Electronic Journal of Statistics
Vol. 18 (2024) 1130–1159
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2224

Dimension-free bounds for sums of dependent matrices and operators with heavy-tailed distributions

Shogo Nakakita

The University of Tokyo
e-mail: nakakita@g.ecc.u-tokyo.ac.jp

Pierre Alquier

ESSEC Business School, Asia-Pacific campus, Singapore
e-mail: alquier@essec.edu

Masaaki Imaizumi

The University of Tokyo, RIKEN Center for Advanced Intelligence Project
e-mail: imaizumi@g.ecc.u-tokyo.ac.jp

Abstract: We prove deviation inequalities for sums of high-dimensional random matrices and operators with dependence and heavy tails. Estimation of high-dimensional matrices is a concern for numerous modern applications. However, most results are stated for independent observations. Therefore, it is critical to derive results for dependent and heavy-tailed matrices. In this paper, we derive a dimension-free upper bound on the deviation of the sums. Thus, the bound does not depend explicitly on the dimension of the matrices but rather on their effective rank. Our result generalizes several existing studies on the deviation of sums of matrices. It relies on two techniques: (i) a variational approximation of the dual of moment generating functions, and (ii) robustification through the truncation of the eigenvalues of the matrices. We reveal that our results are applicable to several problems, such as covariance matrix estimation, hidden Markov models, and overparameterized linear regression.

MSC2020 subject classifications: Primary 62H12; secondary 40C05, 60G51.
Keywords and phrases: Dependent process, heavy-tailed distribution, high-dimension, random matrix.

Received August 2023.

Contents

arXiv: 2210.09756

1. Introduction

We study non-asymptotic upper bounds on the deviations of the sums of multiple random matrices (or operators) from its expectation. Assume that we observe a sequence of n random, symmetric matrices $M_1, \dots, M_n$ that are potentially high-dimensional, dependent, and heavy-tailed, but have a common expectation $\Sigma := \mathbb{E}[M_\ell]$. We are interested in evaluating the deviation of their empirical mean from the expectation Σ, measured in terms of the operator norm $\|\cdot\|$ for matrices. Specifically, we want to derive an upper bound of the following value for each integer n:

$$\left\| \frac{1}{n} \sum_{\ell=1}^{n} M_\ell - \Sigma \right\|.$$

This problem is foundational and important; moreover, it has a variety of applications, the most typical example being the estimation of covariance matrices. Let $Y_1, \dots, Y_n$ be a sequence of random vectors; then, we can estimate its covariance matrix $\Sigma = \mathbb{E}[Y_1 Y_1^\top]$ using the empirical mean $n^{-1}\sum_{\ell=1}^{n} M_\ell$ by defining $M_\ell = Y_\ell Y_\ell^\top$. This setup can easily be applied for estimating Fisher information matrices, for example. Other applications include estimation of adjacency matrices of random graphs [26], signal recovery in compressed sensing [14], and linear regression under overparameterization [6]. Considering the increasing variety of data in modern data science, it is expedient to study the upper bound in various settings, including dependent or heavy-tailed observations M_ℓ.

This problem has been actively investigated in various directions. The first study [28] derived upper bounds on the operator norm of the deviation. In the high-dimensional case, several studies [7, 25, 29, 21] derived upper bounds that do not depend on the dimensionality of the matrices M_ℓ, referred to as *dimension-free* bounds, using instead the effective rank of the matrices. Particularly, another study [16] investigated an infinite-dimensional version of the problem. These results enable us to estimate high-dimensional matrices without assuming sparsity [8] or specifying the distribution of the matrix [2, 17, 32]. The exact asymptotic risk is also studied by [19]. A bootstrap method and dimension-free bound are developed by [24] for high-dimensional operators in this setting. In the case of heavy-tailed matrices M_ℓ, a study [23] derived a dimension-free upper bound that clarifies how the tail property affects the bound. The tightness of the bound is further improved by [31, 20] in the heavy-tailed setting. In the case of dependent matrices, a study [18] derived a bound in expectation, following the approach of [30].

1.1. Focus and result

We aim to derive a dimension-free upper bound on the deviations of the empirical mean of random matrices that are dependent and heavy-tailed. This setting is a generalization of the aforementioned studies. To handle the setup, we first introduce notation and assumptions. Let $\mathbb{H}$ be a Hilbert space. The first time, we only consider $\mathbb{H} = \mathbb{R}^p$. We then extend the results to an infinite-dimensional $\mathbb{H}$.

Let M be a symmetric linear operator from $\mathbb{H}$ to itself. In dimension-free bounds, the dimension of $\mathbb{H}$ is replaced by the *effective rank*, as shown by [21]. It is defined as follows:

Definition 1 (Effective Rank)**.** For a symmetric positive semi-definite trace-class operator $M : \mathbb{H} \to \mathbb{H}$, the effective rank is defined as

$$\mathbf{r}(M) := \frac{\mathrm{Tr}(M)}{\|M\|},$$

where $\mathrm{Tr}(M)$ denotes the trace of M and $\|M\|$ is its operator norm.

It can be interpreted as measuring the effective dimension of the image of M, which can be smaller than the actual dimension of $\mathbb{H}$.

To measure the dependence of a sequence of matrices $M_1, \dots, M_n$, we consider a coefficient $\Gamma_{\ell,n}$ for $\ell = 1, \dots, n$ that bounds the martingale increment

$$|\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{F}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)]| \leq \Gamma_{\ell,n},$$

for any Lipschitz-continuous function $g(.)$, where $\mathcal{F}_\ell = \sigma(M_1, \dots, M_\ell)$ is the σ-algebra generated by $M_1, \dots, M_\ell$. The formal definition of the coefficient from [27, 12], which has been used in many papers on dependent variables, is provided below. This coefficient leads to a general notion of dependence that includes many dependent processes, such as causal Bernoulli shifts and chains with infinite memory. We discuss this point below.

We introduce functions $S_\ell : \mathbb{R}_+ \to \mathbb{R}_+$ for $\ell = 1, \dots, n$ to measure the tail of the distribution of $\|M_\ell\|$ as $S_\ell(t) = \mathbb{P}(\|M_\ell\| \geq t)$. We also define

$$G(t) := \max_{1 \leq \ell \leq n} \int_t^\infty S_\ell(u) \mathrm{d}u.$$

Our general bounds are given in terms of $S_\ell(\cdot)$ and $G(\cdot)$. We can then study how the tail probability of $\|M_\ell\|$ affects the order of magnitude of the upper bound. Particularly, when $\|M_\ell\|$ is bounded, we obtain the rates proven in [21, 32] for independent observations, but in a broader dependent framework. In the case where $S_\ell(t)$ and $G(t)$ decay exponentially fast in t, we recover the same rates up to a $\log n$ factor. Our results also provide a rate of convergence in the case where $G(t)$ decays polynomially in t (in this case, the rate is slower).

Our main result takes the form of an upper bound in probability on the deviation of the empirical mean of the matrices. Therefore, for any $t, \tau > 0$, the following inequality holds with probability at least $1 - \exp(-t) - \sum_{\ell=1}^n S_\ell(\tau)$:

$$\left\| \frac{1}{n} \sum_{\ell=1}^n M_\ell - \Sigma \right\| \leq 2\sqrt{2} \, \|\Sigma\| \left(2\tau + \max_{\ell=1,\dots,n} \Gamma_{\ell,n} \right) \sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}} + G(\tau).$$

The results suggest that (i) we can obtain a dimension-free upper bound under quite general conditions of heavy-tail and dependence; (ii) the dependence property affects the bound through a factor $\Gamma_{\ell,n}$; and (iii) the heavy-tail property appears to affect the bound via a factor 2τ and an additional term $G(\tau)$, where τ is a free parameter that can be adjusted to balance both terms. Particularly, even under a slow, polynomial decay of $G(\tau)$, we can still select $\tau = \tau_n$ such that $\sum_{\ell=1}^n S_\ell(\tau) = o(1)$ and $G(\tau) = o(1)$ and then obtain an upper bound that converges to 0, but at a rate slower than $1/\sqrt{n}$.

From a technical perspective, this study makes two contributions. The first is the evaluation of the moment-generating function using a variational inequality, following [11]. We provide an upper bound the deviation of the sum of matrices using its moment-generating function. This approach was employed by [32] and others. We extend it to our setting with dependent random matrices using an inequality due to [27]. The second is the truncation technique that addresses heavy tails. This technique is classical in addressing unbounded losses in machine learning and has been used in the context of time series by [4]. It is also related to the influence function used in works on robust statistics [10, 11, 32, 1]. We apply this technique in our setting by truncating the eigenvalues of the dependent random matrices. Specifically, we control the effect of the heavy tails by decomposing the deviation of the empirical mean into two parts: the deviations of the truncated mean and the deviations between the truncated and standard means.

1.2. Organization

Section 2 introduces the setting of the problem and also provides assumptions and examples of situations where they are satisfied. Section 3 presents the main results. We begin with the case of dependent but bounded matrices in Theorem 4. We then extend this result to the unbounded case in Corollary 5. Section 4 describes several applications in which we apply our bound. Section 5 contains the proofs of the main results. Section 6 concludes the paper. The appendix provides the rest of the proofs.

1.3. Notation

Let $\mathbb{H}$ be a Hilbert space equipped with the scalar product $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ be the corresponding norm. Let $\mathcal{S}$ be the set of symmetric linear continuous operators $\mathbb{H}\to\mathbb{H}$, that is, for any $(u,v)\in\mathbb{H}^2$, and for any $M\in\mathcal{S}$, $\langle Mu,v\rangle=\langle u,Mv\rangle<\infty$. In the special case $\mathbb{H}=\mathbb{R}^p$, $\mathcal{S}$ is simply the set of symmetric matrices. For any $M\in\mathcal{S}$ we let $\|M\|$ denote its operator norm $\|M\|=\sup_{u\in\mathbb{H},\|u\|=1}\|Mu\|$. Throughout this paper, $\mathbf{M}=(M_\ell)_{\ell=1,\dots,n}$ is a finite random sequence of elements of $\mathcal{S}$, whose expectation is constant with $\Sigma=\mathbb{E}[M_\ell]$. Note that an expectation of a random operator M_ℓ is defined as a linear operator $\Sigma:\mathbb{H}\to\mathbb{H}$ satisfying $\langle u,\Sigma v\rangle=\mathbb{E}[\langle u,M_\ell v\rangle]$ for any $u,v\in\mathbb{H}$. This paper aims to examine the estimation of Σ. For any $\ell\in\{0,\dots,n\}$, let $\mathcal{F}_\ell=\sigma(M_1,\dots,M_\ell)$. For probability measures P,P', $P\ll P'$ means that P is absolutely continuous with respect to P', and $\mathrm{KL}(P\|P')=\int\log(dP/dP')dP$ denotes the Kullback–Leibler divergence. For a sequence of sets $A_1,A_2,\ldots$, we define $\bigtimes_{i=1}^{\infty}A_i:=A_1\times A_2\times\cdots$.

2. Dependent matrices with heavy-tailed distributions

2.1. Setup

First, we introduce some assumptions on $\mathbf{M}=(M_\ell)_{\ell=1,\dots,n}$. The first is the dependence between the M_ℓ's, which is quantified through a weak dependence coefficient. The second is the tail probability of the distribution of $\|M_\ell\|$.

2.1.1. Dependence

We introduce a coefficient to measure the dependence of the process of operators/matrices, which is essentially from [27]. It is also discussed in [12]. First, we define the set of Lipschitz functions on ℓ operators/matrices for $\ell \in \{1, \ldots, n\}$.

Definition 2 (Lipschitz function on ℓ elements)**.** Let E be a space equipped with the norm $\|\cdot\|_E$. For any $\ell \in \mathbb{N}$ and $L > 0$, we let $\mathrm{Lip}_\ell(E, L)$ denote the set of all functions $h : E^\ell \to \mathbb{R}$ such that for any $(a_1, \ldots, a_\ell, b_1, \ldots, b_\ell) \in E^{2\ell}$,

$$|h(a_1, \ldots, a_\ell) - h(b_1, \ldots, b_\ell)| \le L \sum_{i=1}^{\ell} \|a_i - b_i\|_E.$$

Owing to this definition, we can introduce our weak dependence condition:

Assumption 1 (Weak dependence)**.** *For any $\ell \in \{1, ..., n-1\}$ and function $g \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$, $\mathbb{E}[g(M_{\ell+1}, \ldots, M_n) \mid \mathcal{F}_\ell]$ and $\mathbb{E}[g(M_{\ell+1}, \ldots, M_n)]$ exist. Further, there exist real numbers $(\Gamma_{\ell,n})_{1\le\ell\le n-1}$ such that for any $\ell \in \{1, \ldots, n-1\}$ and for any function $g \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$, we have*

$$|\mathbb{E}[g(M_{\ell+1}, \ldots, M_n) \mid \mathcal{F}_\ell] - \mathbb{E}[g(M_{\ell+1}, \ldots, M_n)]| \le \Gamma_{\ell,n}, \tag{1}$$

almost surely. We set $\Gamma_n = \max_{1\le\ell\le n-1} \Gamma_{\ell,n}$.

This assumption has several noteworthy points: (i) The coefficient used in the assumption is a generalization of the uniform mixing coefficient for bounded processes; (ii) the coefficient quantifies the dependence of the M_ℓ's: the larger it is, the more the matrices are dependent, while $\Gamma_n = 0$ for independent matrices; and (iii) it is not comparable with the α/β-mixing property: examples of non-mixing processes with small Γ_n are known. A classical real-valued example from [5] is given by $M_{\ell+1} = (M_\ell + \varepsilon_\ell)/2$ where the (ε_ℓ) are i.i.d. from a Bernoulli distribution with parameter 1/2. It is proven in Section 1.5 page 8 of [12] that this process is not strongly mixing. On the other hand, it is quite direct to check that $\Gamma_{\ell,n} \le 1$. Essentially, when Γ_n remains bounded for large n, we recover the same rates of estimation as for independent matrices. This includes linear auto-regressive moving-average (ARMA) processes and a causal Bernoulli shifts (CBS), that are described below. For more details, we refer the reader to [27, 13, 12, 4].

In previous studies such as [27], Assumption 1 is used for bounded processes; that is, for any ℓ, $\|M_\ell\|$ is bounded almost surely. As aforementioned, when working with unbounded processes, we begin by studying a truncated and bounded version of the process. A fact that we often use in this paper is that when $\mathbf{M} = (M_1, \ldots, M_n)$ satisfies Assumption 1, then so does $(f(M_1), \ldots, f(M_n))$ where $f : E \to E$ is an adequate truncation function (that is, f is 1-Lipschitz).

Proposition 1. *Assume that $\mathbf{M} = (M_1, \ldots, M_n)$ satisfies Assumption 1 and that $f : \mathcal{S} \to \mathcal{S}$ is 1-Lipschitz. Then $(f(M_1), \ldots, f(M_n))$ also satisfies Assumption 1.*

2.1.2. Tail probability of the random matrices

We introduce an assumption on the tail probability of $\|M_\ell\|$ that includes heavy-tailed matrices/operators.

Assumption 2 (Tail probability)**.** *We define, for any $t \geq 0$, $S_\ell(t) = \mathbb{P}(\|M_\ell\| > t)$ the tail function of $\|M_\ell\|$. We assume that $\int_0^\infty S_\ell(t)\mathrm{d}t < \infty$, and we define*

$$G(t) = \max_{1\leq \ell \leq n} \int_t^\infty S_\ell(u)\mathrm{d}u.$$

2.2. Examples

We provide examples in which Assumptions 1 and 2 are satisfied. A recurring case of interest is the one of a stationary $\mathbb{H}$-valued stochastic process $(Y_\ell)_{\ell\in\mathbb{Z}}$, with $M_\ell = Y_\ell Y_\ell^\top$. The estimation of $\Sigma = \mathbb{E}[M_\ell]$ corresponds to the estimation of the covariance matrix of $(Y_\ell)_{\ell\in\mathbb{Z}}$.

2.2.1. Independent matrices

Before diving into time dependence, we study the simple case where the matrices M_ℓ are independent and identically distributed. Therefore, Assumption 1 is trivially satisfied with $\Gamma_n = 0$. Moreover, in Assumption 2, we have $S_\ell = S_1$ for any ℓ and thus $G(t) = \int_t^\infty S_1(u)\mathrm{d}u$.

We then consider the special case where $M_\ell = Y_\ell Y_\ell^\top$ and the Y_ℓ are i.i.d. Then, $S_\ell(t) = \mathbb{P}(\|M_\ell\| > t) = \mathbb{P}(\|Y_\ell\|^2 > t)$ and thus Assumption 2 can be checked by the study of the tails of $\|Y_\ell\|^2$. We detail three cases of interest.

Bounded case: if $\|Y_\ell\| \leq C$ almost surely, then $S_\ell(t) = 0$ for any $t \geq C^2$, and thus Assumption 2 is satisfied with $G(t) = 0$ for $t \geq C^2$.

Exponential tails: let us start with a specific example: $Y_\ell \sim \mathcal{N}(0,\Sigma)$ in $\mathbb{R}^p$. Then, by (the proof of) Lemma 1 of [22], for all $s \in (0, 1/(2||\Sigma||))$, we have $\log \mathbb{E}\exp(s\|Y_\ell\|^2) \leq \frac{s^2 \mathrm{Tr}(\Sigma^2)}{1-2s\|\Sigma\|}$, and thus it holds that

$$\mathbb{P}(\|Y_\ell\|^2 > t) \leq \frac{\mathbb{E}[\exp(s\|Y_\ell\|^2)]}{\exp(st)} \leq \exp\left(\frac{s^2\mathrm{Tr}(\Sigma^2)}{1-2s\|\Sigma\|} - st\right).$$

We then put $s = \frac{t}{2\mathrm{Tr}(\Sigma^2)+2\|\Sigma\|t}$ and obtain:

$$S_\ell(t) = \mathbb{P}(\|Y_\ell\|^2 > t) \leq \exp\left(-\frac{t^2}{4(\mathrm{Tr}(\Sigma^2)+\|\Sigma\|t)}\right).$$

Particularly, we set $t \geq \frac{\mathrm{Tr}(\Sigma^2)}{\|\Sigma\|}$ and obtain

$$S_\ell(t) \leq \exp\left(-\frac{t^2}{4(\mathrm{Tr}(\Sigma^2)+\|\Sigma\|t)}\right) \leq \exp\left(-\frac{t}{2\|\Sigma\|}\right)$$

and thus $G(t) \leq 2\|\Sigma\| \exp(-\frac{t}{2\|\Sigma\|})$ holds. More generally, we consider examples where $S_\ell(t) \leq \exp(-at)$ and thus $G(t) \leq \exp(-at)/a$ for some $a > 0$ for large enough $t > 0$.

Polynomial tails: we consider a more general situation where $Y_\ell = \sqrt{R_\ell} V_\ell$ where V_ℓ is distributed on the unit sphere in $\mathbb{R}^p$, and R_ℓ is a non-negative random variable. In this case,

$$S_\ell(t) = \mathbb{P}(\|Y_\ell\|^2 > t) = \mathbb{P}(R_\ell > t)$$

and thus Assumption 2 is satisfied if $\mathbb{P}(R_\ell > t) = o(1/t)$ when $t \to \infty$, and we have:

$$G(t) = \int_t^\infty \mathbb{P}(R_\ell > u) \mathrm{d}u.$$

This includes exponential tails as above, where $\mathbb{P}(R_\ell > u) \leq \exp(-at)$ for some $a > 0$. This also includes heavier tail probabilities. For example, if R_ℓ is a (shifted) Pareto random variable, $\mathbb{P}(R_\ell > t) = \frac{1}{(t+1)^a}$, Assumption 2 is satisfied if $a > 1$ and we have $G(t) \leq \frac{a-1}{(t+1)^{a-1}}$.

2.2.2. Causal Bernoulli shift

An important category of examples is the class of causal Bernoulli shifts (CBS), which includes a large class of stochastic processes. We consider a bounded CBS first, and then define a class of unbounded processes built on CBSs.

Example 1 (Causal Bernoulli shifts, CBS)**.** Let $\Xi = (\xi_\ell)_{\ell \in \mathbb{Z}}$ be a sequence of bounded i.i.d. $\mathbb{H}$-valued random variables: $\|\xi_\ell\| \leq B_\xi$ almost surely. Let $C : \bigtimes_{i=1}^\infty \mathbb{H} \to \mathbb{H}$ with $C(0, 0, \dots) = 0$. Assume that, for any $(a_1, b_1, a_2, b_2, \dots) \in \bigtimes_{i=1}^\infty \mathbb{H}$ we have

$$\|C(a_1, a_2, \dots) - C(b_1, b_2, \dots)\| \leq \sum_{\ell=1}^\infty \alpha_\ell \|a_\ell - b_\ell\| \text{ and } \mathcal{A} := \sum_{\ell=1}^\infty \alpha_\ell < \infty.$$

Then, we can define the stationary process $(X_\ell)_{\ell \in \mathbb{Z}}$ given by

$$X_\ell = C(\xi_\ell, \xi_{\ell-1}, \xi_{\ell-2}, \dots).$$

This process $(X_\ell)_{\ell \in \mathbb{Z}}$ is called a CBS. Note that $\|X_\ell\| \leq B := \mathcal{A} B_\xi$ almost surely.

CBSs include many well-known stationary and ergodic processes, such as causal ARMA. We have the following result:

Proposition 2. *Let $(Y_\ell)_{\ell \in \mathbb{Z}}$ be a CBS and let $M_\ell = Y_\ell Y_\ell^\top$ for any $\ell \in \mathbb{Z}^2$. Then, $\mathbf{M} = (M_\ell)_{\ell=1,\dots,n}$ satisfies Assumption 1 with $\Gamma_{\ell,n} = 4BB_\xi \sum_{i=\ell+1}^\infty \min(i,n)\alpha_i$ and Assumption 2 with $G(t) = \mathbf{1}_{\{t \leq 4B^2\}}$.*

In this result, we do not consider heavy tails, because CBSs are bounded processes. The proof of Proposition 2 is included in that of Proposition 3, which is about a more general class of unbounded processes.

2.2.3. Application to unbounded processes

Proposition 3. *We assume that $(X_\ell)_{\ell\in\mathbb{Z}}$ is a CBS. Let $\mathcal{E} = (\varepsilon_\ell)_{\ell\in\mathbb{Z}}$ be a sequence of centered i.i.d. $\mathbb{H}$-valued random variables with $S_\varepsilon(t) := \mathbb{P}(\|\varepsilon_\ell\|^2 \geq t)$ such that $\int_0^\infty S_\varepsilon(t)dt < \infty$ holds, all independent from $(X_\ell)_{\ell\in\mathbb{Z}}$. We define the process $(Y_\ell)_{\ell\in\mathbb{Z}}$ as*

$$Y_\ell = X_\ell + \varepsilon_\ell,$$

and $M_\ell = Y_\ell Y_\ell^\top$ for any $\ell \in \mathbb{Z}$. Then, $\mathbf{M} = (M_\ell)_{\ell=1,\dots,n}$ satisfies Assumption 1 with $\Gamma_{\ell,n} = 4BB_\xi \sum_{i=\ell+1}^\infty \min(i,n)\alpha_i$ and Assumption 2 with $S_\ell(t) \leq \mathbf{1}_{\{t\leq 4B^2\}} + S_\varepsilon(t/4)$ [removed some words].

Furthermore, if $\Gamma := 4BB_\xi \sum_{i=2}^\infty i\alpha_i < \infty$, then $\max_{1\leq\ell\leq n}\Gamma_{\ell,n} \leq \Gamma$ holds.

2.2.4. Application to chains with infinite memory

We finally discuss chains with infinite memory, which turn out to be special cases of CBSs.

Example 2 (Chain with Infinite Memory, CIM)**.** Let $\Xi = (\xi_\ell)_{\ell\in\mathbb{Z}}$ be a sequence of bounded i.i.d. $\mathbb{H}$-valued random variables: $\|\xi_\ell\| \leq B_\xi$ almost surely. Let $D : \bigtimes_{i=1}^\infty \mathbb{H} \to \mathbb{H}$ with $D(0,0,\dots) = 0$. Assume that, for any $(a_0, b_0, a_1, b_1, a_2, b_2, \dots) \in \mathbb{H}^\infty$ we have

$$\|D(a_0,a_1,a_2,\dots) - D(b_0,b_1,b_2,\dots)\| \leq \sum_{\ell=0}^\infty \beta_\ell\|a_\ell - b_\ell\| \text{ and } \mathcal{B} := \sum_{\ell=1}^\infty \beta_\ell < 1.$$

Then, there is a stationary solution $(X_\ell)_{\ell\in\mathbb{Z}}$ to the equation [15]:

$$X_\ell = D(\xi_\ell, X_{\ell-1}, X_{\ell-2}, X_{\ell-3}, \dots).$$

The process $(X_\ell)_{\ell\in\mathbb{Z}}$ is called a chain with infinite memory (CIM).

There is a simple connection between CBSs and CIMs. Using Proposition 4.1 of [4], a CIM can be rewritten as a CBS as

$$X_\ell = C(\xi_\ell, \xi_{\ell-1}, \xi_{\ell-2}, \dots) \text{ with } \alpha_\ell = \beta_0\mathcal{B}^{\ell-1}.$$

Remark 1. Let us briefly discuss vector auto-regression (VAR) in this framework: $X_\ell = AX_{\ell-1} + \xi_\ell$, with $A \in \mathbb{R}^p \otimes \mathbb{R}^p$. A VAR with bounded noise terms ξ_ℓ is obviously a CIM, and thus Assumptions 1 and 2 are satisfied by such a process. They even remain satisfied by $Y_\ell = X_\ell + \varepsilon_\ell$ for heavy-tailed ε_ℓ's, by using Proposition 3. However, when the noise ξ_ℓ in the VAR is unbouded, we need to apply another technique to handle it. Therefore, we have to approximate the VAR by a finite-order moving-average (MA) process and show it satisfies Assumptions 1 and 2. As the approximation error vanishes when the order k of the MA grows, we can apply our result without difficulty.

3. Main results

We introduce our main result in stages. First, we consider the case where M_ℓ is a $p \times p$ matrix with the bounded property, and then we extend it to the unbounded and heavy-tailed cases. Finally, we extend the result to the case where M_ℓ is an operator between infinite-dimensional spaces.

3.1. Result on p-dimensional matrix

3.1.1. Bounded case

We first consider the case $\mathbb{H} = \mathbb{R}^p$, with $p \in \mathbb{N}$, and the matrices $\|M_\ell\|$ are bounded for all $\ell = 1, \ldots, n$. Obviously, M_ℓ is not heavy-tailed in this case; thus, the main contribution here is to handle the dependence in $\mathbf{M}$.

The derivation of this result starts with the variational inequality: with a probability measure μ on a parameter space Θ, and with a random parameter θ in Θ and a random variable X, it holds that with probability at least $1-\exp(-t)$, for any probability measure $\rho \ll \mu$ and any measurable function h,

$$\mathbb{E}_\rho[h(X,\theta)] \leq \mathbb{E}_\rho\left[\log \mathbb{E}_X\left[\exp\left(h(X,\theta)\right)\right]\right] + \mathrm{KL}\left(\rho \| \mu\right) + t$$

where $\mathbb{E}_\rho$ denotes the expectation with respect to θ under the distribution ρ ; note that the left-hand side is still random. This result is taken from [11] and [32]. In our setting, $X = \mathbf{M}$ and we control $\mathbb{E}_X\left[\exp\left(h(X,\theta)\right)\right]$ thanks to an inequality by [27] for dependent matrices (these steps are detailed in the proofs below). We obtain

$$\mathbb{E}\left[\exp\left(\lambda h(\mathbf{M}) - \lambda \mathbb{E}[h(\mathbf{M})]\right)\right] \leq \exp\left(\frac{\lambda^2 L^2 \sum_{\ell=1}^n \left(2\kappa + \Gamma_{\ell,n}\right)^2}{8n^2}\right).$$

An adequate choice of h leads to our main result: the concentration bound for the estimation of Σ using the empirical mean of $\mathbf{M}$.

Theorem 4. *Assume that $\mathbf{M}$ is a sequence of positive semi-definite symmetric random $p \times p$ matrices such that, for some $\kappa > 0$, for all $\ell = 1, \ldots, n$, $\mathbb{E}[M_\ell] = \Sigma$ and $\|M_\ell\| \leq \kappa^2$ almost surely. Under Assumption 1, for all $t > 0$, with probability at least $1 - \exp(-t)$, we have*

$$\left\| \frac{1}{n} \sum_{\ell=1}^n M_\ell - \Sigma \right\| \leq 4\sqrt{2}\, \|\Sigma\| \left(\kappa^2 + \Gamma_n\right) \sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}}.$$

Let us comment briefly on this result. First, this is a dimension-free bound in which p does not appear. The statistical dimension is instead described by the effective rank $\mathbf{r}(\Sigma)$. This is identical to the statistical dimension of the independent case of [21] and others. Then, the effect of this dependence appears as Γ_n in the factor $(2\kappa^2 + \Gamma_n)$ of the upper bound. If $\mathbf{M}$ is independent, we have $\Gamma_n = 0$. More generally, we described above a large class of processes where Γ_n is bounded from above by a constant Γ. In both cases, our upper bound matches the one of [21] up to constants.

3.1.2. Heavy-tailed case

We then extend Theorem 4 to unbounded, possibly heavy-tailed matrices M_ℓ.

The general idea is to apply Theorem 4 to a sequence of transformed matrices $\{f(M_1), \dots, f(M_n)\}$ where $f : \mathcal{S} \to \mathcal{S}$ is a bounded function, such that $\sup_{M\in\mathcal{S}} \|f(M)\| \leq \tau$. This application yields a bound on $\|\frac{1}{n}\sum_{\ell=1}^{n} f(M_\ell) - \mathbb{E}[f(M_\ell)]\|$. Then, we handle the effect of f, that is, $\|\frac{1}{n}\sum_{\ell=1}^{n} f(M_\ell) - \frac{1}{n}\sum_{\ell=1}^{n} M_\ell\|$ and $\|\mathbb{E}[f(M_\ell)] - \Sigma\|$, to obtain an upper bound on $\|\frac{1}{n}\sum_{\ell=1}^{n} M_\ell - \Sigma\|$. This results in the introduction of an additional term depending on τ in the upper bound. This technique leads to the following results.

Corollary 5. *Assume that* $\mathbf{M}$ *is a sequence of* $p \times p$ *positive semi-definite, symmetric, random matrices with* $\mathbb{E}[M_\ell] = \Sigma$*, which satisfies Assumptions 1 and 2. For any* $\tau > 0$ *and for all* $t > 0$*, with probability at least* $1 - \exp(-t) - \sum_{\ell=1}^{n} S_\ell(\tau)$ *it holds that*

$$\left\| \frac{1}{n}\sum_{\ell=1}^{n} M_\ell - \Sigma \right\| \leq 4\sqrt{2}\,\|\Sigma\|\,(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}} + G(\tau).$$

First, note that the bound holds with probability $1 - \exp(-t) - \sum_{\ell=1}^{n} S_\ell(\tau)$. If τ is constant and $S_\ell(\tau) > 0$, then $\sum_{\ell=1}^{n} S_\ell(\tau)$ can grow to ∞ when $n \to \infty$, and the statement becomes vacuous for large n. However, by letting $\tau = \tau_n \to \infty$, and if the S_ℓ's decrease fast enough, we are able to keep $\sum_{\ell=1}^{n} S_\ell(\tau)$ small enough (for example, smaller than $1/n$).

The effect of the heavy-tailed $G(\tau)$ also appears additively in the second term of the derived upper bound. Here again, by letting $\tau = \tau_n \to \infty$, we can make the term $G(\tau)$ small enough. The tightness of the bound, of course, depends on how far we are from the boundedness assumption, that is, on how fast the function G decreases. We provide the following examples:

Bounded Case: Assume that $\|M_i\| \leq \kappa$ almost surely for some $\kappa > 0$, then Assumption 2 is satisfied with $G(\tau) = 0$ for $\tau \geq \kappa$. Thus, we can take $\tau = \kappa$ and recover exactly Theorem 4.

Exponential-Tail Case: Assume that $S_\ell(\cdot)$ has an exponential decay, that is, there is an $a > 0$ such that for any ℓ, $S_\ell(t) \leq \exp(-at)$. Notably, $G(t) \leq \exp(-at)/a$. Thus, Corollary 5 states that with probability at least $1 - \exp(-t) - n\exp(-a\tau)$,

$$\left\| \frac{1}{n}\sum_{\ell=1}^{n} M_\ell - \Sigma \right\| \leq 4\sqrt{2}\,\|\Sigma\|\,(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}} + \frac{\exp(-a\tau)}{a}.$$

For some $\alpha > 1$, we put $\tau = \tau_n = \frac{\alpha \log n}{a}$ which implies that $n\sum_{\ell=1}^{n} S_\ell(\tau_n) \leq \frac{1}{n^{\alpha-1}}$, and we set $t = \log(\delta^{-1})$. Subsequently, for every $\delta \in (0, 1)$, with probability at least $1 - \delta - \frac{1}{n^{\alpha-1}}$, we have

$$\left\| \frac{1}{n}\sum_{\ell=1}^{n} M_\ell - \Sigma \right\| \leq 4\sqrt{2}\,\|\Sigma\| \left(\frac{\alpha \log n}{a} + \Gamma_n \right) \sqrt{\frac{4\mathbf{r}(\Sigma) + \log(\delta^{-1})}{n}} + \frac{1}{an^{\alpha}}.$$

In this upper bound, the effect of the heavy-tail appears in the second term in $1/(an^{\alpha})$, which is negligible with respect to the first term (because α is chosen > 1). The main difference with the bounded case is the factor $(\frac{\alpha \log(n)}{a} + \Gamma_n)$ in the first term, which increases in $\log(n)$. Thus, the dependence in n and in the statistical dimension are similar to the ones in [21, 32] up to an additional $\log(n)$ factor.

Polynomial-Tail Case: Assume that S_ℓ has a polynomial decay $S_\ell(t) \leq a t^{-b}$ with $a > 0$ and $b > 2$. Then, $G(t) \leq \frac{a}{b-1} t^{1-b}$. Thus, with $\tau = \tau_n$, the bound is, with probability at least $1 - \exp(-t) - n a \tau_n^{-b}$,

$$\left\| \frac{1}{n} \sum_{\ell=1}^{n} M_\ell - \Sigma \right\| \leq 4\sqrt{2} \left\| \Sigma \right\| (\tau_n + \Gamma_n) \sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}} + \frac{a}{b-1} \tau_n^{1-b}.$$

Here, for some $\alpha > 1$, we take $\tau_n = a^{1/b} (n)^{\alpha/b}$, and also set $t = \log \delta^{-1}$. Then, for any $\delta \in (0,1)$, we obtain that with probability at least $1 - \delta - 1/n^{\alpha-1}$,

$$\left\| \frac{1}{n} \sum_{\ell=1}^{n} M_\ell - \Sigma \right\|$$

$$\leq 4\sqrt{2} \left\| \Sigma \right\| \left(a^{1/b} n^{\alpha/b} + \Gamma_n \right) \sqrt{\frac{4\mathbf{r}(\Sigma) + \log(2\delta^{-1})}{n}} + \frac{1}{(b-1) n^{\alpha-1}}.$$

The rate in the first term is more seriously deteriorated. However, we still have convergence as soon as $1 < \alpha < b/2$, which is possible only in the case $b > 2$. Our rate is not as sharp as the one in [29] for heavy-tailed matrices in the independent case. We are not aware of how to extend the work of [29] to dependent matrices, and claim that our result is the first rate obtained on matrices that are simultaneously heavy-tailed and dependent.

Remark 2 (Comparison)**.** We discuss the comparison between Corollary 5 and the analysis of the case with independent matrices by [32, 1]. Corollary 5 does not always recover the rates that are known in the i.i.d setting; however, the techniques used [32, 1] strongly rely on the independence assumption. Under specific assumptions, we make the following findings: (i) In the bounded case, we recover the same rates as the previous studies, extending them from i.i.d to the non-i.i.d setting for free. (ii) In the exponential tail case, we recover these rates up to a $\log(n)$ factor, which we interpret as a cost of extending them to the dependent setting. (iii) In the polynomial tail case, we admit that we have a slower rate than the one above. However, we are not aware of any work that tackles simultaneously heavy tails and time dependence. The fact that we obtain a rate of convergence here, even if it is slow, is already a contribution.

3.2. Result on infinite-dimensional operator

Here, we consider the case of an infinite-dimensional separable Hilbert space $\mathbb{H}$, which has also been studied in [21, 16]. Following [16], we extend our result for the p-dimensional setting to the infinite-dimensional case.

The idea is to find a finite-dimensional approximation of the spectral norm of operators using an orthonormal basis. Let $(e_j)_{j\in\mathbb{N}}$ be an orthonormal basis of $\mathbb{H}$ and $\mathbb{H}_k := \operatorname{span}\{e_1,\dots,e_k\}$. For an $\mathbb{H}\otimes\mathbb{H}$-valued random operator M_ℓ, let $(M_\ell^{(j_1,j_2)})_{j_1,j_2=1}^k$ be a sequence of real-valued random variables such that $M_\ell^{(j_1,j_2)} := \langle M_\ell e_{j_1}, e_{j_2}\rangle$. We see that

$$\begin{aligned}&\sup_{u_k\in\mathbb{H}_k:\|u_k\|=1}\left|\left\langle\left(\frac{1}{n}\sum_{\ell=1}^n M_\ell-\Sigma\right)u_k,u_k\right\rangle\right|\\&=\sup_{\substack{u_k^{(j)}\in\mathbb{R},j=1,\dots,k\\ \sum_{j=1}^k(u_k^{(j)})^2=1}}\left|\frac{1}{n}\sum_{i=1}^n\sum_{j_1=1}^k\sum_{j_2=1}^k u_k^{(j_1)}u_k^{(j_2)}\left(M_\ell^{(j_1,j_2)}-\mathbb{E}\left[M_\ell^{(j_1,j_2)}\right]\right)\right|.\end{aligned}$$

Then, the right-hand side is a spectral norm of the difference between the sampled *matrix* and the population one, to which Theorem 4 and Corollary 5 are applicable. Based on this approach, and considering the limit $k\to\infty$, we obtain the following result:

Theorem 6. *Assume that* $\mathbf{M}$ *is a sequence of positive, semi-definite, symmetric,* $\mathbb{H}\otimes\mathbb{H}$*-valued random operators with* $\mathbb{E}[M_\ell]=\Sigma$*, and also satisfies Assumptions 1 and 2. For any* $\tau>0$ *and for all* $t>0$*, with probability at least* $1-\exp(-t)-\sum_{\ell=1}^n S_\ell(\tau)$*, it holds that*

$$\left\|\frac{1}{n}\sum_{\ell=1}^n M_\ell-\Sigma\right\|\le 4\sqrt{2}\,\|\Sigma\|\,(\tau+\Gamma_n)\sqrt{\frac{4\mathbf{r}(\Sigma)+t}{n}}+G(\tau).$$

The obtained upper bound remains the same, even for infinite dimensions. Our approach in the finite-dimensional case cannot be applied directly in the infinite-dimensional case. This is because our proof using variational equalities depends on a density function of p-dimensional Gaussian vector. Thus, we cannot avoid first considering $\mathbb{H}_k$ and subsequently letting $k\to\infty$.

4. Applications

4.1. Covariance operator estimation

We consider the problem of covariance operator estimation using dependent samples with heavy tails under the setting and assumptions of Proposition 3. Let $(X_\ell)_{\ell\in\mathbb{N}}$ be a CBS in $\mathbb{H}$ and consider the strongly stationary process $(Y_\ell)_{\ell\in\mathbb{Z}}$, given by

$$Y_\ell = X_\ell+\varepsilon_\ell,$$

as in Proposition 3. Additionally, assume that $\mathbb{E}[X_1]=0$ and its covariance operator is $\Sigma\in\mathcal{S}$; that is, Σ is defined as $\Sigma u=\mathbb{E}[\langle Y_1,u\rangle Y_1]$ for any $u\in\mathbb{H}$.

Assume that we have n observations $\mathbf{Y} = (Y_\ell)_{\ell=1,\dots,n}$ from the process $(Y_\ell)_{\ell\in\mathbb{Z}}$. Then, we define the empirical covariance operator:

$$M_\ell u := \langle Y_\ell, u\rangle Y_\ell,$$

for any $u \in \mathbb{H}$. Using this notion, we obtain n operators $\mathbf{M}$ from $\mathbf{Y}$ and then obtain the empirical covariance operator as

$$\widehat{\Sigma} := \frac{1}{n}\sum_{\ell=1}^{n} M_\ell. \tag{2}$$

By a direct application of Corollary 5, we obtain the following result, stated without proof.

Proposition 7. *Assume that the sequence* $\mathbf{M}$ *satisfies the setting of Proposition 3. Consider the empirical covariance operator defined in (2). Then, for any* $\tau > 0$ *and* $t > 0$*, the following inequality holds with probability at least* $1 - \exp(-t) - \sum_{\ell=1}^{n} S_\ell(\tau)$*:*

$$\|\widehat{\Sigma} - \Sigma\| \le 4\sqrt{2}\,\|\Sigma\|\,(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}\,(\Sigma) + t}{n}} + G(\tau),$$

where $S_\ell(\tau) = \mathbf{1}_{\{\tau\le 4B^2\}} + S_\varepsilon(\tau/4)$*, and* $G(\tau) = \int_\tau^\infty S_\ell(t)\mathrm{d}t$*.*

4.2. Lagged covariance matrix estimation

We consider the estimation of a lagged covariance matrix, which is also called a cross-covariance matrix. Consider the same process $(Y_\ell)_{\ell\in\mathbb{Z}}$ as in Section 4.1. Here, we aim to estimate

$$\Sigma_1 := \mathbb{E}[Y_\ell Y_{\ell+1}^\top],$$

from n observations $\mathbf{Y} = (Y_\ell)_{\ell=1,\dots,n}$. This problem and the solution discussed below can obviously be extended to $\Sigma_h := \mathbb{E}[Y_\ell Y_{\ell+h}^\top]$ for $h \ge 2$. Note that Σ_1 is not symmetric; hence, our main results cannot be directly applied to a naive estimator, $\widehat{\Sigma}_1 := (n-1)^{-1}\sum_{\ell=1}^{n-1} Y_\ell Y_{\ell+1}^\top$. We still denote $\Sigma = \mathbb{E}[Y_\ell Y_\ell^\top]$, which is shown as Σ_h for $h = 0$, and its empirical estimator $\widehat{\Sigma} := (n-1)^{-1}\sum_{\ell=1}^{n-1} Y_\ell Y_\ell^\top$.

To estimate Σ_1, we define an augmented process and estimator for the covariance matrix of the process. We define the Hilbert space $\mathbb{H}^2$ equipped with the scalar product $\langle (y_1, y_2), (y_1', y_2')\rangle = \langle y_1, y_1'\rangle + \langle y_1, y_2'\rangle$ for $(y_1, y_2), (y_1', y_2') \in \mathbb{H}^2$. Let $\widetilde{Y}_\ell = (Y_\ell, Y_{\ell+1})^\top$ be the $\mathbb{H}^2$-valued augmented process, whose covariance is

$$\Sigma_{0:1} := \mathbb{E}\left[\widetilde{Y}_\ell \widetilde{Y}_\ell^\top\right] = \begin{pmatrix} \mathbb{E}[Y_\ell Y_\ell^\top] & \mathbb{E}[Y_\ell Y_{\ell+1}^\top] \\ \mathbb{E}[Y_{\ell+1} Y_\ell^\top] & \mathbb{E}[Y_{\ell+1} Y_{\ell+1}^\top] \end{pmatrix} = \begin{pmatrix} \Sigma_0 & \Sigma_1 \\ \Sigma_1^\top & \Sigma_0 \end{pmatrix}. \tag{3}$$

The main idea is to estimate $\Sigma_{0:1}$, which directly leads to an estimator of Σ_1.

Using observations $\mathbf{Y}$, we build $\widetilde{Y}_1, \ldots, \widetilde{Y}_{n-1}$ and their sample-wise product matrices $M_1, \ldots, M_{n-1}$ as

$$M_\ell := \widetilde{Y}_\ell \widetilde{Y}_\ell^\top = \begin{pmatrix} Y_\ell Y_\ell^\top & Y_\ell Y_{\ell+1}^\top \\ Y_{\ell+1} Y_\ell^\top & Y_{\ell+1} Y_{\ell+1}^\top \end{pmatrix}.$$

We then construct an estimator

$$\widehat{\Sigma}_{0:1} := \frac{1}{n-1} \sum_{\ell=1}^{n-1} M_\ell = \begin{pmatrix} \widehat{\Sigma} & \widehat{\Sigma}_1 \\ \widehat{\Sigma}_1^\top & \widehat{\Sigma} \end{pmatrix}. \tag{4}$$

We show a concentration inequality for $\widehat{\Sigma}_{0:1}$ and additionally show the convergence of $\widehat{\Sigma}$ and $\widehat{\Sigma}_1$.

Proposition 8. *Assume that $\mathbf{Y}$ is as in Proposition 3. Consider the matrices in (3) and the estimator in (4). Then, for any $\tau > 0$ and $t > 0$, the following inequality holds with probability at least $1 - \exp(-t) - \sum_{\ell=1}^{n-1} S_\ell(\tau)$:*

$$\|\widehat{\Sigma}_{0:1} - \Sigma_{0:1}\| \leq 4\sqrt{2} \left(\|\Sigma_{0:1}\|\right) (\tau + \Gamma_n) \sqrt{\frac{4\mathbf{r}\left(\Sigma_{0:1}\right) + t}{n-1}} + G(\tau),$$

where $S_\ell(\tau) = \mathbf{1}_{\{\tau \leq 4B^2\}} + S_\varepsilon(\tau/4)$, and $G(\tau) = \int_\tau^\infty S_\ell(t)\mathrm{d}t$. Furthermore, with the same probability, we obtain

$$\begin{aligned} &\max\{\|\widehat{\Sigma} - \Sigma\|, \|\widehat{\Sigma}_1 - \Sigma_1\|\} \\ &\leq 4\sqrt{2} \left(\|\Sigma_1\| + \|\Sigma\|\right) (\tau + \Gamma_n) \sqrt{\frac{4\mathbf{r}\left(\Sigma_{0:1}\right) + t}{n-1}} + G(\tau). \end{aligned}$$

The first statement is simply an application of Theorem 4 to the estimator $\widehat{\Sigma}_{0:1}$. The second simply follows from the first statement together with the facts $\|\Sigma_{0:1}\| \leq \|\Sigma_1\| + \|\Sigma\|$. By using the relation $\mathrm{Tr}(\Sigma_{0:1}) = 2\mathrm{Tr}(\Sigma) = 2\|\Sigma\|\mathbf{r}\left(\Sigma\right)$, we obtain the result.

4.3. Linear hidden Markov model

We consider a linear hidden Markov model (HMM) and study estimation in this model. Specifically, we consider a HMM model with a lag order 1 and set $\mathbb{H} = \mathbb{R}^p$. Assume that we observe a sequence of p-dimensional random vectors $\mathbf{Y} = (Y_\ell)_{\ell=0}^n$ which follows the following equations for $\ell \in \mathbb{Z}$:

$$Y_\ell = X_\ell + \varepsilon_\ell, \tag{5}$$

$$X_\ell = AX_{\ell-1} + \xi_\ell \tag{6}$$

where $A \in \mathbb{R}^{p \times p}$ is an unknown parameter matrix such that $\|A\| \in (0, 1)$, and $(X_\ell)_{\ell \in \mathbb{Z}}$ is a latent process such that $\|X_\ell\| \leq B$ almost surely. Here, $(\varepsilon_\ell)_{\ell \in \mathbb{Z}}$ is a sequence of i.i.d. p-dimensional noise variable with zero mean and finite variance, and ξ_ℓ is a sequence of i.i.d. p-dimensional bounded noise variable with zero mean such that $\|\xi_\ell\| \le B_\xi$ almost surely. Under the condition $\|A\| \in (0,1)$ and the boundedness of the ξ_ℓ's, the upper bound B is guaranteed to be finite. For brevity, we assume that $\mathbb{E}[\varepsilon_\ell \varepsilon_\ell^\top] = \mathbb{E}[\xi_\ell \xi_\ell^\top] = I$. We also define the covariance matrix $\Sigma := \mathbb{E}[Y_\ell Y_\ell^\top]$ and the lagged covariance matrix $\Sigma_1 := \mathbb{E}[Y_{\ell+1} Y_\ell^\top]$. Here, we aim to estimate the unknown parameter matrix A.

We study a convenient form of the HMM model. We define noise matrices $\mathsf{E} = (\varepsilon_0, \dots, \varepsilon_{n-1}) \in \mathbb{R}^{p\times n}$, $\mathsf{E}_+ = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^{p\times n}$, and $\mathsf{Z} = (\xi_1, \dots, \xi_n) \in \mathbb{R}^{p\times n}$ and also define matrices $\mathsf{Y} = (Y_0, \dots, Y_{n-1}) \in \mathbb{R}^{p\times n}$ and $\mathsf{Y}_+ = (Y_1, \dots, Y_n) \in \mathbb{R}^{p\times n}$. Then, we rewrite (5) and (6) as

$$(\mathsf{Y}_+ - \mathsf{E}_+) = A(\mathsf{Y} - \mathsf{E}) + \mathsf{Z}.$$

Multiplying $(\mathsf{Y} - \mathsf{E})^\top$ on both sides from the right and taking an expectation yields

$$\begin{aligned} A &= (\mathbb{E}[(\mathsf{Y}_+ - \mathsf{E}_+)(\mathsf{Y} - \mathsf{E})^\top] - \mathbb{E}[\mathsf{Z}(\mathsf{Y} - \mathsf{E})^\top])\mathbb{E}[(\mathsf{Y} - \mathsf{E})(\mathsf{Y} - \mathsf{E})^\top]^{-1} \\ &= \Sigma_1(\Sigma + I)^{-1}. \end{aligned}$$

Here, we utilize the independent properties of the noise, and $\mathbb{E}[\varepsilon_\ell \varepsilon_\ell^\top] = I$.

We then define an estimator of A. Using the estimators $\widehat{\Sigma} := \mathsf{Y}\mathsf{Y}^\top/n = n^{-1}\sum_{\ell=0}^{n-1} Y_\ell Y_\ell^\top$ and $\widehat{\Sigma}_1 := \mathsf{Y}_+\mathsf{Y}^\top/n = n^{-1}\sum_{\ell=0}^{n-1} Y_{\ell+1} Y_\ell^\top$, we define the following estimator:

$$\widehat{A} := \widehat{\Sigma}_1(\widehat{\Sigma} + I)^{-1}. \tag{7}$$

Then, we obtain the following result:

Proposition 9. *Consider the HMM model (5)-(6) and the estimator in (7) for the parameters in the model. Then, for any $t > 0$ and $\tau > 0$, with probability at least $1 - \exp(-t) - \sum_{\ell=1}^{n-1} S_\ell(\tau)$, the following inequality holds:*

$$\begin{aligned} &\|\widehat{A} - A\| \\ &\le 4\sqrt{2}\,(\|\Sigma_1\| + \|\Sigma\|)\,(1 + \|\Sigma_1\|)\,(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}\,(\Sigma_{0:1}) + t}{n-1}} + (1 + \|\Sigma_1\|)G(\tau), \end{aligned}$$

where $\Sigma_{0:1}$ is defined in (3), $S_\ell(\tau) = \mathbf{1}_{\{\tau \le 4B^2\}} + S_\varepsilon(\tau/4)$, and $G(\tau) = \int_\tau^\infty S_\ell(t)\mathrm{d}t$.

It is obtained by bounding the estimation error $\|\widehat{A} - A\|$ with the estimation errors of the covariance matrix Σ and the lagged covariance matrix Σ_1, as described in Proposition 4.2. Note that it is possible to extend the number of lags in this HMM model to more than 1.

4.4. Overparameterized linear regression

Here, we study a linear regression problem with dependent and heavy-tail covariates in the overparameterization framework developed by [6].

Let $(X_\ell)_{\ell\in\mathbb{Z}}$ be a CBS as a $\mathbb{H}$-valued latent process and $(Y_\ell)_{\ell\in\mathbb{Z}}$ be a generated process as a $\mathbb{H}$-valued covariate such that

$$Y_\ell = X_\ell + \varepsilon_\ell, \tag{8}$$

where ε_ℓ is an i.i.d. $\mathbb{H}$-valued noise variable with a mean value of zero. Additionally, we define $\theta^* \in \mathbb{H}$ as a true unknown parameter and a covariance operator $\Sigma = \mathbb{E}[Y_\ell Y_\ell^\top]$. For $\ell \in \mathbb{Z}$, we consider an $\mathbb{R}$-valued random variable Z_ℓ called the response variable, given by:

$$Z_\ell = \langle \theta^*, Y_\ell\rangle + U_\ell, \tag{9}$$

where U_ℓ is an $\mathbb{R}$-valued independent random variable with mean zero and a variance $\sigma^2 > 0$.

The goal of the regression problem is to estimate θ^* from the observations $\{(Z_i, Y_i) : i = 1, \dots, n\}$. We introduce a design matrix and operator as $\mathsf{Z} = (Z_1, \dots, Z_n)^\top \in \mathbb{R}^n$ and $\mathsf{Y} : \mathbb{H} \to \mathbb{R}^n$ such that $\mathsf{Y}\theta = (\langle Y_1, \theta\rangle, \dots, \langle Y_n, \theta\rangle)^\top \in \mathbb{R}^n$ holds for $\theta \in \mathbb{H}$. Similarly, with $E = (e_1, ..., e_n)^\top \in \mathbb{R}^n$, we define an operator $\mathsf{Y}^\top : \mathbb{R}^n \to \mathbb{H}$ such that $\mathsf{Y}^\top E = \sum_{i=1}^n e_i Y_i$. Further, we define an empirical covariance operator $\widehat{\Sigma} : \mathbb{H} \to \mathbb{H}$ as $\widehat{\Sigma} = \mathsf{Y}^\top \mathsf{Y}/n$ and a projection operator $\Pi_{\mathsf{Y}} : \mathbb{H} \to \mathbb{H}$ as $\Pi_{\mathsf{Y}} := \mathsf{Y}^\top(\mathsf{Y}\mathsf{Y}^\top)^{-1}\mathsf{Y}$.

To estimate θ^*, we define the minimum norm estimator as:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{H}}\{\|\theta\|^2 : \mathsf{Y}^\top\mathsf{Y}\theta = \mathsf{Y}^\top\mathsf{Z}\} = \mathsf{Y}^\top(\mathsf{Y}\mathsf{Y}^\top)^\dagger \mathsf{Z}, \tag{10}$$

where † denotes the pseudo-inverse of operators. The excess risk of $\widehat{\theta}$ is measured using

$$R(\widehat{\theta}) := \mathbb{E}_{(Z_*,Y_*)}[(Z_* - \langle Y_*, \widehat{\theta}\rangle)^2 - (Z_* - \langle Y_*, \theta^*\rangle)^2], \tag{11}$$

where (Z_*, Y_*) is an i.i.d. copy of (Z_1, Y_1) from the regression model (9) and $\mathbb{E}_{(Z_*,Y_*)}[\cdot]$ is the expectation with respect to (Z_*, Y_*).

We present a technical assumption that specializes in the overparameterization setting. Let $\Pi_\Sigma^\perp$ be a projection operator onto a linear space spanned by vectors orthogonal to any eigenvector of Σ.

Assumption 3. $\dim(\Pi_\Sigma^\perp(\mathsf{Y})) > n$ *holds almost surely.*

This assumption is identical to Assumption 1 in [6] and is intended to address cases where no degeneracies exist, such as the perfect collinearity between the variables.

With this setting, we bound the risk of the estimator for the overparameterized linear regression model.

Proposition 10. *Consider the linear regression model* (9) *with the process* $(X_\ell)_{\ell\in\mathbb{Z}}$ *in* (8) *being a CBS. Assume that Assumption 3 holds. Consider the estimator* (10) *and its excess risk* (11). *Assume, for any* $t, \tau > 0$, *with probability at least* $1 - \exp(-t) - \sum_{\ell=1}^n S_\ell(\tau)$, *we have*

$$R(\widehat{\theta}) \le 4\sqrt{2}c\|\theta^*\|^2\,\|\Sigma\|\,(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}(\Sigma) + t}{n}} + G(\tau) + ct\sigma^2\mathrm{Tr}(C),$$

where $C = (\mathsf{YY}^\top)^{-1}\mathsf{Y}\Sigma\mathsf{Y}^\top(\mathsf{YY}^\top)^{-1}$, $S_\ell(\tau) = \mathbf{1}_{\{\tau \leq 4B^2\}} + S_\varepsilon(\tau/4)$, *and* $G(\tau) = \int_\tau^\infty S_\ell(t)\mathrm{d}t$.

This result indicates that we can bound the bias term of the risk of the overparameterized linear regression estimator, even in the dependent and heavy-tailed setting. The last term $ct\sigma^2\mathrm{Tr}(C)$ represents the variance of the risk, which converges to zero by removing correlations and controlling for them using different techniques. This goes beyond the scope of this paper, see Lemma 11 in [6] for further details.

5. Proofs for main results in Section 3

5.1. Outline

We first state two lemmas at the core of our proofs in Section 5.2. Lemma 11 appears in many forms in the proofs of the PAC-Bayes bounds [9, 3]. For convenience, we use the version stated in [11, 32]. Lemma 12 is Rio's version of Hoeffding's inequality [27] for weakly dependent random variables, that we applied to matrices.

Then, we prove Theorem 4 in Section 5.3. We essentially follow the techniques developed in [11, 32]. However, both these studies rely on exponential inequalities for independent random variables. Therefore, we use Rio's inequality, which requires the boundedness assumption.

In Section 5.4, we introduce a truncation function that transforms unbounded matrices into bounded ones. We thus apply Theorem 4 to the truncated matrices. We then control the effect of the truncation function to prove Corollary 5.

We mention that we have developed the proof to deal with dependent matrices. For the case with independent matrices, the tools developed in [11, 32] can be used. However, their proofs rely strongly on the independence property, which is why we needed to introduce new arguments for dependent matrices.

5.2. Preliminary results

Lemma 11 ([11])**.** *Assume that* X *is a random variable defined in a measurable space* $(\mathcal{X}, \mathcal{A})$, *and* $(\Theta, \mathcal{F})$ *is a measurable parameter space. Let* μ *be a probability measure on* $(\Theta, \mathcal{F})$ *and* $h : \mathcal{X} \times \Theta \to \mathbb{R}$ *be a real-valued* $\mathcal{A} \otimes \mathcal{F}/\mathcal{B}(\mathbb{R})$*-measurable function such that* $\mathbb{E}_X[\exp(h(X,\theta))] < \infty$ *for* μ*-almost all* θ*. It holds that with probability at least* $1 - \exp(-t)$*, for all probability measures* $\rho \ll \mu$ *simultaneously,*

$$\mathbb{E}_\rho[h(X,\theta)] \leq \mathbb{E}_\rho[\log \mathbb{E}_X[\exp(h(X,\theta))]] + \mathrm{KL}(\rho\|\mu) + t.$$

Proof. The proof is merely a consequence of the duality relationship:

$$\mathbb{E}_X\left[\exp\left\{\sup_{\rho \ll \mu}\left(\mathbb{E}_\rho\left[h(X,\theta) - \log \mathbb{E}_X[\exp(h(X,\theta))]\right] - \mathrm{KL}(\rho\|\mu)\right)\right\}\right]$$

$$
\begin{aligned}
&= \mathbb{E}_X \mathbb{E}_\mu \left[\exp\left(h\left(X,\theta\right) - \log \mathbb{E}_X\left[\exp(h\left(X,\theta\right))\right]\right)\right] \\
&= \mathbb{E}_X \mathbb{E}_\mu \left[\frac{\exp(h\left(X,\theta\right))}{\mathbb{E}_X\left[\exp(h\left(X,\theta\right))\right]}\right] \\
&= \mathbb{E}_\mu \mathbb{E}_X \left[\frac{\exp(h\left(X,\theta\right))}{\mathbb{E}_X\left[\exp(h\left(X,\theta\right))\right]}\right] \\
&= 1.
\end{aligned}
$$

We use Tonelli's theorem to exchange the order of expectations. Then Markov's inequality leads to the inequality that holds with probability at least $1-\exp(-t)$;

$$\sup_{\rho \ll \mu} \left(\mathbb{E}_\rho\left[h\left(X,\theta\right) - \log \mathbb{E}_X\left[\exp(h\left(X,\theta\right))\right]\right] - \mathrm{KL}\left(\rho \| \mu\right)\right) < t.$$

This completes the proof. □

Lemma 12 (Rio's version of Hoeffding's inequality [27], applied to matrices)**.** *Let* $\{M_1,\ldots,M_n\}$ *be a sequence of positive semi-definite symmetric random matrices with* $\max_{\ell=1,..,n}\|M_\ell\| \le \kappa^2$ *almost surely for some* $\kappa > 0$*. Let us assume that Assumption 1 is satisfied. Then, for any function* $h \in \mathrm{Lip}_n(E,L)$ *and for any* $\lambda > 0$ *we have*

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda h(M_1,\ldots,M_n) - \lambda \mathbb{E}[h(M_1,\ldots,M_n)]\right)\right] \\
&\le \exp\left(\frac{\lambda^2 L^2 \sum_{\ell=1}^n \left(2\kappa + \Gamma_{\ell,n}\right)^2}{8}\right).
\end{aligned}
$$

5.3. Bounded case (Theorem 4)

Proof of Theorem 4. The proof consists of truncation of $\rho = \rho_{u,v}$ given by [32] and the lemma above obtained using duality.

(Step 1) Let us assume that Σ is invertible. Otherwise, we only need to consider a lower-dimensional subspace, and the proof is similar to the case with invertible Σ. Let μ denote a $2p$-dimensional product measure of two p-dimensional Gaussian measures with a zero mean and covariance $\left(2\mathbf{r}\left(\Sigma\right)\right)^{-1}\Sigma$. We define $\mathbb{S}^{p-1}$ as the unit ball in $\mathbb{R}^p$. Let us set $u, v \in \Sigma^{1/2}\mathbb{S}^{p-1}$ and define f_u, f_v as probability density functions with respect to the Lebesgue measure such that

$$
\begin{aligned}
f_u\left(x\right) &= \frac{\exp\left(-\mathbf{r}\left(\Sigma\right)\left(x-u\right)^\top \Sigma^{-1}\left(x-u\right)\right)\mathbf{1}\{\|x-u\| \le \sqrt{\|\Sigma\|}\}}{\int \exp\left(-\mathbf{r}\left(\Sigma\right)\left(x'-u\right)^\top \Sigma^{-1}\left(x'-u\right)\right)\mathbf{1}\{\|x'-u\| \le \sqrt{\|\Sigma\|}\}\mathrm{d}x'}, \\
f_v\left(x\right) &= \frac{\exp\left(-\mathbf{r}\left(\Sigma\right)\left(x-u\right)^\top \Sigma^{-1}\left(x-v\right)\right)\mathbf{1}\{\|x-v\| \le \sqrt{\|\Sigma\|}\}}{\int \exp\left(-\mathbf{r}\left(\Sigma\right)\left(x'-v\right)^\top \Sigma^{-1}\left(x'-v\right)\right)\mathbf{1}\{\|x'-v\| \le \sqrt{\|\Sigma\|}\}\mathrm{d}x'}.
\end{aligned}
$$

Here, $\mathbf{1}\{\mathcal{E}\}$ is an indicator function which is 1 if an event $\mathcal{E}$ is true and 0 otherwise. Assume that the independent random vectors θ, η have densities

f_u and f_v. Note that $\mathbb{E}[(\theta,\eta)] = (u,v)$ by the symmetricity of f_u, f_v, and $\max\{\|\theta\|,\|\eta\|\} \leq 2\sqrt{\|\Sigma\|}$ almost surely. Let $\rho_{u,v}$ be a probability measure of θ,η given by $\rho_{u,v}(\mathrm{d}x,\mathrm{d}y) = f_u(x) f_v(y)\,\mathrm{d}x\mathrm{d}y$, $x,y \in \mathbb{R}^d$. In the proof of Theorem 1 of [32], it is shown that

$$\mathrm{KL}\left(\rho_{u,v}\|\mu\right) \leq 2\log 2 + 2\mathbf{r}\left(\Sigma\right).$$

(Step 2) Let $f\left(A,\theta,\eta\right) := \theta^\top A\eta$ for any $A \in \mathbb{R}^p \otimes \mathbb{R}^p$ and $\theta,\eta \in \mathbb{R}^p$. Lemma 12 with $h(M_1,\dots,M_n) = \sum_{\ell=1}^n f(M_\ell,\theta,\eta)$ gives that for any $\lambda > 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{M}}\left[\exp\left(\lambda\sum_{\ell=1}^n f\left(M_\ell,\theta,\eta\right)\right)\right]\\
&= \mathbb{E}_{\mathbf{M}}\left[\exp\left(\lambda\sum_{\ell=1}^n \theta^\top M_\ell\eta\right)\right]\\
&\leq \exp\left(n\lambda\theta^\top\Sigma\eta + \frac{\lambda^2\left\|\theta\right\|^2\left\|\eta\right\|^2 n\left(2\kappa^2 + 2\Gamma_n\right)^2}{8}\right),
\end{aligned}$$

because for any $A_1,\dots,A_n,B_1,\dots,B_n \in \mathbb{R}^p \otimes \mathbb{R}^p$,

$$\begin{aligned}
\left|h\left(A_1,\dots,A_n\right) - h\left(B_1,\dots,B_n\right)\right| &= \left|\theta^\top\left(\sum_{\ell=1}^n\left(A_\ell - B_\ell\right)\right)\eta\right|\\
&\leq \left\|\theta\right\|\left\|\eta\right\|\sum_{\ell=1}^n\left\|A_\ell - B_\ell\right\|.
\end{aligned}$$

It holds that

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\rho_{u,v}}\left[\log\mathbb{E}_{\mathbf{M}}\left[\exp\left(\lambda\sum_{\ell=1}^n f\left(M_\ell,\theta,\eta\right)\right)\right]\right]\\
&\leq \frac{1}{n}\mathbb{E}_{\rho_{u,v}}\left[n\lambda\theta^\top\Sigma\eta + \frac{\lambda^2\left\|\theta\right\|^2\left\|\eta\right\|^2 n\left(\kappa^2 + \Gamma_n\right)^2}{2}\right]\\
&= \mathbb{E}_{\rho_{u,v}}\left[\lambda\theta^\top\Sigma\eta + \frac{\lambda^2\left\|\theta\right\|^2\left\|\eta\right\|^2\left(\kappa^2 + \Gamma_n\right)^2}{2}\right]\\
&\leq \lambda u^\top\Sigma v + \frac{\lambda^2\left(2\sqrt{\|\Sigma\|}\right)^4\left(\kappa^2 + \Gamma_n\right)^2}{2}\\
&= \lambda u^\top\Sigma v + 8\lambda^2\left\|\Sigma\right\|^2\left(\kappa^2 + \Gamma_n\right)^2.
\end{aligned}$$

The last inequality comes from the fact that $\max\{\|\theta\|,\|\eta\|\} \leq 2\sqrt{\|\Sigma\|}$. Therefore, from Lemma 11 with $h\left(M_1,\dots,M_n,\theta,\eta\right) = \lambda\sum_{\ell=1}^n f\left(M_\ell,\theta,\eta\right)$ and the fact that $\log 2 \leq \mathbf{r}\left(\Sigma\right)$ for any Σ, we obtain

$$\frac{1}{n}\sum_{\ell=1}^n \lambda u^\top M_\ell v \leq \lambda u^\top\Sigma v + 8\lambda^2\left\|\Sigma\right\|^2\left(\kappa^2 + \Gamma_n\right)^2 + \frac{4\mathbf{r}\left(\Sigma\right) + t}{n},$$

simultaneously for all u, v with probability at least $1 - \exp(-t)$. By choosing

$$\lambda = \sqrt{\frac{4\mathbf{r}\left(\Sigma\right) + t}{8n\left\|\Sigma\right\|^2\left(\kappa^2 + \Gamma_n\right)^2}},$$

we obtain

$$\left\|\frac{1}{n}\sum_{\ell=1}^{n} M_\ell - \Sigma\right\| \leq 4\sqrt{2}\left\|\Sigma\right\|\left(\kappa^2 + \Gamma_n\right)\sqrt{\frac{4\mathbf{r}\left(\Sigma\right) + t}{n}}.$$

This is our claim. □

5.4. Heavy-tailed case (Corollary 5)

We first present a truncation function, which is necessary to our robustification strategy for heavy-tailed random matrices.

Definition 3. For any $\tau > 0$, we define the truncation function $\psi_\tau : \mathbb{R} \to \mathbb{R}$ as follows:

$$\psi_\tau(x) = \begin{cases} -\tau \text{ if } x < \tau, \\ x \text{ if } |x| \leq \tau, \\ \tau \text{ if } x > \tau. \end{cases}$$

There is a standard method for extending a real function $\mathbb{R} \to \mathbb{R}$ to a function of symmetric matrices $\mathcal{S} \to \mathcal{S}$, by applying the function to the eigenvalues of the matrix. Specifically, given $A \in \mathcal{S}$, A can be written as

$$A = Q\begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} Q^T,$$

for some matrix Q such that $QQ^T = I$, where $(\lambda_1, \dots, \lambda_p)$ are the eigenvalues of A. We then define $\psi_\tau(A)$ by

$$\psi_\tau(A) = Q\begin{pmatrix} \psi_\tau(\lambda_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \psi_\tau(\lambda_p) \end{pmatrix} Q^T.$$

We can now state the first corollary of Theorem 4.

Corollary 13. *Assume that $\{M_1, \dots, M_n\}$ satisfies Assumption 1. Fix $\tau > 0$. Then for all $t > 0$, with probability at least $1 - \exp(-t)$ it holds that*

$$\left\|\frac{1}{n}\sum_{\ell=1}^{n}(\psi_\tau(M_\ell) - \mathbb{E}[\psi_\tau(M_\ell)])\right\| \leq 4\sqrt{2}\left\|\Sigma\right\|(\tau + \Gamma_n)\sqrt{\frac{4\mathbf{r}\left(\Sigma\right) + t}{n}}.$$

Proof of Corollary 13. Because $x \mapsto \psi_\tau(x)$ is 1-Lipschitz, the sequence of matrices $\{\psi_\tau(M_1) - \mathbb{E}[\psi_\tau(M_1)], \ldots, \psi_\tau(M_n) - \mathbb{E}[\psi_\tau(M_n)]\}$ satisfies Assumption 1. As they are bounded by τ and all have the same expectation (zero), therefore, we can apply Theorem 4 to yield the result. □

As stated in the outline of the proof, we now have to understand the difference between the expectation of the truncated matrices and the expectations of the (non-truncated) matrices themselves.

Proposition 14. *Fix $\tau > 0$. Under Assumption 2, we have*

$$\max_{1 \le \ell \le n} \|\mathbb{E}[\psi_\tau(M_\ell)] - \mathbb{E}[M_\ell]\| \le G(\tau).$$

Proof of Proposition 14. For any $\ell = 1, \ldots, n$, we have

$$\begin{aligned}
\|\mathbb{E}[\psi_\tau(M_\ell)] - \mathbb{E}[M_\ell]\| &\le \mathbb{E}[\|\psi_\tau(M_\ell) - M_\ell\|] \\
&= \mathbb{E}[(\|M_\ell\| - \tau)\mathbf{1}_{\{\|M_\ell\| - \tau > 0\}}] \\
&\le \int_0^\infty (u-\tau)\mathbf{1}_{\{u-\tau>0\}} \mathrm{d}\mathbb{P}(\|M_\ell\| - \tau \le u) \\
&= -\int_\tau^\infty (u-\tau)\mathrm{d}\mathbb{P}(\|M_\ell\| - \tau > u) \\
&= [-(u-\tau)\mathbb{P}(\|M_\ell\| - \tau > u)]_\tau^\infty \\
&\quad + \int_\tau^\infty \mathbb{P}(\|M_\ell\| - \tau > u)du.
\end{aligned}$$

The first term is null as $xS_\ell(x) = \int_0^\infty \mathbf{1}_{\{u \le x\}} S_\ell(x)du \le \int_0^\infty S_\ell(u)du$ and the dominated convergence theorem gives $xS_\ell(x) \to 0$ as $x \to \infty$, and thus

$$\begin{aligned}
\|\mathbb{E}[\psi_\tau(M_\ell)] - \mathbb{E}[M_\ell]\| &\le \int_\tau^\infty \mathbb{P}(\|M_\ell\| - \tau > u)du \\
&\le \int_\tau^\infty \mathbb{P}(\|M_\ell\| > u)du \\
&= \int_\tau^\infty S_\ell(u)du,
\end{aligned}$$

which ends the proof. □

Corollary 15. *Assume that $\{M_1, \ldots, M_n\}$ satisfies Assumptions 1 and 2, and $\mathbb{E}[M_\ell] = \Sigma$. Fix $\tau > 0$. For all $t > 0$, with probability at least $1 - \exp(-t)$ it holds that*

$$\left\| \frac{1}{n} \sum_{\ell=1}^n \psi_\tau(M_\ell) - \Sigma \right\| \le 4\sqrt{2}\, \|\Sigma\| \, (\tau + \Gamma_n) \sqrt{\frac{4\mathbf{r}\,(\Sigma) + t}{n}} + G(\tau).$$

Proof of Corollary 15. First, we decompose the norm as

$$\left\| \frac{1}{n} \sum_{\ell=1}^n \psi_\tau(M_\ell) - \Sigma \right\|$$

$$\leq \left\| \frac{1}{n} \sum_{\ell=1}^{n} (\psi_\tau(M_\ell) - \mathbb{E}[\psi_\tau(M_\ell)]) \right\| + \left\| \frac{1}{n} \sum_{\ell=1}^{n} (\mathbb{E}[\psi_\tau(M_\ell)] - \mathbb{E}[M_\ell]) \right\|$$
$$\leq \left\| \frac{1}{n} \sum_{\ell=1}^{n} [\psi_\tau(M_\ell) - \mathbb{E}[\psi_\tau(M_\ell)]] \right\| + \frac{1}{n} \sum_{\ell=1}^{n} \left\| \mathbb{E}[\psi_\tau(M_\ell)] - \Sigma \right\|,$$

where we use the triangle inequality in the first line, and Jensen's inequality and $\mathbb{E}[M_\ell] = \Sigma$ in the second line. As Assumption 1 is satisfied, we can upper bound the first term with probability $1 - \exp(-t)$ by Corollary 13 together with Proposition 1. Because Assumption 2 is also satisfied, we can bound the second term using Proposition 14. □

Note that Corollary 15 already provides an estimation result for Σ when matrices M_ℓ are unbounded. However, in contrast to Corollary 5, not only does the bound depend on τ but the estimator $\frac{1}{n}\sum_{\ell=1}^{n} \psi_\tau(M_\ell)$ does as well. A mistake in the choice of τ can lead to poor estimation in practice.

To control the distance between this estimator $\frac{1}{n}\sum_{\ell=1}^{n} \psi_\tau(M_\ell)$ and the standard estimator $\frac{1}{n}\sum_{\ell=1}^{n} M_\ell$, we prove the following proposition.

Proposition 16. *Under Assumption 2, we have*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{\ell=1}^{n} \psi_\tau(M_\ell) - \frac{1}{n} \sum_{\ell=1}^{n} M_\ell \right\| \neq 0 \right) \leq \sum_{\ell=1}^{n} S_\ell(t).$$

Proof of Proposition 16. We have

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{\ell=1}^{n} \psi_\tau(M_\ell) - \frac{1}{n} \sum_{\ell=1}^{n} M_\ell \right\| \neq 0 \right)$$
$$\leq \mathbb{P}\left(\frac{1}{n} \sum_{\ell=1}^{n} \left\| \psi_\tau(M_\ell) - M_\ell \right\| > 0 \right)$$
$$= \mathbb{P}\left(\exists \ell : \left\| \psi_\tau(M_\ell) - M_\ell \right\| > 0 \right)$$
$$\leq \sum_{\ell=1}^{n} S_\ell(t).$$

□

We can now prove Corollary 5.

Proof of Corollary 5. Using the triangle inequality,

$$\left\| \frac{1}{n} \sum_{\ell=1}^{n} M_\ell - \Sigma \right\| \leq \left\| \frac{1}{n} \sum_{\ell=1}^{n} M_\ell - \frac{1}{n} \sum_{\ell=1}^{n} \psi_\tau(M_\ell) \right\| + \left\| \frac{1}{n} \sum_{\ell=1}^{n} \psi_\tau(M_\ell) - \Sigma \right\|.$$

The assumptions of Corollary 5 include: $\{M_1, \ldots, M_n\}$ satisfy Assumptions 1 and 2, and $\mathbb{E}[M_\ell] = \Sigma$, which enables us to use Corollary 15 to upper bound the second term with probability $1 - \exp(-t)$. This also allows for the use of Proposition 16 to prove that the first term will be null with probability at least $1 - \sum_{\ell=1}^{n} S_\ell(\tau)$. □

5.5. Infinite-dimensional case (Theorem 6)

Proof of Theorem 6. For a sequence of $\mathbb{H}\otimes\mathbb{H}$-valued positive semi-definite symmetric random operators $\{M_1,\dots,M_n\}$ with $\mathbb{E}\left[M_\ell\right]=\Sigma$ and $\max_{1\leq\ell\leq n}\|M_\ell\|\leq\kappa^2$ almost surely for some $\kappa>0$ satisfying Assumption 1,

$$P\left(\sup_{\substack{u_k\in\mathbb{H}_k\\ \|u_k\|=1}}\left|\left\langle\left(\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right)u_k,u_k\right\rangle\right|\geq 4\sqrt{2}\,\|\Sigma\|\left(\kappa^2+\Gamma_n\right)\sqrt{\frac{4\mathbf{r}\left(\Sigma\right)+t}{n}}\right)$$
$$\leq\exp(-t),$$

because for Σ_k such that $\Sigma_k^{(j_1,j_2)}:=\mathbb{E}[M_\ell^{(j_1,j_2)}]$ and $\Sigma:=\mathbb{E}[M_\ell]$, $\|\Sigma_k\|\leq\|\Sigma\|$ and $\mathrm{Tr}\left(\Sigma_k\right)\leq\mathrm{Tr}\left(\Sigma\right)$, and Γ_n is also uniform for each, as evident from the proof. Note that for any $c\geq0$ and $k\in\mathbb{N}$,

$$\left\{\sup_{\substack{u_k\in\mathbb{H}_k\\ \|u_k\|=1}}\left|\left\langle\left(\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right)u_k,u_k\right\rangle\right|\geq c\right\}$$
$$\subset\left\{\sup_{\substack{u_{k+1}\in\mathbb{H}_{k+1}\\ \|u_{k+1}\|=1}}\left|\left\langle\left(\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right)u_{k+1},u_{k+1}\right\rangle\right|\geq c\right\},$$

and

$$\lim_{k\to\infty}\left\{\sup_{\substack{u_k\in\mathbb{H}_k\\ \|u_k\|=1}}\left|\left\langle\left(\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right)u_k,u_k\right\rangle\right|\geq c\right\}=\left\{\left\|\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right\|\geq c\right\}.$$

The continuity of P leads to

$$P\left(\left\|\frac{1}{n}\sum_{\ell=1}^{n}M_\ell-\Sigma\right\|\geq 4\sqrt{2}\,\|\Sigma\|\left(\kappa^2+\Gamma_n\right)\sqrt{\frac{4\mathbf{r}\left(\Sigma\right)+t}{n}}\right)\leq\exp(-t).$$

Then, using the same approach to extend Theorem 4 to Corollary 5, we obtain the statement. □

6. Conclusion

We studied the deviations of the empirical mean of random matrices from its expected value in the dependent, heavy-tailed case. The upper bound derived here is independent of the dimension of the matrices but depends on the trace of the expectation and the tail of the distribution. Additionally, the upper bound increases with the strength of the dependence between the matrices. The proof here is based on a variational inequality and robustification by truncation. Our result is applied to the estimation problem of covariance operators/matrices, parameter estimation in linear hidden Markov models, and linear regression under overparameterization.

A limitation of our result is the tightness of the obtained upper bound. It is difficult to achieve lower bounds when random matrices are dependent and heavy-tailed, while some lower bounds are known when they are independent and each element is Gaussian. Therefore, deriving lower bounds in this case is an interesting subject for future research.

Appendix A: Proof for Examples

Proof of Proposition 1. Let $f : E \to E$ be a 1-Lipschitz function and define $\mathcal{G}_\ell = \sigma(f(M_1), \dots, f(M_\ell))$. We aim to prove that, for any $g \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$, we have

$$|\mathbb{E}[g(f(M_{\ell+1}), \dots, f(M_n)) \mid \mathcal{G}_\ell] - \mathbb{E}[g(f(M_{\ell+1}, \dots, f(M_n))]| \leq \Gamma_{\ell,n}. \tag{12}$$

Let h be defined by $h(a_1, \dots, a_\ell) = g(f(a_1), \dots, f(a_\ell))$. Then $h \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$. Indeed,

$$\begin{aligned}
&|h(a_1, \dots, a_\ell) - h(b_1, \dots, b_\ell)| \\
&= |g(f(a_1), \dots, f(a_\ell)) - g(f(b_1), \dots, f(b_\ell))| \\
&\leq L \sum_{i=1}^{\ell} \|f(a_i) - f(b_i)\|_E \\
&\leq L \sum_{i=1}^{\ell} \|a_i - b_i\|_E,
\end{aligned}$$

where we used respectively the definition of h, the fact that $g \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$ and the fact that f is 1-Lipschitz. Thus, because $(M_1, \dots, M_n)$ satisfies Assumption 1 and $h \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$, then

$$|\mathbb{E}[h(M_{\ell+1}, \dots, M_n) \mid \mathcal{F}_\ell] - \mathbb{E}[h(M_{\ell+1}, \dots, M_n)]| \leq \Gamma_{\ell,n}$$

that we can rewrite as

$$|\mathbb{E}[g(f(M_{\ell+1}), \dots, f(M_n)) \mid \mathcal{F}_\ell] - \mathbb{E}[g(f(M_{\ell+1}, \dots, f(M_n))]| \leq \Gamma_{\ell,n}. \tag{13}$$

This is almost (12); however, the conditional expectation does not hold with respect to the correct σ-algebra. This is easily fixed because $\mathcal{G}_\ell \subseteq \mathcal{F}_\ell$. Thus,

$$\begin{aligned}
&|\mathbb{E}[g(f(M_{\ell+1}), \dots, f(M_n)) \mid \mathcal{G}_\ell] - \mathbb{E}[g(f(M_{\ell+1}, \dots, f(M_n))]| \\
&= |\mathbb{E}[\mathbb{E}[g(f(M_{\ell+1}), \dots, f(M_n)) \mid \mathcal{F}_\ell] \mid \mathcal{G}_\ell] - \mathbb{E}[g(f(M_{\ell+1}, \dots, f(M_n))]| \\
&\leq \mathbb{E}\left[|\mathbb{E}[g(f(M_{\ell+1}), \dots, f(M_n)) \mid \mathcal{F}_\ell] - \mathbb{E}[g(f(M_{\ell+1}, \dots, f(M_n))]| \mid \mathcal{G}_\ell\right] \\
&\leq \Gamma_{\ell,n},
\end{aligned}$$

by using (13). □

Proof of Proposition 3. We define $(\bar{\xi}_\ell)_{\ell\in\mathbb{Z}}$ as an independent copy Ξ. We fix $\ell \in \{1, \dots, n\}$; we verify (1). To do so, we define, for $m > \ell$,

$$\bar{X}_m = C(\xi_m, \xi_{m-1}, \dots, \xi_{\ell+1}, \bar{\xi}_\ell, \bar{\xi}_{\ell-1}, \bar{\xi}_{\ell-2}, \dots),$$

and $\bar{Y}_m = \bar{X}_m + \varepsilon_m$. We put $\mathcal{G}_\ell = \sigma(\xi_\ell, \xi_{\ell-1}, \xi_{\ell-2}, \dots; \varepsilon_\ell, \varepsilon_{\ell-1}, \dots)$. Then, for $g \in \mathrm{Lip}_{n-\ell}(\mathcal{S}, 1)$, we have

$$\begin{aligned}
&\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{F}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)] \\
&\quad = \mathbb{E}[\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{G}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)] \mid \mathcal{F}_\ell],
\end{aligned}$$

and we prove an upper bound on $\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{G}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)]$. Hence, we have

$$\begin{aligned}
&\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{G}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)] \\
&= \mathbb{E}[g(\bar{Y}_{\ell+1}\bar{Y}_{\ell+1}^\top, \dots, \bar{Y}_n\bar{Y}_n^\top) - g(Y_{\ell+1}Y_{\ell+1}^\top, \dots, Y_nY_n^\top) \mid \mathcal{G}_\ell] \\
&\le \sum_{m=\ell+1}^{n} \left\|\mathbb{E}\left[\bar{Y}_m\bar{Y}_m^\top - Y_mY_m^\top \mid \mathcal{G}_\ell\right]\right\| \\
&= \sum_{m=\ell+1}^{n} \left\|\mathbb{E}\left[(\bar{X}_m + \varepsilon_m)(\bar{X}_m + \varepsilon_m)^\top - (X_m + \varepsilon_m)(X_m + \varepsilon_m)^\top \mid \mathcal{G}_\ell\right]\right\| \\
&= \sum_{m=\ell+1}^{n} \left\|\mathbb{E}\left[\bar{X}_m\bar{X}_m^\top - X_mX_m^\top \mid \mathcal{G}_\ell\right]\right\| \\
&= \sum_{m=\ell+1}^{n} \left\|\mathbb{E}\left[\bar{X}_m\bar{X}_m^\top - \bar{X}_mX_m^\top + \bar{X}_mX_m^\top - X_mX_m^\top \mid \mathcal{G}_\ell\right]\right\| \\
&\le \sum_{m=\ell+1}^{n} \Big(\left\|\mathbb{E}\left[\bar{X}_m\bar{X}_m^\top - \bar{X}_mX_m^\top \mid \mathcal{G}_\ell\right]\right\| + \left\|\mathbb{E}\left[\bar{X}_mX_m^\top - X_mX_m^\top \mid \mathcal{G}_\ell\right]\right\|\Big) \\
&\le \sum_{m=\ell+1}^{n} B\Big(\left\|\mathbb{E}\left[\bar{X}_m^\top - X_m^\top \mid \mathcal{G}_\ell\right]\right\| + \left\|\mathbb{E}\left[\bar{X}_m - X_m \mid \mathcal{G}_\ell\right]\right\|\Big).
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&\left\|\mathbb{E}\left[\bar{X}_m - X_m \mid \mathcal{G}_\ell\right]\right\| \\
&= \left\|\mathbb{E}\left[C(\xi_m, \dots, \xi_{\ell+1}, \bar{\xi}_\ell, \bar{\xi}_{\ell-1}, \dots) - C(\xi_m, \dots, \xi_{\ell+1}, \xi_\ell, \xi_{\ell-1}, \dots) \mid \mathcal{G}_\ell\right]\right\| \\
&\le \sum_{i=m-\ell}^{\infty} \alpha_i \mathbb{E}\left[\|\bar{\xi}_{m-i} - \xi_{m-i}\| \mid \mathcal{G}_\ell\right] \le 2 \sum_{i=m-\ell}^{\infty} \alpha_i B_\xi,
\end{aligned}$$

and thus,

$$\begin{aligned}
\mathbb{E}[g(M_{\ell+1}, \dots, M_n) \mid \mathcal{G}_\ell] - \mathbb{E}[g(M_{\ell+1}, \dots, M_n)] &\le \sum_{m=\ell+1}^{n} \left[4B \sum_{i=m-\ell}^{\infty} \alpha_i B_\xi\right] \\
&\le 4B \sum_{i=\ell+1}^{\infty} \min(i, n)\alpha_i B_\xi.
\end{aligned}$$

Thus, (1) is satisfied with $\Gamma_{\ell,n} = 4BB_\xi \sum_{i=\ell+1}^{\infty} \min(i,n)\alpha_i$. Let us now verify Assumption 2. We have

$$\begin{aligned}\mathbb{P}(\|M_\ell\| \geq t) &= \mathbb{P}(\|(X_\ell + \varepsilon_\ell)(X_\ell + \varepsilon_\ell)^\top\| \geq t)\\ &= \mathbb{P}(\|X_\ell + \varepsilon_\ell\|^2 \geq t)\\ &= \mathbb{P}(\|X_\ell + \varepsilon_\ell\| \geq \sqrt{t})\\ &\leq \mathbb{P}(\|X_\ell\| \geq \sqrt{t}/2) + \mathbb{P}(\|\varepsilon_\ell\| \geq \sqrt{t}/2)\\ &\leq \mathbf{1}_{\{t \leq 4B^2\}} + \mathbb{P}(\|\varepsilon_\ell\| \geq \sqrt{t}/2),\end{aligned}$$

which ends the proof. □

Proof of Proposition 8. Since $(X_\ell)_{\ell\in\mathbb{N}}$ is a CBS, $X_\ell = C(\xi_\ell, \xi_{\ell-1}, \xi_{\ell-2}, \dots)$ with

$$\|C(a_1, a_2, \dots) - C(b_1, b_2, \dots)\| \leq \sum_{\ell=1}^{\infty} \alpha_\ell \|a_\ell - b_\ell\| \text{ and } \mathcal{A} := \sum_{\ell=1}^{\infty} \alpha_\ell < \infty.$$

Using the form, we show that $(\widetilde{X}_\ell)_{\ell\in\mathbb{N}} := ((X_\ell, X_{\ell+1})^\top)_{\ell\in\mathbb{N}}$ is also a CBS, since we have

$$\widetilde{X}_\ell = (C(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots), C(\xi_{t-1}, \xi_{t-2}, \xi_{t-3}, \dots)) = D(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots)$$

with some function D which satisfies

$$\|D(a_1, a_2, \dots) - D(b_1, b_2, \dots)\| \leq \sum_{\ell=1}^{\infty} (\alpha_\ell + \alpha_{\ell+1}) \|a_\ell - b_\ell\|.$$

Since $(\widetilde{X}_\ell)_{\ell\in\mathbb{N}}$ is a CBS, $(M_1, \dots, M_{\ell-1})$ satisfies Assumption 1 with $\Gamma_{\ell,n} = 8BB_\xi \sum_{i=\ell+1}^{n} \min(i,n)\alpha_i$ and $\Gamma_n := 8BB_\xi \sum_{i=2}^{n} \min(i,n)\alpha_i$ by Proposition 3.

For Assumption 2, we utilize the fact that the largest eigenvalue of a matrix is no more than a sum of the largest eigenvalues of its submatrices and obtain

$$\begin{aligned}&\mathbb{P}(\|M_\ell\| \geq t)\\ &\leq \mathbb{P}(\|(X_\ell + \varepsilon_\ell)(X_\ell + \varepsilon_\ell)^\top\| \geq t/4)\\ &\quad + \mathbb{P}(\|(X_\ell + \varepsilon_\ell)(X_{\ell+1} + \varepsilon_{\ell+1})^\top\| \geq t/2)\\ &\quad + \mathbb{P}(\|(X_{\ell+1} + \varepsilon_{\ell+1})(X_{\ell+1} + \varepsilon_{\ell+1})^\top\| \geq t/4)\\ &= 2\mathbb{P}(\|X_\ell + \varepsilon_\ell\|^2 \geq t/4) + \mathbb{P}(\|(X_\ell + \varepsilon_\ell)(X_{\ell+1} + \varepsilon_{\ell+1})^\top\| \geq t/2)\\ &\leq 2\mathbb{P}(\|X_\ell + \varepsilon_\ell\| \geq \sqrt{t}/2) + \mathbb{P}(\|X_\ell + \varepsilon_\ell\|\|X_{\ell+1} + \varepsilon_{\ell+1}\| \geq t/2)\\ &\leq 2\mathbb{P}(\|X_\ell + \varepsilon_\ell\| \geq \sqrt{t}/2) + 2\mathbb{P}(\|X_\ell + \varepsilon_\ell\| \geq \sqrt{t/2})\\ &\leq 4\mathbb{P}(\|X_\ell\| \geq \sqrt{t/2}) + 4\mathbb{P}(\|\varepsilon_\ell\| \geq \sqrt{t/2})\\ &\leq 4\mathbf{1}_{\{t \leq 4B^2\}} + 4\mathbb{P}(\|\varepsilon_\ell\| \geq \sqrt{t/2}).\end{aligned}$$

Hence, Assumption 2 holds $G(\tau) = 4\mathbf{1}_{\{t \leq 2B^2\}} + 4\mathbb{P}(\|\varepsilon_\ell\| \geq \sqrt{t/2})$. Thus, Corollary 5 shows the first statement.

Finally, the fact $\|\Sigma_{0:1}\| \leq \|\Sigma_1\| + \|\Sigma\|$ yields the second statement. □

Proof of Proposition 9. First, we confirm that $(X_\ell)_{\ell\in\mathbb{Z}}$ is a CBS in Example 1 by its definition. Hence, by Proposition 2, a sequence of matrices generated by $Y_\ell Y_\ell^\top$ satisfies Assumption 1 and 2.

We show that the estimation error $\|\widehat{A}-A\|$ is bounded by the estimation error of $\widehat{\Sigma}$ and $\widehat{\Sigma}_1$. We bound the error as

$$\begin{aligned}\|\widehat{A}-A\| &= \|(\widehat{\Sigma}_1-\Sigma_1)(\widehat{\Sigma}+I)^{-1}+\Sigma_{Y,1}((\widehat{\Sigma}+I)^{-1}-(\Sigma+I)^{-1})\|\\ &\leq \|\widehat{\Sigma}_1-\Sigma_1\|\|(\widehat{\Sigma}+I)^{-1}\|+\|\Sigma_1(\Sigma+I)^{-1}(\Sigma-\widehat{\Sigma})(\widehat{\Sigma}+I)^{-1}\|\\ &\leq \|\widehat{\Sigma}_1-\Sigma_1\|+\|\Sigma-\widehat{\Sigma}\|\|\Sigma_1\|. \end{aligned}\tag{14}$$

Here, we use the facts $\|(\widehat{\Sigma}+I)^{-1}\|\leq 1$ and $\|(\Sigma+I)^{-1}\|\leq 1$.

We combine the above results. Using the same discussion for Proposition 8 in Section 4.2, we have

$$\begin{aligned}&\max\{\|\widehat{\Sigma}-\Sigma\|,\|\widehat{\Sigma}_1-\Sigma_1\|\}\\ &\leq 4\sqrt{2}\left(\|\Sigma_1\|+\|\Sigma\|\right)\left(\tau+\Gamma_n\right)\sqrt{\frac{4\mathbf{r}\left(\Sigma_{0:1}\right)+t}{n}}+G(\tau),\end{aligned}$$

where the definition of $\Sigma_{0:1}$ follows Section 4.2. We combine this inequality with the result (14), and we obtain the statement. □

Proof of Proposition 10. By Lemma 7 in [6], the risk $R(\widehat{\theta})$ is evaluated as

$$\begin{aligned}R(\widehat{\theta}) &\leq 2(\theta^*)^\top(I-\Pi_{\mathsf{Y}})\Sigma(I-\Pi_{\mathsf{Y}})\theta^*+\sigma^2\mathrm{Tr}((\mathsf{Y}\mathsf{Y}^\top)^{-1}\mathsf{Y}\Sigma\mathsf{Y}^\top(\mathsf{Y}\mathsf{Y}^\top)^{-1})\\ &=2(\theta^*)^\top B\theta^*+ct\sigma^2\mathrm{Tr}(C),\end{aligned}$$

where $B=(I-\Pi_{\mathsf{Y}})\Sigma(I-\Pi_{\mathsf{Y}})$. We bound the first term as

$$\begin{aligned}(\theta^*)^\top B\theta^* &= (\theta^*)^\top(I-\Pi_{\mathsf{Y}})\Sigma(I-\Pi_{\mathsf{Y}})\theta^*\\ &=(\theta^*)^\top(I-\Pi_{\mathsf{Y}})(\Sigma-n^{-1}\mathsf{Y}^\top\mathsf{Y})(I-\Pi_{\mathsf{Y}})\theta^*\\ &\leq\|\theta^*\|^2\|I-\Pi_{\mathsf{Y}}\|\|\Sigma-n^{-1}\mathsf{Y}^\top\mathsf{Y}\|\\ &\leq\|\theta^*\|^2\|\Sigma-n^{-1}\mathsf{Y}^\top\mathsf{Y}\|,\end{aligned}$$

where the second equality follows $(I-\Pi_{\mathsf{Y}})\mathsf{Y}^\top=(I-\mathsf{Y}^\top(\mathsf{Y}\mathsf{Y}^\top)^{-1}\mathsf{Y})\mathsf{Y}^\top=\mathsf{Y}^\top-\mathsf{Y}^\top(\mathsf{Y}\mathsf{Y}^\top)^{-1}(\mathsf{Y}\mathsf{Y}^\top)=0$, and the second inequality follows $\|I-\Pi_{\mathsf{Y}}\|\leq 1$ from the non-expansive property of projection operators. Recalling that $n^{-1}\mathsf{Y}^\top\mathsf{Y}=\widehat{\Sigma}$ as in (2), Proposition 7 yields the statement. □

Funding

The first author's work is supported by Japan Science and Technology Agency CREST (JPMJCR21D2). The second author's work is partially funded by CY Initiative of Excellence (grant "Investissements d'Avenir" ANR-16-IDEX-0008), Project "EcoDep" PSI-AAP2020-0000000013. The third author's work is supported by Japan Society for the Promotion of Science KAKENHI (21K11780) and Japan Science and Technology Agency FOREST (JPMJFR216I).

References

[1] Abdalla, P. and Zhivotovskiy, N. (2022) Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails, *arXiv preprint arXiv:2205.08494*.

[2] Adamczak, R., Litvak, A., Pajor, A. and Tomczak-Jaegermann, N. (2010) Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles, *Journal of the American Mathematical Society*, **23**, 535–561. MR2601042

[3] Alquier, P. (2024) User-friendly introduction to PAC-Bayes bounds, *Foundations and Trends in Machine Learning*, **17**, 174–303.

[4] Alquier, P. and Wintenberger, O. (2012) Model selection for weakly dependent time series forecasting, *Bernoulli*, **18**, 883–913. MR2948906

[5] Andrews, D. W. (1984) Non-strong mixing autoregressive processes, *Journal of Applied Probability*, **21**, 930–934. MR0766830

[6] Bartlett, P. L., Long, P. M., Lugosi, G. and Tsigler, A. (2020) Benign overfitting in linear regression, *Proceedings of the National Academy of Sciences*, **117**, 30063–30070. MR4263288

[7] Bunea, F. and Xiao, L. (2015) On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca, *Bernoulli*, **21**, 1200–1230. MR3338661

[8] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010) Optimal rates of convergence for covariance matrix estimation, *The Annals of Statistics*, **38**, 2118–2144. MR2676885

[9] Catoni, O. (2007) *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, Institute of Mathematical Statistics Lecture Notes – Monograph Series, 56, Institute of Mathematical Statistics, Beachwood, OH, Ohio. MR2483528

[10] Catoni, O. (2012) Challenging the empirical mean and empirical variance: a deviation study, in *Annales de l'IHP Probabilités et statistiques*, vol. 48, pp. 1148–1185. MR3052407

[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] Dedecker, J., Doukhan, P., Lang, G., José Rafael, L. R., Louhichi, S. and Prieur, C. (2007) Weak dependence, in *Weak dependence: With examples and applications*, Springer, pp. 9–20. MR2338725

[13] Dedecker, J. and Prieur, C. (2005) New dependence coefficients. examples and applications to statistics, *Probability Theory and Related Fields*, **132**, 203–236. MR2199291

[14] Donoho, D. L. (2006) Compressed sensing, *IEEE Transactions on information theory*, **52**, 1289–1306. MR2241189

[15] Doukhan, P. and Wintenberger, O. (2008) Weakly dependent chains with infinite memory, *Stochastic Processes and their Applications*, **118**, 1997–2013. MR2462284

[16] Giulini, I. (2018) Robust dimension-free gram operator estimates,

Bernoulli, **24**, 3864–3923. MR3788191

[17] Guédon, O. and Rudelson, M. (2007) Lp-moments of random vectors via majorizing measures, *Advances in Mathematics*, **208**, 798–823. MR2304336

[18] Han, F. and Li, Y. (2020) Moment bounds for large autocovariance matrices under dependence, *Journal of Theoretical Probability*, **33**, 1445–1492. MR4125963

[19] Han, Q. (2022) Exact spectral norm error of sample covariance, *arXiv preprint arXiv:2207.13594*.

[20] Jeong, H., Li, X., Plan, Y. and Yilmaz, O. (2022) Sub-gaussian matrices on sets: Optimal tail dependence and applications, *Communications on Pure and Applied Mathematics*, **75**, 1713–1754. MR4465901

[21] Koltchinskii, V. and Lounici, K. (2017) Concentration inequalities and moment bounds for sample covariance operators, *Bernoulli*, **23**, 110–133. MR3556768

[22] Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection, *Annals of statistics*, pp. 1302–1338. MR1805785

[23] Liaw, C., Mehrabian, A., Plan, Y. and Vershynin, R. (2017) A simple tool for bounding the deviation of random matrices on geometric sets, in *Geometric aspects of functional analysis*, Springer, New York, pp. 277–299. MR3645128

[24] Lopes, M. E., Erichson, N. B. and Mahoney, M. W. (2023) Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching, *Bernoulli*, **29**, 428–450. MR4497253

[25] Mendelson, S. and Paouris, G. (2014) On the singular values of random matrices, *Journal of the European Mathematical Society*, **16**, 823–834. MR3191978

[26] Oliveira, R. I. (2009) Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges, *arXiv preprint arXiv: 0911.0600*. MR2447295

[27] Rio, E. (2000) Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes, *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, **330**, 905–908. MR1771956

[28] Rudelson, M. (1999) Random vectors in the isotropic position, *Journal of Functional Analysis*, **164**, 60–72. MR1694526

[29] Srivastava, N. and Vershynin, R. (2013) Covariance estimation for distributions with 2+ ϵ moments, *The Annals of Probability*, **41**, 3081–3111. MR3127875

[30] van Handel, R. (2017) Structured random matrices, *Convexity and concentration*, pp. 107–156. MR3837269

[31] Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, Cambridge. MR3837109

[32] Zhivotovskiy, N. (2024) Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle, *Electronic Journal of Probability*, **29**, 1–28. MR4693860