1}{2}\frac{1}{\theta_\sigma}$$

for all $j \ge 0$. Let $\mathcal{E}_0$ be the event of probability at least $1 - (4d + 1)e^{-t}$ on which the inequalities of Lemmas 5.1 and 5.2 hold. It follows from (5.5), Lemma 5.1 and Lemma 5.2 that on the event $\mathcal{E}_0$, for all $j \ge 1$

$$\begin{aligned}\left\| U_n^{(j)} - \mathbb{E}H \right\| &\le L_n\left(\left\| U_n^{(j-1)} - \mathbb{E}H \right\| \right) + \left\| \frac{1}{\theta_\sigma} \frac{1}{n!} \sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H; \theta_\sigma) \right\| \\ &\le \left(\mathrm{r}_H \frac{26t}{k} + \frac{1}{2} \right) \delta_{j-1} + \frac{3(1 + 2\sqrt{2})}{2} \sigma \sqrt{\frac{t}{k}} \le \delta_j\end{aligned}$$

given that $\mathrm{r}_H \frac{26t}{k} \le \frac{1}{4}$; we have also used the numerical bound $\frac{3(1+2\sqrt{2})}{2} \le 5.75$.

Finally, it is easy to see that for all $j \ge 1$ and $\gamma = \mathrm{r}_H \frac{26t}{k} + \frac{1}{2} \le \frac{3}{4}$,

$$\delta_j = \delta_0 \gamma^j + \sum_{l=0}^{j-1} \gamma^l \cdot 5.75\sigma \sqrt{\frac{t}{k}} \le \sum_{l \ge 0} (3/4)^l \cdot 5.75\sigma \sqrt{\frac{t}{k}} \le 23\sigma \sqrt{\frac{t}{k}}. \tag{5.6}$$

Since $U_n^{(j)} \to \widehat{U}_n^\star$ pointwise as $j \to \infty$, the result follows. Alternatively, if the bounds of Lemmas 5.1 and 5.2 that depend on the effective rank $\mathrm{r}(\mathbb{E}(H - \mathbb{E}H)^2)$ are used instead, the second claim can be deduced in exactly similar way.

5.4. Proof of Lemma 5.1

Recall that $\sigma^2 \geq \|\mathbb{E}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2\|$, $\theta_\sigma := \frac{1}{\sigma}\sqrt{\frac{2t}{k}}$, and

$$\psi(\theta_\sigma x) = \begin{cases} \theta_\sigma x - \operatorname{sign}(x)\dfrac{\theta_\sigma^2 x^2}{2}, & x \in [-1/\theta_\sigma, 1/\theta_\sigma], \\ \dfrac{1}{2}\operatorname{sign}(x), & |x| > 1/\theta_\sigma. \end{cases}$$

The idea of the proof is to exploit the fact that $\psi(\theta_\sigma x)$ is "almost linear" whenever $x \in [-1/\theta_\sigma, 1/\theta_\sigma]$, and its nonlinear part is active only for a small number of multi-indices $(i_1, \dots, i_m) \in I_n^m$. Let

$$\chi_{i_1,\dots,i_m} = I\left\{\|H_{i_1,\dots,i_m} - \mathbb{E}H\| \leq \frac{1}{2\theta_\sigma}\right\}.$$

Note that by Chebyshev's inequality, and taking into account the fact that

$$\begin{aligned} \|H_{i_1,\dots,i_m} - \mathbb{E}H\| &\leq \|H_{i_1,\dots,i_m} - \mathbb{E}H\|_{\mathrm{F}}, \\ \Pr(\chi_{i_1,\dots,i_m} = 0) &\leq 4\theta_\sigma^2 \mathbb{E}\|H_{i_1,\dots,i_m} - \mathbb{E}H\|_{\mathrm{F}}^2 \\ &\leq \frac{8t}{k}\frac{\operatorname{tr}(\mathbb{E}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2)}{\|\mathbb{E}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2\|} = \mathrm{r}_H \frac{8t}{k}. \end{aligned} \tag{5.7}$$

Define the event

$$\mathcal{E} = \left\{ \sum_{(i_1,\dots,i_m)\in I_n^m} (1 - \chi_{i_1,\dots,i_m}) \leq \mathrm{r}_H \frac{8t}{k}\frac{n!}{(n-m)!} \cdot \left(1 + \sqrt{\frac{3}{8\mathrm{r}_H}}\right)\right\}.$$

We will apply a version of Chernoff bound to the $\mathbb{R}$-valued U-statistic

$$\frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} (1 - \chi_{i_1,\dots,i_m}).$$

A combination of Fact 6, Fact 4 applied in the scalar case $d = 1$, and Fact 5 implies that

$$\Pr\left(\frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} (1 - \chi_{i_1,\dots,i_m}) \geq \mathrm{r}_H \frac{8t}{k} \cdot (1 + \tau)\right) \leq e^{-\tau^2 8t\mathrm{r}_H/3} \tag{5.8}$$

for $0 < \tau < 1$. Indeed, applying decomposition (5.2) and Fact 6 to the scalar-valued $\tilde{H}_{i_1,\dots,i_m} := 1 - \chi_{i_1,\dots,i_m}$, we get for the corresponding $\tilde{W}_{i_1,\dots,i_n}$ that

$$\Pr\left(\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}(1-\chi_{i_1,\dots,i_m}) \geq \mathrm{r}_H\frac{8t}{k}\cdot(1+\tau)\right)$$
$$\leq \max_{(i_1,\dots,i_n)\in\pi_n}\left[\inf_{\theta>0}\mathbb{E}\exp\left(\theta\tilde{W}_{i_1,\dots,i_n} - \theta\mathrm{r}_H\frac{8t}{k}\cdot(1+\tau)\right)\right].$$

Followed by application of Fact 5, this implies (5.8). The choice of $\tau = \sqrt{\frac{3}{8\mathrm{r}_H}}$ yields that $\Pr(\mathcal{E}) \geq 1 - e^{-t}$.

By triangle inequality, whenever $\chi_{i_1,\dots,i_m} = 1$ and $\delta \leq \frac{1}{2}\frac{1}{\theta_\sigma}$, it holds that $\|H_{i_1,\dots,i_m} - U\| \leq \frac{1}{\theta_\sigma}$ for any U such that $\|U - \mathbb{E}H\| \leq \delta$, and consequently

$$\frac{1}{\theta_\sigma}\psi\big(\theta_\sigma(H_{i_1,\dots,i_m} - U)\big) = (H_{i_1,\dots,i_m} - U) - \frac{\theta_\sigma}{2}\operatorname{sign}(H_{i_1,\dots,i_m} - U)(H_{i_1,\dots,i_m} - U)^2.$$

Denoting

$$S_{i_1,\dots,i_m}(U) := \operatorname{sign}(H_{i_1,\dots,i_m} - U)(H_{i_1,\dots,i_m} - U)^2$$

for brevity, we deduce that

$$\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n}\big(W_{i_1,\dots,i_n}(U;\theta_\sigma) - W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\big) - (\mathbb{E}H - U)$$
$$= \frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}\left(\frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(\mathbb{E}H) - \frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(U)\right)\chi_{i_1,\dots,i_m}$$
$$+ \frac{1}{\theta_\sigma}\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}(1-\chi_{i_1,\dots,i_m})\big(Y_{i_1,\dots,i_m}(U;\theta_\sigma) - Y_{i_1,\dots,i_m}(\mathbb{E}H;\theta_\sigma)\big)$$
$$- \frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}(1-\chi_{i_1,\dots,i_m})(\mathbb{E}H - U).$$

We will separately control the terms on the right hand side of the equality above. First, note that on the event $\mathcal{E}$,

$$\left\|\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}(1-\chi_{i_1,\dots,i_m})(\mathbb{E}H - U)\right\| \leq \mathrm{r}_H\frac{8t}{k}\cdot\left(1+\sqrt{\frac{3}{8\mathrm{r}_H}}\right)\delta \leq \mathrm{r}_H\frac{13t}{k}\delta \tag{5.9}$$

since $\|\mathbb{E}H - U\| \leq \delta$. Next, recalling that $\psi(\cdot)$ is operator Lipschitz (by Fact 3), wee see that for any $(i_1,\dots,i_m) \in I_n^m$

$$\frac{1}{\theta_\sigma}\big\|Y_{i_1,\dots,i_m}(U;\theta_\sigma) - Y_{i_1,\dots,i_m}(\mathbb{E}H;\theta_\sigma)\big\| \leq \|\mathbb{E}H - U\| \leq \delta,$$

hence on event $\mathcal{E}$,

$$\frac{1}{\theta_\sigma}\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}(1-\chi_{i_1,\dots,i_m})\big(Y_{i_1,\dots,i_m}(U;\theta_\sigma)-Y_{i_1,\dots,i_m}(\mathbb{E}H;\theta_\sigma)\big)\right\|$$
$$\le \mathrm{r}_H\frac{8t}{k}\cdot\left(1+\sqrt{\frac{3}{8\mathrm{r}_H}}\right)\delta\le \mathrm{r}_H\frac{13t}{k}\delta. \tag{5.10}$$

Finally, it remains to control the term

$$\mathcal{Q}(\delta):=\sup_{\|U-\mathbb{E}H\|\le\delta}\left\|\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}\left(\frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(\mathbb{E}H)-\frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(U)\right)\chi_{i_1,\dots,i_m}\right\|.$$

Lemma 5.3. *With probability at least* $1-2de^{-t}$,

$$\mathcal{Q}(\delta)\le\frac{3(1+\sqrt{2})}{2}\sigma\sqrt{\frac{t}{k}}+\frac{\delta}{2}.$$

Moreover, the same bound holds with probability at least $1-14\mathrm{r}(\mathbb{E}(H-\mathbb{E}H)^2)e^{-t}$ *for* $t\ge 1$.

Proof. Observe that for all $U\in\mathbb{H}^d$ and $(i_1,\dots,i_m)\in I_n^m$,

$$-(H_{i_1,\dots,i_m}-U)^2\preceq\operatorname{sign}(H_{i_1,\dots,i_m}-U)(H_{i_1,\dots,i_m}-U)^2\preceq(H_{i_1,\dots,i_m}-U)^2,$$

hence

$$\left\|\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}\left(\frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(\mathbb{E}H)-\frac{\theta_\sigma}{2}S_{i_1,\dots,i_m}(U)\right)\chi_{i_1,\dots,i_m}\right\|$$
$$\le\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}\frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m}-U)^2\chi_{i_1,\dots,i_m}\right\|$$
$$+\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}\frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m}-\mathbb{E}H)^2\chi_{i_1,\dots,i_m}\right\|.$$

Moreover,

$$(H_{i_1,\dots,i_m}-U)^2\preceq 2(H_{i_1,\dots,i_m}-\mathbb{E}H)^2+2(U-\mathbb{E}H)^2,$$

implying that

$$\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}\frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m}-U)^2\chi_{i_1,\dots,i_m}\right\|$$
$$\le 2\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}\frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m}-\mathbb{E}H)^2\chi_{i_1,\dots,i_m}\right\|+\theta_\sigma\|U-\mathbb{E}H\|^2.$$

Hence, we have shown that

$$\mathcal{Q}(\delta) \le 3\frac{(n-m)!}{n!}\left\| \sum_{(i_1,\dots,i_m)\in I_n^m} \frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m} \right\| + \theta_\sigma \delta^2. \tag{5.11}$$

Since $\delta \le \frac{1}{2\theta_\sigma}$,

$$\theta_\sigma \delta^2 \le \frac{\delta}{2}. \tag{5.12}$$

Next, we will estimate the first term in (5.11) as follows:

$$\begin{aligned} &3\frac{(n-m)!}{n!}\left\| \sum_{(i_1,\dots,i_m)\in I_n^m} \frac{\theta_\sigma}{2}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m} \right\| \\ &\quad \le 3\frac{(n-m)!}{n!}\left\| \sum_{(i_1,\dots,i_m)\in I_n^m} \frac{\theta_\sigma}{2}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m} - \mathbb{E}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big]\big] \right\| \\ &\qquad + \frac{3\theta_\sigma}{2}\big\|\mathbb{E}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big]\big\|. \end{aligned}$$

Clearly, $\|\mathbb{E}[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}]\| \le \sigma^2$, hence

$$\frac{3\theta_\sigma}{2}\big\|\mathbb{E}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big]\big\| \le \frac{3\sigma}{2}\sqrt{\frac{2t}{k}}. \tag{5.13}$$

The remaining part will be estimated using the Matrix Bernstein's inequality (Fact 7).

To this end, note that by the definition of $\chi_{i_1,\dots,i_m}$,

$$\big\|(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m} - \mathbb{E}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big]\big\| \le \left(\frac{1}{2\theta_\sigma}\right)^2$$

almost surely. Moreover,

$$\begin{aligned} &\big\|\mathbb{E}\big((H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m} - \mathbb{E}\big[(H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big]\big)^2\big\| \\ &\quad \le \big\|\mathbb{E}\big((H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big)^2\big\| \le \left(\frac{1}{2\theta_\sigma}\right)^2 \big\|\mathbb{E}(H_{i_1,\dots,i_m} - \mathbb{E}H)^2\big\|, \end{aligned}$$

where we used the fact that

$$\big((H_{i_1,\dots,i_m} - \mathbb{E}H)^2 \chi_{i_1,\dots,i_m}\big)^2 \preceq \left(\frac{1}{2\theta_\sigma}\right)^2 (H_{i_1,\dots,i_m} - \mathbb{E}H)^2.$$

Applying the Matrix Bernstein inequality (Fact 7), we get that with probability at least $1-2de^{-t}$

$$3\frac{(n-m)!}{n!}\left\|\sum_{(i_1,\dots,i_m)\in I_n^m}\frac{\theta_\sigma}{2}\left[(H_{i_1,\dots,i_m}-\mathbb{E}H)^2\chi_{i_1,\dots,i_m}-\mathbb{E}(H_{i_1,\dots,i_m}-\mathbb{E}H)^2\chi_{i_1,\dots,i_m}\right]\right\|$$
$$\le\frac{3\theta_\sigma}{2}\left[\frac{2}{2\theta_\sigma}\left\|\mathbb{E}(H_{i_1,\dots,i_m}-\mathbb{E}H)^2\right\|^{1/2}\sqrt{\frac{t}{k}}\bigvee\frac{4}{3}\frac{t}{k}\frac{1}{(2\theta_\sigma)^2}\right]\le\frac{3}{2}\sigma\sqrt{\frac{t}{k}}. \tag{5.14}$$

Alternatively, the second claim of Fact 7 implies that the same upper bound holds with probability at least $1-14\mathrm{r}(\mathbb{E}(H-\mathbb{E}H)^2)e^{-t}$ for $t\ge 1$.

The bound of Lemma 5.3 now follows from the combination of bounds (5.12), (5.13), (5.14) and (5.11).

Combining the bound of Lemma 5.3 with (5.9) and (5.10), we get the desired result of Lemma 5.1. □

5.5. Proof of Lemma 5.2

Fact 4 implies that for all $s>0$,

$$\Pr\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n}W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right)\ge s\right)\le\inf_{\theta>0}\left[e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,e^{(\theta/\theta_\sigma)W_{1,\dots,n}(\mathbb{E}H,\theta_\sigma)}\right]$$
$$\le e^{-\theta_\sigma sk}\mathbb{E}\,\mathrm{tr}\,e^{kW_{1,\dots,n}(\mathbb{E}H,\theta_\sigma)}. \tag{5.15}$$

Since

$$W_{1,\dots,n}(\mathbb{E}H,\theta_\sigma)=\frac{1}{k}\left(\psi\left(\theta_\sigma(H_{1,\dots,m}-\mathbb{E}H)\right)+\dots+\psi\left(\theta_\sigma(H_{(k-1)m+1,\dots,km}-\mathbb{E}H)\right)\right)$$

is a sum of k independent random matrices, we can apply the first inequality of Fact 8 to deduce that

$$\mathbb{E}\,\mathrm{tr}\,e^{kW_{1,\dots,n}(\mathbb{E}H,\theta_\sigma)}\le\mathrm{tr}\exp\left(k\theta_\sigma^2\mathbb{E}(H-\mathbb{E}H)^2\right)\le d\exp\left(k\theta_\sigma^2\sigma^2\right),$$

where we used the fact that $\mathrm{tr}(A)\le d\|A\|$ for $\mathbb{H}^{d\times d}\ni A\succeq 0$. Finally, setting $s=\frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k}}$, we obtain from (5.15) that

$$\Pr\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n}W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right)\ge s\right)\le de^{-t}.$$

Similarly, since $-\lambda_{\min}(A) = \lambda_{\max}(-A)$ for $A \in \mathbb{H}^{d\times d}$, it follows from the second inequality of Fact 8 that

$$\begin{aligned}
&\Pr\left(\lambda_{\min}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \le -s\right) \\
&\quad = \Pr\left(\lambda_{\max}\left(-\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \ge s\right) \\
&\quad \le e^{-\theta_\sigma sk}\,\mathbb{E}\,\mathrm{tr}\exp\bigl(kW_{1,\dots,n}(\mathbb{E}H,\theta_\sigma)\bigr) \\
&\quad \le de^{-\theta_\sigma sk}\exp\bigl(k\theta_\sigma^2\sigma^2\bigr) \le de^{-t}
\end{aligned}$$

for $s = \frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k}}$.

To establish the second bound that depends only on the effective rank of $\mathbb{E}(H-\mathbb{E}H)^2$, we modify the argument as follows: let $\phi(x) = \max(e^x - 1, 0)$, and note that $\phi(x)$ is convex and nonnegative. Fix $\tau > 0$, and note that for all $s > 0$,

$$\begin{aligned}
&\Pr\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \ge s\right) \\
&\quad = \Pr\left(\phi\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{\tau}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right)\right) \ge \phi(\tau s)\right) \\
&\quad \le \frac{1}{\phi(\tau s)}\mathbb{E}\,\mathrm{tr}\,\phi\left(\frac{\tau}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \\
&\quad \le \frac{1}{\phi(\tau s)}\mathbb{E}\,\mathrm{tr}\,\phi\left(\frac{\tau}{\theta_\sigma}W_{1,\dots,n}(\mathbb{E}H;\theta_\sigma)\right) \\
&\quad = \frac{1}{\phi(\tau s)}\mathbb{E}\left(\mathrm{tr}\exp\left(\frac{\tau}{\theta_\sigma}W_{1,\dots,n}(\mathbb{E}H;\theta_\sigma)\right) - I\right).
\end{aligned}$$

Setting $\tau := \theta_\sigma \cdot k$, we obtain that

$$\Pr\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \ge s\right) \le \frac{1}{e^{\theta_\sigma sk} - 1}\mathbb{E}\bigl(\mathrm{tr}\exp\bigl(kW_{1,\dots,n}(\mathbb{E}H;\theta_\sigma)\bigr) - I\bigr).$$

As before, Fact 8 implies that $\mathbb{E}\,\mathrm{tr}\exp(kW_{1,\dots,n}(\mathbb{E}H;\theta_\sigma)) \le \mathrm{tr}\exp(k\theta_\sigma^2\mathbb{E}(H-\mathbb{E}H)^2)$. It is easy to show that (see the proof of Theorem 3.2 in [34]) that

$$\mathrm{tr}\exp\bigl(k\theta_\sigma^2\mathbb{E}(H-\mathbb{E}H)^2\bigr) - I \le \mathrm{r}\bigl(\mathbb{E}(H-\mathbb{E}H)^2\bigr)\bigl(\exp\bigl(k\theta_\sigma^2\bigl\|\mathbb{E}(H-\mathbb{E}H)^2\bigr\|\bigr) - 1\bigr).$$

Note that the factor in front of the exponent is the effective rank of $\mathbb{E}(H - \mathbb{E}H)^2$ instead of dimension d. Hence,

$$\begin{aligned}\Pr\left(\lambda_{\max}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \geq s\right) \\ \leq \frac{\mathrm{r}(\mathbb{E}(H-\mathbb{E}H)^2)}{e^{\theta_\sigma sk}-1}\exp\bigl(k\theta_\sigma^2\sigma^2\bigr) \\ \leq \mathrm{r}\bigl(\mathbb{E}(H-\mathbb{E}H)^2\bigr)\exp\bigl(k\theta_\sigma^2\sigma^2 - \theta_\sigma sk\bigr)\left(1+\frac{1}{\theta_\sigma sk}\right),\end{aligned}$$

where we used the inequality $\frac{e^x}{e^x-1} \leq 1 + \frac{1}{x}$ that holds for positive x in the last step. For $s = \frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k}}$, the right-hand side of the previous bound is equal to $\mathrm{r}(\mathbb{E}(H - \mathbb{E}H)^2)(1 + \frac{1}{3t})e^{-t}$, where we used the fact that $\theta_\sigma = \frac{1}{\sigma}\sqrt{\frac{2t}{k}}$. If $t \geq 1$, the latter is bounded from above by $\frac{4}{3}\mathrm{r}(\mathbb{E}(H - \mathbb{E}H)^2)e^{-t}$. Similar argument yields that

$$\Pr\left(\lambda_{\min}\left(\frac{1}{\theta_\sigma}\frac{1}{n!}\sum_{\pi_n} W_{i_1,\dots,i_n}(\mathbb{E}H;\theta_\sigma)\right) \leq -\frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k}}\right) \leq \mathrm{r}\bigl(\mathbb{E}(H-\mathbb{E}H)^2\bigr)\left(1+\frac{1}{3t}\right)e^{-t},$$

concluding the proof.

5.6. Proof of Lemma 3.2

Part (a) follows from a well-known result (e.g., [5]) which states that, given a convex, differentiable function $G : \mathbb{R}^D \to \mathbb{R}$ such that its gradient satisfies $\|\nabla G(U_1) - \nabla G(U_2)\|_2 \leq L\|U_1 - U_2\|_2$, the j-th iteration $U^{(j)}$ of the gradient descent algorithm run with step size $\alpha \leq \frac{1}{L}$ satisfies

$$G\bigl(U^{(j)}\bigr) - G(U_*) \leq \frac{\|U^{(0)} - U_*\|_2^2}{2\alpha j},$$

where $U_* = \operatorname{argmin} G(U)$.

The proof of part (b) follows the lines of the proof of Theorem 3.1: more specifically, the claim follows from equation (5.6).

5.7. Proof of Corollary 3.1

Note that

$$\bigl\|\mathbb{E}\mathcal{D}(H_{i_1\dots i_m})^2\bigr\| = \max\bigl(\bigl\|\mathbb{E}H_{i_1\dots i_m}H_{i_1\dots i_m}^*\bigr\|, \bigl\|\mathbb{E}H_{i_1\dots i_m}^*H_{i_1\dots i_m}\bigr\|\bigr).$$

We apply Theorem 3.1 applied to self-adjoint random matrices

$$\mathcal{D}(H_{i_1\dots i_m}) \in \mathbb{C}^{(d_1+d_2)\times(d_1+d_2)},\ (i_1,\dots,i_m) \in I_n^m,$$

and obtain that

$$\left\|\bar{U}_n^\star - \mathcal{D}(\mathbb{E}H)\right\| \le 15\sigma\sqrt{\frac{t}{k}}$$

with probability $\ge 1 - \min(4(d_1+d_2), 18\tilde{r}_H)e^{-t}$. It remains to apply Fact 9:

$$\begin{aligned}\left\|\bar{U}_n^\star - \mathcal{D}(\mathbb{E}H)\right\| &= \left\|\begin{pmatrix} \hat{U}_{11}^\star & \hat{U}_{12}^\star - \mathbb{E}H \\ (\hat{U}_{12}^\star)^* - \mathbb{E}H^* & \hat{U}_{22}^\star \end{pmatrix}\right\| \\ &\ge \left\|\begin{pmatrix} 0 & \hat{U}_{12}^\star - \mathbb{E}H \\ (\hat{U}_{12}^\star)^* - \mathbb{E}H^* & 0 \end{pmatrix}\right\| = \left\|\hat{U}_{12}^\star - \mathbb{E}H\right\|,\end{aligned}$$

and the claim follows.

5.8. Proof of Lemma 4.1

Recall that $\mu = \mathbb{E}Y$.

(a) Observe that

$$\begin{aligned}\left\|\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\right\| &= \sup_{\|v\|_2=1} \mathbb{E}\langle v, Y-\mu\rangle^2 \|Y-\mu\|_2^2 \\ &= \sup_{\|v\|_2=1} \left[\sum_{j=1}^d \langle v, Y-\mu\rangle^2 \big(Y^{(j)} - \mu^{(j)}\big)^2\right].\end{aligned}$$

Next, for $j = 1, \dots, d$,

$$\begin{aligned}\mathbb{E}\langle v, Y-\mu\rangle^2\big(Y^{(j)}-\mu^{(j)}\big)^2 &\le \mathbb{E}^{1/2}\langle v, Y-\mu\rangle^4 \mathbb{E}^{1/2}\big(Y^{(j)}-\mu^{(j)}\big)^4 \\ &\le K\mathbb{E}\langle v, Y-\mu\rangle^2 \mathbb{E}\big(Y^{(j)}-\mu^{(j)}\big)^2,\end{aligned}$$

hence

$$\left\|\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\right\| \le K \sup_{\|v\|_2=1} \mathbb{E}\langle v, Y-\mu\rangle^2 \sum_{j=1}^d \mathbb{E}\big(Y^{(j)}-\mu^{(j)}\big)^2,$$

and the result follows.

(b) Note that

$$\begin{aligned}&\mathrm{tr}\big[\mathbb{E}\big((Y-\mu)(Y-\mu)^T\big)^2\big] \\ &\quad= \sum_{j=1}^d \mathbb{E}\big(Y^{(j)}-\mu^{(j)}\big)^2 \|Y-\mu\|_2^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{d} \mathbb{E}\big(Y^{(j)} - \mu^{(j)}\big)^4 + \sum_{i \neq j} \mathbb{E}\big[\big(Y^{(i)} - \mu^{(i)}\big)^2 \big(Y^{(j)} - \mu^{(j)}\big)^2\big] \\
&\le \sum_{j=1}^{d} \mathbb{E}\big(Y^{(j)} - \mu^{(j)}\big)^4 + \sum_{i \neq j} \mathbb{E}^{1/2}\big(Y^{(i)} - \mu^{(i)}\big)^4 \mathbb{E}^{1/2}\big(Y^{(j)} - \mu^{(i)}\big)^4 \\
&= \left(\sum_{j=1}^{d} \mathbb{E}^{1/2}\big(Y^{(j)} - \mu^{(j)}\big)^4\right)^2 \le K' \left(\sum_{j=1}^{d} \mathbb{E}\big(Y^{(j)} - \mu^{(j)}\big)^2\right)^2 = K'\big(\mathrm{tr}(\Sigma)\big)^2.
\end{aligned}
$$

(c) The inequality follows from Corollary 5.1 in [35].

5.9. Proof of Corollary 4.2

It is easy to see ((e.g., see the proof of Theorem 1 in [31]) that $\widetilde{\Sigma}_{\star}^{\tau}$ can be equivalently represented as

$$
\widetilde{\Sigma}_{\star}^{\tau} = \underset{S \in \mathbb{R}^{d \times d}, S = S^T}{\operatorname{argmin}} \big[\|S - \widetilde{\Sigma}_{\star}\|_{\mathrm{F}}^2 + \tau \|S\|_1\big]. \tag{5.16}
$$

The remaining proof is based on the following lemma.

Lemma 5.4. *Inequality* (4.4) *holds on the event* $\mathcal{E} = \{\tau \ge 2\|\widetilde{\Sigma}_{\star}^{\tau} - \Sigma\|\}$.

To verify this statement, it is enough to repeat the steps of the proof of Theorem 1 in [31], replacing each occurrence of the sample covariance $\hat{S}_{2n}$ by its robust counterpart $\widetilde{\Sigma}_{\star}^{\tau}$. Result of Corollary 4.2 then follows from the combination of Theorem 3.2 and Lemma 4.1 which imply that

$$
\Pr(\mathcal{E}) \ge 1 - \min\big(4d + 1, 18K'\mathrm{r}(\Sigma)\big)e^{-t}
$$

whenever $\tau \ge \gamma \cdot 138\sqrt{K}\|\Sigma\|\sqrt{\frac{\mathrm{r}(\Sigma)(t+\Xi)}{\lfloor n/2 \rfloor}}$.

Acknowledgements

Authors gratefully acknowledge support by the National Science Foundation grants DMS-1712956 and CCF-1908905. We appreciate the insightful comments of the Referees and the Associate Editor that helped to improve some results and their presentation.

References

[1] Aleksandrov, A.B. and Peller, V.V. (2016). Operator Lipschitz functions. *Russian Math. Surveys* **71** 605. MR3588921 https://doi.org/10.4213/rm9729

[2] Alon, N., Matias, Y. and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing* (*Philadelphia*, *PA*, 1996) 20–29. New York: ACM. MR1427494 https://doi.org/10.1145/237814.237823

[3] Angluin, D. and Valiant, L.G. (1979). Fast probabilistic algorithms for Hamiltonian circuits and matchings. *J*. *Comput*. *System Sci*. **18** 155–193. MR0532174 https://doi.org/10.1016/0022-0000(79)90045-X

[4] Arcones, M.A. and Giné, E. (1993). Limit theorems for U-processes. *Ann*. *Probab*. **21** 1494–1542. MR1235426

[5] Bertsekas, D.P. (2009). *Convex Optimization Theory*. Nashua, NH: Athena Scientific. MR2830150

[6] Bhatia, R. (2013). *Matrix Analysis* **169**. Springer Science & Business Media. MR1477662 https://doi.org/10.1007/978-1-4612-0653-8

[7] Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J*. *ACM* **58** Art. 11, 37. MR2811000 https://doi.org/10.1145/1970392.1970395

[8] Carlen, E. (2010). Trace inequalities and quantum entropy: An introductory course. In *Entropy and the Quantum*. *Contemp*. *Math*. **529** 73–140. Providence, RI: Amer. Math. Soc. MR2681769 https://doi.org/10.1090/conm/529/10428

[9] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann*. *Inst*. *Henri Poincaré Probab*. *Stat*. **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[10] Catoni, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. arXiv preprint arXiv:1603.05229.

[11] Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. arXiv preprint arXiv:1712.02747.

[12] Chen, R.Y., Gittens, A. and Tropp, J.A. (2012). The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Inf*. *Inference* **1** 2–20. MR3311439 https://doi.org/10.1093/imaiai/ias001

[13] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R.I. (2016). Sub-Gaussian mean estimators. *Ann*. *Statist*. **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

[14] de la Peña, V.H. and Giné, E. (1999). *Decoupling*: *From Dependence to Independence*, *Randomly Stopped Processes*. *U-Statistics and Processes*. *Martingales and Beyond*. *Probability and Its Applications* (*New York*). New York: Springer. MR1666908 https://doi.org/10.1007/978-1-4612-0537-1

[15] Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2017). FARM-Test: factor-adjusted robust multiple testing with false discovery control. arXiv preprint arXiv:1711.05386.

[16] Fan, J., Wang, W. and Zhu, Z. (2016). Robust low-rank matrix recovery. arXiv preprint arXiv:1603.08315.

[17] Giné, E., Latała, R. and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability*, *II* (*Seattle*, *WA*, 1999). *Progress in Probability* **47** 13–38. Boston, MA: Birkhäuser. MR1857312

[18] Giulini, I. (2016). Robust principal component analysis in Hilbert spaces. arXiv preprint arXiv:1606.00187.

[19] Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* **23** 23–57. MR3556765 https://doi.org/10.3150/15-BEJ702

[20] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann*. *Math*. *Stat*. **19** 293–325. MR0026294 https://doi.org/10.1214/aoms/1177730196

[21] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J*. *Amer*. *Statist*. *Assoc*. **58** 13–30. MR0144363

[22] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann*. *Math*. *Stat*. **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[23] Huber, P.J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science* 1248–1251. Springer.

[24] Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.* **23** 92–119. MR2431867 https://doi.org/10.1214/088342307000000087

[25] Joly, E. and Lugosi, G. (2016). Robust estimation of U-statistics. *Stochastic Process. Appl.* **126** 3760–3773. MR3565476 https://doi.org/10.1016/j.spa.2016.04.021

[26] Kabanava, M. and Rauhut, H. (2017). Masked Toeplitz covariance estimation. arXiv preprint arXiv:1709.09377.

[27] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[28] Lepskiĭ, O.V. (1991). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697. MR1147167 https://doi.org/10.1137/1136085

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

[30] Levina, E. and Vershynin, R. (2012). Partial estimation of covariance matrices. *Probab. Theory Related Fields* **153** 405–419. MR2948681 https://doi.org/10.1007/s00440-011-0349-4

[31] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437 https://doi.org/10.3150/12-BEJ487

[32] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

[33] Minsker, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statist. Probab. Lett.* **127** 111–119. MR3648301 https://doi.org/10.1016/j.spl.2017.03.020

[34] Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

[35] Minsker, S. and Wei, X. Estimation of the covariance structure of heavy-tailed distributions. In *NIPS* 2017.

[36] Nemirovsky, A.S. and Yudin, D.B. (1983). *Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics*. New York: Wiley. Translated from the Russian and with a preface by E.R. Dawson. MR0702836

[37] Polyak, B.T. and Khlebnikov, M.V. (2017). Principle component analysis: Robust versions. *Autom. Remote Control* **78** 490–506. MR3666090

[38] Tropp, J.A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459 https://doi.org/10.1007/s10208-011-9099-z

[39] Tyler, D.E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734 https://doi.org/10.1214/aos/1176350263

[40] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[41] Wegkamp, M. and Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812 https://doi.org/10.3150/14-BEJ690

[42] Zhang, T., Cheng, X. and Singer, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. *J. Multivariate Anal.* **149** 114–123. MR3507318 https://doi.org/10.1016/j.jmva.2016.03.010

Received February 2018 and revised July 2019