Fundamental Limits of Distributed Covariance Matrix Estimation Under Communication Constraints

Mohammad Reza Rahmani[1] **Mohammad Hossein Yassaee**[1 2] **Mohammad Ali Maddah-Ali**[3]
Mohammad Reza Aref[1]

Abstract

Estimating high-dimensional covariance matrices is crucial in various domains. This work considers a scenario where two collaborating agents access disjoint dimensions of m samples from a high–dimensional random vector, and they can only communicate a limited number of bits to a central server, which wants to accurately approximate the covariance matrix. We analyze the fundamental trade–off between communication cost, number of samples, and estimation accuracy. We prove a lower bound on the error achievable by any estimator, highlighting the impact of dimensions, number of samples, and communication budget. Furthermore, we present an algorithm that achieves this lower bound up to a logarithmic factor, demonstrating its near-optimality in practical settings.

1. Introduction

Estimating the covariance matrix of a random vector from its i.i.d samples is one of the primary problems in various fields, such as financial mathematics, statistics, and machine learning (Hotelling, 1933; Dahmen et al., 2000; Ledoit & Wolf, 2003). Let $\{\mathbf{Z}^{(i)}\}_{i=1}^{m} = \{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}\}$ be m i.i.d. samples of a random vector $\mathbf{Z}$. Then its covariance matrix can be estimated using *sample covariance estimator*, as:

$$\widehat{\mathbf{C}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Z}^{(i)}\mathbf{Z}^{(i)\top}. \tag{1}$$

For sub–Gaussian random vector $\mathbf{Z}$ (see Definition 5.2), (Vershynin, 2018, Theorem 4.7.1) establishes some bounds on the operator norm of the estimation error of estimator (1). For alternative assumptions on the covariance matrix $\mathbf{C} = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$, such as sparsity, low-rankness, and Toeplitz–structure, various estimators have been reported in the literature (Huang et al., 2006; Furrer & Bengtsson, 2007; Bickel et al., 2008; Bickel & Levina, 2008; El Karoui, 2008; Wu & Pourahmadi, 2009; Chen et al., 2012). These estimators differ from the traditional sample covariance estimator. Also in some studies, such as (Cai et al., 2010; 2013), the optimality of some covariance matrix estimators is investigated.

In distributed settings, like federated learning (McMahan et al., 2017), the data may be distributed among multiple agents, with each agent having access to only a subset of the data. One can imagine two classes of problems: (i) *sample–split* (or horizontal split), where each agent has access to a subset of samples. (ii) *feature–split (or vertical split)*, where each agent has access to a subset of dimensions (or features) of all samples.

Extensive research papers explore the extension of core machine learning algorithms to distributed scenarios with sample-split settings. For example, the problem of distributed principal component analysis for dimension reduction in sample-split settings has been investigated in (Qu et al., 2002; Bai et al., 2005; Balcan et al., 2014; Kannan et al., 2014). In addition, the distributed gradient descent algorithm in this setting has been studied in (Langford et al., 2009; Zinkevich et al., 2010; Niu et al., 2011). Moreover, the distributed support vector machine is studied in (Navia-Vázquez et al., 2006; Zhu et al., 2007; Lu et al., 2008; Forero et al., 2010).

In contrast, in the feature split setting each agent has access to a subset of dimensions for all data points. This situation can arise in medical data, where a part of the health data of each patient is stored in a different database (Allaart et al., 2022). Another example is when some weather stations collect the weather information of various regions of a country, and we want to estimate the correlations between them, without sending all of the information to a central server. Some studies extend some machine learning tasks to the vertical split setting, such as (Yang et al., 2019; Shen et al., 2019; Hadar et al., 2019; Hadar & Shayevitz, 2019; Wu

[1]Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran [2]Sharif Center for Information Systems and Data Science, Sharif Institute for Convergence Science & Technology, Tehran, Iran [3]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA. Correspondence to: Mohammad Hossein Yassaee <yassaee@sharif.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

et al., 2020).

In this paper, we consider the problem of estimating the covariance matrix in a vertical-split setting, under communication constraints. In particular, we consider a distributed system consisting of two agents and a central server, where Agent 1 and Agent 2 have access to d_1 and d_2 dimensions of m i.i.d. samples of a random vector, respectively. The goal is to estimate the covariance matrix on the central server. Due to a limited communication budget, Agent 1 can send messages with at most B_1 bits to the central server. Similarly, Agent 2 has a communication budget of B_2 bits to communicate with the master. Consequently, the central server estimates the covariance matrix by processing the received messages. The main questions here are twofold: (1) Considering the agents' limited communication budgets and the restricted number of samples, what is the ultimate accuracy in the estimation? (2) How can this ultimate accuracy be achieved?

This paper answers both of these questions. We will find the fundamental information–theoretic lower–bound on the accuracy of the covariance estimation. Additionally, we will introduce a scheme for estimating the covariance matrix while respecting the communication limits inherent in the problem. We will prove that the estimation error of the proposed scheme matches with the obtained lower bound within a logarithmic factor in practical settings.

Prior works: Several research papers have studied the problem of distributed covariance estimation with limited communication budgets, with some focusing on the horizontal-split case (Zhang et al., 2013; Braverman et al., 2016; Han et al., 2018), while others concentrate on vertical-split cases (Hadar et al., 2019; Hadar & Shayevitz, 2019). In particular, (Hadar et al., 2019) investigates the problem of estimating the correlation $\rho = \mathbb{E}[XY]$ between two *scalar* ($d_1 = d_2 = 1$) *Gaussian* or *binary* random variables X, Y in the vertical split settings, where only Agent 1 has limited communication budget (i.e., $B_2 = \infty$) and the number of samples is unbounded ($m = \infty$). For this set-up, (Hadar et al., 2019) characterizes the exact order of the optimal communication budget for any estimation accuracy. (Hadar & Shayevitz, 2019) proposes a solution for the case where the objective is to estimate correlation $\mathbb{E}[X_k Y]$ between *a vector* $\mathbf{X} = [X_1, \cdots, X_d]^\top$ and a *scalar* Y ($d_1 > d_2 = 1$), without any claim on its optimality. The proposed solution in (Hadar & Shayevitz, 2019) outperforms the solutions based on estimating the correlation $\mathbb{E}[X_k Y]$, for each k, separately.

Our contributions: In this paper, we address the problem of distributed covariance matrix estimation, for the general family of *sub–Gaussian random vectors* with *finite number of samples*, and *limited communication budget between agents and the central server*.

Our main contributions are:

- We derive a near optimal trade–off curve between the number of samples, communication budgets, the number of dimensions each agent has access to, and the expected estimation error in the distributed covariance matrix estimation problem.

- We prove that any estimation algorithm with parameters (m, d_1, d_2, B_1, B_2) has the error of $\Omega\left(\max\left\{\sqrt{d/m}, \sqrt{d_1 d_2/B_{\min}}, 2^{-\min\{\frac{B_1}{d_1^2}, \frac{B_2}{d_2^2}\}}\right\}\right)$, where $B_{\min} = \min\{B_1, B_2\}$.

- Interestingly, to achieve a satisfactory approximation, it is necessary to increase the strength of the communication link between the poor agent (the agent with low-dimensional input) in proportion to the dimension of the rich agent.

- We also propose a scheme for achieving an expected operator norm of the error matrix $\widetilde{\mathcal{O}}\left(\sqrt{d/m} + \sqrt{\max\left\{d_1^2/B_1, d_2^2/B_2, d_1 d_2/B_{\min}\right\}}\right)$.

- We extend the method used in (Hadar et al., 2019; Hadar & Shayevitz, 2019) to a similar setting with our problem, and show that the obtained result from their method is *weaker* than ours.

The paper is structured as follows: In Section 2, we review the notations. In Section 3, we present the problem formulation formally. Section 4 is dedicated to reviewing the main results. Section 5 reviews some definitions and lemmas used in establishing the results. In Section 6, we prove the lower-bound and compare it with the results of (Hadar et al., 2019; Hadar & Shayevitz, 2019). In Section 7, we state the achievable scheme. Finally, we conclude the paper in Section 8.

2. Notations

We use uppercase bold symbols, like $\mathbf{A}$, to denote matrices and lowercase bold symbols, like $\mathbf{v}$, for denoting vectors. For any vector $\mathbf{v} = [v_1, v_2, \cdots, v_d]^\top$, we define the ℓ_p-norm as $\|\mathbf{v}\|_p = \left(\sum_{i=1}^d v_i^p\right)^{1/p}$. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the operator norm and the Frobenius norm are denoted by $\|\mathbf{A}\|_{\mathsf{op}}$ and $\|\mathbf{A}\|_{\mathsf{F}}$, respectively. A generic norm, $\|.\|_{\mathsf{dist}}$, is defined, encompassing both operator and Frobenius norms for flexibility.

3. Problem Formulation

We consider a system including a central server and two agents, as shown in Fig. 1. The central server is interested in estimating the covariance matrix $\mathbf{C} = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$ of a d-dimensional σ sub–Gaussian random vector $\mathbf{Z} \sim P$ (see Definition 5.2), from m i.i.d. samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}$. However, the central server does not have direct access to these samples. Rather, Agent 1 has full knowledge of the first d_1 dimensions of all m samples, denoted by $\{\mathbf{X}^{(i)}\}_{i=1}^m = \{\mathbf{Z}^{(i)}_{[1:d_1]}\}_{i=1}^m$, and Agent 2 is aware of the remaining $d_2 = d - d_1$ dimensions, denoted by $\{\mathbf{Y}^{(i)}\}_{i=1}^m = \{\mathbf{Z}^{(i)}_{[d_1+1:d]}\}_{i=1}^m$. The central server aims to estimate $\mathbf{C}$ by receiving up to B_1 and B_2 from agents one and two respectively. We also define $B_{\min} = \min\{B_1, B_2\}$ and $d_{\min} = \min\{d_1, d_2\}$ and will use these notations through this paper.

We refer to this problem of distributed covariance matrix estimation (DCME) with parameters $(\sigma, m, d_{1:2}, B_{1:2})$ as $\mathsf{DCME}(\sigma, m, B_{1:2}, d_{1:2})$. More formally, this problem consists of two encoder functions and one decoder function as follows:

- Two encoder functions $\mathcal{E}_1 : \mathbb{R}^{d_1 \times m} \mapsto [1 : 2^{B_1}]$ and $\mathcal{E}_2 : \mathbb{R}^{d_2 \times m} \mapsto [1 : 2^{B_2}]$, where encoder one maps $\{\mathbf{X}^{(i)}\}_{i=1}^m$ to $M_1 = \mathcal{E}_1(\{\mathbf{X}^{(i)}\}_{i=1}^m)$ and encoder two maps $\{\mathbf{Y}^{(i)}\}_{i=1}^m$ to $M_2 = \mathcal{E}_2(\{\mathbf{Y}^{(i)}\}_{i=1}^m)$.
- A decoder function $\mathcal{D} : [1 : 2^{B_1}] \times [1 : 2^{B_2}] \mapsto \mathsf{S}_+^{d\times d}$, where $\mathsf{S}_+^{d\times d}$ is the set of positive semi-definite matrices of dimension $d \times d$. The decoder function maps (M_1, M_2) to $\widehat{\mathbf{C}} = \mathcal{D}(M_1, M_2)$.

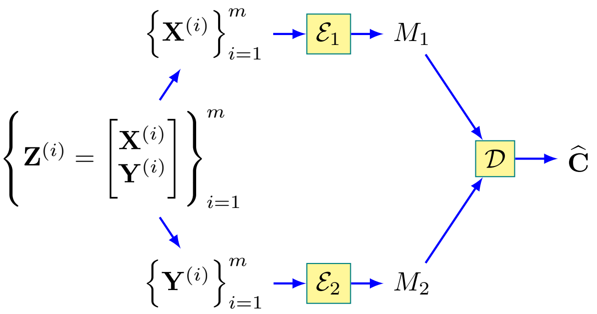
The distortion of a DCME scheme, under the dist norm, where dist can be either the operator norm or Frobenius norm, is quantified by the dist norm of the difference between the estimated covariance matrix $\widehat{\mathbf{C}}$ and the true covariance matrix $\mathbf{C}$, in other words, $\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C}) = \left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{dist}}$. The expected distortion of a DCME scheme is assessed by:

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right] = \mathbb{E}_{\{\mathbf{Z}^{(i)}\}_{i=1}^m \sim P^{\otimes m}}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{dist}}\right]. \quad (2)$$

The objective is to design the encoding functions $\mathcal{E}_1(.)$ and $\mathcal{E}_2(.)$ and the decoding function $\mathcal{D}(.,.)$, minimizing $\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right]$ and characterize $\min \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right]$ as a function of the parameters $(\sigma, m, d_{1:2}, B_{1:2})$.

4. Main Results

In this paper, we state two main theorems about the expected distortion of $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ scheme. The first theorem presents a general lower bound for any $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ scheme and the second one proposes a DCME scheme which has an expected distortion that is matched with the derived lower bound, in certain practical regimes.



$\left\{\mathbf{X}^{(i)}\right\}_{i=1}^m$
$\mathcal{E}_1$
M_1
$\left\{\mathbf{Z}^{(i)} = \begin{bmatrix}\mathbf{X}^{(i)}\\ \mathbf{Y}^{(i)}\end{bmatrix}\right\}_{i=1}^m$
$\mathcal{D}$
$\widehat{\mathbf{C}}$
$\left\{\mathbf{Y}^{(i)}\right\}_{i=1}^m$
$\mathcal{E}_2$
M_2

Figure 1. Setting of the problem $\mathsf{DCME}(\sigma, m, B_{1:2}, d_{1:2})$. $\mathbf{Z} \in \mathbb{R}^d$ is a σ–sub–Gaussian random vector with covariance matrix $\mathbf{C}$. $\mathbf{X}, \mathbf{Y}$ contain the first d_1 and the reminder d_2 dimensions of $\mathbf{Z}$, respectively. The $\widehat{\mathbf{C}}$ is an estimation of $\mathbf{C}$, with the constraint that $H(M_1) \leq B_1$ and $H(M_2) \leq B_2$.

4.1. The Lower Bound

In the first theorem, we use a min–max argument to find a lower bound on the expected distortion $\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right]$ in any DCME scheme with the parameters $(\sigma, m, d_{1:2}, B_{1:2})$.

Let $\mathcal{P} = \mathsf{subG}^{(d)}(\sigma)$ denote the family of σ–sub–Gaussian d–dimensional distributions. Then the min–max error metric under dist norm is defined as follows:

$$\mathcal{M}_{\mathsf{dist}}(\sigma, m, d_{1:2}, B_{1:2}) := \inf_{\mathcal{E}_1, \mathcal{E}_2, \mathcal{D}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right].$$

The main theorem on the lower bound is as follows:

Theorem 4.1. *Consider the problem of* $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$. *Then, for any choice of the encoder functions* $\mathcal{E}_1, \mathcal{E}_2$ *and the decoder function* $\mathcal{D}$, $\mathcal{M}_{\mathsf{dist}}(\sigma, m, d_{1:2}, B_{1:2})$ *is lower-bounded as:*

$$\mathcal{M}_{\mathsf{op}} \geq \frac{\sigma^2}{32} \min\left\{ \max\left\{ \sqrt{\frac{d_1 d_2}{2B_{\min}}}, \sqrt{\frac{d}{3m}}, 8 \cdot 2^{\frac{-16B_1}{d_1^2}}, 8 \cdot 2^{\frac{-16B_2}{d_2^2}} \right\}, 2 \right\}, \quad (3)$$

and

$$\mathcal{M}_{\mathsf{F}} \geq \frac{\sigma^2}{32} \min\left\{ \max\left\{ \sqrt{\frac{d_1 d_2 d_{\min}}{14B_{\min}}}, \sqrt{\frac{d^2}{42m}}, \frac{4\sqrt{d_1}}{7} \cdot 2^{\frac{-16B_1}{d_1^2}}, \frac{4\sqrt{d_2}}{7} \cdot 2^{\frac{-16B_2}{d_2^2}} \right\}, \frac{\sqrt{d}}{7} \right\}. \quad (4)$$

The above theorem states that given parameters $(\sigma, m, d_{1:2}, B_{1:2})$, there is no DCME scheme that can achieve an error with operator norm less than $\mathcal{O}\left(\sigma^2 \max\left\{\sqrt{\frac{d_1 d_2}{B_{\min}}}, \sqrt{\frac{d}{m}}, 2^{-\min\{\frac{B_1}{d_1^2}, \frac{B_2}{d_2^2}\}}\right\}\right)$, for any distribution $P \in \mathsf{subG}^{(d)}(\sigma)$.

The proof of Theorem 4.1 can be found in Section 6.1. Here, we highlight the main steps of the proof. To prove Theorem 4.1, we first reduce the estimation problem to a finite hypothesis testing problem between a family of distributions $\{P_v\}_{v\in\mathcal{V}}$ with the corresponding covariance matrices $\{\mathbf{C}_v\}_{v\in\mathcal{V}}$, where $\mathcal{V}$ is a set with finite cardinality. Since $\mathbf{Z}$ is a sub–Gaussian random vector, we use Gaussian distributions to *p*ack the set of σ–sub–Gaussian distributions. For any covariance matrix $\overline{\mathbf{C}}$, we consider the following decomposition of it:

$$\overline{\mathbf{C}} = \begin{bmatrix} \overline{\mathbf{C}}_{\mathbf{XX}} & \overline{\mathbf{C}}_{\mathbf{XY}} \\ \overline{\mathbf{C}}_{\mathbf{XY}}^{\top} & \overline{\mathbf{C}}_{\mathbf{YY}} \end{bmatrix}, \tag{5}$$

where $\overline{\mathbf{C}}_{\mathbf{XX}} = \overline{\mathbf{C}}_{[1:d_1,1:d_1]}$, $\overline{\mathbf{C}}_{\mathbf{XY}} = \overline{\mathbf{C}}_{[1:d_1,d_1+1:d]}$, and $\overline{\mathbf{C}}_{\mathbf{YY}} = \overline{\mathbf{C}}_{[d_1+1:d,d_1+1:d]}$. Our lower bound is based on finding separate lower bounds for approximating each part of the *true* covariance matrix $\mathbf{C}$. To do this, we introduce two different families of Gaussian distributions.

Family with varying cross–covariance: The first family of Gaussian distributions, denoted by $\{P_v\}_{v\in\mathcal{V}}$, has the corresponding covariance matrices $\{\mathbf{C}_v\}_{v\in\mathcal{V}}$. Through this family, the self–covariance matrices $\mathbf{C}_{v,\mathbf{XX}}$ and $\mathbf{C}_{v,\mathbf{YY}}$ are fixed for all $v \in \mathcal{V}$. However, cross–covariance $\mathbf{C}_{v,\mathbf{XY}}$ is varying (See Lemma 6.2). In this setup, we apply some strong data processing inequality (See Definition 5.6) to obtain a lower bound on the error probability of the hypothesis testing problem among the members $\{P_v\}_{v\in\mathcal{V}}$. The first two terms, within $\max$ in (3) and (4), correspond to the lower bound obtained by this packing.

Family with varying self–covariance: The second family of Gaussian distributions, denoted by $\{P_u\}_{u\in\mathcal{U}}$, has corresponding covariance matrices $\{\mathbf{C}_u\}_{v\in\mathcal{U}}$. Through this family, $\mathbf{C}_{v,\mathbf{XY}}$ and $\mathbf{C}_{v,\mathbf{YY}}$ are fixed for all $u \in \mathcal{U}$, but, $\mathbf{C}_{v,\mathbf{XX}}$ is varying. (See Lemma 6.3). In this setup, we apply the classical data processing inequality (See Theorem 5.5) to obtain a lower bound for the error probability of the hypothesis testing problem. The last two terms within $\max$ in (3) and (4), correspond to the lower bound obtained by this family of distributions.

In the next step, we derive a lower bound on the expected distortion in terms of the separation of the distribution family, i.e. $\rho_{\mathsf{dist}} := \inf_{\substack{v,v'\in\mathcal{V} \\ v\neq v'}} \left\{ \|\mathbf{C}_v - \mathbf{C}_{v'}\|_{\mathsf{dist}} \right\}$, and the error probability of the hypothesis testing problem.

Corollary 4.2. *Theorem 4.1 implies that any* DCME *with distortion less than or equal to ε requires at least $m = \Omega(\frac{\sigma^4 d}{\varepsilon^2})$ samples and $B_1 = \Omega(\max\{\frac{\sigma^4 d_1 d_2}{\varepsilon^2}, d_1^2 \log\frac{\sigma^2}{\varepsilon}\})$ and $B_2 = \Omega(\max\{\frac{\sigma^4 d_1 d_2}{\varepsilon^2}, d_2^2 \log\frac{\sigma^2}{\varepsilon}\})$ bits of communication from Agents 1 and 2, respectively.*

Remark 4.3 (Extension to more than two users). Consider a scenario with $K > 2$ agents labeled as $1, 2, \ldots, K$. We define a subset of users, denoted as $\mathcal{S} \subset [K]$. To establish a lower bound, we assume that the agents in $\mathcal{S}$ collude, and similarly, the agents in $\mathcal{S}^c = [K]\backslash\mathcal{S}$ also collude. This allows us to create two super-agents, denoted as A and B. The super-agent A has access to $\sum_{i\in\mathcal{S}} d_i$ dimensions and $\sum_{i\in\mathcal{S}} B_i$ bits of communication budget, while the super-agent B has access to $\sum_{i\in\mathcal{S}^c} d_i$ dimensions and $\sum_{i\in\mathcal{S}^c} B_i$. Then Theorem 4.1 gives a lower bound for the colluded scenario which itself is a lower bound for the non-colluded scenario. Maximizing such lower bounds over the choice of subset $\mathcal{S}$ implies:

$$\mathcal{M}_{\mathsf{op}} \geq \frac{\sigma^2}{32}\sqrt{\max_{\substack{\mathcal{S}\subset[K] \\ \mathcal{S}\neq\emptyset}} \left\{ \frac{(\sum_{i\in\mathcal{S}} d_i)(\sum_{i\in\mathcal{S}^c} d_i)}{2\min\{\sum_{i\in\mathcal{S}} B_i, \sum_{i\in\mathcal{S}^c} B_i\}} \right\}}.$$

4.2. The Achievable Scheme

In the second theorem, we propose a $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ scheme and find an upper bound on its expected distortion. The details of the achievable scheme can be found in Section 7. Here we present the main idea.

Consider the decomposition (5) of $\mathbf{C}$. Agent 1 can estimate $\mathbf{C}_{\mathbf{XX}}$ using data points $\{\mathbf{X}^{(i)}\}_{i=1}^m$, and Agent 2 can estimate $\mathbf{C}_{\mathbf{YY}}$ using data points $\{\mathbf{Y}^{(i)}\}_{i=1}^m$. Therefore they spend parts of their communication budgets on reporting quantized versions of $\mathbf{C}_{\mathbf{XX}}$ and $\mathbf{C}_{\mathbf{YY}}$ to the central server. They can allocate the rest of their communication budgets to report some quantized versions of $\{\mathbf{X}^{(i)}\}_{i=1}^m$ and $\{\mathbf{Y}^{(i)}\}_{i=1}^m$ to the central server. Then, the central server can estimate $\mathbf{C}_{\mathbf{XY}}$ with this received information and form some estimation $\widehat{\mathbf{C}}$.

Theorem 4.4. *Assume that $m \geq 9d$, $B_1 \geq 15d_1 \max\{d_1, d_2\} \log_2(\max\{d_1, d_2\})$, and $B_2 \geq 15d_2 \max\{d_1, d_2\} \log_2(\max\{d_1, d_2\})$, then there exists a* DCME *whose expected distortion satisfies the following inequality:*

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})\right]$$
$$= \sigma^2\mathcal{O}\left(\sqrt{\frac{d}{m}} + \log_2\left(\frac{B_{\min}}{d_1 d_2}\right)\sqrt{\max\left\{\frac{d_1^2}{B_1}, \frac{d_2^2}{B_2}, \frac{d_1 d_2}{B_{\min}}\right\}}\right).$$

We state the following corollary, which is a direct consequence of Theorem 4.4.

Corollary 4.5. *Consider* $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$*. Then, for any distortion ε, $\varepsilon \leq \sigma^2/2$, there exists a* DCME *scheme with the expected distortion $\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \leq \varepsilon$, if $m \geq \tau\frac{d\sigma^4}{\varepsilon^2}$ and*

$$B_k \geq \tau' \frac{\sigma^4 d_k \max\{d_1, d_2\}}{\varepsilon^2} \cdot \log_2^2\left(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\}\right),$$

for $k = 1, 2$ and some constants τ, τ'.

We call a $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ problem as η–*homogeneous* if the $\frac{1}{\eta}\frac{B_2}{d_2} \leq \frac{B_1}{d_1} \leq \eta\frac{B_2}{d_2}$, for some constant η, meaning they are approximately equal. This category is relevant in real-world scenarios where the communication load dedicated to each agent is proportional to the number of dimensions it handles. Comparing Theorems 4.1 and 4.4, we conclude the following corollary:

Corollary 4.6 (Tightness for the homogeneous case.)**.** *In an η–homogeneous $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ setting, the expected distortion of the proposed achievable scheme attains:*

$$\mathbb{E}[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})] = \sigma^2 \mathcal{O}\left(\sqrt{\frac{d}{m}} + \log_2\left(\frac{B_{\min}}{d_1 d_2}\right)\sqrt{\eta\frac{d_1 d_2}{B_{\min}}}\right).$$

This expression matches with the established lower bound, up to a logarithmic term, implying that the proposed DCME *scheme performs near-optimum.*

5. Preliminaries

5.1. Sub–Gaussian Random Variables

Sub–Gaussian random variables, as formally defined in Definition 5.1, are a family of random variables whose tails decay faster than the tail of a Gaussian distribution.

Definition 5.1 (Sub–Gaussian Random Variable (Wainwright, 2019, Definition 2.2))**.** A random variable X is said to be σ–sub–Gaussian if:

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \quad \text{for all } \lambda \in \mathbb{R}.$$

The definition of sub–Gaussian variables can be extended to random vectors as follows:

Definition 5.2 (Sub–Gaussian Random Vector (Wainwright, 2019, Section 6.3))**.** A random vector $\mathbf{X} \in \mathbb{R}^d$ is called sub–Gaussian with parameter σ if for all $\mathbf{v} \in \mathbb{S}^{d-1}$, $\mathbf{v}^\top\mathbf{X}$ is a σ–sub–Gaussian random variable, where $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ is the d-dimensional unit sphere.

Some properties of sub–Gaussian random variables are listed in Appendix A.2.

5.2. Packing and Covering Numbers

Definition 5.3 (Covering Number (Wainwright, 2019, Definition 5.1))**.** A set $\{x_1, x_2, \ldots, x_N\} \subseteq \mathcal{K}$ is called a ϵ–covering set with respect to a metric d if for all $x \in \mathcal{K}$, there exists some $j \in [N]$ such that $\mathsf{d}(x, x_j) \leq \epsilon$. The covering number $\mathcal{N}(\mathcal{K}, \mathsf{d}, \epsilon)$ is the cardinality of the smallest ϵ–covering set of set $\mathcal{K}$, for the metric d.

Definition 5.4 (Packing Number (Wainwright, 2019, Definition 5.4))**.** A set $\{x_1, x_2, \ldots, x_M\} \subseteq \mathcal{K}$ is called a ϵ–packing set with respect to a metric d if $\mathsf{d}(x_i, x_j) > \epsilon$ for all distinct $i, j \in [M]$. The packing number $\mathcal{M}(\mathcal{K}, \mathsf{d}, \epsilon)$ is the cardinality of the largest ϵ–packing set of set $\mathcal{K}$, with respect to the metric d.

5.3. Strong Data Processing Inequality

Theorem 5.5 (Data Processing Inequality (Cover, 1999, Theorem 2.8.1))**.** *If $U \multimap X \multimap Y$ forms a Markov chain, then:*

$$I(U;Y) \leq I(U;X).$$

Strong Data Processing Inequality is a refined version of the data processing inequality.

Definition 5.6 (Strong Data Processing (SDPI) Coefficient or Rate of Information Bottleneck (Anantharam et al., 2013))**.** Let X and Y be random variables with joint distribution $(X, Y) \sim p_{X,Y}(x, y)$. We define:

$$s(X;Y) = \sup_{\substack{U: U \multimap X \multimap Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)}.$$

The SDPI constant has tensorization property, which is stated in (Polyanskiy & Wu, 2023, Proposition 33.11):

$$s(X^{\otimes n}; Y^{\otimes n}) = s(X;Y). \tag{6}$$

In (Kim et al., 2017), the SDPI constant is derived for multivariate normal distribution.

Lemma 5.7 ((Kim et al., 2017, Section 2.6))**.** *If $(\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{C} = \begin{bmatrix} \mathbf{C}_{\mathbf{XX}} & \mathbf{C}_{\mathbf{XY}} \\ \mathbf{C}_{\mathbf{YX}} & \mathbf{C}_{\mathbf{YY}} \end{bmatrix}\right)$, then we have:*

$$s(\mathbf{X};\mathbf{Y}) = \left\|\mathbf{C}_{\mathbf{XX}}^{-1/2}\mathbf{C}_{\mathbf{XY}}\mathbf{C}_{\mathbf{YY}}^{-1/2}\right\|_{\mathsf{op}}^2.$$

6. Proof of the Lower Bound

6.1. Proof of Theorem 4.1

We use Fano's method to lower bound the $\mathcal{M}_{\mathsf{dist}}(\sigma, B_1, B_2, d_1, d_2, m)$. This method, first introduced in (Khas' minskii, 1979), has undergone extensive development in various papers (Ibragimov & Has' Minskii, 1981; Birgé, 1983; Yu et al., 1997; Yang & Barron, 1999; Birgé, 2005; Raskutti et al., 2011; Guntuboyina, 2011; Candes & Davenport, 2013; Duchi & Wainwright, 2013; Polyanskiy & Wu, 2023). We adopt the version described in (Duchi, 2021, Section 7.4) and adapt it to the distributed covariance matrix estimation (DCME) problem.

We consider a family of distributions $\{P_v\}_{v\in\mathcal{V}} \subset \mathsf{subG}^{(d)}(\sigma)$ indexed by a finite set $\mathcal{V}$. For each $v \in \mathcal{V}$, let $\mathbf{C}_v := \mathbb{E}_{\mathbf{X}\sim P_v}[\mathbf{X}\mathbf{X}^\top]$ denote the corresponding covariance matrix. For this set, we define the separation ρ with respect to the dist norm metric on the space of covariance matrices as:

$$\rho_{\mathsf{dist}} := \inf_{\substack{v,v'\in\mathcal{V}\\ v\neq v'}} \left\{ \left\| \mathbf{C}_v - \mathbf{C}_{v'} \right\|_{\mathsf{dist}} \right\}.$$

Lemma 6.1, a direct consequence of (Duchi, 2021, Proposition 7.10), establishes a fundamental bound in the context of distributed covariance matrix estimation:

Lemma 6.1. *Consider a set $\mathcal{V}$ with separation ρ, and a corresponding set of distributions $\{P_v\}_{v\in\mathcal{V}}$. Assume a random variable $V\in\mathcal{V}$ is chosen uniformly, and given $V=v$, samples $\{\mathbf{Z}^{(i)}\}_{i=1}^m$ are drawn i.i.d. from P_v. In addition, assume that Agents 1 and 2 access $\{\mathbf{X}^{(i)} = \mathbf{Z}^{(i)}_{[1:d_1]}\}_{i=1}^m$ and $\{\mathbf{Y}^{(i)} = \mathbf{Z}^{(i)}_{[d_1+1:d]}\}_{i=1}^m$, respectively. For any* DCME *scheme with parameters $(\sigma, m, d_{1:2}, B_{1:2})$, we have:*

$$\inf_{\mathcal{E}_1,\mathcal{E}_2,\mathcal{D}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}},\mathbf{C})\right] \geq \frac{\rho_{\mathsf{dist}}}{2}\left[1 - \frac{I(V;M_1,M_2)+1}{\log_2(|\mathcal{V}|)}\right].$$

We proceed by first establishing two lemmas that utilize Lemma 6.1 with specific distribution families $\{P_v\}_{v\in\mathcal{V}}$. These lemmas will then pave the way for deriving the main theorem regarding the min–max lower bound.

Lemma 6.2. *Consider a set $\mathcal{V}$ and a corresponding set of distributions $\{P_v\}_{v\in\mathcal{V}}$, where $P_v = \mathcal{N}(\mathbf{0},\mathbf{C}_v)$ and:*

$$\mathbf{C}_v = \begin{bmatrix} \frac{\sigma^2}{2}\mathbf{I}_{d_1} & \delta\mathbf{D}_v \\ \delta\mathbf{D}_v^\top & \frac{\sigma^2}{2}\mathbf{I}_{d_2} \end{bmatrix},$$

and $\mathbf{D}_v$ is some matrix in $\mathbb{R}^{d_1\times d_2}$. Define:

$$\beta_{\mathsf{dist}}(\{\mathbf{D}_v\}_{v\in\mathcal{V}}) = \frac{\sqrt{1+\mathbb{1}_{\{\mathsf{dist=F}\}}}\inf_{v,v':v\neq v'}\|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}}}{\max_{\mathbf{v}\in\mathcal{V}}\left\{\|\mathbf{D}_v\|_{\mathsf{op}}\right\}}.$$

Then, the following lower bound on $\mathcal{M}_{\mathsf{dist}}$ hold:

$$\mathcal{M}_{\mathsf{dist}} \geq \frac{\sigma^2}{16}\beta_{\mathsf{dist}}(\{\mathbf{D}_v\}_{v\in\mathcal{V}}) \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{B_{\min}}}, 2\right\}.$$

Furthermore, considering a set of distributions $\{P_v\}_{v\in\mathcal{V}}$, where $P_v = \mathcal{N}(\mathbf{0},\mathbf{C}'_v)$ and:

$$\mathbf{C}'_v = \begin{bmatrix} \frac{\sigma^2}{2}\mathbf{I}_{d/2} & \delta\mathbf{D}'_v \\ \delta\mathbf{D}'^\top_v & \frac{\sigma^2}{2}\mathbf{I}_{d/2} \end{bmatrix},$$

and $\mathbf{D}'_v$ is some matrix in $\mathbb{R}^{d/2\times d/2}$, we have the following lower bound on $\mathcal{M}_{\mathsf{dist}}$:

$$\mathcal{M}_{\mathsf{dist}} \geq \frac{\sigma^2}{8\sqrt{2}}\beta_{\mathsf{dist}}(\{\mathbf{D}'_v\}_{v\in\mathcal{V}}) \min\left\{\sqrt{\frac{2\log_2(|\mathcal{V}|)}{3md}}, 1\right\}.$$

Proof. The complete proof is stated in Appendix B.1. Here, we highlight some fundamental steps of the proof, briefly.

- $\mathbf{C}_v$ is the covariance matrix of a σ^2–sub–Gaussian random vector $\mathbf{Z}$, therefore $\mathbf{0} \preceq \mathbf{C}_v \preceq \sigma^2\mathbf{I}$. This forces δ to satisfy the condition $\delta \leq \frac{\sigma^2}{2\max_{v\in\mathcal{V}}\{\|\mathbf{D}_v\|_{\mathsf{op}}\}}$.

- The separation of set $\mathcal{V}$ is:

$$\rho_{\mathsf{dist}} = \sqrt{1+\mathbb{1}_{\{\mathsf{dist=F}\}}}\,\delta \inf_{v,v':v\neq v'}\|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}}.$$

- The vector $\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{C}_v)$ has the same marginal distribution over the first d_1 dimensions and the second d_2 dimensions, for all $v\in\mathcal{V}$. Therefore $\mathsf{X} = \left\{\mathbf{X}^{(i)}\right\}_{i=1}^m$ is independent from V. Similarly $\mathsf{Y} = \left\{\mathbf{Y}^{(i)}\right\}_{i=1}^m$ is independent from V. Subsequently, M_1 and M_2 are also independent from V. This implies $I(V;M_1,M_2) \leq I(M_1;M_2|V)$.

- **Using SDPI:** We now utilize SDPI to bound $I(M_1;M_2|V)$. For any $V=v$, $M_1 -\!\!\circ\!\!- \mathsf{X} -\!\!\circ\!\!- \mathsf{Y} -\!\!\circ\!\!- M_2$ forms a Markov Chain. Then

$$\begin{aligned} I(M_1;M_2|V=v) &\leq I\left(M_1;\mathsf{Y}|\,V=v\right) \\ &\overset{\text{(a)}}{\leq} s_v\left(\mathsf{X};\mathsf{Y}\right) I\left(M_1;\mathsf{X}|\,V=v\right) \\ &\overset{\text{(b)}}{\leq} s_v\left(\mathbf{X};\mathbf{Y}\right) B_1, \end{aligned}$$

where in (a), $s_v(\mathsf{X};\mathsf{Y})$ is the SDPI coefficient (Definition 5.6) for the joint distribution $p_{\mathsf{X},\mathsf{Y}}$, given $V=v$. In addition, (b) holds due to (6). From Lemma 5.7 we have:

$$s_v\left(\mathbf{X};\mathbf{Y}\right) = \left\|\mathbf{C}_{v,\mathbf{XX}}^{-1/2}\mathbf{C}_{v,\mathbf{XY}}\mathbf{C}_{v,\mathbf{YY}}^{-1/2}\right\|_{\mathsf{op}}^2 = \left(\frac{2\delta}{\sigma^2}\right)^2\|\mathbf{D}_v\|_{\mathsf{op}}^2.$$

We obtain a similar upper bound on $I(M_1;M_2|V=v)$ w.r.t. B_2. In summary, we have:

$$I(M_1;M_2|V=v) \leq \left(\frac{2\delta}{\sigma^2}\right)^2 B_{\min}\max_{\mathbf{v}\in\mathcal{V}}\left\{\|\mathbf{D}_v\|_{\mathsf{op}}^2\right\}.$$

- Setting $\delta = \frac{\sigma^2}{4\max_{\mathbf{v}\in\mathcal{V}}\{\|\mathbf{D}_v\|_{\mathsf{op}}\}}\min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{B_{\min}}},2\right\}$, gives the first term the in lower bound.

- Assuming $\delta \leq \frac{\sigma^2}{2\sqrt{2}\max_{v\in\mathcal{V}}\{\|\mathbf{D}'_v\|_{\mathsf{op}}\}}$, and defining $\mathsf{X}' = \left\{\mathbf{Z}^{(i)}_{[1:d/2]}\right\}_{i=1}^m$ and $\mathsf{Y}' = \left\{\mathbf{Z}^{(i)}_{[d/2+1:d]}\right\}_{i=1}^m$, we can derive another upper bound on $I(V;M_1,M_2)$ due to data processing inequality and the Markov Chain $V -\!\!\circ\!\!- (\mathsf{X},\mathsf{Y}) = (\mathsf{X}',\mathsf{Y}') -\!\!\circ\!\!- (M_1,M_2)$:

$$\begin{aligned} I(V;M_1,M_2) &\leq I\left(V;\mathsf{X}',\mathsf{Y}'\right) \\ &\leq \frac{2md\delta^2}{\ln(2)\sigma^4}\max_{\mathbf{v}\in\mathcal{V}}\left\{\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2\right\}. \end{aligned}$$

If we set $\delta = \frac{\sigma^2}{2\sqrt{2}\max_{\mathbf{v}\in\mathcal{V}}\{\|\mathbf{D}'_v\|_{\text{op}}\}} \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{12md}}, 1\right\}$, the second lower bound is derived. □

Lemma 6.3. *For the set $\mathcal{U}$, we consider the set of distributions $\{P_u\}_{u\in\mathcal{U}}$, where $P_u = \mathcal{N}(\mathbf{0}, \mathbf{C}_u)$ and:*

$$\mathbf{C}_u = \begin{bmatrix} \frac{\sigma^2}{2}\mathbf{I}_{d_1/2} & \delta\mathbf{D}_u & \mathbf{0} & \mathbf{0} \\ \delta\mathbf{D}_u^\top & \frac{\sigma^2}{2}\mathbf{I}_{d_1/2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{\sigma^2}{2}\mathbf{I}_{d_2/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{\sigma^2}{2}\mathbf{I}_{d_2/2} \end{bmatrix},$$

where $\mathbf{D}_u$ is some matrix in $\mathbb{R}^{d_1/2\times d_1/2}$. If we define:

$$\beta_{\text{dist}}(\{\mathbf{D}_u\}_{u\in\mathcal{U}}) = \frac{\sqrt{1+\mathbb{1}_{\{\text{dist}=\text{F}\}}}\inf_{u,u':u\neq u'}\|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\text{dist}}}{\max_{\mathbf{u}\in\mathcal{U}}\left\{\|\mathbf{D}_u\|_{\text{op}}\right\}},$$

then we have this lower bound on $\mathcal{M}_{\text{dist}}$:

$$\mathcal{M}_{\text{dist}} \geq \frac{\sigma^2}{2}\beta_{\text{dist}}\big(\{\mathbf{D}_v\}_{v\in\mathcal{V}}\big)\left[1 - \frac{B_1+1}{\log_2(|\mathcal{U}|)}\right].$$

Proof. The proof is presented in Appendix B.2. □

Now we are ready to state the proof of Theorem 4.1.

Proof. We use Lemmas 6.2 and 6.3 with three appropriate sets $\{\mathbf{D}_v\}_{v\in\mathcal{V}}$, $\{\mathbf{D}'_v\}_{v\in\mathcal{V}}$, and $\{\mathbf{D}_u\}_{u\in\mathcal{U}}$ to prove the theorem. In Appendix A.5 we introduce the $\vert\!\vert\!\vert.\vert\!\vert\!\vert_{\text{dist}}$ of a vectorized matrix and the packing and covering sets of the unit $\vert\!\vert\!\vert.\vert\!\vert\!\vert_{\text{op}}$ ball of matrices under the dist norm. We set the $\{\mathbf{D}_v\}_{v\in\mathcal{V}}$ as the ϵ–packing points of $\mathcal{B}^{(d_1d_2)}_{\vert\!\vert\!\vert\cdot\vert\!\vert\!\vert_{\text{op}}}(1)$ (see Equation (21)), under $\vert\!\vert\!\vert.\vert\!\vert\!\vert_{\text{dist}}$ norm. Thus $\inf_{v,v':v\neq v'}\|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\text{dist}} \geq \epsilon$, $\max_{v\in\mathcal{V}}\{\|\mathbf{D}_v\|^2_{\text{op}}\} \leq 1$, and from (22) and (28), $\log_2(|\mathcal{V}|) \geq d_1d_2\log_2\left(\frac{\nu_{\text{dist}}}{\epsilon}\right)$, where:

$$\nu^{(d_1,d_2)}_{\text{dist}} = \begin{cases} 1 & \text{if dist} = \text{op} \\ \frac{\sqrt{d_{\min}}}{14} & \text{if dist} = \text{F} \end{cases}.$$

We set $\epsilon = \nu^{(d_1,d_2)}_{\text{dist}}/2$. Using Lemma 6.2, we have the following min–max lower bound:

$$\mathcal{M}_{\text{dist}} \geq \frac{\sigma^2}{16}\kappa^{(d_1,d_2)}_{\text{dist}}\min\left\{\frac{1}{2}\sqrt{\frac{d_1d_2}{B_{\min}}}, 1\right\},$$

where $\kappa^{(d_1,d_2)}_{\text{dist}} = \sqrt{1+\mathbb{1}_{\{\text{dist}=\text{F}\}}}\nu^{(d_1,d_2)}_{\text{dist}}$. We set the $\{\mathbf{D}'_v\}_{v\in\mathcal{V}}$ as the ϵ'–packing points of $\mathcal{B}^{(d^2/4)}_{\vert\!\vert\!\vert\cdot\vert\!\vert\!\vert_{\text{op}}}(1)$ (see Equation (21)), under $\vert\!\vert\!\vert.\vert\!\vert\!\vert_{\text{dist}}$ norm. Thus $\inf_{v,v':v\neq v'}\|\mathbf{D}'_v - \mathbf{D}'_{v'}\|_{\text{dist}} \geq \epsilon'$, $\max_{v\in\mathcal{V}}\{\|\mathbf{D}'_v\|^2_{\text{op}}\} \leq 1$, and from (22) and (28), $\log_2(|\mathcal{V}|) \geq \frac{d^2}{4}\log_2\left(\frac{\nu'_{\text{dist}}}{\epsilon'}\right)$, where:

$$\nu'^{(d)}_{\text{dist}} = \begin{cases} 1 & \text{if dist} = \text{op} \\ \frac{\sqrt{d/2}}{14} & \text{if dist} = \text{F} \end{cases}.$$

We set $\epsilon' = \nu'^{(d)}_{\text{dist}}/2$. Using Lemma 6.2, we have the following min–max lower bound:

$$\begin{aligned}\mathcal{M}_{\text{dist}} &\geq \frac{\sigma^2}{16}\kappa'^{(d)}_{\text{dist}}\min\left\{\sqrt{\frac{d^2/4}{3md}}, \frac{1}{\sqrt{2}}\right\} \\ &\geq \frac{\sigma^2}{32}\kappa'^{(d)}_{\text{dist}}\min\left\{\sqrt{\frac{d}{3m}}, \sqrt{2}\right\},\end{aligned}$$

where $\kappa'^{(d)}_{\text{dist}} = \sqrt{1+\mathbb{1}_{\{\text{dist}=\text{F}\}}}\nu'^{(d)}_{\text{dist}}$. If we define the set $\{\mathbf{D}_u\}_{u\in\mathcal{U}}$ as the ϵ–packing points of $\mathcal{B}^{(d_1^2/4)}_{\vert\!\vert\!\vert\cdot\vert\!\vert\!\vert_{\text{op}}}(1)$, under $\vert\!\vert\!\vert.\vert\!\vert\!\vert_{\text{dist}}$ norm, we have $\inf_{u,u':u\neq u'}\|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\text{dist}} \geq \epsilon$, $\max_{u\in\mathcal{U}}\left\{\|\mathbf{D}_u\|_{\text{op}}\right\} \leq 1$, and $\log_2(|\mathcal{U}|) \geq \frac{d_1^2}{4}\log_2\left(\frac{\nu^{(d_1/2)}_{\text{dist}}}{\epsilon}\right)$, where:

$$\nu^{(d_1/2)}_{\text{dist}} = \begin{cases} 1 & \text{if dist} = \text{op} \\ \frac{\sqrt{d_1}}{14\sqrt{2}} & \text{if dist} = \text{F} \end{cases}.$$

Now if we set $\epsilon = \nu^{(d_1/2)}_{\text{dist}}\cdot 2^{\frac{-16B_1}{d_1^2}}$ and use Lemma 6.3, we have this min–max lower bound:

$$\mathcal{M}_{\text{dist}} \geq \frac{\sigma^2}{4}\cdot\kappa^{(d_1/2)}_{\text{dist}}\cdot 2^{\frac{-16B_1}{d_1^2}},$$

where $\kappa^{(d_1/2)}_{\text{dist}} = \sqrt{1+\mathbb{1}_{\{\text{dist}=\text{F}\}}}\nu^{(d_1/2)}_{\text{dist}}$. Similarly, we write:

$$\mathcal{M}_{\text{dist}} \geq \frac{\sigma^2}{4}\cdot\kappa^{(d_2/2)}_{\text{dist}}\cdot 2^{\frac{-16B_2}{d_2^2}}.$$

The final result is obtained. □

6.2. Comparison of the Proof Methods with a Naïve Extension of (Hadar et al., 2019)

In (Hadar et al., 2019), the authors consider the special case where $m = \infty$ and $B_2 = \infty$, thus the central server can access the stream $\mathsf{Y} = \{Y_i\}_{i=1}^{\infty}$. For simplicity, we assume $\sigma = 1$, throughout this section. The goal of that work is to approximate the covariance $c = \mathbb{E}[XY]$ between jointly normal random variables X and Y. They considered the expected squared error function as the distortion. The derivation of the lower bound on the error is based on the combination of the Bayesian Cramer-Rao (BCR) lower bound and the strong data processing inequality. We review a brief description of their converse method (as given in (Polyanskiy & Wu, 2023, Chapter 33.6)) before comparing the methods. Let $P^{(c)} = \mathcal{N}\left(0, \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}\right)$. Also let $\mathcal{L}(\widehat{c}, c) = (\widehat{c} - c)^2$. It is known that the min–max error satisfies

$$\min_{\widehat{c}}\max_{c}\mathbb{E}[\mathcal{L}(\widehat{c}, c)] \geq \frac{1+o(1)}{J_F(0)}$$

where $J_F(c)$ is the Fisher information of the family $\{P^{(c)} : c \in [-1,1]\}$. Then they proceed by observing that the Fisher information can be bounded by using the strong data processing inequality and Taylor approximation of the Kullback–Leibler (KL) divergence.

In particular, one can infer the following inequality from the calculation in (Polyanskiy & Wu, 2023):

$$c^2 J_F(0) + o(c^2) \le \kappa \cdot s(\mathsf{X};\mathsf{Y}|c)B_1, \tag{7}$$

where κ is a constant. This inequality and the identity $s(\mathsf{X};\mathsf{Y}|c) = c^2$ implies $\min_{\widehat{c}} \max_c \mathbb{E}[\mathcal{L}(\widehat{c}, c)] = \Omega(B_1^{-1})$.

One can readily extend this method to the case where d_1 is arbitrary and $d_2 = 1$, using the multivariate BCR. To this end, let $\mathbf{c} = \mathbb{E}[Y\mathbf{X}]$ and $\mathcal{L}(\widehat{\mathbf{c}}, \mathbf{c}) = \|\widehat{\mathbf{c}} - \mathbf{c}\|_2^2$. In this case, the multi-variate counterpart of (7) is:

$$\mathbf{c}^\top \mathbf{J}_F(\mathbf{0})\mathbf{c} + o(\|\mathbf{c}\|^2) \le \kappa \cdot s(\mathbf{X};\mathsf{Y}|\mathbf{c})B_1, \tag{8}$$

where $\mathbf{J}_F$ is the Fisher information matrix. Lemma 5.7 yields $s(\mathbf{X};\mathsf{Y}|\mathbf{c}) = \|\mathbf{c}\|^2$. This implies $\mathbf{J}_F(0) \preceq \kappa B_1 \mathbf{I}_{d_1}$. The multivariate BCR (Polyanskiy & Wu, 2023, Theorem 29.4) states that:

$$\begin{aligned}\min_{\widehat{\mathbf{c}}} \max_{\mathbf{c}} \mathbb{E}[\mathcal{L}(\widehat{\mathbf{c}}, \mathbf{c})] &\ge (1+o(1))\mathrm{tr}\left[\mathbf{J}_F(\mathbf{0})^{-1}\right] \\ &\ge (1+o(1))\frac{d_1}{\kappa B_1}\end{aligned} \tag{9}$$

Ignoring the little difference between the square error here and the distortion based on the operator norm (which is the root of the square error) in our setting, the order of error using both approaches is matched.

The main difficulty arises in the case $d_1 > 1, d_2 > 1$. In this regime, we aim to approximate the cross-covariance $\mathbf{C}_{\mathbf{XY}}$, which is a matrix. Here, to apply multivariate BCR, we need to vectorize that matrix. Let $\mathbf{c}_{\mathbf{XY}} = \mathrm{vec}(\mathbf{C}_{\mathbf{XY}})$ and $\widehat{\mathbf{c}}_{\mathbf{XY}} = \mathrm{vec}(\widehat{\mathbf{C}}_{\mathbf{XY}})$. Also, BCR gives a lower bound on the squared error loss. In the matrix space, this loss is the squared Frobenius norm $\mathcal{L}_{\mathsf{F}}(\widehat{\mathbf{C}}_{\mathbf{XY}}, \mathbf{C}_{\mathbf{XY}}) := \left\|\widehat{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{F}}^2 = \left\|\widehat{\mathbf{c}}_{\mathbf{XY}} - \mathbf{c}_{\mathbf{XY}}\right\|^2$. We also consider the family of normal distributions $P^{\mathbf{C}} = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{I}_{d_1} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{I}_{d_2}\end{bmatrix}\right)$ characterized with cross-covariance matrix $\mathbf{C}$. In this case, the counterpart of (8) is:

$$\begin{aligned}\mathbf{c}_{\mathbf{XY}}^\top \mathbf{J}_F(\mathbf{0})\mathbf{c}_{\mathbf{XY}} + o(\|\mathbf{c}_{\mathbf{XY}}\|^2) &\le \kappa \cdot s(\mathbf{X};\mathbf{Y}|\mathbf{c}_{\mathbf{XY}})B_1 \\ &\overset{(a)}{\le} \kappa \cdot \left\|\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}^2 B_1 \le \kappa \cdot \left\|\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{F}}^2 B_1,\end{aligned} \tag{10}$$

where (a) is true due to Lemma 5.7 and the Fisher information matrix is defined with respect to the vectorization of the cross–covariance matrix. The inequality (10) results in $\mathbf{J}_F(\mathbf{0}) \preceq \kappa B_1 \mathbf{I}_{d_1 d_2}$. Then similar calculation to (9) implies:

$$\min_{\widehat{\mathbf{C}}_{\mathbf{XY}}} \max_{\mathbf{C}_{\mathbf{XY}}} \mathbb{E}[\mathcal{L}_{\mathsf{F}}(\widehat{\mathbf{C}}_{\mathbf{XY}}, \mathbf{C}_{\mathbf{XY}})] = \Omega\left(\frac{d_1 d_2}{\kappa B_1}\right), \tag{11}$$

which is *weaker* than our lower bound on Frobenius distortion (see the first term in (4), which corresponds to approximation error for *cross covariance matrix*) by a factor of $d_{\min}$.

7. Statement of Achievable Scheme and Proof of Theorem 4.4

In this section, we propose a near-optimal achievable DCME scheme and find an upper bound on its expected distortion.

Means do not matter. If the random vectors $\mathbf{Z}^{(i)}$ do not have a zero mean, we can redefine them as $\mathbf{Z}'^{(i)} = \frac{1}{\sqrt{2}}(\mathbf{Z}^{(2i-1)} - \mathbf{Z}^{(2i)})$, which will have zero mean and retain the same covariance matrix as $\mathbf{Z}^{(i)}$. Therefore, we can use the samples $\{\mathbf{Z}'^{(i)}\}_{i=1}^{m/2}$ in place of $\{\mathbf{Z}^{(i)}\}_{i=1}^{m}$. Hence, we can assume without loss of generality that $\mathbb{E}[\mathbf{Z}] = 0$.

The scheme is divided into two parts, in which we separately approximate the self–covariance matrices $\mathbf{C}_{\mathbf{XX}}, \mathbf{C}_{\mathbf{YY}}$ and the cross–covariance matrix $\mathbf{C}_{\mathbf{XY}}$ (see (5)).

Empirical estimation for Self–covariance matrices. Each agent can estimate its self-covariance matrix from its data using an empirical covariance estimator. More precisely, Agent 1 estimate $\mathbf{C}_{\mathbf{XX}}$ using $\widetilde{\mathbf{C}}_{\mathbf{XX}} = \frac{1}{m}\sum_{i=1}^m \mathbf{X}^{(i)}\mathbf{X}^{(i)\top}$ and similarly, Agent 2 can estimate $\mathbf{C}_{\mathbf{YY}}$ using $\widetilde{\mathbf{C}}_{\mathbf{YY}} = \frac{1}{m}\sum_{i=1}^m \mathbf{Y}^{(i)}\mathbf{Y}^{(i)\top}$.

Quantization of estimated self-covariance matrices. The empirical self–covariance matrix $\mathbf{C}_{\mathbf{XX}}$ lies in the ball $\mathcal{B}^{d_1^2}_{|||\cdot|||_{\mathsf{op}}}(\tau\sigma^2)$ with high probability for some constant $\tau > 0$. To quantize it, Agent 1 finds an ϵ-covering of this ball with $2^{B_1/2}$ points with smallest possible ϵ. Then if its empirical estimation $\mathbf{C}_{\mathbf{XX}}$ lies in the ball $\mathcal{B}^{d_1^2}_{|||\cdot|||_{\mathsf{op}}}(\tau\sigma^2)$, the Agent 1 quantizes $\mathbf{C}_{\mathbf{XX}}$ to $B_1/2$ bits, by finding the nearest point in the covering to the empirical estimation. If the empirical estimation $\widetilde{\mathbf{C}}_{\mathbf{XX}}$ lies outside the ball, Agent 1 declares an error. Agent 2 similarly quantized its empirical estimation.

Quantization of Data for approximating the cross–covariance. Choosing number $n = \min\left\{\min\{\frac{B_1}{d_1}, \frac{B_2}{d_2}\}/\log_2\left(\frac{\min\{B_1,B_2\}}{d_1 d_2}\right), m\right\}$, we define matrices $\mathsf{X} \in \mathbb{R}^{d_1 \times n}$ and $\mathsf{Y} \in \mathbb{R}^{d_2 \times n}$ as $\mathsf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}]$ and $\mathsf{Y} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}]$. The *empirical* estimator for $\mathbf{C}_{\mathbf{XY}}$ using the first n samples is $\widetilde{\mathbf{C}}_{\mathbf{XY}} = \frac{1}{n}\mathsf{X}\mathsf{Y}^\top$. This guides agent 1 to quantize the *whole block* of its data X to $B_1/2$ bits, and Agent 2 does a similar quantization on its data. To do this, it is well-known that $\mathsf{\mathbf{X}}$ lies in the ball $\mathcal{B}^{nd_1}_{|||\cdot|||_{\mathsf{op}}}(\tau\sigma\sqrt{d_1+n})$ with high probability. To quantize it, Agent 1 finds an ϵ-covering of this ball with $2^{B_1/2}$ points with smallest possible ϵ. Then if $\mathsf{\mathbf{X}}$ lies in the ball $\mathcal{B}^{nd_1}_{|||\cdot|||_{\mathsf{op}}}(\tau\sigma\sqrt{d_1+n})$, the agent 1 quantizes $\mathsf{\mathbf{X}}$ to $B_1/2$ bits, by finding the nearest point $\widehat{\mathsf{\mathbf{X}}}$ in the covering to the empirical estimation. If the empirical estimation $\mathsf{\mathbf{X}}$ lies outside the ball, agent 1 declares an error. Similarly, agent 2 finds a quantization $\widehat{\mathsf{\mathbf{Y}}}$ of $\mathsf{\mathbf{Y}}$ using $B_2/2$ bits.

Estimation of the cross–covariance at the central server. upon receiving $\widehat{\mathsf{\mathbf{X}}}$ and $\widehat{\mathsf{\mathbf{Y}}}$, the central server estimates $\mathbf{C}_{\mathbf{XY}}$ as $\widehat{\mathbf{C}}_{\mathbf{XY}} = \frac{1}{n}\widehat{\mathsf{\mathbf{X}}}\widehat{\mathsf{\mathbf{Y}}}^\top$. The central server returns $\widehat{\mathbf{C}} = \mathbf{0}$ if it receives any error. Otherwise, it computes:

$$\widehat{\mathbf{C}}^* = \begin{bmatrix} \widehat{\mathbf{C}}_{\mathbf{XX}} & \frac{1}{n}\hat{\mathsf{\mathbf{X}}}\hat{\mathsf{\mathbf{Y}}}^\top \\ \frac{1}{n}\hat{\mathsf{\mathbf{Y}}}\hat{\mathsf{\mathbf{X}}}^\top & \widehat{\mathbf{C}}_{\mathbf{YY}} \end{bmatrix}.$$

If $\widehat{\mathbf{C}}^*$ is not positive semi-definite, we modify it accordingly. By decomposing $\widehat{\mathbf{C}}^*$ into its spectral form, $\widehat{\mathbf{C}}^* = \sum_{i=1}^r \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, we define $\widehat{\mathbf{C}}^*+$ as: $\widehat{\mathbf{C}}^*_+ = \sum_{i=1}^r \lambda_i \mathbb{1}_{\{\lambda_i \geq 0\}} \mathbf{v}_i \mathbf{v}_i^\top$.

The analysis of this DCME scheme is based on the concentration inequalities for the operator norm of random matrices (e.g. (Vershynin, 2018)) and is deferred to Appendix D.

8. Conclusion

This paper studied the problem of estimating the covariance matrix in a vertical split setting, with a constrained communication budget. We established a min—max lower bound for the expected distortion of a DCME problem in Theorem 4.1, which we defined in Section 3. We also proposed a scheme to solve the DCME problem and derived an upper bound for its expected distortion in Theorem 4.4. We noted that in some realistic scenarios, the proposed scheme achieves the min—max error, up to a logarithmic factor.

Impact Statement

This work delves into the machine learning domain, specifically focusing on improving covariance matrix estimation. Covariance matrices play a pivotal role in diverse applications like principal component analysis, offering valuable insights into data relationships. Our approach advocates for **communication efficiency**. Estimating the covariance matrix within finite communication constraints not only conserves bandwidth but also contributes to lower energy consumption by reducing the computational burden on electronic chips. These environmental and economic benefits constitute additional valuable impacts of our work.

References

Allaart, C. G., Keyser, B., Bal, H., and Van Halteren, A. Vertical split learning-an exploration of predictive performance in medical and other use cases. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Anantharam, V., Gohari, A., Kamath, S., and Nair, C. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *arXiv preprint arXiv:1304.6133*, 2013.

Bai, Z.-J., Chan, R. H., and Luk, F. T. Principal component analysis for distributed data sets with updating. In *International Workshop on Advanced Parallel Processing Technologies*, pp. 471–483. Springer, 2005.

Balcan, M.-F., Kanchanapally, V., Liang, Y., and Woodruff, D. Improved distributed principal component analysis. *arXiv preprint arXiv:1408.5823*, 2014.

Bickel, P. J. and Levina, E. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008. ISSN 00905364. URL http://www.jstor.org/stable/25464728.

Bickel, P. J., Levina, E., et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1): 199–227, 2008.

Birgé, L. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:181–237, 1983.

Birgé, L. A new lower bound for multiple hypothesis testing. *IEEE transactions on information theory*, 51(4):1611–1615, 2005.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.com/books?id=koNqWRluhP0C.

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1011–1020, 2016.

Cai, T. T., Zhang, C.-H., Zhou, H. H., et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Cai, T. T., Ren, Z., and Zhou, H. H. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143, 2013.

Candes, E. J. and Davenport, M. A. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

Chen, R. Y., Gittens, A., and Tropp, J. A. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Dahmen, J., Keysers, D., Pitz, M., and Ney, H. Structured covariance matrices for statistical image object recognition. In *Mustererkennung 2000*, pp. 99–106. Springer, 2000.

Duchi, J. Lecture notes for statistics 311/electrical engineering 377. https://web.stanford.edu/class/stats311/lecture-notes.pdf, 2021. Accessed: 2024-01-15.

Duchi, J. C. and Wainwright, M. J. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

El Karoui, N. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6), December 2008.

Forero, P. A., Cano, A., and Giannakis, G. B. Consensus-based distributed linear support vector machines. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 35–46, 2010.

Furrer, R. and Bengtsson, T. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

Guntuboyina, A. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

Hadar, U. and Shayevitz, O. Distributed estimation of gaussian correlations. *IEEE Transactions on Information Theory*, 65(9):5323–5338, 2019.

Hadar, U., Liu, J., Polyanskiy, Y., and Shayevitz, O. Communication complexity of estimating correlations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 792–803, 2019.

Han, Y., Özgür, A., and Weissman, T. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pp. 3163–3188. PMLR, 2018.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

Ibragimov, I. A. and Has' Minskii, R. Z. *Statistical Estimation: Asymptotic Theory*. Springer Science & Business Media, New York, NY, 1981. ISBN 978-1-4899-0029-6 978-1-4899-0027-2. doi: 10.1007/978-1-4899-0027-2. URL http://link.springer.com/10.1007/978-1-4899-0027-2.

Johnson, C. R. and Horn, R. A. *Matrix analysis*. Cambridge university press Cambridge, 1985.

Kannan, R., Vempala, S., and Woodruff, D. Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory*, pp. 1040–1057. PMLR, 2014.

Khas' minskii, R. Z. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

Kim, H., Gao, W., Kannan, S., Oh, S., and Viswanath, P. Discovering potential correlations via hypercontractivity. *Advances in Neural Information Processing Systems*, 30, 2017.

Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(3), 2009.

Ledoit, O. and Wolf, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5): 603–621, 2003.

Lu, Y., Roychowdhury, V., and Vandenberghe, L. Distributed parallel support vector machines in strongly connected networks. *IEEE Transactions on Neural Networks*, 19(7):1167–1178, 2008.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Navia-Vázquez, A., Gutierrez-Gonzalez, D., Parrado-Hernández, E., and Navarro-Abellan, J. Distributed support vector machines. *IEEE Transactions on Neural Networks*, 17(4):1091, 2006.

Niu, F., Recht, B., Ré, C., and Wright, S. J. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *arXiv preprint arXiv:1106.5730*, 2011.

Polyanskiy, Y. and Wu, Y. Information theory: From coding to learning. *Book draft*, 2023.

Qu, Y., Ostrouchov, G., Samatova, N., and Geist, A. Principal component analysis for dimension reduction in massive distributed data sets. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, volume 1318, pp. 1788, 2002.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Shen, M., Zhang, J., Zhu, L., Xu, K., and Tang, X. Secure svm training over vertically-partitioned datasets using consortium blockchain for vehicular social networks. *IEEE Transactions on Vehicular Technology*, 69(6):5773–5783, 2019.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wu, W. B. and Pourahmadi, M. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pp. 1755–1768, 2009.

Wu, Y., Cai, S., Xiao, X., Chen, G., and Ooi, B. C. Privacy preserving vertical federated learning for tree-based models. *arXiv preprint arXiv:2008.06170*, 2020.

Yang, K., Fan, T., Chen, T., Shi, Y., and Yang, Q. A quasi-newton method based vertical federated learning framework for logistic regression. *arXiv preprint arXiv:1912.00513*, 2019.

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pp. 1564–1599, 1999.

Yu, B., Assouad, F., and Le Cam, L. Festschrift for lucien le cam. In *Assouad, Fano, and Le Cam*, volume 423, pp. 435. Springer, 1997.

Zhang, Y., Duchi, J., Jordan, M. I., and Wainwright, M. J. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., Cui, H., and Chang, E. Parallelizing support vector machines on distributed computers. *Advances in neural information processing systems*, 20, 2007.

Zinkevich, M., Weimer, M., Smola, A. J., and Li, L. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 4, pp. 4. Citeseer, 2010.

A. Some Preliminary Lemmas, Corollaries, and Propositions

A.1. A Lemma from Linear Algebra

Lemma A.1. *Consider the matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and define matrix* $\mathbf{B} \in \mathbb{R}^{(m+n)\times(m+n)}$ *as follows:*

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix}.$$

If we denote the singular value decomposition of $\mathbf{A}$ *as* $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$*, then the eigenvalues and eigenvectors of* $\mathbf{B}$ *are:*

$$\{\pm\sigma_i\}_{i=1}^{r}, \qquad \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} \pm\mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \right\}_{i=1}^{r}$$

Proof. From the singular value decomposition of $\mathbf{A}$, we have:

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \qquad \mathbf{A}^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i.$$

We write:

$$\begin{aligned} \frac{1}{\sqrt{2}}\mathbf{B}\begin{bmatrix} \pm\mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} &= \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \pm\mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{A}\mathbf{v}_i \\ \pm\mathbf{A}^\top\mathbf{u}_i \end{bmatrix} \\ &= \frac{\sigma_i}{\sqrt{2}} \begin{bmatrix} \mathbf{u}_i \\ \pm\mathbf{v}_i \end{bmatrix} \\ &= \frac{\pm\sigma_i}{\sqrt{2}} \begin{bmatrix} \pm\mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}. \end{aligned}$$

This completes the proof. □

A.2. Some Properties of Sub–Gaussian Random Variables

To study some of the properties of sub–Gaussian random variables, it is necessary to be familiar with another family of random variables. This family of random variables is an extension of the class of sub–Gaussian random variables and is called sub–Gamma random variables.

Definition A.2 ((Boucheron et al., 2013, Chapter 2.4))**.** A random variable X is called (σ, α)–sub–Gamma, if:

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-\alpha|\lambda|)}\right),$$

for all λ, $|\lambda| < \frac{1}{\alpha}$.

We state and prove some properties of sub–Gaussian and sub–Gamma random variables.

Lemma A.3 ((Boucheron et al., 2013))**.** *Consider an independent sequence* $\{X_i\}_{i=1}^{m}$ *of random variables,*

- *if* X_i*,* $i \in [m]$ *is a* σ_i*–sub–Gaussian random variable, then* $\sum\limits_{i=1}^{n} X_i$ *is* $\sqrt{\sum\limits_{i=1}^{n} \sigma_i^2}$*–sub–Gaussian.*
- *if* X_i*,* $i \in [m]$ *is a* (σ_i, α_i)*–sub–Gamma random variable, then* $\sum\limits_{i=1}^{n} X_i$ *is* $\left(\sqrt{\sum\limits_{i=1}^{n} \sigma_i^2}, \max\limits_i\{\alpha_i\}\right)$*–sub–Gamma.*

Lemma A.4. *Any (σ, α)–sub–Gamma random variable X satisfies the following inequality:*

$$\begin{aligned}\mathbb{P}\left[X \geq t\right] &\leq \exp\left(\frac{-t^2}{2(\sigma^2+\alpha t)}\right)\\ &\leq \exp\left(\frac{1}{2(\sigma^2+\alpha)}\min\{t, t^2\}\right)\end{aligned}$$

Proof. Some variations of this lemma are presented in different papers. For completeness, we prove it here. We write:

$$\begin{aligned}\mathbb{P}[X \geq t] &\overset{\text{(a)}}{=} \mathbb{P}\left[e^{\lambda X} \geq e^{\lambda t}\right]\\ &\overset{\text{(b)}}{\leq} e^{-\lambda t}\,\mathbb{E}\left[e^{\lambda X}\right]\\ &\overset{\text{(c)}}{\leq} \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2(1-\alpha|\lambda|)}\right).\end{aligned}$$

Note that (a) holds when $\lambda > 0$, (b) is derived from Markov's inequality, and (c) follows from Definition A.2, assuming $|\lambda| < \frac{1}{\alpha}$. Now we set $\lambda = \frac{t}{\sigma^2+t\alpha}$, which satisfies the criteria $0 < \lambda < \frac{1}{\alpha}$. Thus:

$$\begin{aligned}\mathbb{P}[X \geq t] &\leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2(1-\alpha|\lambda|)}\right)\Bigg|_{\lambda=\frac{t}{\sigma^2+t\alpha}}\\ &= \exp\left(\frac{-t^2}{2(\sigma^2+\alpha t)}\right).\end{aligned}$$

Note that if $t \leq 1$, we have: $\sigma^2 + \alpha \geq \sigma^2 + \alpha t$, therefore:

$$\frac{t^2}{2(\sigma^2+\alpha t)} \geq \frac{t^2}{2(\sigma^2+\alpha)} \qquad (0 < t \leq 1).$$

On the other hand, if $t \geq 1$, we have: $t(\sigma^2 + \alpha) \geq \sigma^2 + \alpha$, therefore:

$$\frac{t^2}{2(\sigma^2+\alpha t)} \geq \frac{t}{2(\sigma^2+\alpha)} \qquad (t \geq 1).$$

Thus:

$$\frac{t^2}{2(\sigma^2+\alpha t)} \geq \frac{1}{2(\sigma^2+\alpha)}\min\{t, t^2\}.$$

and the second inequality is proved. □

Lemma A.5 (A maximal inequality for sub–Gamma Random Variables (Boucheron et al., 2013, Corollary 2.6))**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of centered sub–Gamma random variables with the same parameters (σ, α). Then:*

$$\mathbb{E}\Big[\max_{i\in[n]} X_i\Big] \leq \sigma\sqrt{2\ln(n)} + \alpha\ln(n). \tag{12}$$

Lemma A.6. *Assume that X, Y are centered sub–Gaussian random variables with parameters σ_1 and σ_2, respectively. Then $XY - \mathbb{E}[XY]$ is a sub–Gamma random variable with parameters $(5\sigma_1\sigma_2, 2.5\sigma_1\sigma_2)$.*

Proof. We write:

$$\begin{align*}
\mathbb{E}\left[e^{\lambda(XY-\mathbb{E}[XY])}\right] &= 1+\lambda\,\mathbb{E}\left[(XY-\mathbb{E}[XY])\right]+\sum_{k=2}^{+\infty}\frac{\lambda^k\,\mathbb{E}\left[(XY-\mathbb{E}[XY])^k\right]}{k!} \\
&= 1+\sum_{k=2}^{+\infty}\frac{\lambda^k}{k!}\,\mathbb{E}\left[(XY-\mathbb{E}[XY])^k\right] \\
&\leq 1+\sum_{k=2}^{+\infty}\frac{|\lambda|^k}{k!}\,\mathbb{E}\left[|XY-\mathbb{E}[XY]|^k\right] \\
&= 1+\sum_{k=2}^{+\infty}\frac{|\lambda|^k}{k!}\|XY-\mathbb{E}[XY]\|_k^k \tag{13} \\
&\leq 1+\sum_{k=2}^{+\infty}\frac{|\lambda|^k}{k!}\left(\|XY\|_k+\|\,\mathbb{E}[XY]\|_k\right)^k \tag{14} \\
&= 1+\sum_{k=2}^{+\infty}\frac{|\lambda|^k}{k!}\left(\left(\mathbb{E}\left[|XY|^k\right]\right)^{\frac{1}{k}}+|\,\mathbb{E}[XY]|\right)^k \\
&\leq 1+\sum_{k=2}^{+\infty}\frac{|\lambda|^k}{k!}\left(\left(\mathbb{E}\left[X^{2k}\right]\mathbb{E}\left[Y^{2k}\right]\right)^{\frac{1}{2k}}+\sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}\right)^k \tag{15} \\
&\leq 1+\sum_{k=2}^{+\infty}(|\lambda|\sigma_1\sigma_2)^k\frac{\left(\left(2^{k+1}k!\right)^{\frac{1}{k}}+1\right)^k}{k!} \tag{16} \\
&\leq 1+\sum_{k=2}^{+\infty}(|\lambda|\sigma_1\sigma_2)^k.2.(2.5)^k \tag{17} \\
&= 1+\frac{25(\lambda\sigma_1\sigma_2)^2}{2(1-2.5|\lambda|\sigma_1\sigma_2)} \tag{18} \\
&\leq \exp\left(\frac{25(\lambda\sigma_1\sigma_2)^2}{2(1-2.5|\lambda|\sigma_1\sigma_2)}\right), \tag{19}
\end{align*}$$

where

- in (13) for a random variable Z, $\|Z\|_k := \mathbb{E}^{1/k}[|Z|^k]$ is the L_k norm of the random variable Z,
- (14) follows directly from the application of Minkowski's inequality (also known as the triangle inequality) to the L_k norm.
- (15) follows from Cauchy–Schwarz inequality,
- in (16), we use the following upper bound for the $2k$-th moment of a σ–sub–Gaussian random variable Z (see (Boucheron et al., 2013, Theorem 2.1)),
$$\mathbb{E}[Z^{2k}]\leq 2(2\sigma^2)^k k!,$$
- (17) follows from the fact that the function $h[k] := \frac{\left(\left(2^{k+1}k!\right)^{\frac{1}{k}}+1\right)^k}{(2.5)^k k!}$ is a decreasing function on $\{2,3,\cdots\}$ and takes its maximum at $k=2$, which is equal to 2 (see Figure 2).

Finally, (19) implies that $XY-\mathbb{E}[XY]$ is a $(5\sigma_1\sigma_2, 2.5\sigma_1\sigma_2)$–sub–Gamma random variable. □

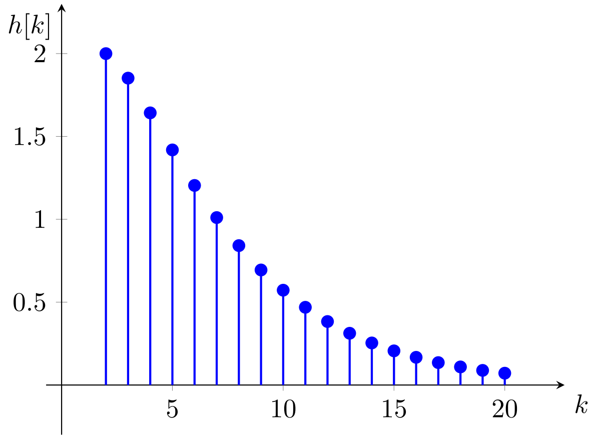


Figure 2. Diagram of the function $h[k] := \frac{\left(\left(2^{k+1}k!\right)^{\frac{1}{k}}+1\right)^k}{(2.5)^k k!}$.

Corollary A.7. *Let $\{(X_i, Y_i)\}_{i=1}^m$ be a sequence of i.i.d. pairs of random variables where X_i's and Y_i's are σ_1 sub–Gaussian and σ_2 sub–Gaussian, respectively. If we define $Z_i = X_iY_i - \mathbb{E}[X_iY_i]$, then we have:*

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i \geq 10\sigma_1\sigma_2 t\right] \leq \exp\left(-m.\min\{t, t^2\}\right).$$

Proof. We know from Lemma A.6 that $Z_i = X_iY_i - \mathbb{E}[X_iY_i]$ is a $(5\sigma_1\sigma_2, 2.5\sigma_1\sigma_2)$–sub–Gamma random variable. Therefore, using Lemma A.3, we conclude that $\sum_{i=1}^{m} Z_i$ is a $(5\sigma_1\sigma_2\sqrt{m}, 2.5\sigma_1\sigma_2)$–sub–Gamma random variable. Thus,

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m}(X_iY_i - \mathbb{E}[X_iY_i]) \geq 10\sigma_1\sigma_2 t\right] &= \mathbb{P}\left[\sum_{i=1}^{m} Z_i \geq 10m\sigma_1\sigma_2 t\right] \\
&\leq \exp\left(\frac{-100m^2\sigma_1^2\sigma_2^2t^2}{2(25\sigma_1^2\sigma_2^2 m + 25\sigma_1^2\sigma_2^2 mt)}\right) \\
&= \exp\left(\frac{-2mt^2}{1+t}\right) \\
&\leq \exp\left(-m.\min\{t, t^2\}\right).
\end{aligned} \tag{20}$$

□

A.3. An Important Relation Between the Packing and the Covering Numbers of a Set

The packing and covering numbers are defined in Section 5.2. There is an important relation between the packing and the covering numbers of a set, which is stated in the following lemma:

Lemma A.8 ((Wainwright, 2019, Lemma 5.5))**.** *For all $\epsilon > 0$, the packing and covering numbers are related as follows:*

$$\mathcal{M}(\mathcal{K}, \mathsf{d}, 2\epsilon) \leq \mathcal{N}(\mathcal{K}, \mathsf{d}, \epsilon) \leq \mathcal{M}(\mathcal{K}, \mathsf{d}, \epsilon).$$

A.4. Finding Upper Bound on Operator Norm of Matrices, Using Covering Nets

The following lemma is useful in finding an upper bound for the operator norm of a random matrix.

Lemma A.9 ((Vershynin, 2018, Exercise 4.4.3))**.** *Let $\mathbf{A}$ be a $m \times n$ matrix. We define the sets $\mathcal{S}^{m-1} = \{\mathbf{u} \in \mathbb{R}^{m-1} : \|\mathbf{u}\| = 1\}, \mathcal{S}^{n-1} = \{\mathbf{v} \in \mathbb{R}^{n-1} : \|\mathbf{v}\| = 1\}$. We fix an arbitrary $\epsilon > 0$ and denote ϵ–covering set of $\mathcal{S}^{m-1}$ by $\mathcal{N}_\epsilon^{(m)}$ and*

ϵ–covering set of $\mathcal{S}^{n-1}$ by $\mathcal{N}_\epsilon^{(n)}$. We have:

$$\|\mathbf{A}\|_{\mathsf{op}} \le \frac{1}{1-2\epsilon} \max_{\mathbf{u}\in\mathcal{N}_\epsilon^{(m)},\mathbf{v}\in\mathcal{N}_\epsilon^{(n)}} \left\{\mathbf{u}^\top \mathbf{A}\mathbf{v}\right\}.$$

A.5. Packing and Covering in Matrix Spaces

Consider the family of matrices defined as follows:

$$\mathcal{A} = \{\mathbf{A} \in \mathbb{R}^{m\times n} : \|\mathbf{A}\|_{\mathsf{op}} \le r\}.$$

We vectorize each member of this family as:

$$\mathbf{a} = \mathrm{vec}(\mathbf{A}) = [A_{11}, A_{12}, \dots, A_{1n}, A_{21}, \dots, A_{mn}]^\top.$$

We convert the dist norm on the matrix space $\mathbb{R}^{m\times n}$ to a norm on $\mathbb{R}^{mn}$ via $\|\|\mathbf{a}\|\|_{\mathsf{dist}} = \|\mathbf{A}\|_{\mathsf{dist}}$. Now we define the ball of radius r under norm $\|\|.\|\|_{\mathsf{op}}$ as:

$$\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r) = \left\{\mathbf{x} \in \mathbb{R}^{mn} : \|\|\mathbf{x}\|\|_{\mathsf{op}} \le r\right\} = \left\{\mathrm{vec}(\mathbf{A}) : \mathbf{A} \in \mathbb{R}^{m\times n}, \|\mathbf{A}\|_{\mathsf{op}} \le r\right\}. \tag{21}$$

We consider an ϵ–covering net for $\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r)$ under the norm $\|\|.\|\|_{\mathsf{dist}}$, where dist can denote Frobenius or operator norm.

- Consider the case $\mathsf{dist} = \mathsf{op}$, in this case, from (Wainwright, 2019, Lemma 5.7), we have:

$$\left(\frac{r}{\epsilon}\right)^{mn} \le \mathcal{N}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r), \|\|.\|\|_{\mathsf{op}}, \epsilon) \le \left(1 + \frac{2r}{\epsilon}\right)^{mn} \le \left(\frac{3r}{\epsilon}\right)^{mn}.$$

From Lemma A.8 we conclude:

$$\left(\frac{r}{\epsilon}\right)^{mn} \le \mathcal{N}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r), \|\|.\|\|_{\mathsf{op}}, \epsilon) \le \mathcal{M}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r), \|\|.\|\|_{\mathsf{op}}, \epsilon) \le \mathcal{N}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r), \|\|.\|\|_{\mathsf{op}}, \epsilon) \le \left(1 + \frac{4r}{\epsilon}\right)^{mn} \tag{22}$$

Matrix quantization scheme: We quantize matrix $\mathbf{A}$, whose operator norm is at most r, under the norm $\|\|.\|\|_{\mathsf{op}}$, with the matrices corresponding to the covering points of $\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r)$. Note that the number of these points is less than $\left(\frac{3r}{\epsilon}\right)^{mn}$, so we can send the index of the quantized matrix using at most $mn\log_2(\frac{3r}{\epsilon})$ bits. Furthermore, if we denote the output of the quantization with $Q_{\mathsf{op}}(\mathbf{A})$, we have:

$$\left\|\mathbf{A} - Q_{\mathsf{op}}(\mathbf{A})\right\|_{\mathsf{op}} \le \epsilon.$$

- In the case $\mathsf{dist} = \mathsf{F}$, we only find a lower bound on the packing number $\mathcal{M}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(r), \|\|.\|\|_{\mathsf{F}}, \epsilon)$.

Lemma A.10. *For $\mathcal{M}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(1), \|\|.\|\|_{\mathsf{F}}, \epsilon)$, we have:*

$$\mathcal{M}(\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(1), \|\|.\|\|_{\mathsf{F}}, \epsilon) \ge \left(\frac{\sqrt{\min\{m,n\}}}{14\epsilon}\right)^{mn}$$

Proof. Let $\mathcal{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_M\}$ be a *maximal ϵ–packing* of the ball $\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(1)$. Then $\mathcal{A}$ is also an ϵ–covering. In particular:

$$\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{op}}}(1) \subseteq \bigcup_{i=1}^{M} \mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{F}}}(\mathbf{A}_i; \epsilon),$$

where $\mathcal{B}^{(mn)}_{\|\|\cdot\|\|_{\mathsf{F}}}(\mathbf{A}; \epsilon) = \{\mathrm{vec}(\mathbf{B}) : \left\|\mathbf{B} - \mathbf{A}\right\|_{\mathsf{F}} \le \epsilon\}$ is the Frobenius ball with center $\mathbf{A}$ and radius ϵ. The proof follows a probabilistic argument which is similar to the volume argument usually used to prove packing numbers.

Let $\boldsymbol{G} = [g_{ij}]_{m \times n}$ be a random matrix with independent $g_{ij} \sim \mathcal{N}\left(0, \frac{1}{4(\sqrt{m}+\sqrt{n})^2}\right)$ elements. It follows from (Vershynin, 2018, Theorem 7.3.1) that $\mathbb{E}\left[\|\boldsymbol{G}\|_{\mathsf{op}}\right] \leq \frac{1}{2}$. Thus Markov inequality yields:

$$\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{op}}}(1)\right] = \mathbb{P}\left[\|\boldsymbol{G}\|_{\mathsf{op}} \leq 1\right] = 1 - \mathbb{P}\left[\|\boldsymbol{G}\|_{\mathsf{op}} > 1\right] \geq 1 - \frac{1}{2} = \frac{1}{2}. \tag{23}$$

On the other side, union bound gives:

$$\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{op}}}(1)\right] \leq \sum_{i=1}^{M} \mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)\right]. \tag{24}$$

We now proceed to find an upper bound on the inner term in the summation. Observe:

$$\begin{aligned}
\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)\right] &= \int_{\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)} \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} \exp\left(-2(\sqrt{m}+\sqrt{n})^2 \|\mathbf{G}-\mathbf{A}_i\|_{\mathsf{F}}^2\right) d\mathbf{G} \\
&\leq \int_{\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}} \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} d\mathbf{G} \\
&= \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} \mathsf{Vol}\left(\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)\right) \\
&= \left(\frac{2\epsilon^2(\sqrt{m}+\sqrt{n})^2}{\pi}\right)^{\frac{mn}{2}} \mathsf{Vol}\left(\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;1)\right),
\end{aligned} \tag{25}$$

where we have used the density formula for normal distribution. Now we view $\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;1)$ as a mn-dimensional euclidean ball. It is well known that the volume of this ball is given by

$$\mathsf{Vol}\left(\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;1)\right) = \frac{\pi^{\frac{mn}{2}}}{\Gamma(1+\frac{mn}{2})}.$$

Using the bound $\Gamma(1+x) >> (\frac{x}{e})^x$, in (25), we obtain:

$$\begin{aligned}
\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)\right] &\leq \left(\frac{4e\epsilon^2(\sqrt{m}+\sqrt{n})^2}{mn}\right)^{\frac{mn}{2}} \\
&\leq \left(\frac{16e\epsilon^2 \max\{m,n\}}{mn}\right)^{\frac{mn}{2}} \\
&= \left(\frac{16e\epsilon^2}{\min\{m,n\}}\right)^{\frac{mn}{2}}.
\end{aligned} \tag{26}$$

Putting (23), (24) and (26) together yields:

$$M \geq \frac{1}{2}\left(\frac{\min\{m,n\}}{16e\epsilon^2}\right)^{\frac{mn}{2}} \geq \left(\frac{\sqrt{\min\{m,n\}}}{8\sqrt{e}\epsilon}\right)^{mn} \geq \left(\frac{\sqrt{\min\{m,n\}}}{14\epsilon}\right)^{mn}. \tag{27}$$

This concludes the proof. □

From Lemma A.10, we conclude that:

$$\mathcal{M}(\mathcal{B}^{(mn)}_{|\!|\!|\cdot|\!|\!|_{\mathsf{op}}}(r), |\!|\!|.|\!|\!|_{\mathsf{F}}, \epsilon) \geq \left(\frac{r \cdot \sqrt{\min\{m,n\}}}{14\epsilon}\right)^{mn}. \tag{28}$$

B. Materials for Completing the Proof of Theorem 4.1

B.1. The Detailed Proof of Lemma 6.2

Proof. First, we know that $\mathbf{C}_v$ is the covariance matrix of a σ^2–sub–Gaussian random vector $\mathbf{Z}$. So we must have $\mathbf{C}_v \succeq \mathbf{0}$. Also from Definition 5.2, for all vectors $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$, the random variable $\mathbf{u}^\top \mathbf{Z}$ is σ^2–sub–Gaussian, therefore $\mathrm{Var}[\mathbf{u}^\top \mathbf{Z}] \leq \sigma^2$. This implies that for all $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$:

$$\mathrm{Var}[\mathbf{u}^\top \mathbf{Z}] = \mathbb{E}[\mathbf{u}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{u}] = \mathbf{u}^\top \mathbf{C}_v \mathbf{u} \leq \sigma^2.$$

Therefore we must have $\|\mathbf{C}_v\|_{\mathsf{op}} \leq \sigma^2$.

We write $\mathbf{C}_v$ as $\mathbf{C}_v = \frac{\sigma^2}{2}\mathbf{I} + \begin{bmatrix} \mathbf{0} & \delta \mathbf{D}_v \\ \delta \mathbf{D}_v^\top & \mathbf{0} \end{bmatrix}$. From Lemma A.1, the eigenvalues of $\mathbf{C}_v$ are $\frac{\sigma^2}{2} \pm \delta\sigma_i(\mathbf{D}_v)$. Therefore it suffices to impose the constraint $\delta \leq \frac{\sigma^2}{2\max\limits_{v\in\mathcal{V}}\{\|\mathbf{D}_v\|_{\mathsf{op}}\}}$, to ensure the constraints $\mathbf{C}_v \succeq \mathbf{0}$ and $\|\mathbf{C}_v\|_{\mathsf{op}} \leq \sigma^2$are satisfied.

Then, we write:

$$\begin{aligned}
\rho &= \inf_{v,v':v\neq v'} \|\mathbf{C}_v - \mathbf{C}_{v'}\|_{\mathsf{dist}} \\
&= \inf_{v,v':v\neq v'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_v - \mathbf{D}_{v'}) \\ \delta(\mathbf{D}_v - \mathbf{D}_{v'})^\top & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\
&\overset{\text{(a)}}{=} \begin{cases} \delta \inf\limits_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}} & \text{when } \mathsf{dist} = \mathsf{op} \\ \sqrt{2}\delta \inf\limits_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}} & \text{when } \mathsf{dist} = \mathsf{F} \end{cases} \\
&= \sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\,\delta \inf_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}}\,,
\end{aligned}$$

where (a) is true due to Lemma A.1. Also we derive an upper bound for $I(V; M_1, M_2)$:

$$\begin{aligned}
I(V; M_1, M_2) &= I(V; M_1) + I(V; M_2|M_1) \\
&\overset{\text{(a)}}{=} I(V; M_2|M_1) \\
&\leq I(V; M_2|M_1) + I(M_1; M_2) \\
&= I(V; M_2) + I(M_1; M_2|V) \\
&\overset{\text{(b)}}{=} I(M_1; M_2|V),
\end{aligned} \tag{29}$$

where (a) and (b) are true because for all $v \in \mathcal{V}$, the vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v)$ has the same marginal distribution over the first d_1 dimensions and the second d_2 dimensions. Therefore $\mathsf{X} = \left\{\mathbf{X}^{(i)}\right\}_{i=1}^m$ is independent from V and similarly, $\mathsf{Y} = \left\{\mathbf{Y}^{(i)}\right\}_{i=1}^m$ is also independent from V. We conclude that M_1 and M_2 are independent from V.

Now we upper-bound the $I(M_1; M_2|V = v)$. Note that the Markov chain of the problem is: $M_1 -\!\!\circ\!\!- \mathsf{X} -\!\!\circ\!\!- \mathsf{Y} -\!\!\circ\!\!- M_2$. We write:

$$\begin{aligned}
I(M_1; M_2|V = v) &\leq I\left(M_1; \mathsf{Y}\middle|\, V = v\right) \\
&\overset{\text{(a)}}{\leq} s\left(\mathsf{X}; \mathsf{Y}\middle|\, V = v\right) I\left(M_1; \mathsf{X}\middle|\, V = v\right) \\
&\overset{\text{(b)}}{\leq} s\left(\mathsf{X}; \mathsf{Y}\middle|\, V = v\right) B_1,
\end{aligned}$$

where in (a), $s\left(\mathsf{X}; \mathsf{Y}\middle|\, V = v\right)$ is the SDPI constant (Definition 5.6) for the joint distribution $p_{\mathsf{X},\mathsf{Y}}$, when $V = v$. In addition, (b) holds due to (6). From Lemma 5.7 we have:

$$\begin{aligned}
s\left(\mathbf{X}; \mathbf{Y}\middle|\, V = v\right) &= \left\| \mathbf{C}_{v,\mathbf{XX}}^{-1/2} \mathbf{C}_{v,\mathbf{XY}} \mathbf{C}_{v,\mathbf{YY}}^{-1/2} \right\|_{\mathsf{op}}^2 \\
&= \left(\frac{2\delta}{\sigma^2}\right)^2 \|\mathbf{D}_v\|_{\mathsf{op}}^2\,.
\end{aligned}$$

Therefore:

$$I(M_1; M_2 | V = v) \leq \left(\frac{2\delta}{\sigma^2}\right)^2 B_1 \|\mathbf{D}_v\|_{\mathsf{op}}^2 . \tag{30}$$

The second upper bound is very similar to the first one:

$$I(M_1; M_2 | V = v) \leq \left(\frac{2\delta}{\sigma^2}\right)^2 B_2 \|\mathbf{D}_v\|_{\mathsf{op}}^2 . \tag{31}$$

Therefore, from (29), (30), and (31) we have:

$$\begin{aligned} I(V; M_1, M_2) &\leq \left(\frac{2\delta}{\sigma^2}\right)^2 \min\{B_1, B_2\} \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}\in\mathcal{V}} \|\mathbf{D}_v\|_{\mathsf{op}}^2 \\ &\leq \left(\frac{2\delta}{\sigma^2}\right)^2 \min\{B_1, B_2\} \max_{\mathbf{v}\in\mathcal{V}} \left\{\|\mathbf{D}_v\|_{\mathsf{op}}^2\right\} . \end{aligned}$$

Then from Lemma 6.1 we write:

$$\begin{aligned} \mathcal{M}_{\mathsf{dist}} &\geq \frac{\rho}{2}\left[1 - \frac{I(V; M_1, M_2) + 1}{\log_2(|\mathcal{V}|)}\right] \\ &\geq \frac{\sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\delta \inf\limits_{v,v':v\neq v'} \|\mathbf{D}_\mathbf{v} - \mathbf{D}_{\mathbf{v}'}\|_{\mathsf{dist}}}{2} \left[1 - \frac{2\left(\frac{2\delta}{\sigma^2}\right)^2 \min\{B_1, B_2\} \max\limits_{\mathbf{v}\in\mathcal{V}} \left\{\|\mathbf{D}_v\|_{\mathsf{op}}^2\right\}}{\log_2(|\mathcal{V}|)}\right] . \end{aligned}$$

We set $\delta = \frac{\sigma^2}{4 \max_{\mathbf{v}\in\mathcal{V}} \{\|\mathbf{D}_v\|_{\mathsf{op}}\}} \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{\min\{B_1, B_2\}}}, 2\right\}$, obviously this value of δ satisfies the criteria $\delta \leq \frac{\sigma^2}{2 \max_{v\in\mathcal{V}} \{\|\mathbf{D}_v\|_{\mathsf{op}}\}}$. Therefore we have:

$$\mathcal{M}_{\mathsf{dist}} \geq \frac{\sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\sigma^2}{16} \cdot \frac{\inf\limits_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}}}{\max\limits_{\mathbf{v}\in\mathcal{V}} \left\{\|\mathbf{D}_v\|_{\mathsf{op}}\right\}} \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{\min\{B_1, B_2\}}}, 2\right\} .$$

The proof of the first inequality is completed.

For the second inequality, first note that we have:

$$\begin{aligned} \rho' &= \inf_{v,v':v\neq v'} \left\|\mathbf{C}'_v - \mathbf{C}'_{v'}\right\|_{\mathsf{dist}} \\ &= \inf_{v,v':v\neq v'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}'_v - \mathbf{D}'_{v'}) \\ \delta(\mathbf{D}'_v - \mathbf{D}'_{v'})^\top & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\ &\overset{\text{(a)}}{=} \begin{cases} \delta \inf\limits_{v,v':v\neq v'} \left\|\mathbf{D}'_v - \mathbf{D}'_{v'}\right\|_{\mathsf{dist}} & \text{when } \mathsf{dist} = \mathsf{op} \\ \sqrt{2}\delta \inf\limits_{v,v':v\neq v'} \left\|\mathbf{D}'_v - \mathbf{D}'_{v'}\right\|_{\mathsf{dist}} & \text{when } \mathsf{dist} = \mathsf{F} \end{cases} \\ &= \sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\delta \inf_{v,v':v\neq v'} \left\|\mathbf{D}'_v - \mathbf{D}'_{v'}\right\|_{\mathsf{dist}} , \end{aligned} \tag{32}$$

Now we define $\mathbf{Z} = \left\{\mathbf{Z}^{(i)}\right\}_{i=1}^m$, $\mathbf{X}' = \left\{\mathbf{Z}^{(i)}_{[1:d/2]}\right\}_{i=1}^m$ and $\mathbf{Y}' = \left\{\mathbf{Z}^{(i)}_{[d/2+1:d]}\right\}_{i=1}^m$, then try find another upper bound for $I(V; M_1, M_2)$, due to the Markov chain of the problem, which is $M_1 \multimap \mathbf{X} \multimap \mathbf{Y} \multimap M_2$. Note that with the second set of distributions, $\mathbf{X}'$ and $\mathbf{Y}'$ are independent from V, and we have:

$$\begin{aligned}
I(V;M_1,M_2) &\le I(V;\mathbf{X},\mathbf{Y}) \\
&= I(V;\mathbf{Z}) \\
&= I(V;\mathbf{X}',\mathbf{Y}') \\
&= I(V;\mathbf{X}') + I(V;\mathbf{Y}'|\mathbf{X}') \\
&= I(V;\mathbf{Y}'|\mathbf{X}') \\
&\le I(V;\mathbf{Y}'|\mathbf{X}') + I(\mathbf{X}';\mathbf{Y}') \\
&= I(V;\mathbf{Y}') + I(\mathbf{X}';\mathbf{Y}'|V) \\
&= I(\mathbf{X}';\mathbf{Y}'|V).
\end{aligned} \tag{33}$$

We can write:

$$\begin{aligned}
I(\mathbf{X}';\mathbf{Y}'\big|\, V=v) &= m\, I\left(\mathbf{Z}_{[1:d/2]};\mathbf{Z}_{[d/2+1:d]}|V=v\right) \\
&= m\left[h(\mathbf{Z}_{[1:d/2]}|V=v) + h(\mathbf{Z}_{[d/2+1:d]}|V=v) - h(\mathbf{Z}_{[1:d/2]},\mathbf{Z}_{[d/2+1:d]}|V=v)\right] \\
&\overset{\text{(a)}}{=} \frac{m}{2}\log_2\left(\frac{\det\left\{\frac{\sigma^2}{2}\mathbf{I}_{d/2}\right\}\det\left\{\frac{\sigma^2}{2}\mathbf{I}_{d/2}\right\}}{\det\left\{\mathbf{C}'_v\right\}}\right) \\
&= \frac{m}{2}\left(2r_v\log_2(\frac{\sigma^2}{2}) - \sum_{i=1}^{r_v}\log_2\left(\frac{\sigma^4}{4} - \delta^2\sigma_i^2(\mathbf{D}'_v)\right)\right) \\
&\overset{\text{(b)}}{\le} \frac{-mr_v}{2}\log_2\left(1 - \frac{4\delta^2\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2}{\sigma^4}\right) \\
&\overset{\text{(c)}}{\le} \frac{4mr_v\delta^2\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2}{\ln(2)\sigma^4} \\
&\le \frac{2md\delta^2\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2}{\ln(2)\sigma^4},
\end{aligned} \tag{34}$$

where (a) is true duo to the (Cover, 1999, Theorem 8.4.1) and $r_v = \operatorname{rank}(\mathbf{D}'_v)$. Also (b) and (c) are true because $g(x) = -\log_2(1-x) = \log_2(\frac{1}{1-x})$ is an increasing function, and simply we have for all $x \in [0,1/2]$: $\log_2(\frac{1}{1-x}) < \frac{2x}{\ln(2)}$, and we assume that $\delta \le \frac{\sigma^2}{2\sqrt{2}\max\limits_{v\in\mathcal{V}}\{\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}\}}$.

Therefore, from (33) and (34) we have:

$$\begin{aligned}
I(V;M_1,M_2) &\le \frac{2md\delta^2}{\ln(2)\sigma^4}\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2 \\
&\le \frac{2md\delta^2}{\ln(2)\sigma^4}\max_{\mathbf{v}\in\mathcal{V}}\left\{\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2\right\}.
\end{aligned}$$

Then from Lemma 6.1 we write:

$$\begin{aligned}
\mathcal{M}_{\mathsf{dist}} &\ge \frac{\rho}{2}\left[1 - \frac{I(V;M_1,M_2)+1}{\log_2(|\mathcal{V}|)}\right] \\
&\ge \frac{\sqrt{1+\mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\,\delta\inf\limits_{v,v':v\neq v'}\left\|\mathbf{D}'_v - \mathbf{D}'_{v'}\right\|_{\mathsf{dist}}}{2}\left[1 - \frac{6md\delta^2}{\sigma^4\log_2(|\mathcal{V}|)}\max_{\mathbf{v}\in\mathcal{V}}\left\{\left\|\mathbf{D}'_v\right\|_{\mathsf{op}}^2\right\}\right].
\end{aligned}$$

We set $\delta = \frac{\sigma^2}{2\sqrt{2}\max\limits_{\mathbf{v}\in\mathcal{V}}\{\|\mathbf{D}'_v\|_{\mathsf{op}}\}} \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{12md}}, 1\right\}$, obviously this value of δ satisfies the criteria $\delta \leq \frac{\sigma^2}{2\sqrt{2}\max\limits_{v\in\mathcal{V}}\{\|\mathbf{D}'_v\|_{\mathsf{op}}\}}$. Therefore we have:

$$\mathcal{M}_{\mathsf{dist}} \geq \frac{\sqrt{1+\mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\sigma^2}{8} \cdot \frac{\inf\limits_{v,v':v\neq v'} \|\mathbf{D}'_v - \mathbf{D}'_{v'}\|_{\mathsf{dist}}}{\max\limits_{v\in\mathcal{V}}\left\{\|\mathbf{D}'_v\|_{\mathsf{op}}\right\}} \min\left\{\sqrt{\frac{\log_2(|\mathcal{V}|)}{3md}}, \frac{1}{\sqrt{2}}\right\}.$$

The proof of the second part is completed. □

B.2. Proof of Lemma 6.3

Proof. Same as Lemma 6.2, we must have $\mathbf{C}_u \succeq \mathbf{0}$ and $\|\mathbf{C}_u\|_{\mathsf{op}} \leq \sigma^2$. Note that the eigenvalues of $\mathbf{C}_u$ are $\left\{\frac{\sigma^2}{2} \pm \delta\sigma_i(\mathbf{D}_u)\right\}_{i=1}^{\mathrm{rank}(\mathbf{D}_u)}$. Thus for satisfying conditions $\mathbf{C}_u \succeq \mathbf{0}$ and $\|\mathbf{C}_u\|_{\mathsf{op}} \leq \sigma^2$, we must have $\delta\left\|\mathbf{D}_u\right\|_{\mathsf{op}} \leq \frac{\sigma^2}{2}$, which results in $\delta \leq \frac{\sigma^2}{2\max\limits_{u\in\mathcal{U}}\left\{\left\|\mathbf{D}_u\right\|_{\mathsf{op}}\right\}}$.

Then, we write:

$$\begin{aligned}
\rho &= \inf_{u,u':u\neq u'} \|\mathbf{C}_u - \mathbf{C}_{u'}\|_{\mathsf{dist}} \\
&= \inf_{u,u':u\neq u'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_u - \mathbf{D}_{u'}) & \mathbf{0} & \mathbf{0} \\ \delta(\mathbf{D}_u - \mathbf{D}_{u'})^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\
&= \inf_{u,u':u\neq u'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_u - \mathbf{D}_{u'}) \\ \delta(\mathbf{D}_u - \mathbf{D}_{u'})^\top & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\
&\overset{\text{(a)}}{=} \sqrt{1+\mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\delta \inf_{u,u':u\neq u'} \|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\mathsf{dist}},
\end{aligned} \tag{35}$$

where (a) is true due to Lemma A.1.

We also derive an upper bound for $I(U; M_1, M_2)$:

$$\begin{aligned}
I(U; M_1, M_2) &= I(U; M_1) + I(U; M_2|M_1) \\
&\leq I(U; M_1) + I(U; M_2|M_1) + I(M_1; M_2) \\
&= I(U; M_1) + I(U, M_1; M_2) \\
&= I(U; M_1) + I(U; M_2) + I(M_1; M_2|U) \\
&\leq I(U; M_1) + I(U; \mathbf{Y}) + I(\mathbf{X}; \mathbf{Y}|U) \\
&= I(U; M_1) \\
&\leq B_1
\end{aligned} \tag{36}$$

Therefore, we set $\delta = \frac{\sigma^2}{2\max\limits_{u\in\mathcal{U}}\left\{\left\|\mathbf{D}_u\right\|_{\mathsf{op}}\right\}}$ ans from Lemma 6.1 we have:

$$\begin{aligned}
\mathcal{M}_{\mathsf{dist}}(\sigma, B_1, B_2, d_1, d_2, m) &\geq \frac{\rho}{2}\left[1 - \frac{I(U; M_1, M_2) + 1}{\log_2(|\mathcal{U}|)}\right] \\
&\geq \frac{\sqrt{1+\mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\sigma^2}{2}\left[1 - \frac{B_1 + 1}{\log_2(|\mathcal{U}|)}\right] \frac{\inf\limits_{u,u':u\neq u'} \|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\mathsf{op}}}{\max\limits_{u\in\mathcal{U}}\left\{\left\|\mathbf{D}_u\right\|_{\mathsf{op}}\right\}}.
\end{aligned} \tag{37}$$

□

C. Materials for Completing the Proof of Theorem 4.4

C.1. Two Lemmas and One Proposition That Are Useful in Proving Theorem 4.4

Lemma C.1. *Assume that $\mathbf{X} \in \mathbb{R}^{d_1}$ is a zero mean, sub–Gaussian vector with parameter $sigma_1$, and we have m i.i.d. samples from $\mathbf{X}$ as $\{\mathbf{X}^{(i)}\}_{i=1}^m$. Also assume that $\mathbf{Y} \in \mathbb{R}^{d_2}$ is a zero mean, sub–Gaussian vector with parameter $sigma_2$, and we have m i.i.d. samples from $\mathbf{Y}$ as $\{\mathbf{Y}^{(i)}\}_{i=1}^m$. Consider the cross–covariance matrix $\mathbf{C}_{\mathbf{XY}} \in \mathbb{R}^{d_1 \times d_2}$ as $\mathbf{C} = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top]$ and assume that we use the estimator $\widetilde{\mathbf{C}}_{\mathbf{XY}} = \frac{1}{m}\sum\limits_{i=1}^m \mathbf{X}^{(i)}\mathbf{Y}^{(i)\top}$. Then we have:*

$$\begin{aligned}
&\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 10\sigma_1\sigma_2 t\Big] \\
&\qquad \leq (9)^{d_1+d_2} \exp\Big(- m.\min\big\{t, t^2\big\}\Big),
\end{aligned}$$

and:

$$\begin{aligned}
&\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 11\sigma_1\sigma_2\Big] \\
&\qquad \leq \min\Big\{1, \exp\big(3(d_1+d_2) - m\big)\Big\}.
\end{aligned}$$

Proof. We use Lemma A.9 with $\epsilon = \frac{1}{4}$ and write:

$$\begin{aligned}
\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq t\Big] &\leq \mathbb{P}\Big[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}} \mathbf{u}^\top(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v} \geq \frac{t}{2}\Big] \\
&\leq \sum_{j=1}^{|\mathcal{N}_{1/4}^{(d_1)}|}\sum_{k=1}^{|\mathcal{N}_{1/4}^{(d_2)}|} \mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big],
\end{aligned} \tag{38}$$

where we denote the $1/4$–covering points of $\mathcal{S}^{d_1-1}$ by $\{\mathbf{u}^{(j)}\}_{j=1}^{|\mathcal{N}_{1/4}^{(d_1)}|}$ and the $1/4$–covering points of $\mathcal{S}^{d_2-1}$ by $\{\mathbf{v}^{(k)}\}_{k=1}^{|\mathcal{N}_{1/4}^{(d_2)}|}$. We also know from (Vershynin, 2018, Corollary 4.2.13) that $\mathcal{N}_{1/4}^{(d)} \leq 9^d$.

We have:

$$\begin{aligned}
&\mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] = \mathbb{P}\Big[\mathbf{u}^{(j)\top}(\frac{1}{m}\sum_{i=1}^m \mathbf{X}^{(i)}\mathbf{Y}^{(i)\top} - \mathbb{E}[\mathbf{X}\mathbf{Y}^\top])\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] \\
&= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^m (\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}) - \mathbb{E}\Big[(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)})\Big] \geq \frac{t}{2}\Big].
\end{aligned} \tag{39}$$

We know that $\mathbf{X}^{(i)}$ is a σ_1–sub–Gaussian vector, therefore, from Definition 5.2, we conclude that $U_i = \mathbf{u}^{(j)\top}\mathbf{X}^{(i)}$ is a σ_1–sub–Gaussian random variable. Similarly we conclude that $V_i = \mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}$ is a σ_2–sub–Gaussian random variable. Therefore, from Lemma A.6, $U_iV_i - \mathbb{E}[U_iV_i]$ is a $(\sigma = 5\sigma_1\sigma_2, \alpha = 2.5\sigma_1\sigma_2)$–sub–Gamma random variable. Corollary A.7 yields:

$$\begin{aligned}
&\mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] \\
&= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^m (\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}) - \mathbb{E}\Big[(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)})\Big] \geq \frac{t}{2}\Big] \\
&= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^m (U_iV_i - \mathbb{E}[U_iV_i]) \geq \frac{t}{2}\Big] \\
&\leq \exp\Big(- m.\min\Big\{\frac{t}{10\sigma_1\sigma_2}, (\frac{t}{10\sigma_1\sigma_2})^2\Big\}\Big).
\end{aligned} \tag{40}$$

Therefore we combine (38), (39), and (53) and write:

$$\begin{aligned}\mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\geq t\right] &\leq \sum_{j=1}^{|\mathcal{N}_{1/4}^{(d_1)}|}\sum_{k=1}^{|\mathcal{N}_{1/4}^{(d_2)}|}\mathbb{P}\left[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)}\geq\frac{t}{2}\right] \\ &\leq (9)^{d_1+d_2}\exp\left(-m.\min\left\{\frac{t}{10\sigma_1\sigma_2},(\frac{t}{10\sigma_1\sigma_2})^2\right\}\right).\end{aligned} \tag{41}$$

Thus:

$$\begin{aligned}\mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}\right\|_{\mathsf{op}}\geq 11\sigma_1\sigma_2\right] &\leq \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}+\left\|\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\geq 11\sigma_1\sigma_2\right] \\ &\leq \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\geq 10\sigma_1\sigma_2\right] \\ &\leq \min\left\{1,\exp\left((d_1+d_2)\ln(9)-m\right)\right\} \\ &\leq \min\left\{1,\exp\left(3(d_1+d_2)-m\right)\right\}.\end{aligned} \tag{42}$$

□

We have the following proposition, directly from Lemma C.1.

Proposition C.2. *Assume that $\mathbf{X}\in\mathbb{R}^{d_1}$ is a zero mean, σ_1^2–sub–Gaussian vector, and we have m i.i.d. samples from $\mathbf{X}$ as $\{\mathbf{X}^{(i)}\}_{i=1}^m$. Also assume that $\mathbf{Y}\in\mathbb{R}^{d_2}$ is a zero mean, σ_2^2–sub–Gaussian vector, and we have m i.i.d. samples from $\mathbf{Y}$ as $\{\mathbf{Y}^{(i)}\}_{i=1}^m$. Consider the cross–covariance matrix $\mathbf{C}_{\mathbf{XY}}\in\mathbb{R}^{d_1\times d_2}$ as $\mathbf{C}=\mathbb{E}[\mathbf{XY}^\top]$ and assume that we use the estimator $\widetilde{\mathbf{C}}_{\mathbf{XY}}=\frac{1}{m}\sum_{i=1}^m\mathbf{X}^{(i)}\mathbf{Y}^{(i)\top}$. Then we have:*

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right]\leq 32\sigma_1\sigma_2\max\left\{\sqrt{\frac{d_1+d_2}{m}},\frac{d_1+d_2}{m}\right\}. \tag{43}$$

Proof. Lemma A.9 with $\epsilon=\frac{1}{4}$ implies that:

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right]\leq 2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)},\mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}}\mathbf{u}^\top(\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}})\mathbf{v}\right] \tag{44}$$

$$=2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)},\mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}}\frac{1}{m}\sum_{i=1}^m\left\{(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)})-\mathbb{E}\left[(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)})\right]\right\}\right] \tag{45}$$

Let

$$Z_{\mathbf{u},\mathbf{v}}=\frac{1}{m}\sum_{i=1}^m\left\{(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)})-\mathbb{E}\left[(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)})\right]\right\}$$

Using similar reasoning to the one used in establishing (40), we conclude that $Z_{\mathbf{u},\mathbf{v}}$ is a $(\frac{5\sigma_1\sigma_2}{\sqrt{m}},\frac{2.5\sigma_1\sigma_2}{m})$–sub–Gamma random variable. Now, we invoke Lemma A.5 to obtain:

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XY}}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right]\leq 2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)},\mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}}Z_{\mathbf{u},\mathbf{v}}\right] \tag{46}$$

$$\leq 10\sigma_1\sigma_2\sqrt{\frac{2\ln\left(|\mathcal{N}_{1/4}^{(d_1)}|.|\mathcal{N}_{1/4}^{(d_2)}|\right)}{m}}+5\sigma_1\sigma_2\frac{\ln\left(|\mathcal{N}_{1/4}^{(d_1)}|.|\mathcal{N}_{1/4}^{(d_2)}|\right)}{m} \tag{47}$$

$$\begin{aligned} &\leq 10\sigma_1\sigma_2\sqrt{\frac{2(d_1+d_2)\ln(9)}{m}} + 5\sigma_1\sigma_2\frac{(d_1+d_2)\ln(9)}{m} && (48)\\ &\leq 32\sigma_1\sigma_2 \max\left\{\sqrt{\frac{d_1+d_2}{m}}, \frac{d_1+d_2}{m}\right\} && (49)\end{aligned}$$

□

Lemma C.3. *Let $\mathbf{A}$ be a $d \times n$ random matrix whose columns $\mathbf{A}_i$ are independent, mean zero, σ^2–sub–Gaussian random vectors. Then,*

$$\|\mathbf{A}\|_{\mathsf{op}} \leq 6\sigma\sqrt{d+n},$$

with probability at least $1 - \exp\big(-2(d+n)\big)$.

Proof. We use Lemma A.9 with $\epsilon = \frac{1}{4}$ and write:

$$\begin{aligned} \mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq t\Big] &\leq \mathbb{P}\Bigg[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(n)}} \mathbf{u}^\top\mathbf{A}\mathbf{v} \geq \frac{t}{2}\Bigg] \\ &\leq \sum_{i=1}^{|\mathcal{N}_{1/4}^{(d)}|}\sum_{j=1}^{|\mathcal{N}_{1/4}^{(n)}|} \mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big], \end{aligned} \tag{50}$$

where we denote the $1/4$–covering points of $\mathcal{S}^{d-1}$ by $\{\mathbf{u}^{(i)}\}_{i=1}^{|\mathcal{N}_{1/4}^{(d)}|}$ and $1/4$–covering points of $\mathcal{S}^{n-1}$ by $\{\mathbf{v}^{(j)}\}_{j=1}^{|\mathcal{N}_{1/4}^{(n)}|}$. We rewrite $\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)}$ as:

$$\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} = \sum_{k=1}^{n} v_k^{(j)}\mathbf{u}^{(i)\top}\mathbf{A}_k, \tag{51}$$

where $v_k^{(j)}$ is the k–th element of $\mathbf{v}^{(j)}$. Therefore, from Definition 5.2 and Lemma A.3, $\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)}$ is a sub–Gaussian random variable with parameter σ. Therefore we write:

$$\mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big] \leq \exp\Big(\frac{-t^2}{8\sigma^2}\Big). \tag{52}$$

We know from (Wainwright, 2019, Lemma 5.7) that:

$$\mathcal{N}_{1/4}^{(d)} \leq 9^d, \qquad \mathcal{N}_{1/4}^{(n)} \leq 9^n. \tag{53}$$

Then from (50), we have:

$$\begin{aligned} \mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq t\Big] &\leq \sum_{i=1}^{|\mathcal{N}_{1/4}^{(d)}|}\sum_{j=1}^{|\mathcal{N}_{1/4}^{(n)}|} \mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big] \\ &\leq 9^{n+d}\exp\Big(\frac{-t^2}{8\sigma^2}\Big). \end{aligned} \tag{54}$$

Setting $t = 6\sigma\sqrt{d+n}$ implies,

$$\begin{aligned} \mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d+n}\Big] &\leq 9^{n+d}\exp\Big(\frac{-9(d+n)}{2}\Big) \\ &\leq \exp\big(-2(n+d)\big). \end{aligned} \tag{55}$$

□

D. Detailed Proof of Theorem 4.4 and Corollary 4.5

Proof. We define events $\mathcal{E}_{1,1}$, $\mathcal{E}_{2,1}$, $\mathcal{E}_{1,2}$, and $\mathcal{E}_{2,2}$ as "Receiving error from Agent 1, when $\left\|\widetilde{\mathbf{C}}_{\mathbf{XX}}\right\|_{\mathsf{op}} > 11\sigma^2$.", "Receiving error from Agent 2, when $\left\|\widetilde{\mathbf{C}}_{\mathbf{YY}}\right\|_{\mathsf{op}} > 11\sigma^2$.", "Receiving error from Agent 1, when $\|\mathbf{X}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}$.", and "Receiving error from Agent 2, when $\|\mathbf{Y}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}$.", respectively. We also define event $\mathcal{E}$ as $\mathcal{E} = \mathcal{E}_{1,1} \vee \mathcal{E}_{2,1} \vee \mathcal{E}_{1,2} \vee \mathcal{E}_{2,2}$. We write:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}}\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right]\mathbb{P}\left[\mathcal{E}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^c\right]\mathbb{P}\left[\mathcal{E}^c\right] \tag{56}$$

We find an upper bound for every term of (56).

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] &= \mathbb{E}\left[\left\|\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] \\ &\leq \sigma^2.\end{aligned} \tag{57}$$

From Lemma C.1, we have:

$$\begin{aligned}\mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XX}}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] &\leq \min\left\{1, \exp\left(6d_1 - m\right)\right\}, \\ \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{YY}}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] &\leq \min\left\{1, \exp\left(6d_2 - m\right)\right\}.\end{aligned}$$

Also Lemma C.3 yields:

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{X}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}\right] &\leq \exp\left(-2(d_1+n)\right), \\ \mathbb{P}\left[\|\mathbf{Y}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}\right] &\leq \exp\left(-2(d_2+n)\right)\end{aligned}$$

Therefore we can upper-bound $\mathbb{P}\left[\mathcal{E}\right]$:

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}\right] &\leq \mathbb{P}[\mathcal{E}_{1,1}] + \mathbb{P}[\mathcal{E}_{2,1}] + \mathbb{P}[\mathcal{E}_{1,2}] + \mathbb{P}[\mathcal{E}_{2,2}] \\ &= \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XX}}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] + \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{YY}}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] \\ &\quad + \mathbb{P}\left[\|\mathbf{X}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}\right] \\ &\quad + \mathbb{P}\left[\|\mathbf{Y}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}\right] \\ &\leq \exp\left(6d_1 - m\right) + \exp\left(6d_2 - m\right) \\ &\quad + \exp\left(-2(d_1+n)\right) + \exp\left(-2(d_2+n)\right) \\ &\leq 2\exp\left(6d - m\right) + 2\exp(-2n).\end{aligned} \tag{58}$$

Note that if $m \geq 9d$, then $2\exp\left(6d-m\right) \leq \frac{1}{4}$. Also if $n \geq 2$, then $2\exp(-2n) \leq \frac{1}{4}$. If both of these constraints are met, we have:

$$\begin{aligned}&\mathbb{P}\left[\mathcal{E}\right] \leq 2\,.\,(9)^d\exp\left(\frac{-m}{100}\right) + 2\exp(-2n) \leq \frac{1}{2} \\ &\mathbb{P}\left[\mathcal{E}^c\right] \geq \frac{1}{2}.\end{aligned} \tag{59}$$

Now we find an upper bound for $\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^c\right]$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^c\right] &\le \frac{\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+-\mathbf{C}\right\|_{\mathsf{op}}\right]}{\mathbb{P}\left[\mathcal{E}^c\right]} \\
&\le 2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+-\mathbf{C}\right\|_{\mathsf{op}}\right] \\
&= 2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+-\widehat{\mathbf{C}}^*+\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right] \\
&\le 2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+-\widehat{\mathbf{C}}^*\right\|_{\mathsf{op}}\right]+2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right] \\
&= 2\,\mathbb{E}\left[\left|\lambda_{\min}(\widehat{\mathbf{C}}^*)\right|\mathbb{1}_{\{\lambda_{\min}(\widehat{\mathbf{C}}^*)<0\}}\right]+2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right] \\
&\le 2\,\mathbb{E}\left[\left|\lambda_{\min}(\widehat{\mathbf{C}}^*)-\lambda_{\min}(\mathbf{C})\right|\mathbb{1}_{\{\lambda_{\min}(\widehat{\mathbf{C}}^*)<0\}}\right] \\
&\quad +2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right] \\
&\overset{\text{(a)}}{\le} 4\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right],
\end{aligned} \tag{60}$$

where (a) is a consequence of Weyl's inequality (Johnson & Horn, 1985, Section 4.3). Then:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{XX}} & \frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^\top \\ \frac{1}{n}\hat{\mathbf{Y}}\hat{\mathbf{X}}^\top & \widehat{\mathbf{C}}_{\mathbf{YY}}\end{bmatrix}-\begin{bmatrix}\mathbf{C}_{\mathbf{XX}} & \mathbf{C}_{\mathbf{XY}} \\ \mathbf{C}_{\mathbf{XY}}^\top & \mathbf{C}_{\mathbf{YY}}\end{bmatrix}\right\|_{\mathsf{op}}\right] \\
&= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}} & \frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^\top-\mathbf{C}_{\mathbf{XY}} \\ \frac{1}{n}\hat{\mathbf{Y}}\hat{\mathbf{X}}^\top-\mathbf{C}_{\mathbf{XY}}^\top & \widehat{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\end{bmatrix}\right\|_{\mathsf{op}}\right] \\
&\le \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right]+\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right]+\mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^\top-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right]
\end{aligned} \tag{61}$$

We use matrix quantization scheme defined in Appendix A.5 to quantize matrices $\widetilde{\mathbf{C}}_{\mathbf{XX}}$, $\widetilde{\mathbf{C}}_{\mathbf{YY}}$, $\mathbf{X}$, and $\mathbf{Y}$. Therefore, we can use the relation between communication load and the resolution of this quantization, which is stated in Appendix A.5.

- Quantization of $\widetilde{\mathbf{C}}_{\mathbf{XX}} \in \mathbb{R}^{d_1\times d_1}$: $r = 11\sigma^2$, therefore:

$$d_1^2\log_2\left(\frac{6\sigma^2}{\epsilon_1'}\right) = B_1' = \frac{B_1}{2} \Rightarrow \epsilon_1' = 33\sigma^2\cdot 2^{\frac{-B_1}{2d_1^2}}. \tag{62}$$

- Quantization of $\widetilde{\mathbf{C}}_{\mathbf{YY}} \in \mathbb{R}^{d_2\times d_2}$: $r = 11\sigma^2$, therefore:

$$d_2^2\log_2\left(\frac{6\sigma^2}{\epsilon_2'}\right) = B_2' = \frac{B_2}{2} \Rightarrow \epsilon_2' = 33\sigma^2\cdot 2^{\frac{-B_2}{2d_2^2}}. \tag{63}$$

- Quantization of $\mathbf{X} \in \mathbb{R}^{d_1\times n}$: $r = 6\sigma\sqrt{d_1+n}$, therefore:

$$nd_1\log_2\left(\frac{18\sigma\sqrt{d_1+n}}{\epsilon_1''}\right) = B_1'' = \frac{B_1}{2} \Rightarrow \epsilon_1'' = 18\sigma\sqrt{d_1+n}\cdot 2^{\frac{-B_1}{2nd_1}}. \tag{64}$$

- Quantization of $\mathbf{Y} \in \mathbb{R}^{d_2\times n}$: $r = 6\sigma\sqrt{d_2+n}$, therefore:

$$nd_2\log_2\left(\frac{18\sigma\sqrt{d_2+n}}{\epsilon_2''}\right) = B_2'' = \frac{B_2}{2} \Rightarrow \epsilon_2'' = 18\sigma\sqrt{d_2+n}\cdot 2^{\frac{-B_2}{2nd_2}}. \tag{65}$$

From Proposition C.2, we have:

$$\begin{aligned} \mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right] &\leq 32\sigma^2\sqrt{\frac{2d_1}{m}}\max\left\{\sqrt{\frac{2d_1}{m}},1\right\}, \\ \mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right] &\leq 32\sigma^2\sqrt{\frac{2d_2}{m}}\max\left\{\sqrt{\frac{2d_2}{m}},1\right\}. \end{aligned} \tag{66}$$

We also have:

$$\begin{aligned} \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{XY}^\top-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] &\leq 32\sigma^2\max\left\{\begin{matrix}\sqrt{\frac{d_1+d_2}{n}}, \\ \frac{d_1+d_2}{n}\end{matrix}\right\} \\ &= 32\sigma^2\sqrt{\frac{d}{n}}\max\left\{\sqrt{\frac{d}{n}},1\right\}. \end{aligned} \tag{67}$$

From (66) and (62) we write:

$$\begin{aligned} \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{XX}}-\widetilde{\mathbf{C}}_{\mathbf{XX}}+\widetilde{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{XX}}-\widetilde{\mathbf{C}}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right]+\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{XX}}-\mathbf{C}_{\mathbf{XX}}\right\|_{\mathsf{op}}\right] \\ &\leq \epsilon_1'+32\sigma^2\sqrt{\frac{2d_1}{m}}\max\left\{\sqrt{\frac{2d_1}{m}},1\right\} \\ &= 33\sigma^2\cdot 2^{\frac{-B_1}{2d_1^2}}+32\sigma^2\sqrt{\frac{2d_1}{m}}\max\left\{\sqrt{\frac{2d_1}{m}},1\right\}. \end{aligned} \tag{68}$$

From (66) and (63) we write:

$$\begin{aligned} \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{YY}}-\widetilde{\mathbf{C}}_{\mathbf{YY}}+\widetilde{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{YY}}-\widetilde{\mathbf{C}}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right]+\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{YY}}-\mathbf{C}_{\mathbf{YY}}\right\|_{\mathsf{op}}\right] \\ &\leq \epsilon_2'+32\sigma^2\sqrt{\frac{2d_2}{m}}\max\left\{\sqrt{\frac{2d_2}{m}},1\right\} \\ &= 33\sigma^2\cdot 2^{\frac{-B_2}{2d_2^2}}+32\sigma^2\sqrt{\frac{2d_2}{m}}\max\left\{\sqrt{\frac{2d_2}{m}},1\right\}. \end{aligned} \tag{69}$$

From (67), (64), and (65) we write:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^{\top}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^{\top}-\frac{1}{n}\hat{\mathbf{X}}\mathbf{Y}^{\top}+\frac{1}{n}\hat{\mathbf{X}}\mathbf{Y}^{\top}-\frac{1}{n}\mathbf{X}\mathbf{Y}^{\top}+\frac{1}{n}\mathbf{X}\mathbf{Y}^{\top}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] \\
&\leq \mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}(\hat{\mathbf{Y}}-\mathbf{Y})^{\top}\right\|_{\mathsf{op}}\right]+\mathbb{E}\left[\left\|\frac{1}{n}(\hat{\mathbf{X}}-\mathbf{X})\mathbf{Y}^{\top}\right\|_{\mathsf{op}}\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}\mathbf{Y}^{\top}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\left\|\hat{\mathbf{X}}\right\|_{\mathsf{op}}\right]\mathbb{E}\left[\left\|\hat{\mathbf{Y}}-\mathbf{Y}\right\|_{\mathsf{op}}\right]+\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{Y}\right\|_{\mathsf{op}}\right]\mathbb{E}\left[\left\|\hat{\mathbf{X}}-\mathbf{X}\right\|_{\mathsf{op}}\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}\mathbf{Y}^{\top}-\mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] \\
&\leq \frac{6\sigma\sqrt{d_1+n}}{n}\epsilon_2''+\frac{6\sigma\sqrt{d_2+n}}{n}\epsilon_1''+32\sigma^2\sqrt{\frac{d}{n}}\max\left\{\sqrt{\frac{d}{n}},1\right\} \\
&= \frac{108\sigma^2\sqrt{(d_1+n)(d_2+n)}}{n}\left(2^{\frac{-B_1}{2nd_1}}+2^{\frac{-B_2}{2nd_2}}\right)+32\sigma^2\sqrt{\frac{d}{n}}\max\left\{\sqrt{\frac{d}{n}},1\right\} \\
&\leq \frac{108\sigma^2\sqrt{(d_1+n)(d_2+n)}}{n}\left(2^{\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}\right)+32\sigma^2\sqrt{\frac{d}{n}}\max\left\{\sqrt{\frac{d}{n}},1\right\}.
\end{aligned}
\tag{70}
$$

If we set:

$$
n=\min\left\{\frac{\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)},m\right\},
\tag{71}
$$

We have:

$$
\begin{aligned}
2^{\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}} &= \exp\left(\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}\ln(2)\right) \\
&= \min\left\{\exp\left(\frac{-1}{2}\ln\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\right),\exp\left(\frac{-1}{2m}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}\ln(2)\right)\right\} \\
&\leq \exp\left(\frac{-1}{2}\ln\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\right) \\
&= \sqrt{\frac{d_1d_2}{\min\{B_1,B_2\}}}.
\end{aligned}
\tag{72}
$$

$$
\begin{aligned}
\frac{d_1+n}{n} &= 1+\frac{d_1}{n} \\
&= 1+d_1\max\left\{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\max\{\frac{d_1}{B_1},\frac{d_2}{B_2}\},\frac{1}{m}\right\} \\
&\leq 1+\max\left\{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\max\{\frac{d_1^2}{B_1},\frac{d_1d_2}{B_2}\},\frac{d}{m}\right\} \\
&= 1+\max\left\{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\frac{d_1^2}{B_1},\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)\frac{d_1d_2}{B_2},\frac{d}{m}\right\} \\
&\leq 1+\max\left\{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right),1\right\} \\
&\leq 2\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right),
\end{aligned}
\tag{73}
$$

Similarly:

$$\frac{d_1+n}{n} \leq 2\log_2\Big(\frac{\min\left\{B_1,B_2\right\}}{d_1 d_2}\Big), \tag{74}$$

Thus:

$$\begin{aligned} \frac{\sqrt{(d_1+n)(d_2+n)}}{n} &= \sqrt{\frac{d_1+n}{n}.\frac{d_2+n}{n}} \\ &\leq 2\log_2\Big(\frac{\min\left\{B_1,B_2\right\}}{d_1 d_2}\Big). \end{aligned} \tag{75}$$

$$\begin{aligned} \sqrt{\frac{d}{n}} &= \max\left\{\sqrt{\frac{d\log_2\big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\big)}{\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}}, \sqrt{\frac{d}{m}}\right\} \\ &\leq \sqrt{d\cdot\max\left\{\frac{d_1}{B_1},\frac{d_2}{B_2}\right\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\Big)} + \sqrt{\frac{d}{m}} \\ &\leq \sqrt{2\max\left\{\frac{d_1^2}{B_1},\frac{d_1 d_2}{B_1},\frac{d_2^2}{B_2},\frac{d_1 d_2}{B_2}\right\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\Big)} + \sqrt{\frac{d}{m}} \\ &= \sqrt{2\max\left\{\frac{d_1^2}{B_1},\frac{d_2^2}{B_2},\frac{d_1 d_2}{\min\{B_1,B_2\}}\right\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\Big)} + \sqrt{\frac{d}{m}}. \end{aligned} \tag{76}$$

Note that $B_1 \geq 15d_1\max\{d_1,d_2\}\log_2(\max\{d_1,d_2\})$, $B_2 \geq 15d_2\max\{d_1,d_2\}\log_2(\max\{d_1,d_2\})$. The function $g(B) = \frac{B}{\log_2(B)}$ is an increasing function in the interval $[e,+\infty)$, Therefore we write:

$$\begin{aligned} \frac{B_1}{\log_2(B_1)} &\geq \frac{B_1}{\log_2(B_1)}\bigg|_{B_1=15d_1\max\{d_1,d_2\}\log_2(\max\{d_1,d_2\}} \\ &= \frac{15d_1\max\{d_1,d_2\}\log_2(\max\{d_1,d_2\}}{\log_2(15)+\log_2(d_1)+\log_2(\max\{d_1,d_2\})+\log_2\big(\log_2(\max\{d_1,d_2\})\big)} \\ &\geq \frac{15d_1\max\{d_1,d_2\}\log_2(\max\{d_1,d_2\}}{\log_2(15)+3\log_2(\max\{d_1,d_2\})} \\ &\geq 2d_1\max\{d_1,d_2\} \\ &\geq d_1(d_1+d_2) \\ &= d_1 d. \end{aligned} \tag{77}$$

This yields $\frac{B_1}{d_1} \geq d\log_2(B_1)$. Similarly we conclude that $\frac{B_2}{d_2} \geq d\log_2(B_2)$. Therefore we have:

$$\begin{aligned} \min\left\{\frac{B_1}{d_1},\frac{B_2}{d_2}\right\} &\geq d\cdot\log_2\big(\min\{B_1,B_2\}\big) \\ &\geq d\cdot\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\Big). \end{aligned} \tag{78}$$

Therefore $\frac{\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}{\log_2\left(\frac{\min\{B_1,B_2\}}{d_1d_2}\right)} \geq d$, also we have $m \geq 9d \geq d$, thus we conclude that $n \geq d$. Now we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^\top - \mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] &\leq \frac{108\sigma^2\sqrt{(d_1+n)(d_2+n)}}{n}\Big(2^{\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}\Big) + 32\sigma^2\sqrt{\frac{d}{n}}\max\left\{\sqrt{\frac{d}{n}},1\right\} \\
&\leq 216\sigma^2\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)\sqrt{\frac{d_1d_2}{\min\{B_1,B_2\}}} \\
&\quad + 32\sigma^2\sqrt{2\max\Big\{\frac{d_1^2}{B_1},\frac{d_2^2}{B_2},\frac{d_1d_2}{\min\{B_1,B_2\}}\Big\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)} \\
&\quad + 32\sigma^2\sqrt{\frac{d}{m}}.
\end{aligned} \tag{79}$$

From assumption $m \geq 9d$ we have:

$$\begin{aligned}
\mathbb{E}\left[\big\|\widehat{\mathbf{C}}^* - \mathbf{C}\big\|_{\mathsf{op}}\right] &\leq \mathbb{E}\left[\big\|\widehat{\mathbf{C}}_{\mathbf{XX}} - \mathbf{C}_{\mathbf{XX}}\big\|_{\mathsf{op}}\right] + \mathbb{E}\left[\big\|\widehat{\mathbf{C}}_{\mathbf{YY}} - \mathbf{C}_{\mathbf{YY}}\big\|_{\mathsf{op}}\right] + \mathbb{E}\left[\left\|\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{Y}}^\top - \mathbf{C}_{\mathbf{XY}}\right\|_{\mathsf{op}}\right] \\
&\leq 33\sigma^2\cdot 2^{\frac{-B_1}{2d_1^2}} + 32\sigma^2\sqrt{\frac{2d_1}{m}}\max\left\{\sqrt{\frac{2d_1}{m}},1\right\} \\
&\quad + 33\sigma^2\cdot 2^{\frac{-B_2}{2d_2^2}} + 32\sigma^2\sqrt{\frac{2d_2}{m}}\max\left\{\sqrt{\frac{2d_2}{m}},1\right\} \\
&\quad + 216\sigma^2\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)\sqrt{\frac{d_1d_2}{\min\{B_1,B_2\}}} + 32\sigma^2\sqrt{\frac{d}{m}} \\
&\quad + 32\sigma^2\sqrt{2\max\Big\{\frac{d_1^2}{B_1},\frac{d_2^2}{B_2},\frac{d_1d_2}{\min\{B_1,B_2\}}\Big\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)} \\
&\leq 96\sigma^2\sqrt{\frac{d}{m}} + 33\sigma^2\cdot 2^{\frac{-1}{2}\min\{\frac{B_1}{d_1^2},\frac{B_2}{d_2^2}\}} \\
&\quad + 216\sigma^2\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)\sqrt{\frac{d_1d_2}{\min\{B_1,B_2\}}} \\
&\quad + 32\sigma^2\sqrt{2\max\Big\{\frac{d_1^2}{B_1},\frac{d_2^2}{B_2},\frac{d_1d_2}{\min\{B_1,B_2\}}\Big\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)}.
\end{aligned} \tag{80}$$

Finally:

$$\begin{aligned}
\mathbb{E}\left[\big\|\widehat{\mathbf{C}} - \mathbf{C}\big\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\big\|\widehat{\mathbf{C}} - \mathbf{C}\big\|_{\mathsf{op}} \mid \mathcal{E}\right]\mathbb{P}\left[\mathcal{E}\right] + \mathbb{E}\left[\big\|\widehat{\mathbf{C}} - \mathbf{C}\big\|_{\mathsf{op}} \mid \mathcal{E}^c\right]\mathbb{P}\left[\mathcal{E}^c\right] \\
&\leq \sigma^2\,\mathbb{P}\left[\mathcal{E}\right] + 4\,\mathbb{E}\left[\big\|\widehat{\mathbf{C}}^* - \mathbf{C}\big\|_{\mathsf{op}}\right] \\
&\leq 2\sigma^2\exp\big(6d-m\big) + 2\sigma^2\exp(-2n) \\
&\quad + 384\sigma^2\sqrt{\frac{d}{m}} + 132\sigma^2\cdot 2^{\frac{-1}{2}\min\{\frac{B_1}{d_1^2},\frac{B_2}{d_2^2}\}} \\
&\quad + 864\sigma^2\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)\sqrt{\frac{d_1d_2}{\min\{B_1,B_2\}}} \\
&\quad + 128\sigma^2\sqrt{2\max\Big\{\frac{d_1^2}{B_1},\frac{d_2^2}{B_2},\frac{d_1d_2}{\min\{B_1,B_2\}}\Big\}\log_2\Big(\frac{\min\{B_1,B_2\}}{d_1d_2}\Big)}
\end{aligned} \tag{81}$$

Note that $m \geq 9d$, therefore $\exp(6d - m) \leq \sqrt{\frac{d}{m}}$. We also have:

$$
\begin{aligned}
\exp(-2n) &= \max\left\{\exp\Big(\frac{-2\min\{\frac{B_1}{d_1}, \frac{B_2}{d_2}\}}{\log_2\big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\big)}\Big), \exp(-2m)\right\} \\
&\leq \sqrt{\max\{\frac{d_1}{B_1}, \frac{d_2}{B_2}\}\log_2\big(\frac{\min\{B_1,B_2\}}{d_1 d_2}\big)} + \sqrt{\frac{d}{m}}.
\end{aligned}
\tag{82}
$$

Thus we have:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] \leq\ & 388\sigma^2\sqrt{\frac{d}{m}} + 132\sigma^2 \cdot 2^{\frac{-1}{2}\min\{\frac{B_1}{d_1^2}, \frac{B_2}{d_2^2}\}} \\
&+ 864\sigma^2 \log_2\Big(\frac{\min\{B_1, B_2\}}{d_1 d_2}\Big)\sqrt{\frac{d_1 d_2}{\min\{B_1, B_2\}}} \\
&+ 184\sigma^2\sqrt{\max\left\{\frac{d_1^2}{B_1}, \frac{d_2^2}{B_2}, \frac{d_1 d_2}{\min\{B_1, B_2\}}\right\}\log_2\Big(\frac{\min\{B_1, B_2\}}{d_1 d_2}\Big)}.
\end{aligned}
\tag{83}
$$

From the assumption $B_1 \geq 15 d_1 \max\{d_1, d_2\}\log_2(\max\{d_1, d_2\})$, and $B_2 \geq 15 d_2 \max\{d_1, d_2\}\log_2(\max\{d_1, d_2\})$, we conclude:

$$
\min\{B_1, B_2\} \geq 15\min\{d_1, d_2\}\max\{d_1, d_2\}\log_2(\max\{d_1, d_2\}) = 15 d_1 d_2 \log_2(\max\{d_1, d_2\}),
\tag{84}
$$

Therefore $\log_2\left(\frac{\min\{B_1,B_2\}}{d_1 d_2}\right) \geq 1$. Thus we can simplify the upper bound more:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] \leq\ & 388\sigma^2\sqrt{\frac{d}{m}} + 132\sigma^2 \cdot 2^{\frac{-1}{2}\min\{\frac{B_1}{d_1^2}, \frac{B_2}{d_2^2}\}} \\
&+ 1048\sigma^2 \log_2\Big(\frac{\min\{B_1, B_2\}}{d_1 d_2}\Big)\sqrt{\max\left\{\frac{d_1^2}{B_1}, \frac{d_2^2}{B_2}, \frac{d_1 d_2}{\min\{B_1, B_2\}}\right\}}.
\end{aligned}
\tag{85}
$$

Assuming $\varepsilon \leq \frac{\sigma^2}{2}$, if we set:

$$
\begin{aligned}
m &= \tau\frac{d\sigma^4}{\varepsilon^2}, \\
B_1 &= \tau'\frac{\sigma^4 d_1 \max\{d_1, d_2\}}{\varepsilon^2}\log_2^2\Big(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\}\Big), \\
B_2 &= \tau'\frac{\sigma^4 d_2 \max\{d_1, d_2\}}{\varepsilon^2}\log_2^2\Big(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\}\Big),
\end{aligned}
\tag{86}
$$

then we have:

$$
\sigma^2\sqrt{\frac{d}{m}} = \frac{\varepsilon}{\sqrt{\tau}}.
\tag{87}
$$

$$
\begin{aligned}
\sigma^2 \cdot 2^{\frac{-1}{2}\min\{\frac{B_1}{d_1^2}, \frac{B_2}{d_2^2}\}} &= \sigma^2 \cdot 2^{\frac{-1}{2}\min\left\{\tau'\frac{\sigma^4\max\{d_1,d_2\}}{d_1\varepsilon^2}\log_2^2(\frac{\sigma^2}{\varepsilon}\max\{d_1,d_2\}), \tau'\frac{\sigma^4\max\{d_1,d_2\}}{d_2\varepsilon^2}\log_2^2(\frac{\sigma^2}{\varepsilon}\max\{d_1,d_2\})\right\}} \\
&\leq \sigma^2 \cdot 2^{\frac{-\tau'\sigma^4}{2\varepsilon^2}\log_2\left(\frac{\sigma^2}{\varepsilon}\max\{d_1,d_2\}\right)} \\
&\leq \sigma^2\Big(\frac{\varepsilon}{\sigma^2\max\{d_1, d_2\}}\Big)^{\frac{\tau'\sigma^4}{\varepsilon^2}} \\
&\leq \varepsilon.
\end{aligned}
\tag{88}
$$

$$
\max\left\{\frac{d_1^2}{B_1}, \frac{d_2^2}{B_2}, \frac{d_1 d_2}{B_{\min}}\right\} \leq \frac{\varepsilon^2}{\tau'\sigma^4\log_2^2(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\})}.
\tag{89}
$$

$$\begin{aligned}\sigma^2 \log_2\left(\frac{B_{\min}}{d_1 d_2}\right)\sqrt{\max\left\{\frac{d_1^2}{B_1}, \frac{d_2^2}{B_2}, \frac{d_1 d_2}{B_{\min}}\right\}} &= \sigma^2 \log_2\left(\frac{\sigma^4 \tau_1}{\varepsilon^2}\log_2^2(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\})\right)\sqrt{\frac{\varepsilon^2}{\tau'\sigma^4 \log_2^2(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\})}} \\ &= \frac{\varepsilon}{\sqrt{\tau'}}\frac{2\log_2\left(\frac{\sigma^2\sqrt{\tau_1}}{\varepsilon}\right) + 2\log_2\left(\log_2(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\})\right)}{\log_2(\frac{\sigma^2}{\varepsilon}\max\{d_1, d_2\})} \\ &\leq \frac{4\varepsilon}{\sqrt{\tau'}}.\end{aligned} \tag{90}$$

Therefore, we can write:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] \leq 388\frac{\varepsilon}{\sqrt{\tau}} + 132\varepsilon + 4192\frac{\varepsilon}{\sqrt{\tau'}}. \tag{91}$$

The proof of Theorem 4.4 and Corollary 4.5 is completed. □