# Fundamental Limits of Distributed Covariance Matrix Estimation via a Conditional Strong Data Processing Inequality

Mohammad Reza Rahmani  Mohammad Hossein Yassaee

Mohammad Reza Aref*

**Abstract**

Estimating high–dimensional covariance matrices is a key task across many fields. This paper explores the theoretical limits of distributed covariance estimation in a feature–split setting, where communication between agents is constrained. Specifically, we study a scenario in which multiple agents each observe different components of i.i.d. samples drawn from a sub–Gaussian random vector. A central server seeks to estimate the complete covariance matrix using a limited number of bits communicated by each agent. We obtain a nearly tight minimax lower bound for covariance matrix estimation under operator norm and Frobenius norm. Our main technical tool is a novel generalization of the strong data processing inequality (SDPI), termed the "*Conditional Strong Data Processing Inequality (C-SDPI) coefficient*", introduced in this work. The C-SDPI coefficient shares key properties—such as tensorization—with the conventional SDPI. Crucially, it quantifies the average contraction in a state-dependent channel and can be significantly lower than the worst-case SDPI coefficient over the state input. Utilizing the doubling trick of Geng–Nair and an operator Jensen inequality, we compute this coefficient for Gaussian mixture channels. We then employ it to establish minimax lower bounds on estimation error, capturing the trade-offs among sample size, communication cost, and data dimensionality. Building on this, we present a nearly optimal estimation protocol whose sample and communication requirements match the lower bounds up to logarithmic factors. Unlike much of the existing literature, our framework does not assume infinite samples or Gaussian distributions, making it broadly applicable. Finally, we extend our analysis to interactive protocols, showing interaction can significantly reduce communication requirements compared to non–interactive schemes.

## Contents



*The authors are with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. Emails: mohammadreza_rahmani@ee.sharif.edu,{yassaee,aref}@sharif.edu

# 1 Introduction

The estimation of the covariance matrix of a random vector from independent and identically distributed (i.i.d.) samples is a cornerstone problem with pervasive applications across diverse quantitative disciplines, including financial mathematics, classical and high–dimensional statistics, and modern machine learning methodologies [Hot33, DKPN00, LW03]. A particularly salient extension of this fundamental problem involves its analysis within distributed computational environments. In such settings, typified by emerging paradigms like federated learning [MMR+17], data are inherently partitioned and dispersed among multiple autonomous agents, with each agent having access to only a local subset of the overall data. These distributed data architectures can be broadly classified into two principal categories: (i) the *sample–split (or horizontal split)* scenario, where each agent possesses a distinct subset of the data samples; and (ii) the *feature–split (or vertical split)* scenario, where each agent has access to a specific subset of the dimensions (or features) for all available samples.

This paper investigates the challenging problem of covariance matrix estimation specifically within a feature–split distributed setting, under the critical constraint of limited communication. Our system model comprises a central server and multiple agents. Each Agent $k$ observes $d_k$ dimensions of $m$ i.i.d. samples of a $d$–dimensional random vector $\mathbf{Z} \in \mathbb{R}^d$. The objective is for the central server to accurately estimate the true covariance matrix $\mathbf{C} = \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]$. A defining characteristic of this problem is the restricted communication budget: each Agent $k$ is permitted to transmit messages of at most $B_k$ bits to the central server. Consequently, the central server must synthesize its estimate of the covariance matrix solely from these bandwidth–constrained messages. This setting prompts two pivotal questions:

1. What are the fundamental limits on estimation accuracy, considering the interplay between the agents' constrained communication budgets and the finite number of available samples?

2. What estimation schemes can effectively achieve these ultimate accuracy limits?

This work provides comprehensive answers to both questions. We establish the information–theoretic lower bounds on the accuracy of distributed covariance estimation, thereby delineating the inherent performance bottlenecks imposed by communication and sample limitations. Building upon these fundamental insights, we derive explicit lower bounds for both the sample complexity and communication complexity that any estimation scheme must satisfy to attain a desired accuracy. Furthermore, we develop and analyze a novel estimation scheme designed to operate effectively under the stipulated communication constraints. We rigorously demonstrate that the proposed scheme's sample and communication complexities are optimal, aligning with the derived lower bounds within a logarithmic factor.

A cornerstone of our theoretical analysis, particularly for deriving lower bounds in estimation problems subjected to communication constraints, is the Strong Data Processing Inequality (SDPI) [AG76]. SDPI refines the classical Data Processing Inequality (DPI), which asserts that mutual information cannot increase through a Markov chain. SDPI quantifies this information degradation using contraction coefficients [AG76]. In this paper, we extend the classical SDPI to encompass state–dependent channels, introducing a novel concept termed the Conditional Strong Data Processing Inequality (C–SDPI) coefficient. This new quantity precisely measures the conditional mutual information loss between the input and output of a channel whose characteristics depend on an external state. We formally introduce C–SDPI, prove several of its key properties, and meticulously compute this coefficient for Gaussian mixture channels. This theoretical tool is then directly applied to derive minimax lower bounds on the estimation error for the distributed covariance matrix estimation problem.

## 1.1 Prior works

This section reviews prior research relevant to our study, broadly categorized into distributed covariance matrix estimation and works pertaining to the Strong Data Processing Inequality.

### 1.1.1 Prior Works on Distributed Covariance Matrix Estimation

For $m$ i.i.d. samples $\{\mathbf{Z}^{(i)}\}_{i=1}^{m} = \{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}\}$ of a random vector $\mathbf{Z}$, the canonical sample covariance estimator is given by:

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

For sub–Gaussian random vectors, bounds on the operator norm of the estimation error for this estimator are well–established [Ver18, Theorem 4.7.1]. Beyond this standard estimator, a rich body of literature addresses covariance matrix estimation under various structural assumptions, such as sparsity [BL$^+$08b, BL08a, EK08], low–rankness [FB07], and Toeplitz–structure [HLPL06, WP09, CGT12], often deviating from the traditional sample covariance approach. The optimality of certain covariance matrix estimators has also been a focus of investigation [CZZ$^+$10, CRZ13].

The extension of fundamental machine learning algorithms to distributed settings, particularly with sample–split data, has been widely explored. Examples include distributed principal component analysis for dimension reduction [QOSG02, BCL05, BKLW14, KVW14], distributed gradient descent algorithms [LLZ09, ZWSL10, NRRW11], and distributed support vector machines [NVG$^+$06, ZWB$^+$07, LRV08, FCG10].

Statistical inference in distributed environments, especially under communication constraints, has garnered significant research interest. Zhang et al. [ZDJW13] provided information–theoretic lower bounds for distributed parameter estimation, demonstrating the minimum communication required for specific performance levels. While not directly focused on covariance estimation, their methodologies offer a general framework. In distributed mean estimation, Suresh et al. [SFKM17] proposed a communication–efficient algorithm that achieves a mean squared error of $\Theta(d/n)$ with constant bits per dimension per client. Subsequent work by Cai and Wei [CW24] characterized minimax convergence rates for Gaussian mean estimation under communication constraints. Beyond estimation, the closely related problem of distributed hypothesis testing has also been explored. Notably, Szabó et al. [SVVZ23] investigate minimax testing errors in a sample–split distributed framework, where limited communication budget to a central machine. They consider both high–dimensional and infinite–dimensional signal detection problems within a Gaussian white noise model, deriving minimax lower bounds on error probabilities and proposing distributed testing algorithms that achieve these theoretical limits.

Further work in distributed estimation includes Braverman et al. [BGM+16], who derived a minimax lower bound for sparse Gaussian vector mean estimation. Han et al. [HÖW18] offered a geometric approach, distinct from SDPI, for analyzing communication budgets in estimation problems, complementing the information–theoretic methods used in [ZDJW13].

In contrast to the sample–split paradigm, the feature–split setting presents unique challenges and arises in applications such as distributed medical databases where different health data dimensions of a patient reside in separate locations [AKB+22], or in environmental monitoring where sensor stations collect partial weather information for correlation analysis without full data centralization. Prior work extending machine learning tasks to this vertical–split setting includes [YFC+19, SZZ+19, HLPS19, HS19, WCX+20].

Within distributed covariance estimation with communication limits, some studies address the horizontal–split case [ZDJW13, BGM+16, HÖW18], while others focus on vertical–split scenarios [HLPS19, HS19]. Specifically, [HLPS19] investigates estimating the correlation $\rho = \mathbb{E}[XY]$ between two *scalar* ($d_1 = d_2 = 1$) *Gaussian* or *binary* random variables $X, Y$ in a vertical–split setup, assuming a one–sided communication constraint ($B_2 = \infty$) and infinite samples ($m = \infty$). For this specific setting, they characterize the exact order of the optimal communication budget for achieving a specific estimation accuracy. [HS19] proposes a solution for estimating the 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. Their proposed solution outperforms the solutions based on estimating the correlation $\mathbb{E}[X_k Y]$, for each $k$, separately. Our research significantly advances this line of inquiry by considering multi–dimensional random vectors, finite sample regimes, and communication constraints on all agents, providing a more general and practical framework.

### 1.1.2 Prior Works on Strong Data Processing Inequality

The Strong Data Processing Inequality (SDPI) is a pivotal tool for our lower bound proofs. It refines the classical Data Processing Inequality (DPI) [Cov99], which states that mutual information cannot increase along a Markov chain. SDPI quantifies this information loss via contraction coefficients [AG76]. Introduced by Ahlswede and Gács in 1976 [AG76], SDPI has revealed connections to other information–theoretic measures, such as maximal correlation [Wit75]. Further theoretical developments and refinements of SDPI properties have been explored in various works [CIR+93, CRS94, Mic97, DMLM03, SVL13]. It has been established that the SDPI constant for any channel is upper–bounded by the Dobrushin contraction coefficient [CIR+93], another established measure of a channel's noise level [Dob56a, Dob56b], a result rediscovered in the machine learning community [BK13]. Relations between SDPI constants for different $f$–divergences have also been investigated [CRS94].

SDPI has found diverse applications, including studying the existence and uniqueness of Gibbs measures, establishing log–Sobolev inequalities, and analyzing performance limits of noisy circuits [ES99, PW25]. In recent years, strong data processing inequalities have gained considerable attention in the information theory community [KA12, AGKN13, Cou13, Rag13, LCV14, PW17, MZ15, CPW15, Rag16]. In [AGKN13], a new geometric characterization of the Hirschfeld–Gebelein–Rényi maximal correlation [Wit75] is provided and its relation to SDPI is studied. In [Cou13], a relation between rate distortion function and SDPI is obtained. In [Rag13], it is shown that the problem of finding SDPI constsnt and input distributions that achieve it can be addressed using so–called logarithmic Sobolev inequalities, which relate input relative entropy to certain measures of input–output correlation. In [CPW15], SDPI for the power–constrained additive Gaussian channel is studied and the amount of decrease of mutual information under convolution with Gaussian noise is bounded.

In many classical techniques for deriving lower bounds in statistical inference—such as the Cramér method, Fano's method, and Assouad's lemma—a key step entails bounding the mutual information between two components of the statistical model. In scenarios subject to structural

constraints, including communication or privacy limitations, these components are often connected through communication channels that impose a specific Markov structure. Consequently, a natural approach to bounding the mutual information is to investigate how it contracts through these channels, typically by employing strong data processing inequalities (SDPIs).

A substantial body of work has examined the use of SDPIs in distributed and sample–splitting estimation contexts. For instance, [GMN14] applied SDPIs to analyze distributed estimation of the mean of a high–dimensional Gaussian distribution. Similarly, [BGM+16] introduced a distributed variant of SDPI to establish nearly tight (up to logarithmic factors) trade–offs between estimation error and communication budget in sparse Gaussian mean estimation. More recently, [CW24] leveraged SDPI to derive sharp lower bounds for mean estimation of multivariate Gaussian distributions under an independent distributed protocol, where each machine communicates with a central server through separate channels without interaction, covering the full range of communication budgets. The work [SVVZ23] further explored mean testing by comparing global and local $\chi^2$ divergences, an approach closely related to SDPI. Beyond these, SDPIs have also been instrumental in deriving lower bounds for other distributed estimation problems [XR15] and in differentially private estimation [DJW13].

The work of [HS19] was the first to apply SDPI in a vertically partitioned setting, utilizing SDPI and its generalization—symmetric SDPI—to establish lower bounds on the estimation of correlation between two scalar random variables in both one–way and interactive protocols. Subsequently, [ST21] employed SDPI to study testing of correlation between a vector and a scalar in a vertical–split context. Our work can be viewed as a generalization of [HS19]; specifically, we use a further generalization of SDPI to derive lower bounds for covariance matrix estimation in vertical–split data settings.

It is important to note that alternative methodologies have also been developed to establish lower bounds under information constraints, in the horizontal–split setting. For example, [ACST23] extended the framework of [ACT20b, ACT20a] to develop information contraction bounds, which can be interpreted as variants of SDPI, and demonstrated their effectiveness in problems such as sparse Gaussian mean estimation. In addition, geometric techniques introduced by [HÖW18], [BHO20], and [BCÖ20] have been successfully applied to derive lower bounds in distributed parameter estimation and density estimation.

### 1.2 Main Contributions

In this paper, we address the problem of estimating the covariance matrix in a vertical–split setting under communication constraints. Our model considers a distributed system with $K$ agents and a central server. Each Agent $k \in [K]$ possesses $d_k$ dimensions of $m$ i.i.d. samples of a $d$–dimensional sub–Gaussian random vector $\mathbf{Z}$. The central server's objective is to estimate the covariance matrix, with each agent $k$ limited to sending messages of at most $B_k$ bits. The central server then forms its estimate from these received messages. We seek to answer two fundamental questions: (1) What is the ultimate estimation accuracy given limited communication and samples? (2) How can this optimal accuracy be achieved?

Our principal contributions are as follows:

- We introduce and rigorously develop a novel theoretical framework termed the *Conditional Strong Data Processing Inequality (C—SDPI)*, and establish its fundamental properties. Furthermore, we derive explicit expressions for the C-–SDPI constant in the setting of Gaussian mixture channels. Our analysis leverages a range of advanced technical tools, including techniques for establishing Gaussian optimality in certain classes of optimization problems [GN14, AJN22], as well as key properties of specific *operator convex functions* [Tro15].

- We formally define the Distributed Covariance Matrix Estimation (DCME) problem and derive a near–optimal trade–off that captures the interplay between the number of samples

$(m)$, communication budgets $(B_1, B_2, \ldots, B_K)$, the data dimensionality available to each agent $(d_1, d_2, \ldots, d_K)$, and the resulting estimation error.

A central contribution of our work is the relaxation of several restrictive assumptions commonly found in prior studies:

- We allow the underlying random vectors $\mathbf{Z}$ to be general sub–Gaussian, thereby extending beyond the conventional Gaussian setting.
- Our analysis explicitly accommodates the practically relevant case of a finite number of i.i.d. samples $(m)$, in contrast to earlier work that often assumes access to an infinite sample regime.

- By leveraging the *C—SDPI*, a variant of Fano's method, and a symmetrization argument, we establish lower bounds on the *minimax* estimation error achievable by any algorithm parameterized by $(m, d_1, d_2, B_1, B_2)$, measured with respect to both the operator norm and the Frobenius norm. In particular, to estimate the covariance matrix up to an error of $\varepsilon$ in the operator norm, the sample complexity and communication budgets must satisfy $m = \Omega\Big(\frac{d}{\varepsilon^2}\Big)$ and $B_k = \Omega\Big(\frac{dd_k}{\varepsilon^2}\Big), \quad k = 1, 2$, where $d = d_1 + d_2$.
- We construct an explicit estimation scheme that approximates the covariance matrix to within $\varepsilon$ error, with sample complexity and communication budgets satisfying $m = \widetilde{\mathcal{O}}\Big(\frac{d}{\varepsilon^2}\Big), \quad B_k = \widetilde{\mathcal{O}}\Big(\frac{dd_k}{\varepsilon^2}\Big), \quad k = 1, 2$. This establishes the near–optimality of our approach, as it matches the minimax lower bounds up to logarithmic factors.
- Using our derived minimax lower bound, we extend it to the general case with $K > 2$ agents. We then present a new scheme that meets this bound, proving that it is achievable.
- We extend our analysis to the interactive setting of the DCME problem, providing both a minimax lower bound and an achievable scheme. We demonstrate that the total communication budget required to estimate the covariance matrix can be reduced from $B := B_1 + B_2 = \widetilde{\Theta}\big(\frac{d^2}{\varepsilon^2}\big)$ in the non–interactive setting to $B = \widetilde{\Theta}\big(\frac{d_1 d_2}{\varepsilon^2}\big)$ in the interactive setting. This reduction can be substantial, particularly when there is significant imbalance between $d_1$ and $d_2$; for example, when $d_1 = 1$ and $d_2 = d-1$. To the best of our knowledge, this improvement is only known in the non–parametric estimation for the vertical–split problems in the literature, cf. [Liu23].

  Furthermore, our results reveal that the statement in [ST21, Theorem 6], which claims that the communication budget must be $\Omega\big(\frac{d^2}{\varepsilon^2}\big)$ in the case $d_1 = 1, d_2 = d-1$, is incorrect.

### 1.3 Paper Structure

The remainder of this paper is organized as follows: Section 2 provides necessary preliminaries. Section 3 introduces the Conditional Strong Data Processing Inequality (C–SDPI) and its properties, serving as the primary theoretical tool. In Section 4, we formally define the Distributed Covariance Matrix Estimation (DCME) problem as an application of C–SDPI, presenting theorems on both error lower bounds and achievable schemes. Section 5 is dedicated to the complete proof of the lower bound theorems. Section 6 details the proposed estimation algorithm and proves the theorems related to its achievable performance. Section 7 defines the interactive DCME problem, presents an achievable scheme for this scenario, and proves its optimality. Finally, Section 8 concludes the paper.

### 1.4 Notations

In this paper, we use specific notation to represent mathematical entities. Matrices are represented by uppercase bold symbols (e.g., $\mathbf{A}$), while vectors are denoted by lowercase bold symbols

(e.g., $\mathbf{v}$). When referring to sets of vectors or matrices formed by concatenating vectors or matrices, we use a bold sans-serif font (e.g., $\mathbf{A}$). For any vector $\mathbf{v} = [v_1, v_2, \cdots, v_d]^\top$, its $\ell_p$–norm is defined 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 is denoted by $\|\mathbf{A}\|_{\mathsf{op}}$, and the Frobenius norm is denoted by $\|\mathbf{A}\|_{\mathsf{F}}$. We also define a general norm, $\|.\|_{\mathsf{dist}}$, which encompasses both the operator and Frobenius norms for flexibility. We use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. Lastly, the notation $[N]$ represents the set of integers $\{1, 2, \ldots, N\}$.

# 2 Preliminaries

## 2.1 Signed Permutation Matrices

**Definition 2.1** (Signed Permutation Matrix)**.** A signed permutation matrix $\mathbf{A}$ is a square matrix with all the entries belong to $\{-1, 0, 1\}$, possessing exactly one *non–zero* entry of $\pm 1$ in each row and each column, with all other entries being 0.

Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ be the standard basis for $\mathbb{R}^d$. A signed permutation matrix $\mathbf{A}$ of size $d$ can be represented by a permutation $\pi$ on $\{1, \cdots, d\}$ and a signed vector $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_d]^\top$, where $\sigma_i \in \{-1, 1\}$, as follows:

$$\mathbf{A} = \sum_{i=1}^{d} \sigma_i \mathbf{e}_i \mathbf{e}_{\pi(i)}^\top. \tag{2}$$

Each signed permutation matrix $\mathbf{A}$ is an orthogonal matrix, satisfying $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$. Let $\mathcal{P}_d$ be the set of all signed permutation matrices of size $d$. The representation (2) implies that $|\mathcal{P}_d| = 2^d d!$. Furthermore, the signed permutation matrices in $\mathcal{P}_d$ form a group under *matrix multiplication*.

The subsequent lemma is crucial for simplifying certain computations within this work.

**Lemma 2.2.** *Let $\mathbf{A}$ be a random matrix drawn uniformly from the set of signed permutation matrices $\mathcal{P}_d$. Then, for any fixed matrix $\mathbf{B}$, the following holds:*

$$\mathbb{E}\left[\mathbf{A}^\top \mathbf{B} \mathbf{A}\right] = \frac{\mathsf{Tr}\{\mathbf{B}\}}{d} \mathbf{I}_d. \tag{3}$$

The proof of Lemma 2.2 is stated in Appendix A.1.

## 2.2 Sub–Gaussian Random Variables

Sub–Gaussian random variables, formally defined in Definition 2.3, constitute a family of random variables whose tail behavior exhibits faster decay than that of a Gaussian distribution.

**Definition 2.3** (Sub–Gaussian Random Variable [Wai19, Definition 2.2])**.** A random variable $X$ is defined as $\sigma$–sub–Gaussian if it satisfies:

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

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

**Definition 2.4** (Sub–Gaussian Random Vector [Wai19, Section 6.3])**.** A random vector $\mathbf{X} \in \mathbb{R}^d$ is called sub–Gaussian with parameter $\sigma$ if for all $\mathbf{v} \in \mathbb{S}^{d-1}$, $\mathbf{v}^\top \mathbf{X}$ is a $\sigma$–sub–Gaussian random variable, where $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ represents the $d$–dimensional unit sphere.

Several properties of sub–Gaussian random variables are detailed in Appendix A.3.

### 2.3 Packing and Covering Numbers

Packing and covering numbers are two widely used notions for quantifying the "size" or complexity of a subset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ within a metric space $\mathcal{K}$. We present their formal definitions below:

**Definition 2.5** (Covering Number [Wai19, Definition 5.1])**.** A set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subseteq \mathcal{K}$ is defined as an $\epsilon$–covering set with respect to a metric $\mathsf{d}$ if for all $x \in \mathcal{K}$, there exists some $j \in [N]$ such that $\mathsf{d}(\mathbf{x}, \mathbf{x}_j) \leq \epsilon$. The covering number $\mathcal{N}(\mathcal{K}, \mathsf{d}, \epsilon)$ is defined as the cardinality of the smallest $\epsilon$–covering set of set $\mathcal{K}$, with respect to the metric $\mathsf{d}$.

**Definition 2.6** (Packing Number [Wai19, Definition 5.4])**.** A set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\} \subseteq \mathcal{K}$ is called a $\epsilon$–packing set with respect to a metric $\mathsf{d}$ if $\mathsf{d}(\mathbf{x}_i, \mathbf{x}_j) > \epsilon$ for all distinct $i, j \in [M]$. The packing number $\mathcal{M}(\mathcal{K}, \mathsf{d}, \epsilon)$ defined as the cardinality of the largest $\epsilon$–packing set of set $\mathcal{K}$, with respect to the metric $\mathsf{d}$.

## 3 Conditional Strong Data Processing Inequality

The Data Processing Inequality is a well-established result. However, for the sake of completeness, we present its definition here:

**Theorem 3.1** (Data Processing Inequality [PW25, Theorem 2.15])**.** *Let $P_Y = T_{Y|X} \circ P_X$ and $Q_Y = T_{Y|X} \circ Q_X$, Then:*

$$D_{\mathrm{KL}}(Q_Y \| P_Y) \leq D_{\mathrm{KL}}(Q_X \| P_X).$$

*Strong Data Processing Inequality* constitutes a refinement of the Data Processing Inequality.

**Definition 3.2** (Strong Data Processing Inequality (SDPI) or Contraction Coefficient [PW25, Definition 33.10])**.** For any input distribution $P_X$ and Markov kernel $T_{Y|X}$, we define the contraction coefficient as:

$$s(P_X, T_{Y|X}) := \sup_{Q_X : 0 < D_{\mathrm{KL}}(Q_X \| P_X) < \infty} \frac{D_{\mathrm{KL}}(Q_Y \| P_Y)}{D_{\mathrm{KL}}(Q_X \| P_X)},$$

where $P_Y = T_{Y|X} \circ P_X$ and $Q_Y = T_{Y|X} \circ Q_X$. If either $X$ or $Y$ is a constant, we define $s(P_X, T_{Y|X})$ to be 0.

Furthermore SDPI coefficient can be equivalently represented using the following mutual information characterization [Rag16, Theorem V.2]:

$$s(P_X, T_{Y|X}) := \sup_{P_{U|X} : U \multimap X \multimap Y} \frac{I(U;Y)}{I(U;X)}.$$

We now present a generalization of the Strong Data Processing Inequality for state-dependent channels.

**Definition 3.3** (Conditional Strong Data Processing (C–SDPI) Coefficient)**.** For any input distribution $P_X$, *state* distribution $P_V$ and Markov kernel $T_{Y|X,V}$, we define the Conditional Strong Data Processing Inequality (C–SDPI) coefficient as:

$$\mathsf{s}(P_X, T_{Y|X,V} \mid P_V) := \sup_{Q_X : 0 < D_{\mathrm{KL}}(Q_X \| P_X) < \infty} \frac{D_{\mathrm{KL}}(Q_{Y|V} \| P_{Y|V} \mid P_V)}{D_{\mathrm{KL}}(Q_X \| P_X)},$$

where $P_{Y|V} = T_{Y|X,V} \circ P_X$ and $Q_{Y|V} = T_{Y|X,V} \circ Q_X$. If either $X$ or $Y$ is a constant, we define $\mathsf{s}(P_X, T_{Y|X,V} \mid P_V)$ to be 0.

It is readily apparent that:

$$\mathsf{s}(P_X, T_{Y|X,V}|P_V) \le \sup_v \left\{ s(P_X, T_{Y|X,V=v}) \right\}. \tag{4}$$

Also a more careful examination of the definition implies:

$$\mathsf{s}(P_X, T_{Y|X,V}|P_V) \le \mathbb{E}_{\bar{V}\sim P_V}\left[ s(P_X, T_{Y|X,V=\bar{V}}) \right]. \tag{5}$$

**Remark 3.4.** The C–SDPI coefficient $\mathsf{s}(P_X, T_{Y|X,V} \mid P_V)$ can be interpreted as SDPI coefficient of the Markov kernel $\widetilde{T}_{Y,V|X} = P_V T_{Y|X,V}$, that is:

$$\mathsf{s}(P_X, T_{Y|X,V} \mid P_V) = s(P_X, \widetilde{T}_{Y,V|X}).$$

In particular, this implies the following mutual information characterization holds for the C–SDPI coefficient:

$$\mathsf{s}(P_X, T_{Y|X,V} \mid P_V) = \sup_{\substack{P_{U|X}: U \multimap (X,V) \multimap Y \\ (U,X) \perp\!\!\!\perp V}} \frac{I(U;Y|V)}{I(U;X)}.$$

The reason we distinguish between the SDPI and C–SDPI is the following *conditional* tensorization property:

**Theorem 3.5** (Tensorization of C–SDPI coefficient)**.** *For any independent input distributions $P_{X_1}, P_{X_2}, P_V$ and Markov kernels $P_{Y_1|X_1,V}$ and $P_{Y_2|X_2,V}$ conditioned on the same random variable $V$, the following equality holds:*

$$\mathsf{s}\left(P_{X_1}P_{X_2}, T_{Y_1|X_1V}T_{Y_2|X_2,V} \mid P_V\right) = \max_{k\in\{1,2\}} \left\{ \mathsf{s}(P_{X_k}, T_{Y_k|X_k,V}|P_V) \right\}. \tag{6}$$

*In particular:*

$$\mathsf{s}(P_X^{\otimes n}, T_{Y|X,V}^{\otimes n} \mid P_V) = \mathsf{s}(P_X, T_{Y|X,V} \mid P_V), \tag{7}$$

*where we have employed the abuse notation $T_{Y|X,V}^{\otimes n}(y^n|x^n, v) := \prod_{i=1}^n T_{Y|XV}(y_i|x_i, v)$.*

**Remark 3.6.** Using the tensorization property of the SDPI [PW25, Proposition 33.11], we obtain:

$$\mathsf{s}(P_X^{\otimes n}, \widetilde{T}_{Y|X,V}^{\otimes n}|P_V^{\otimes n}) = s(P_X^{\otimes n}, \widetilde{T}_{Y,V|X}^{\otimes n}) = s(P_X, \widetilde{T}_{Y,V|X}) = \mathsf{s}(P_X, T_{Y|X,V} \mid P_V)$$

Here, the product channel is defined as: $\widetilde{T}_{Y|X,V}^{\otimes n}(y^n|x^n, v^n) = \prod_{i=1}^n \widetilde{T}_{Y|X,V}(y_i|x_i, v_i)$. This result contrasts with the conditional tensorization property (Theorem 3.5): in the unconditional setting, the channel state varies across instances, whereas in the conditional setting, it remains fixed.

The subsequent identity is crucial for the proof of Theorem 3.5 and for establishing Gaussian optimality in the subsequent section.

**Lemma 3.7.** *For any joint distributions $Q_{X_1,X_2}$, product channel $T_{Y_1|X_1,V}T_{Y_2|X_2,V}$, distribution $P_V$ and arbitrary conditional distributions $P_{Y_k|V}$ for $k = 1, 2$, the following identity holds:*

$$\begin{aligned} D_{\mathrm{KL}}(Q_{Y_1,Y_2|V}\|P_{Y_1|V}P_{Y_2|V}|P_V) = D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2|X_1,V}\|P_{Y_2|V}|Q_{X_1,V}) \\ - I_Q(X_1; Y_2|Y_1, V), \end{aligned} \tag{8}$$

*where $Q_{Y_1,Y_2|V}$ and the mutual information are computed with respect to the following joint distribution:*

$$Q_{V,X_1,X_2,Y_1,Y_2} = P_V Q_{X_1,X_2} T_{Y_1|X_1,V} T_{Y_2|X_2,V}.$$

*Proof.* The proof is stated in Appendix B. □

We now proceed with the proof of Theorem 3.5.

*Proof of Theorem 3.5.* For brevity, Let $\mathsf{s}_k = \mathsf{s}(P_{X_k}, T_{Y_k|X_k,V}|P_V)$ and $\mathsf{s} = \max\{\mathsf{s}_1, \mathsf{s}_2\}$. Observing that, given the product input $P_{X_1X_2} = P_{X_1}P_{X_2}$, we have $P_{Y_1,Y_2|V} = P_{Y_1|V}P_{Y_2|V}$, we invoke Lemma 3.7 to obtain:

$$\begin{aligned} D_{\mathrm{KL}}(Q_{Y_1,Y_2|V}\|P_{Y_1|V}P_{Y_2|V}|P_V) &\overset{(a)}{\leq} D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2|X_1,V}\|P_{Y_2|V}|Q_{X_1,V}) \\ &\overset{(b)}{\leq} \mathsf{s}_1 D_{\mathrm{KL}}(Q_{X_1}\|P_{X_1}) + \mathbb{E}_{Q_{X_1}}\left[D_{\mathrm{KL}}(Q_{Y_2|X_1,V}\|P_{Y_2|V}|P_V)\right] \\ &\overset{(c)}{\leq} \mathsf{s}_1 D_{\mathrm{KL}}(Q_{X_1}\|P_{X_1}) + \mathbb{E}_{Q_{X_1}}\left[\mathsf{s}_2 D_{\mathrm{KL}}(Q_{X_2|X_1}\|P_{X_2})\right] \\ &\overset{(d)}{\leq} \mathsf{s} D_{\mathrm{KL}}(Q_{X_1,X_2}\|P_{X_1}P_{X_2}), \end{aligned} \tag{9}$$

where (a) follows from Lemma 3.7, (b) follows from the definition of C–SDPI coefficient, (c) also follows from the definition of the C–SDPI coefficient by noting that for a fixed $X_1 = x_1$, the output distribution associated with $Q_{X_2|X_1=x_1}$ is $Q_{Y_2|V,X_1=x_1}$ and (d) follows from the definition of $\mathsf{s}$ and the chain rule for KL–divergence.

Thus, we have established that $\mathsf{s}\left(P_{X_1}P_{X_2}, T_{Y_1|X_1V}T_{Y_2|X_2,V} \mid P_V\right) \leq \max_{k\in\{1,2\}}\left\{\mathsf{s}(P_{X_k}, T_{Y_k|X_k,V}|P_V)\right\}$. The converse inequality, $\mathsf{s}\left(P_{X_1}P_{X_2}, T_{Y_1|X_1V}T_{Y_2|X_2,V} \mid P_V\right) \geq \max_{k\in\{1,2\}}\mathsf{s}(P_{X_k}, T_{Y_k|X_k,V}|P_V)$ is readily follows from the definition. □

### 3.1 SDPI Coefficient for Normal Distribution

In [KGK+17], the SDPI coefficient is derived for multivariate normal distribution.

**Lemma 3.8** ([KGK+17, Section 2.6]). *If* $\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} \sim \mathcal{N}\Big(\boldsymbol{\mu}, \mathbf{C} = \begin{bmatrix}\mathbf{C_{XX}} & \mathbf{C_{XY}}\\ \mathbf{C_{XY}}^\top & \mathbf{C_{YY}}\end{bmatrix}\Big)$, *then the following holds:*

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

### 3.2 C–SDPI Coefficient for (Mixture) Normal Distribution

Consider a joint distribution of $(\mathbf{X}, \mathbf{Y}, V)$ such that the conditional distribution of $(\mathbf{X}, \mathbf{Y})$ given $V = v$ is a normal distribution, and $(\mathbf{X}, V)$ are mutually independent. Consequently, $\mathbf{X}$ is normally distributed, and $(\mathbf{X}, \mathbf{Y})$ follows a mixture normal distribution. More specifically, consider the mixture normal distribution defined by the following conditional normal distribution:

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} |\{V = v\} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{I}_{d_\mathbf{X}} & \mathbf{A}_v^\top\\ \mathbf{A}_v & \mathbf{I}_{d_\mathbf{Y}}\end{bmatrix}\right), \tag{10}$$

where $\mathbf{A}_v$ denotes the *cross* covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, conditioned on $V = v$. It is also assumed that $\mathbf{A}_v\mathbf{A}_v^\top \preceq \mathbf{I}_{d_Y}$. The conditional relationship between $\mathbf{X}$ and $\mathbf{Y}$ given $V = v$ can be expressed by the following Gaussian channel:

$$\mathbf{Y} = \mathbf{A}_v\mathbf{X} + \mathbf{Z}_v, \tag{11}$$

where $\mathbf{Z}_v \sim \mathcal{N}(0, \mathbf{I}_{d_Y} - \mathbf{A}_v\mathbf{A}_v^\top)$ is independent of $\mathbf{X}$. Let $P_\mathbf{X}$ denotes the distribution of $\mathbf{X}$, $P_V$ denotes the distribution of $V$, and $T_{\mathbf{Y}|\mathbf{X},V}$ represents the Markov kernel (channel) corresponding to (11).

We are interested in determining the conditional strong data processing coefficient for this joint distribution. A straightforward upper bound, derived using (5), is given by:

$$\mathsf{s}(P_{\mathbf{X}}, T_{\mathbf{Y}|\mathbf{X},V}|P_V) \leq \mathbb{E}_{\bar{V}\sim P_V}\left[s(P_X, T_{Y|X,V=\bar{V}})\right] = \mathbb{E}_V\left[\|\mathbf{A}_v^\top \mathbf{A}_v\|_{\mathsf{op}}\right], \tag{12}$$

where the final equality follows from Lemma 3.8.

To establish a lower bound, let $Q_{\mathbf{X}} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_{d_X})$. Then, given $V = v$, and using (11), we have $Q_{\mathbf{Y}|V=v} = \mathcal{N}(\mathbf{A}_v\boldsymbol{\mu}, \mathbf{I}_{d_Y})$. For all $\mathsf{s} \geq 0$ we have:

$$\begin{aligned}\mathsf{s}D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}}) - D_{\mathrm{KL}}(Q_{\mathbf{Y}|V}\|P_{\mathbf{Y}|V}|P_V) &= \frac{1}{2}\left(\mathsf{s}\|\boldsymbol{\mu}\|^2 - \mathbb{E}_V\left[\|\mathbf{A}_V\boldsymbol{\mu}\|^2\right]\right) \\ &= \frac{1}{2}\boldsymbol{\mu}^\top\left(\mathsf{s}\mathbf{I}_{d_X} - \mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\right)\boldsymbol{\mu}.\end{aligned} \tag{13}$$

Now for $\mathsf{s} < \|\,\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}}$, the infimum of the above expression over all $\boldsymbol{\mu}$ is $-\infty$, Consequently, the C–SDPI coefficient admits the following lower bound:

$$\mathsf{s}(P_{\mathbf{X}}, T_{\mathbf{Y}|\mathbf{X},V}|P_V) \geq \|\,\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}}. \tag{14}$$

As will be demonstrated, this lower bound is indeed the C–SDPI coefficient. Before formally stating this result, we compare the upper bound (12) and the lower bound (14) through a simple example.

*Example.* Let $d_{\mathbf{X}} = d$ and $d_Y = 1$. Assume that $V$ is drawn uniformly from the set $\{1, \ldots, d\}$. Furthermore, let $\mathbf{A}_v = \mathbf{e}_v^\top$, where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ constitutes the standard basis for $\mathbb{R}^d$. In this instance, it is straightforward to verify that $\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V] = \frac{1}{d}\mathbf{I}_d$. Moreover, $\|\mathbf{A}_v^\top\mathbf{A}_v\|_{\mathsf{op}} = 1$. Consequently, (12) and (14) yield the bounds $\frac{1}{d} \leq \mathsf{s}(P_{\mathbf{X}}, T_{Y|\mathbf{X},V}|P_V) \leq 1$. This demonstrates that the upper and lower bounds can exhibit a significant difference.

**Theorem 3.9.** *The conditional strong data processing coefficient for the mixture normal distribution defined in* (10) *is given by:*

$$\mathsf{s}(P_{\mathbf{X}}, T_{\mathbf{Y}|\mathbf{X},V}|P_V) = \left\|\,\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\right\|_{\mathsf{op}}. \tag{15}$$

We highlight the principal steps of the proof, deferring the detailed derivations to the Supplementary Material.

*Proof.* Let $\mathsf{s}^\star = \|\,\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}}$. Define the following functional:

$$\mathsf{g}(Q_{\mathbf{X}}) := \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}}) - D_{\mathrm{KL}}(Q_{\mathbf{Y}|V}\|P_{\mathbf{Y}|V}|P_V), \tag{16}$$

where $P_{\mathbf{X}} = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{X}}})$ and $P_{\mathbf{Y}|V} = P_{\mathbf{Y}} = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{Y}}})$. It was established in (14) that $\mathsf{s}^\star$ constitutes a lower bound for $\mathsf{s}(P_{\mathbf{X}}, T_{\mathbf{Y}|\mathbf{X},V}|P_V)$. Thus, to proceed, it suffices to prove that:

$$\inf_{Q_{\mathbf{X}}: D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty} \mathsf{g}(Q_{\mathbf{X}}) = 0. \tag{17}$$

The key step is to show that the optimization problem (17) can be restricted to the family of centered normal distributions. More precisely, we have:

**Lemma 3.10.** *Let $\mathcal{F}_{\mathsf{G}}$ denote the set of centered normal distributions $Q_{\mathbf{X}}$ with the property that $D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})$ is finite. Then:*

$$\inf_{Q_{\mathbf{X}}: D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty} \mathsf{g}(Q_{\mathbf{X}}) = \inf_{Q_{\mathbf{X}}\in\mathcal{F}_{\mathsf{G}}} \mathsf{g}(Q_{\mathbf{X}}). \tag{18}$$

The proof scheme closely follows the approach proposed by Geng and Nair [GN14] and subsequently employed by Anantharam et al. [AJN22] for similar functionals. Specifically, we demonstrate that the optimizer must satisfy a certain rotational invariance, thereby implying that it is a normal distribution. The proof of Lemma 3.10 is provided in Supplementary Material C.

Based on Lemma 3.10, it suffices to prove:

$$\inf_{Q_{\mathbf{X}} \in \mathcal{F}_{\mathsf{G}}} \mathsf{g}(Q_{\mathbf{X}}) = 0 \tag{19}$$

We observe that $\mathsf{g}(P_{\mathbf{X}}) = 0$. Thus, it remains to show that $\mathsf{g}(Q_{\mathbf{X}}) \geq 0$ for any $Q_{\mathbf{X}} \neq P_{\mathbf{X}}$. Our proof relies on the concept of operator convexity [Tro15] applied to a specific function. In particular, we utilize the following lemma:

**Lemma 3.11.** *1. [Tro15, Proposition 8.4.8] The logarithm function $f(\mathbf{B}) = \log \mathbf{B}$ is an operator concave function on the set of positive definite matrices.*

*2. (**Operator Jensen inequality**, [Tro15, Theorem 8.5.2]) Let $f$ be an operator concave function defined on the set of positive definite matrices. Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be two (rectangular) matrices that decompose the identity matrix as follows:*

$$\mathbf{K}_1\mathbf{K}_1^\top + \mathbf{K}_2\mathbf{K}_2^\top = \mathbf{I} \tag{20}$$

*Then, for any two positive definite matrices $\mathbf{D}_1$ and $\mathbf{D}_2$, the following Jensen-type inequality holds:*

$$f\left(\mathbf{K}_1\mathbf{D}_1\mathbf{K}_1^\top + \mathbf{K}_2\mathbf{D}_2\mathbf{K}_2^\top\right) \succeq \mathbf{K}_1 f(\mathbf{D}_1)\mathbf{K}_1^\top + \mathbf{K}_2 f(\mathbf{D}_2)\mathbf{K}_2^\top \tag{21}$$

We now proceed to complete the proof.

Let $Q_{\mathbf{X}} = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{X}}} + \mathbf{B})$, where $\mathbf{B} \succ -\mathbf{I}$. Algebraic manipulation reveals that $Q_{\mathbf{Y}|V} = \mathcal{N}(0, \mathbf{I} + \mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top)$. Applying the KL divergence formula for Gaussian distributions [PW25, Example 2.2], we obtain:

$$D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) = \frac{1}{2}\mathsf{Tr}\left\{\mathbf{B} - \log(\mathbf{I} + \mathbf{B})\right\}, \tag{22}$$

$$D_{\mathrm{KL}}(Q_{\mathbf{Y}|V} \| P_{\mathbf{Y}}|P_V) = \frac{1}{2}\,\mathbb{E}_V\left[\mathsf{Tr}\left\{\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top - \log(\mathbf{I} + \mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top)\right\}\right], \tag{23}$$

where $\log(\mathbf{A})$ denotes the matrix logarithm of $\mathbf{A}$. Applying (21) to the logarithm function with $\mathbf{D}_1 = \mathbf{I} + \mathbf{B}$, $\mathbf{D}_2 = \mathbf{I}$, $\mathbf{K}_1 = \mathbf{A}_v$ and an appropriate $\mathbf{K}_2$ such that (20) is satisfied (such $\mathbf{K}_2$ exists, because $\mathbf{A}_v\mathbf{A}_v^\top \preceq \mathbf{I}_{d_Y}$ for any $v$) yields:

$$\log(\mathbf{I} + \mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top) \succeq \mathbf{A}_v \log(\mathbf{I} + \mathbf{B})\mathbf{A}_v^\top.$$

This implies:

$$\mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top - \log(\mathbf{I} + \mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top) \preceq \mathbf{A}_v(\mathbf{B} - \log(\mathbf{I} + \mathbf{B}))\mathbf{A}_v^\top.$$

Taking the trace on both sides implies:

$$\mathsf{Tr}\left\{\mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top - \log(\mathbf{I} + \mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top)\right\} \leq \mathsf{Tr}\left\{\mathbf{A}_v^\top\mathbf{A}_v(\mathbf{B} - \log(\mathbf{I} + \mathbf{B}))\right\}.$$

Consequently, substituting this into (23) results in:

$$D_{\mathrm{KL}}(Q_{\mathbf{Y}|V} \| P_{\mathbf{Y}}|P_V) \leq \frac{1}{2}\mathsf{Tr}\left\{\mathbb{E}\left[\mathbf{A}_V^\top\mathbf{A}_V\right](\mathbf{B} - \log(\mathbf{I} + \mathbf{B}))\right\} \tag{24}$$

$$\leq \frac{\mathsf{s}^\star}{2}\mathsf{Tr}\left\{\mathbf{B} - \log(\mathbf{I} + \mathbf{B})\right\} \tag{25}$$

$$= \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}), \tag{26}$$

where (25) follows from the fact that $\mathbf{B} - \log(\mathbf{I} + \mathbf{B})$ is always positive semi-definite, and the inequality $\mathsf{Tr}\{\mathbf{A}\mathbf{B}\} \leq \|\mathbf{A}\|_{\mathsf{op}}\,\mathsf{Tr}\{\mathbf{B}\}$ for any pair $\mathbf{A}, \mathbf{B} \succeq \mathbf{0}$ [VN37]. This concludes the proof of Theorem 3.9. □

### 3.3 SDPI coefficient of a Gaussian mixture channel

Let $V$ be a latent random variable with distribution $P_V$ taking values in some index set $\mathcal{V}$, and for each $v \in \mathcal{V}$ let:

$$T_{\mathbf{Y}|\mathbf{X},V=v} \;=\; \mathcal{N}\big(\mathbf{A}_v\mathbf{X},\; \mathbf{I} - \mathbf{A}_v\mathbf{A}_v^\top\big),$$

so that the joint channel is $T_{\mathbf{Y},V|\mathbf{X}}(\mathbf{y}, v|\mathbf{x}) = P_V(v)\, T_{\mathbf{Y}|\mathbf{X},V=v}(\mathbf{y}|\mathbf{x})$. By marginalizing out $V$, one obtains the *Gaussian mixture channel*:

$$\widetilde{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \;=\; \int_{\mathcal{V}} \mathcal{N}\big(\mathbf{A}_v\mathbf{x},\; \mathbf{I} - \mathbf{A}_v\mathbf{A}_v^\top\big)\, dP_V(v). \tag{27}$$

Following Remark 3.4 and leveraging the monotonicity of the SDPI coefficient [PW25], we obtain the inequality $s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) \le \big\| \mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\big\|_{\mathsf{op}}$. We show that this upper bound for the SDPI of the Gaussian mixture channel is *indeed* tight.

**Proposition 3.12.** *The SDPI coefficient of the Gaussian mixture channel $\widetilde{T}_{\mathbf{Y}|\mathbf{X}}$ defined in (27) is given by:*

$$s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) = \| \mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}} \tag{28}$$

The proof of Proposition 3.12 is stated in Appendix D.

## 4 Application: Distributed Covariance Matrix Estimation (DCME)

In this section, as an application of the conditional SDPI, we introduce the DCME problem and provide a solution.

### 4.1 Problem formulation

We consider a system including a central server and $K$ agents, as depicted in Fig. 1. The central server aims to estimate the covariance matrix $\mathbf{C} = \mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top]$ of a $d$–dimensional $\sigma$–sub–Gaussian random vector $\mathbf{Z} \sim P$ (see Definition 2.4), based on $m$ independent and identically distributed (i.i.d.) samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}$. However, the central server lacks direct access to these samples. Instead, each Agent $k$, for all $k \in [K]$, possesses complete knowledge of certain dimensions of all $m$ samples. Specifically, Agent $k$ has access to the dimensions $[1 + \sum_{i=1}^{k-1} d_i : \sum_{i=1}^{k} d_i]$ of each data vector. The information accessible to Agent $k$ can be represented by $\{\mathbf{X}_k^{(i)}\}_{i=1}^m = \{\mathbf{Z}^{(i)}_{[1+\sum_{i=1}^{k-1} d_i : \sum_{i=1}^{k} d_i]}\}_{i=1}^m$. The central server's objective is to estimate $\mathbf{C}$ by receiving up to $B_k$ bits from Agent $k$.

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

- $K$ encoder functions $\mathcal{E}_k : \mathbb{R}^{d_k \times m} \mapsto [1 : 2^{B_k}]$ where encoder $k$ maps $\{\mathbf{X}_k^{(i)}\}_{i=1}^m$ to $M_k = \mathcal{E}_k(\{\mathbf{X}_k^{(i)}\}_{i=1}^m)$.
- A decoder function $\mathcal{D} : [1 : 2^{B_1}] \times [1 : 2^{B_2}] \times \cdots \times [1 : 2^{B_K}] \mapsto \mathsf{S}_+^{d\times d}$, where $\mathsf{S}_+^{d\times d}$ represents the set of positive semi–definite matrices of dimension $d \times d$. The decoder function maps $(M_1, M_2, \ldots, M_K)$ to $\widehat{\mathbf{C}} = \mathcal{D}(M_1, M_2, \ldots, M_K)$.

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 $\mathbf{C}$; that is, $\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C}) = \big\|\widehat{\mathbf{C}} - \mathbf{C}\big\|_{\mathsf{dist}}$. The expected distortion of a DCME scheme is given by:

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \;=\; \mathbb{E}_{\{\mathbf{Z}^{(i)}\}_{i=1}^m \sim P^{\otimes m}}\left[\big\|\widehat{\mathbf{C}} - \mathbf{C}\big\|_{\mathsf{dist}}\right]. \tag{29}$$

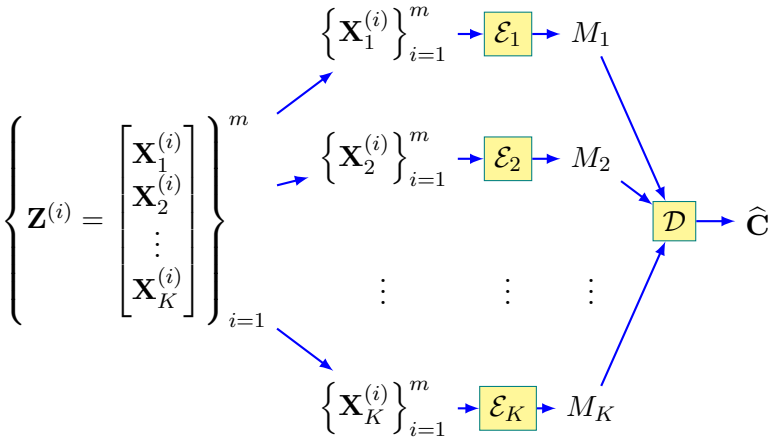


Figure 1: Setting of the problem $\mathsf{DCME}(\sigma, m, B_{1:K}, d_{1:K})$. $\mathbf{Z} \in \mathbb{R}^d$ is a $\sigma$–sub–Gaussian random vector with covariance matrix $\mathbf{C}$. for each $k \in [K]$, $\mathbf{X}_k$ contains the dimensions $[1 + \sum_{i=1}^{k-1} d_i : \sum_{i=1}^{k} d_i]$ of $\mathbf{Z}$. The $\widehat{\mathbf{C}}$ is an estimate of $\mathbf{C}$, subject to the constraint that $H(M_k) \leq B_k$.

The objective is to design the encoding functions $\{\mathcal{E}_k\}_{k=1}^K$ and the decoding function $\mathcal{D}$, to minimize the expected distortion (29) in the worst–case scenario. In other words, we aim to characterize the *m*inimax distortion, defined as:

$$\mathcal{M}_{\mathsf{dist}}(\sigma, m, d_{1:K}, B_{1:K}) := \inf_{\mathcal{E}_1,\cdots,\mathcal{E}_K,\mathcal{D}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right], \tag{30}$$

where $\mathcal{P} = \mathsf{subG}^{(d)}(\sigma)$ represents the family of $\sigma$–sub–Gaussian $d$–dimensional distributions.

#### 4.1.1 Distributed Cross Covariance Matrix Estimation (DCCME)

A key component of the analysis of the bounds on the minimax distortion (30) involves investigating cross–covariance estimation for the case of *two* users. Specifically, we consider the scenario where there are $K = 2$ agents, and the objective is to estimate the *c*oss–covariance matrix $\mathbf{C}_{21} := \mathbb{E}[(\mathbf{X}_2 - \mathbb{E}[\mathbf{X}_2])(\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1])^\top]$ using the same communication setup as in DCME. In this scenario, the encoders remain as previously defined. The decoder function is given by $\mathcal{D}_{21} : [1 : 2^{B_1}] \times [1 : 2^{B_2}] \mapsto \mathbb{R}^{d_2 \times d_1}$, where $\mathbb{R}^{d_2\times d_1}$ is the set of $d_2 \times d_1$–matrices. The decoder maps $(M_1, M_2)$ to $\widehat{\mathbf{C}}_{21}$. Again, the objective is to characterize the minimax distortion, defined as:

$$\mathcal{M}^{(\mathsf{cross})}_{\mathsf{dist}}(\sigma, m, d_{1:2}, B_{1:2}) := \inf_{\mathcal{E}_1,\mathcal{E}_2,\mathcal{D}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}_{21}, \mathbf{C}_{21})\right]. \tag{31}$$

### 4.2 Minimax Lower Bounds for the Expected Distortion of a DCME Problem

We first establish lower bounds for the case $K = 2$ and subsequently derive a lower bound for the general $K$ from the two–user scenario.

We define:

$$
\begin{aligned}
\alpha_{\mathsf{op}}^{(\mathsf{cc})} &:= \sqrt{\frac{d_1(d_1 \vee d_2)}{2B_1}} \bigvee \sqrt{\frac{d_2(d_1 \vee d_2)}{2B_2}}, \\
\alpha_{\mathsf{op}}^{(\mathsf{sc})} &:= \sqrt{\frac{d}{3m}}, \\
\alpha_{\mathsf{F}}^{(\mathsf{cc})} &:= \sqrt{\frac{d_1 d_2}{14}\left(\frac{d_1}{B_1} \vee \frac{d_2}{B_2}\right)}, \\
\alpha_{\mathsf{F}}^{(\mathsf{sc,cross})} &:= \sqrt{\frac{d d_{\min}}{42m}}, \\
\alpha_{\mathsf{F}}^{(\mathsf{sc})} &:= \sqrt{\frac{d^2}{42m}}, \\
\alpha_{\mathsf{F}}^{(\mathsf{cc,self})} &:= \frac{4}{7}\left(\sqrt{d_1} \cdot 2^{\frac{-16B_1}{d_1^2}} \bigvee \sqrt{d_2} \cdot 2^{\frac{-16B_2}{d_2^2}}\right),
\end{aligned}
\tag{32}
$$

The minimax errors of the DCME and DCCME problems (denoted by $\mathcal{M}_{\mathsf{dist}}$ and $\mathcal{M}_{\mathsf{dist}}^{(\mathsf{cross})}$) can be expressed using the previously defined quantities. In these definitions, op refers to the operator norm, and F to the Frobenius norm. The terms cc and sc represent communication complexity and sample complexity, respectively.

Accordingly, $\alpha_{\mathsf{op}}^{(\mathsf{cc})}$ quantifies the impact of communication constraints on the minimax error under the operator norm, while $\alpha_{\mathsf{op}}^{(\mathsf{sc})}$ captures the effect of having a limited number of samples under the same norm. These quantities provide lower bounds on $\mathcal{M}_{\mathsf{op}}$ and $\mathcal{M}_{\mathsf{op}}^{(\mathsf{cross})}$.

Similarly, $\alpha_{\mathsf{F}}^{(\mathsf{cc})}$ reflects the influence of communication constraints on the minimax error when measured using the Frobenius norm. The impact of limited samples differs between full covariance estimation and cross–covariance estimation. Specifically, $\alpha_{\mathsf{F}}^{(\mathsf{sc})}$ represents the sample–induced error in estimating the full covariance matrix, while $\alpha_{\mathsf{F}}^{(\mathsf{sc,cross})}$ pertains to the cross–covariance part.

An additional error term, $\alpha_{\mathsf{F}}^{(\mathsf{cc,self})}$, accounts for the effect of communication limitations on the estimation of the self–covariance matrix. This term arises only in the lower bound of $\mathcal{M}_{\mathsf{F}}$.

**Theorem 4.1.** *Consider the* $\mathsf{DCCME}(\sigma, m, d_{1:2}, B_{1:2})$ *problem. Then,* $\mathcal{M}_{\mathsf{dist}}^{(\mathsf{cross})}(\sigma, m, d_{1:2}, B_{1:2})$ *is lower–bounded as follows:*

$$
\begin{aligned}
\mathcal{M}_{\mathsf{op}}^{(\mathsf{cross})} &\geq \frac{\sigma^2}{32}\left(\left(\alpha_{\mathsf{op}}^{(\mathsf{cc})} \bigvee \alpha_{\mathsf{op}}^{(\mathsf{sc})}\right) \bigwedge 2\right) \\
&= \sigma^2 \Omega\left(\left(\sqrt{d\left(\frac{d_1}{B_1} \vee \frac{d_2}{B_2}\right)} \bigvee \sqrt{\frac{d}{m}}\right) \bigwedge 1\right),
\end{aligned}
\tag{33}
$$

*and:*

$$
\begin{aligned}
\mathcal{M}_{\mathsf{F}}^{(\mathsf{cross})} &\geq \frac{\sigma^2}{32}\left(\left(\alpha_{\mathsf{F}}^{(\mathsf{cc})} \bigvee \alpha_{\mathsf{F}}^{(\mathsf{sc,cross})}\right) \bigwedge \frac{\sqrt{d_{\min}}}{7}\right) \\
&= \sigma^2 \Omega\left(\left(\sqrt{d_1 d_2 (\frac{d_1}{B_1} \vee \frac{d_2}{B_2})} \bigvee \sqrt{\frac{d d_{\min}}{m}}\right) \bigwedge \sqrt{d_{\min}}\right).
\end{aligned}
\tag{34}
$$

**Theorem 4.2.** *Consider the* $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ *problem. Then,* $\mathcal{M}_{\mathsf{dist}}(\sigma, m, d_{1:2}, B_{1:2})$ *is lower–bounded as follows:*

$$\begin{aligned}\mathcal{M}_{\mathsf{op}} \geq \mathcal{M}_{\mathsf{op}}^{(\mathsf{cross})} &\geq \frac{\sigma^2}{32}\left(\left(\alpha_{\mathsf{op}}^{(\mathsf{cc})} \bigvee \alpha_{\mathsf{op}}^{(\mathsf{sc})}\right) \bigwedge 2\right) \\ &= \sigma^2 \Omega\left(\left(\sqrt{d\left(\frac{d_1}{B_1} \vee \frac{d_2}{B_2}\right)} \bigvee \sqrt{\frac{d}{m}}\right) \bigwedge 1\right),\end{aligned} \tag{35}$$

*and:*

$$\begin{aligned}\mathcal{M}_{\mathsf{F}} &\geq \frac{\sigma^2}{32}\left(\left(\left(\alpha_{\mathsf{F}}^{(\mathsf{cc})} \bigvee \alpha_{\mathsf{F}}^{(\mathsf{sc})}\right) \bigwedge \frac{\sqrt{d}}{7}\right) \bigvee \alpha_{\mathsf{F}}^{(\mathsf{cc,self})}\right) \\ &= \sigma^2 \Omega\left(\left(\left(\sqrt{d_1 d_2 (\frac{d_1}{B_1} \vee \frac{d_2}{B_2})} \bigvee \sqrt{\frac{d^2}{m}}\right) \bigwedge \sqrt{d}\right) \bigvee \sqrt{d_1} \cdot 2^{\frac{-16B_1}{d_1^2}} \bigvee \sqrt{d_2} \cdot 2^{\frac{-16B_2}{d_2^2}}\right).\end{aligned} \tag{36}$$

**Corollary 4.3.** *Any* DCME *(or* DCCME*) scheme that approximate the covariance matrix within* $\varepsilon$ *expected–operator norm distortion has communication budget* $B_k = \Omega\left(\sigma^4 \frac{d d_k}{\varepsilon^2}\right)$ *and sample complexity* $m = \Omega\left(\sigma^4 \frac{d}{\varepsilon^2}\right)$*. Furthermore, any* DCME *(or* DCCME*) scheme that achieves expected Frobenius norm distortion at most* $\varepsilon$ *requires a communication budget of* $B_k = \Omega\left(\sigma^4 \frac{d_1 d_2 d_k}{\varepsilon^2}\right)$*. In addition, the sample complexity for any* DCME *scheme is* $m = \Omega\left(\sigma^4 \frac{d^2}{\varepsilon^2}\right)$*, while for any* DCCME *scheme it is* $m = \Omega\left(\sigma^4 \frac{d d_{\min}}{\varepsilon^2}\right)$.

**Remark 4.4.** The lower bounds on the sample complexity of distributed covariance matrix estimation match those of centralized covariance matrix estimation. Clearly, the sample complexity in the distributed setting must be at least as large as that of the centralized setting, where all data are collected on a central server. The sample complexity lower bound for the operator norm is folklore, while the corresponding lower bound for the Frobenius norm is derived in [ABD+20, DMR20].

The complete proof for Theorem 4.2 is provided in detail in Section 5. Within that section, we employ a specialized variant of Fano's method, known as the averaged Fano's method, to establish the theorem. The initial step of this averaged Fano's method involves reducing the estimation problem to a finite hypothesis testing problem, a technique also common in the standard Fano's method. This reduction utilizes a family of distributions $\{P_v\}_{v\in\mathcal{V}}$ with associated covariance matrices $\{\mathbf{C}_v\}_{v\in\mathcal{V}}$, where $\mathcal{V}$ is a set of finite cardinality. Subsequently, a minimax lower bound for the estimation can be derived, expressed in terms of the error probability of this hypothesis testing problem and the separation between the distinct $\mathbf{C}_v$s, defined as $\rho_{\mathsf{dist}} := \inf_{\substack{v,v'\in\mathcal{V} \\ v\neq v'}} \left\{ \|\mathbf{C}_v - \mathbf{C}_{v'}\|_{\mathsf{dist}} \right\}$.

### 4.3 Achievability of the Minimax Lower bounds

In this section, we propose a $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ scheme and derive lower bounds on both its sample complexity and communication budgets for $\varepsilon$–approximation of the covariance matrix under $\|.\|_{\mathsf{dist}}$ norm. A description of the achievable scheme is provided in Section 6 with detailed proofs is relegated to Appendix G. Here, we outline the main idea.

Consider the decomposition bellow of matrix$\mathbf{C}$:

$$\overline{\mathbf{C}} = \begin{bmatrix} \overline{\mathbf{C}}_{11} & \overline{\mathbf{C}}_{12} \\ \overline{\mathbf{C}}_{12}^{\top} & \overline{\mathbf{C}}_{22} \end{bmatrix}, \tag{37}$$

where $\overline{\mathbf{C}}_{11} = \overline{\mathbf{C}}_{[1:d_1,1:d_1]}$, $\overline{\mathbf{C}}_{12} = \overline{\mathbf{C}}_{[1:d_1,d_1+1:d]}$, and $\overline{\mathbf{C}}_{22} = \overline{\mathbf{C}}_{[d_1+1:d,d_1+1:d]}$. Agent 1 can estimate $\mathbf{C}_{11}$ using the data points $\{\mathbf{X}_1^{(i)}\}_{i=1}^m$, and Agent 2 can estimate $\mathbf{C}_{22}$ using the data points $\{\mathbf{X}_2^{(i)}\}_{i=1}^m$. Consequently, they allocate portions of their communication budgets to transmit quantized versions of $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ to the central server. The remaining portions of their communication budgets are then used to transmit quantized versions of $\{\mathbf{X}_1^{(i)}\}_{i=1}^m$ and $\{\mathbf{X}_2^{(i)}\}_{i=1}^m$ to the central server. The central server can then estimate $\mathbf{C}_{12}$ using this received information and construct an estimate $\widehat{\mathbf{C}}$. We present the main result of this subsection in terms of communication budgets and sample complexity.

**Theorem 4.5.** *Consider* $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ *and a permissible distortion* $\varepsilon \leq \sigma^2$. *Assume that the number of samples* $m$ *and the communication budgets* $B_k$ *satisfy the following constraints:*

$$\begin{aligned} m &\geq \tau \frac{d\sigma^4}{\varepsilon^2} \\ B_k &\geq \tau' \frac{\sigma^4 d_k d}{\varepsilon^2} \cdot \log\left(\frac{\sigma^2}{\varepsilon}\right), \end{aligned}$$

*for* $k = 1, 2$ *and some constants* $\tau, \tau'$. *Then, there exists a scheme with expected distortions* $\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \leq \varepsilon$ *and* $\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}_{12}, \mathbf{C}_{12})\right] \leq \varepsilon$.

**Theorem 4.6.** *Consider* $\mathsf{DCME}(\sigma, m, d_{1:2}, B_{1:2})$ *and assume that the number of samples* $m$ *and the communication budgets* $B_k$ *satisfy the following constraints:*

$$\begin{aligned} m &\geq \tau \frac{d^2\sigma^4}{\varepsilon^2} \\ B_k &\geq \tau' \frac{\sigma^4 d_k d d_{\min}}{\varepsilon^2} \cdot \log_2\left(\frac{\sigma^2\sqrt{d_{\min}}}{\varepsilon}\right) \bigvee \tau'' d_k^2 \log_2\left(\frac{\sigma^2\sqrt{d}}{\varepsilon}\right), \end{aligned}$$

*for* $k = 1, 2$ *and some constants* $\tau, \tau', \tau''$. *Then, there exists a scheme with expected distortions* $\mathbb{E}\left[\mathcal{L}_{\mathsf{F}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \leq \varepsilon$. *Further if the numbers of sample and communication budget satisfy the following constraint,*

$$\begin{aligned} m &\geq \tau \frac{\sigma^4 d d_{\min}}{\varepsilon^2} \\ B_k &\geq \tau' \frac{\sigma^4 d_k d d_{\min}}{\varepsilon^2} \cdot \log_2\left(\frac{\sigma^2\sqrt{d_{\min}}}{\varepsilon}\right), \end{aligned}$$

*for* $k = 1, 2$ *and some constants* $\tau, \tau', \tau''$. *Then, there exists a scheme with expected distortions* $\mathbb{E}\left[\mathcal{L}_{\mathsf{F}}(\widehat{\mathbf{C}}_{12}, \mathbf{C}_{12})\right] \leq \varepsilon$.

**Remark 4.7.** Comparing Corollary 4.3, Theorem 4.5 and 4.6 yields that the lower bounds and upper bounds are matched up to a logarithmic factor.

**Remark 4.8.** Assume $d_2 = 1, B_2 = \infty$, that is the agent 2 takes the role of central server. In this case, our setting corresponds to the scenario investigated in [HS19, Theorem 4]. That result

implies an upper bound on the Frobenius norm error, specifically $\mathcal{O}\left(\frac{d}{\sqrt{B}}\right)$, under the premise of each agent possessing an infinite number of correlated samples. This bound is achieved via a novel estimator called *maximum estimator*. Consequently, we infer that the logarithmic factors present in Theorem 1 are dispensable, implying the tightness of our communication budget's lower bound without requiring an additional logarithmic term, contingent on the availability of infinite or sufficiently numerous samples. It is imperative to underscore that [HS19] confines its analysis to Gaussian random vectors and infinite samples, whereas our study encompasses the broader category of sub–Gaussian distributions and accommodates finite samples, thereby operating under more weaker assumptions.

**Remark 4.9** (Centralized Covariance Matrix Estimation with Limited Communication Budget)**.** Consider the centralized covariance matrix estimation problem, a simplified version of the distributed case where only one agent is present. This agent can estimate the entire covariance matrix locally using a sample covariance estimator and then transmit a quantized version of this estimate to a central server.

This approach gives rise to the following lower bound on the expected operator norm distortion:

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})\right] = \sigma^2 \mathcal{O}\left(\max\left\{\sqrt{\frac{d}{m}}, \exp\left(\frac{-c.B}{d^2}\right)\right\}\right).$$

Crucially, this result demonstrates that the communication budget for the centralized case, which scales as $B = \mathcal{O}\Big(d^2 \log\left(\frac{1}{\varepsilon}\right)\Big)$, is substantially lower than the total communication budget required for the distributed case, which scales as $B = \Omega\left(\frac{d^2}{\varepsilon^2}\right)$.

### 4.4 Multi–Agent Scenario

In this section, we analyze the multi agent scenario. Let the number of agents be denoted by $K > 2$ and for all $k \in [K]$, Agent $k$ has access to the dimensions $[1 + \sum_{i=1}^{k-1} d_i : \sum_{i=1}^{k} d_i]$ of all $m$ samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}$. The information available to Agent $k$ can be represented as the set $\{\mathbf{X}_k^{(i)}\}_{i=1}^m = \{\mathbf{Z}^{(i)}_{[1+\sum_{i=1}^{k-1} d_i : \sum_{i=1}^{k} d_i]}\}_{i=1}^m$.

The central server aims to estimate $\mathbf{C}$ by receiving up to $B_k$ bits of information from Agent $k$. To derive a lower bound in this context, we define a subset of agents, denoted as $\mathcal{S} \subset [K]$. Then we assume that the agents within $\mathcal{S}$ collude, and similarly, the agents in the complement set $\mathcal{S}^{\mathsf{c}} = [K] \setminus \mathcal{S}$ also collude. This leads to the formation of two super–agents, $A$ and $B$. The super–agent $A$ has access to $\sum_{k\in\mathcal{S}} d_k$ dimensions and a communication budget of $\sum_{k\in\mathcal{S}} B_k$ bits, while super–agent $B$ has access to $\sum_{k\in\mathcal{S}^c} d_k$ dimensions and a communication budget of $\sum_{k\in\mathcal{S}^{\mathsf{c}}} B_k$ bits. Theorem 4.2 provides a lower bound for this colluded scenario, which also serves as a lower bound for the non–colluded case. By maximizing these lower bounds over the possible choices of subset $\mathcal{S}$, we conclude that:

$$\begin{aligned}\mathcal{M}_{\mathsf{op}} &= \sigma^2 \Omega\left(\sqrt{d \cdot \max_{\substack{\mathcal{S}\subset[K]\\ \mathcal{S}\neq\emptyset}} \left\{\frac{\sum_{k\in\mathcal{S}} d_k}{\sum_{k\in\mathcal{S}} B_k}\right\}} \bigvee \sqrt{\frac{d}{m}}\right) \\ &= \sigma^2 \Omega\left(\sqrt{d \cdot \max_{k\in[K]} \frac{d_k}{B_k}} \bigvee \frac{d}{m}\right).\end{aligned} \tag{38}$$

Specifically, assuming that $K = d$ and that for all $k \in [K]$, $d_k = 1$, we get:

$$\mathcal{M}_{\mathsf{op}} = \sigma^2 \Omega\left(\sqrt{\frac{d}{\min_k B_k \wedge m}}\right). \tag{39}$$

This result indicates that the performance of the DCME problem is determined by the agents with the smallest communication budgets. Also (38) implies that the communication budget of agent $k$ for approximating the covariance matrix within $\varepsilon$ error in operator norm, should satisfy $B_k = \Omega\left(\frac{dd_k}{\varepsilon^2}\right)$. The following theorem (proof deferred to Appendix H) shows that this bound is tight up to a logarithmic factor in $\log d$ and $\frac{1}{\varepsilon}$.

**Theorem 4.10.** *Consider* $\mathsf{DCME}(\sigma, m, d_{1:K}, B_{1:K})$ *and a permissible distortion* $\varepsilon \leq \sigma^2$. *Assume that the number of samples* $m$ *and the communication budgets* $B_k$ *satisfy the following constraints:*

$$m \geq \tau \frac{d\sigma^4}{\varepsilon^2}, \qquad B_k \geq \tau' \frac{\sigma^4 d d_k}{\varepsilon^2} \log_2\left(\tau'' \frac{\sigma^4}{\varepsilon^2} \log(d\sigma^2/\varepsilon)\right)$$

*for* $k \in [1:K]$ *and some constants* $\tau, \tau', \tau''$. *Then, there exists a scheme with expected distortions* $\mathbb{E}\left[\mathcal{L}_{\mathsf{op}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \leq \varepsilon$.

Comparing Theorem 4.10 with Theorem 4.5 for $K = 2$, it is worth mentioning that to get achievability for general $K$, the logarithmic factor depends on $\log d$ in addition to $\frac{1}{\varepsilon}$.

# 5 Proof of Lower Bounds via Conditional SDPI

In this section, we present the proof of Theorems 4.1 and 4.2. We first establish the necessary preliminaries and then proceed with the proof, step by step.

## 5.1 Averaged Fano Method

We employ a variant of Fano's method, which we term the averaged Fano's method, to lower bound $\mathcal{M}_{\mathsf{dist}}(\sigma, m, d_{1:2}, B_{1:2})$. Generally, Fano's method reduces an estimation problem to a hypothesis testing problem and subsequently derives a lower bound for the latter through Fano's inequality.

More precisely, let $\mathcal{P}$ be a family of distributions, and let $\theta : \mathcal{P} \mapsto \Theta$ be a parameter of interest of the distributions in $\mathcal{P}$ (e.g., mean, covariance, etc.) residing in a metric space $\Theta$ equipped with the metric $\|.\|$. The objective is to approximate $\theta(P)$ for an unknown $P \in \mathcal{P}$ using a sample $X$ obtained from $P$. Let this approximation be denoted by $\hat{\theta}(X)$. Consider the set $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$ and define $\mathcal{P}_{\mathcal{V}} = \{P_1, \cdots, P_{|\mathcal{V}|}\}$ as a subset of $\mathcal{P}$. The set $\mathcal{P}_{\mathcal{V}}$ is termed $2\delta$–separated, if for each $i, j \in \mathcal{V}$ with $i \neq j$, we have $\|\theta(P_i) - \theta(P_j)\| \geq 2\delta$. For a given $2\delta$–separated $\mathcal{P}_{\mathcal{V}}$, let $V$ be a uniform random variable drawn from $[1 : |\mathcal{V}|]$ and given $V = v$, let the random variable $X$ be a sample from $P_v$ and $\hat{\theta}(X)$ representing the corresponding approximation of $\theta(P_v)$. Thus, we have the Markov chain $V \multimap X \multimap \hat{\theta}$. Fano's method establishes the following lower bound on the minimax error of estimation $\hat{\theta}$ of $\theta$:

$$\min_{\hat{\theta}: \mathcal{X} \mapsto \Theta} \max_{P \in \mathcal{P}} \mathbb{E}_P\left[\left\|\hat{\theta}(X) - \theta(P)\right\|\right] \geq \delta\left(1 - \frac{I(V;X) + \log 2}{\log |\mathcal{V}|}\right). \tag{40}$$

We employ Fano's method in conjunction with the conditional SDPI to establish the lower bounds (Theorem 4.1 and Theorem 4.2). However, to render the computation of the mutual information in (40) tractable, we utilize the following variant of Fano's method. Here, rather than reducing the approximation to a single hypothesis testing problem, we consider multiple reductions and use the average of the Fano lower bounds as a lower bound for the minimax loss; hence, we refer to it as the averaged Fano method. More precisely, let $W \sim \pi_W$ be a random variable taking values in $\mathcal{W}$. For each instance $W = w$, assume that $\mathcal{P}_{\mathcal{V}}^{(w)} = \{P_1^{(w)}, \cdots, P_{|\mathcal{V}|}^{(w)}\}$ is a $2\delta$–separated subset of $\mathcal{P}$. Let $V$ be a uniform random variable, as defined previously,

independent of $W$. Given $V = v$ and $W = w$, the random variable $X$ is drawn from $P_v^{(w)}$. The standard Fano lower bound (40) then yields:

$$\begin{aligned}\min_{\hat{\theta}:\mathcal{X}\mapsto\Theta}\max_{P\in\mathcal{P}}\mathbb{E}_P\left[\left\|\hat{\theta}(X)-\theta(P)\right\|\right] &\geq \delta\sup_{w\in\mathcal{W}}\left(1-\frac{I(V;X|W=w)+\log 2}{\log|\mathcal{V}|}\right)\\ &\geq \delta\left(1-\frac{I(V;X|W)+\log 2}{\log|\mathcal{V}|}\right)\\ &\geq \delta\left(1-\frac{I(V,W;X)+\log 2}{\log|\mathcal{V}|}\right).\end{aligned} \tag{41}$$

The averaged Fano method has been previously employed in [Wai19, Example 15.19] to derive a concise minimax bound for the PCA problem.

### 5.2 Averaged Fano's Method for Covariance Estimation

Let $W \sim \pi_W$ be a random variable taking value in $\mathcal{W}$. For each $w \in \mathcal{W}$, we consider a family of distributions $\mathcal{P}_{\mathcal{V}}^{(w)} = \{P_v^{(w)}\}_{v\in\mathcal{V}} \subset \mathsf{subG}^{(d)}(\sigma)$ indexed by a finite set $\mathcal{V} = [1 : |\mathcal{V}|]$. For each $(w, v)$, let $\mathbf{C}_v^{(w)} := \mathbb{E}_{\mathbf{Z}\sim P_v^{(w)}}[(\mathbf{Z}-\mathbb{E}[\mathbf{Z}])(\mathbf{Z}-\mathbb{E}[\mathbf{Z}])^\top]$ denotes the corresponding covariance matrix. Further, let $\mathbf{X}_1 = \mathbf{Z}_{[1:d_1]}$ and $\mathbf{X}_2 = \mathbf{Z}_{[d_1+1:d]}$, represent the partitioning of the $d$–dimensional vector $\mathbf{Z}$ into a $d_1$–dimensional vector $\mathbf{X}_1$ and a $d_2$–dimensional vector $\mathbf{X}_2$. For each $(w, v)$, let $\mathbf{D}_v^{(w)} := \mathbf{C}_{v,21}^{(w)} = \mathbb{E}_{\mathbf{Z}\sim P_v^{(w)}}[(\mathbf{X}_2 - \mathbb{E}[\mathbf{X}_2])(\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1])^\top]$ denotes the corresponding *cross*–covariance matrix. For this set, we define the separations $\rho$ and $\rho^{(\mathsf{cross})}$ with respect to the $\mathsf{dist}$ norm metric on the space of covariance matrices as:

$$\begin{aligned}\rho_{\mathsf{dist}} &:= \inf_{\substack{w\in\mathcal{W}\\(v,v')\in\mathcal{V}^2, v\neq v'}}\left\{\left\|\mathbf{C}_v^{(w)}-\mathbf{C}_{v'}^{(w)}\right\|_{\mathsf{dist}}\right\},\\ \rho_{\mathsf{dist}}^{(\mathsf{cross})} &:= \inf_{\substack{w\in\mathcal{W}\\(v,v')\in\mathcal{V}^2, v\neq v'}}\left\{\left\|\mathbf{D}_v^{(w)}-\mathbf{D}_{v'}^{(w)}\right\|_{\mathsf{dist}}\right\}.\end{aligned} \tag{42}$$

We now state Lemma 5.1, a direct consequence of the averaged Fano's method (41):

**Lemma 5.1.** *Consider a collection of $\rho_{\mathsf{dist}}$–separated families of distributions $\mathcal{P}_{\mathcal{V}}^{(w)}$ under the $\mathsf{dist}$–norm on covariance matrices, each family consisting of $|\mathcal{V}|$ distributions $P_v^{(w)}$. Assume a random variable $V \in \mathcal{V}$ is chosen uniformly and independently of $W$, and given $(W, V) = (w, v)$, samples $\{\mathbf{Z}^{(i)}\}_{i=1}^m$ are drawn i.i.d. from $P_v^{(w)}$. Additionally, assume that agents 1 and 2 have access to $\{\mathbf{X}_1^{(i)} = \mathbf{Z}_{[1:d_1]}^{(i)}\}_{i=1}^m$ and $\{\mathbf{X}_2^{(i)} = \mathbf{Z}_{[d_1+1:d]}^{(i)}\}_{i=1}^m$, respectively. For any* DCME *(respectively,* DCCME*) scheme with parameters $(\sigma, m, d_{1:2}, B_{1:2})$, we have:*

$$\begin{aligned}\inf_{\mathcal{E}_1,\mathcal{E}_2,\mathcal{D}}\sup_{P\in\mathcal{P}}\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}},\mathbf{C})\right] &\geq \frac{\rho_{\mathsf{dist}}}{2}\left(1-\frac{I(W,V;M_1,M_2)+\log 2}{\log|\mathcal{V}|}\right),\\ \inf_{\mathcal{E}_1,\mathcal{E}_2,\mathcal{D}}\sup_{P\in\mathcal{P}}\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}_{21},\mathbf{C}_{21})\right] &\geq \frac{\rho_{\mathsf{dist}}^{(\mathsf{cross})}}{2}\left(1-\frac{I(W,V;M_1,M_2)+\log 2}{\log|\mathcal{V}|}\right).\end{aligned}$$

We now proceed by first establishing two lemmas that utilize Lemma 5.1 with specific distribution families $\{P_v^{(w)}\}_{v\in\mathcal{V},w\in\mathcal{W}}$. These lemmas will subsequently pave the way for deriving the main theorem concerning the minimax lower bound.

### 5.3 Construction of $\rho_{\mathsf{dist}}$–Separated Families for Cross–Covariance

Consider a set $\mathcal{V} = [1 : |\mathcal{V}|]$ and a corresponding family of distributions $\mathcal{P}_{\mathcal{V}}^{(w)} = \{P_v^{(w)}\}_{v\in\mathcal{V}}$, where $P_v^{(w)} = \mathcal{N}(\mathbf{0}, \mathbf{C}_v^{(w)})$, and:

$$\mathbf{C}_v^{(w)} = \frac{\sigma^2}{2}\begin{bmatrix} \mathbf{I}_{d_1} & \delta(\mathbf{D}_v^{(w)})^\top \\ \delta\mathbf{D}_v^{(w)} & \mathbf{I}_{d_2} \end{bmatrix}, \tag{43}$$

where $\mathbf{D}_v^{(w)}$ is a matrix in $\mathbb{R}^{d_2\times d_1}$ with $\left\|\mathbf{D}_v^{(w)}\right\|_{\mathsf{op}} \le 1$, and $\delta \le 1$ is a parameter to be determined subsequently. Note that $\mathbf{C}_v^{(w)}$ represents the covariance matrix of a $\sigma$–sub–Gaussian random vector $\mathbf{Z}$. Consequently, we must have $\mathbf{C}_v \succeq \mathbf{0}$. Furthermore, from Definition 2.4, for all vectors $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$, the random variable $\mathbf{u}^\top\mathbf{Z}$ is $\sigma$–sub–Gaussian; therefore, $\mathrm{Var}[\mathbf{u}^\top\mathbf{Z}] \le \sigma^2$. This implies that for all $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$:

$$\mathrm{Var}[\mathbf{u}^\top\mathbf{Z}] = \mathbb{E}[\mathbf{u}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{u}] = \mathbf{u}^\top\mathbf{C}_v^{(w)}\mathbf{u} \le \sigma^2. \tag{44}$$

Therefore, we must have $\left\|\mathbf{C}_v^{(w)}\right\|_{\mathsf{op}} \le \sigma^2$.

We can express $\mathbf{C}_v^{(w)}$ as $\mathbf{C}_v^{(w)} = \frac{\sigma^2}{2}\mathbf{I} + \frac{\sigma^2}{2}\begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_v^{(w)})^\top \\ \delta\mathbf{D}_v^{(w)} & \mathbf{0} \end{bmatrix}$. From Lemma A.1, the eigenvalues of $\mathbf{C}_v^{(w)}$ are given by $\frac{\sigma^2}{2}\left(1 \pm \delta\sigma_i(\mathbf{D}_v^{(w)})\right)$. Therefore, if we assume that $\left\|\mathbf{D}_v^{(w)}\right\|_{\mathsf{op}} \le 1$ and $\delta \le 1$, the constraints $\mathbf{C}_v^{(w)} \succeq \mathbf{0}$ and $\left\|\mathbf{C}_v^{(w)}\right\|_{\mathsf{op}} \le \sigma^2$ are satisfied, ensuring that $P_v^{(w)} \in \mathsf{subG}^{(d)}(\sigma)$.

This choice of distributions exhibits the following properties:

- We have $\rho_{\mathsf{op}} = \rho_{\mathsf{op}}^{(\mathsf{cross})}$ and $\rho_{\mathsf{F}} = \sqrt{2}\rho_{\mathsf{F}}^{(\mathsf{cross})}$. Consequently, it suffices to derive a lower bound on the minimax error for cross–covariance matrix estimation.
- More importantly, the vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v^{(w)})$ has the same marginal distribution over the first $d_1$ dimensions and the second $d_2$ dimensions, for all $(w, v)$. Therefore, $\mathbf{X}_1 = \left\{\mathbf{X}_1^{(i)}\right\}_{i=1}^m$ is independent from $(W, V)$. Similarly, $\mathbf{X}_2 = \left\{\mathbf{X}_2^{(i)}\right\}_{i=1}^m$ is independent from $(W, V)$. Subsequently, $M_1$ (and similarly $M_2$) is also independent from $(W, V)$. This implies:

$$\begin{aligned} I(V, W; M_1, M_2) &= I(V, W; M_1) + I(V, W; M_2 \mid M_1) \\ &= I(V, W; M_2 \mid M_1) && \text{since } M_1 \perp\!\!\!\perp (W, V) \\ &\le I(V, W, M_1; M_2) \\ &= I(M_1; M_2 \mid V, W) && \text{since } M_2 \perp\!\!\!\perp (W, V). \end{aligned} \tag{45}$$

### 5.4 Applying the Conditional SDPI

We now derive an upper bound on $I(M_1; M_2|W, V)$ using conditional SDPI. Observe that conditioned on any $(W, V) = (w, v)$, the structures of encoders and decoder impose the following Markov chain: $M_1 \multimap \mathbf{X}_1 \multimap \mathbf{X}_2 \multimap M_2$. Furthermore, $(M_1, \mathbf{X}_1)$ is independent of $(W, V)$; thus, the constraints in the definition of conditional SDPI are satisfied. By the data processing inequality, we have:

$$I(M_1; M_2|V, W) \le I(M_1; \mathbf{X}_2|V, W) \wedge I(M_2; \mathbf{X}_1|V, W). \tag{46}$$

Now, conditioned on $(W,V)=(w,v)$, $(\mathbf{X}_1,\mathbf{X}_2)$ is sampled from the normal distribution $\mathcal{N}(\mathbf{0},\mathbf{C}_v^{(w)})^{\otimes m}$. Thus, the conditional SDPI constant for mixture of Gaussian (Theorem 3.9) and the tensorization property of conditional SDPI constant (Theorem 3.5) yield:

$$\begin{aligned} I(M_1;\mathbf{X}_2|V,W) &\le \delta^2 \left\| \mathbb{E}_{(W,V)}\left[(\mathbf{D}_V^{(W)})^\top \mathbf{D}_V^{(W)}\right]\right\|_{\mathsf{op}} I(M_1;\mathbf{X}_1) \\ &\le \delta^2 \left\| \mathbb{E}_{(W,V)}\left[(\mathbf{D}_V^{(W)})^\top \mathbf{D}_V^{(W)}\right]\right\|_{\mathsf{op}} B_1, \end{aligned} \tag{47}$$

where we have used the fact that $I(M_1;\mathbf{X}_1)\le H(M_1)\le B_1$. Similarly, we have:

$$\begin{aligned} I(M_2;\mathbf{X}_1|V,W) &\le \delta^2 \left\| \mathbb{E}_{(W,V)}\left[\mathbf{D}_V^{(W)}(\mathbf{D}_V^{(W)})^\top\right]\right\|_{\mathsf{op}} I(M_2;\mathbf{X}_2) \\ &\le \delta^2 \left\| \mathbb{E}_{(W,V)}\left[\mathbf{D}_V^{(W)}(\mathbf{D}_V^{(W)})^\top\right]\right\|_{\mathsf{op}} B_2. \end{aligned} \tag{48}$$

### 5.5 Evaluating the Conditional SDPI Constant Using Random Signed Permutation Matrices

Up to this point, $W$ has not played a specific role, and all preceding steps could have been performed without it. However, the primary challenge in what follows is to obtain a concise upper bound on the conditional SDPI constant $\left\| \mathbb{E}_{(W,V)}\left[\mathbf{D}_V^{(W)}(\mathbf{D}_V^{(W)})^\top\right]\right\|_{\mathsf{op}}$. Achieving this without the introduction of $W$ would be arduous, if not infeasible.

For the present analysis, we assume $d_1\ge d_2$. The complementary case follows by symmetry. We note that there exist $2^{d_1}d_1!$ distinct signed permutation matrices in $\mathbb{R}^{d_1}$; thus, we can impose an ordering on these matrices, denoting them as $\left\{\mathbf{A}_j\right\}_{j=1}^{2^{d_1}d_1!}$. Let $W$ be a random variable taking values uniformly at random in the set $\{1,2,\dots,2^{d_1}d_1!\}$. Further, let $\mathcal{P}_{\mathcal{V}}=\{P_v\}_{v\in\mathcal{V}}$ be a $\rho_{\mathsf{dist}}$–separated set of normal distributions, such that $P_v=\mathcal{N}(\mathbf{0},\mathbf{C}_v)$ with:

$$\mathbf{C}_v=\frac{\sigma^2}{2}\begin{bmatrix}\mathbf{I}_{d_1} & \delta\mathbf{D}_v^\top\\ \delta\mathbf{D}_v & \mathbf{I}_{d_2}\end{bmatrix}, \tag{49}$$

where $\left\|\mathbf{D}_v\right\|_{\mathsf{op}}\le 1$.

Now, for each $w\in\{1,2,\dots,2_1^d d_1!\}$, let $\mathbf{D}_v^{(w)}=\mathbf{D}_v\mathbf{A}_w$ in the matrix representation (43) and $\mathcal{P}_{\mathcal{V}}^{(w)}=\{P_v^{(w)}\}_{v\in\mathcal{V}}$. Given that any signed permutation matrix is a unitary matrix, the distance (with respect to either the operator norm or the Frobenius norm) between any two corresponding matrices in different families $\mathcal{P}_{\mathcal{V}}^{(w)}$ and $\mathcal{P}_{\mathcal{V}}^{(w')}$ are identical, that is, $\left\|P_v^{(w)}-P_{v'}^{(w)}\right\|_{\mathsf{dist}}=\left\|P_v^{(w')}-P_{v'}^{(w')}\right\|_{\mathsf{dist}}$. Consequently, all sets $\mathcal{P}_{\mathcal{V}}^{(w)}$ are $\rho_{\mathsf{dist}}$–separated. Now, by Lemma 2.2, the following identity holds for any matrix $\mathbf{D}_v$ satisfying $\|\mathbf{D}_v\|_{\mathsf{op}}\le 1$:

$$\begin{aligned} \mathbb{E}_W\left[(\mathbf{D}_v^{(W)})^\top D_v^{(W)}\right] &= \mathbb{E}_W\left[\mathbf{A}_W^\top\mathbf{D}_v^\top\mathbf{D}_v\mathbf{A}_W\right] \\ &= \frac{1}{d_1}\mathsf{Tr}\left\{\mathbf{D}_v^\top\mathbf{D}_v\right\}\mathbf{I}_{d_1} \\ &= \frac{1}{d_1}\|\mathbf{D}_v\|_{\mathsf{F}}^2\,\mathbf{I}_{d_1} \\ &\preceq \frac{d_1\wedge d_2}{d_1}\mathbf{I}_{d_1}. \end{aligned} \tag{50}$$

Next, consider:

$$\mathbb{E}_W\left[\mathbf{D}_v^{(W)}(\mathbf{D}_v^{(W)})^\top\right]=\mathbf{D}_v^\top\mathbf{D}_v\preceq\mathbf{I}_{d_2}=\frac{d_1\wedge d_2}{d_2}\mathbf{I}_{d_2}. \tag{51}$$

In summary, combining (45), (46), (47), (48), (50), and (51) yields:

$$\begin{aligned} I(V,W;M_1,M_2) &\le I(M_1;M_2 \mid V,W) \\ &\le I(M_1;\mathbf{X}_2|V,W) \wedge I(M_2;\mathbf{X}_1|V,W) \\ &\le \delta^2 \left( \left\| \mathbb{E}_{(W,V)} \left[ (\mathbf{D}_V^{(W)})^\top \mathbf{D}_V^{(W)} \right] \right\|_{\mathsf{op}} B_1 \right) \bigwedge \left( \left\| \mathbb{E}_{(W,V)} \left[ \mathbf{D}_V^{(W)} (\mathbf{D}_V^{(W)})^\top \right] \right\|_{\mathsf{op}} B_2 \right) \\ &\le \delta^2 (d_1 \wedge d_2) \left( \frac{B_1}{d_1} \bigwedge \frac{B_2}{d_2} \right). \end{aligned} \tag{52}$$

### 5.6 Packing Set for the Operator–Norm Unit Ball with Respect to dist–Norm.

The subsequent step is to determine a lower bound on the cardinality of the $\rho_{\mathsf{dist}}$–separated set $\mathcal{P}_{\mathcal{V}}$ defined in the preceding subsection. In Appendix A.6 we introduce the $|||.|||_{\mathsf{dist}}$ norm of a vectorized matrix and discuss the packing and covering sets of the unit $|||.|||_{\mathsf{op}}$ ball of matrices under the dist norm. We define the set $\{\mathbf{D}_v\}_{v\in\mathcal{V}}$ as the $\epsilon$–packing points of $\mathcal{B}^{(d_1 d_2)}_{|||.|||_{\mathsf{op}}}(1)$ (see Equation (85)), under $|||.|||_{\mathsf{dist}}$ norm. Thus, $\inf_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}} \ge \epsilon$, $\max_{v\in\mathcal{V}} \{\|\mathbf{D}_v\|^2_{\mathsf{op}}\} \le 1$, and from (86) and (92), we have $\log_2(|\mathcal{V}|) \ge d_1 d_2 \log_2 \left( \frac{\nu^{(d_1,d_2)}_{\mathsf{dist}}}{\epsilon} \right)$, where $\nu^{(d_1,d_2)}_{\mathsf{dist}} = 1$ if $\mathsf{dist} = \mathsf{op}$ and $\nu^{(d_1,d_2)}_{\mathsf{dist}} = \frac{\sqrt{d_1 \wedge d_2}}{14}$ if $\mathsf{dist} = \mathsf{F}$. We set $\epsilon = \frac{\nu^{(d_1,d_2)}_{\mathsf{dist}}}{4}$ which yields $\log_2 |\mathcal{V}| \ge 2 d_1 d_2$. Furthermore, we note that the set $\mathcal{P}_{\mathcal{V}}$ corresponding to the packing $\{\mathbf{D}_v\}_{v\in\mathcal{V}}$ is $\rho_{\mathsf{dist}}$–separated with:

$$\begin{aligned} \rho_{\mathsf{op}} = \rho^{(\mathsf{cross})}_{\mathsf{op}} &= \delta\sigma^2 \frac{\nu^{(d_1,d_2)}_{\mathsf{op}}}{4} \\ \rho_{\mathsf{F}} = \sqrt{2}\rho^{(\mathsf{cross})}_{\mathsf{F}} &= \delta\sigma^2 \frac{\nu^{(d_1,d_2)}_{\mathsf{F}}}{2\sqrt{2}}. \end{aligned} \tag{53}$$

Setting $\delta^2 = \left( \frac{d_1 \vee d_2}{4} \left( \frac{d_1}{B_1} \bigvee \frac{d_2}{B_2} \right) \right) \bigwedge 1$, and incorporating (52), (53), and the inequality $\log_2 |\mathcal{V}| \ge 2 d_1 d_2$ into Lemma 5.1, we obtain:

$$\begin{aligned} \mathcal{M}_{\mathsf{op}} &\ge \frac{\sigma^2}{32} \left( \alpha^{(\mathsf{cc})}_{\mathsf{op}} \bigwedge 2 \right) & \mathcal{M}^{(\mathsf{cross})}_{\mathsf{op}} &\ge \frac{\sigma^2}{32} \left( \alpha^{(\mathsf{cc})}_{\mathsf{op}} \bigwedge 2 \right) \\ \mathcal{M}_{\mathsf{F}} &\ge \frac{\sigma^2}{32} \left( \alpha^{(\mathsf{cc})}_{\mathsf{F}} \bigwedge \frac{\sqrt{d_1 \wedge d_2}}{7} \right) & \mathcal{M}^{(\mathsf{cross})}_{\mathsf{F}} &\ge \frac{\sigma^2}{32} \left( \alpha^{(\mathsf{cc})}_{\mathsf{F}} \bigwedge \frac{\sqrt{d_1 \wedge d_2}}{7} \right) \end{aligned} \tag{54}$$

In Theorem 4.1 and Theorem 4.2, there exist additional lower bounds pertaining to sample complexity and the limited communication budget for self–covariance estimation. The proofs for these particular lower bounds are deferred to Appendix E.

## 6 Achievable Scheme: Proof Sketch of Theorem 4.5

In this section, we present a near–optimal achievable Distributed Covariance Matrix Estimation (DCME) scheme and establish an upper bound on its expected distortion. The proposed scheme operates in two distinct phases: one dedicated to approximating the self–covariance matrices $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$, and another for the cross–covariance matrix $\mathbf{C}_{12}$ (see (37)).

**Mean Invariance** Should the observed random vectors $\{\mathbf{Z}^{(i)}\}_{i=1}^{m}$ exhibit a non–zero mean, they can be transformed by defining $\mathbf{Z}'^{(i)} = \frac{1}{\sqrt{2}}(\mathbf{Z}^{(2i-1)} - \mathbf{Z}^{(2i)})$. This redefinition ensures that the new vectors possess a zero mean while retaining the identical covariance matrix as the original $\mathbf{Z}^{(i)}$. Consequently, the set of transformed samples $\{\mathbf{Z}'^{(i)}\}_{i=1}^{m/2}$ can be equivalently employed in place of the initial samples $\{\mathbf{Z}^{(i)}\}_{i=1}^{m}$. Therefore, for the purpose of the subsequent analysis, it is permissible to assume, without loss of generality, that $\mathbb{E}[\mathbf{Z}] = 0$.

**Empirical Estimation of Self–Covariance Matrices** Each agent is capable of estimating its respective self–covariance matrix directly from its local data by employing an empirical covariance estimator. Specifically, Agent 1 computes its estimate of $\mathbf{C}_{11}$ as $\widetilde{\mathbf{C}}_{11} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{X}_1^{(i)}\mathbf{X}_1^{(i)\top}$, while Agent 2 similarly estimates $\mathbf{C}_{22}$ using $\widetilde{\mathbf{C}}_{22} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{X}_2^{(i)}\mathbf{X}_2^{(i)\top}$.

**Quantization of Estimated Self–Covariance Matrices** The empirical self–covariance matrix $\widetilde{\mathbf{C}}_{11}$ lies within 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 $\epsilon$–covering of this ball with $2^{B_1/2}$ points with smallest possible $\epsilon$. If the empirical estimate $\widetilde{\mathbf{C}}_{11}$ lies within the ball $\mathcal{B}^{d_1^2}_{\|\cdot\|_{\mathsf{op}}}(\tau\sigma^2)$, Agent 1 quantizes $\widetilde{\mathbf{C}}_{11}$ to $B_1/2$ bits by selecting the nearest point in the covering to the empirical estimate. If $\widetilde{\mathbf{C}}_{11}$ lies outside the ball, Agent 1 declares an error. Agent 2 performs a similar quantization of its empirical estimate.

**Quantization of Estimated Self–Covariance Matrices** The empirical self–covariance matrix $\widetilde{\mathbf{C}}_{11}$ is expected to reside within the ball $\mathcal{B}^{d_1^2}_{\|\cdot\|_{\mathsf{op}}}(\tau\sigma^2)$ with high probability, for some positive constant $\tau$. To quantize this matrix, Agent 1 determines an $\epsilon$–covering of this ball, comprising $2^{B_1/2}$ points, such that $\epsilon$ is minimized. If the empirical estimate $\widetilde{\mathbf{C}}_{11}$ falls within this ball, Agent 1 quantizes $\widetilde{\mathbf{C}}_{11}$ to $B_1/2$ bits by selecting the nearest point from this covering to its empirical value. Conversely, if $\widetilde{\mathbf{C}}_{11}$ lies outside this specified ball, Agent 1 signals an error. Agent 2 executes an analogous quantization procedure for its own empirical estimate.

**Quantization of Data for Approximating the Cross–Covariance** To approximate the cross–covariance, we first select a subset size $n = \min\left\{\min\{\frac{B_1}{d_1}, \frac{B_2}{d_2}\}/2\log_2\left(\frac{6912\sigma^2}{\varepsilon}\right), m\right\}$. We then define data matrices $\mathbf{X}_1 \in \mathbb{R}^{d_1\times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{d_2\times n}$ by concatenating the first $n$ samples from each agent, such that $\mathbf{X}_1 = [\mathbf{X}_1^{(1)}, \mathbf{X}_1^{(2)}, \ldots, \mathbf{X}_1^{(n)}]$ and $\mathbf{X}_2 = [\mathbf{X}_2^{(1)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_2^{(n)}]$. The *empirical* estimator for $\mathbf{C}_{12}$ is then computed using these $n$ samples as $\widetilde{\mathbf{C}}_{12} = \frac{1}{n}\mathbf{X}_1\mathbf{X}_2^{\top}$. For communication, Agent 1 quantizes its entire block of data, $\mathbf{X}_1$, to $B_1/2$ bits. Agent 2 performs a symmetrical quantization on its data block. It is established that $\mathbf{X}_1$ is highly likely to reside within the ball $\mathcal{B}^{nd_1}_{\|\cdot\|_{\mathsf{op}}}(\tau\sigma\sqrt{d_1+n})$. Agent 1 quantizes $\mathbf{X}_1$ by finding an $\epsilon$–covering of this ball with $2^{B_1/2}$ points, aiming for the smallest possible $\epsilon$. If $\mathbf{X}_1$ lies within this ball, Agent 1 selects the closest point $\widehat{\mathbf{X}}_1$ from the covering to represent its quantized data, using $B_1/2$ bits. Otherwise, Agent 1 signals an error. Similarly, Agent 2 determines a quantized representation $\widehat{\mathbf{X}}_2$ for $\mathbf{X}_2$ utilizing $B_2/2$ bits.

**Cross–Covariance Estimation at the Central Server** Upon receiving the quantized data $\widehat{\mathbf{X}}_1$ and $\widehat{\mathbf{X}}_2$, the central server initially estimates $\mathbf{C}_{12}$ as $\widehat{\mathbf{C}}_{12} = \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^{\top}$. Should an error signal be received from either agent, the central server outputs a zero matrix, $\widehat{\mathbf{C}} = \mathbf{0}$. Otherwise, it

proceeds to compute the composite matrix:

$$\widehat{\mathbf{C}}^* = \begin{bmatrix} \widehat{\mathbf{C}}_{11} & \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top \\ \frac{1}{n}\widehat{\mathbf{X}}_2\widehat{\mathbf{X}}_1^\top & \widehat{\mathbf{C}}_{22} \end{bmatrix}. \tag{55}$$

If $\widehat{\mathbf{C}}^*$ is not positive semi–definite, the central server adjusts it to ensure this property. This adjustment is performed by spectrally decomposing $\widehat{\mathbf{C}}^*$ as $\widehat{\mathbf{C}}^* = \sum_{i=1}^r \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, and then defining $\widehat{\mathbf{C}}^*_+$ by retaining only the non–negative eigenvalues: $\widehat{\mathbf{C}}^*_+ = \sum_{i=1}^r \lambda_i \mathbb{1}_{\{\lambda_i \geq 0\}} \mathbf{v}_i \mathbf{v}_i^\top$. The final estimated covariance matrix returned by the central server is then:

$$\widehat{\mathbf{C}} = \widehat{\mathbf{C}}^*_+. \tag{56}$$

The analysis of our DCME scheme relies on concentration inequalities for random matrices, inspired by but not identical to those in [Ver18] (see Appendix F for more details). The full proof of the scheme appears in Appendix G.

# 7 Interactive Cross-Covariance Estimation

In this section, we study the DCME problem in an interactive setting, which generalizes the interactive correlation estimation explored in [HLPS19]. We assume the presence of two agents, Alice and Bob. Alice and Bob observe i.i.d. samples $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. For simplicity and consistency with prior work, we assume that the pair $\mathbf{Z} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top$ is jointly Gaussian with zero mean and covariance matrix:

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_{d_1} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^\top & \mathbf{I}_{d_2} \end{bmatrix},$$

where $\left\|\mathbf{C}_{12}\right\|_{\mathsf{op}} \leq 1$. The objective for both Alice and Bob is to estimate the cross covariance matrix $\mathbf{C}_{12}$ via rate–limited interactive communication. The correlation estimation problem in [HLPS19] corresponds to the special case where $d_1 = d_2 = 1$. Additionally, the case $(d_1 = d, d_2 = 1)$ is examined in [ST21].

More formally, consider a shared board on which Alice and Bob post their messages during communication. The interactive protocol proceeds as follows: Alice first writes message $M_1 = \mathcal{E}_1(\mathbf{X}_1)$, based on her data. Bob then responds with $M_2 = \mathcal{E}_2(\mathbf{X}_2, M_1)$. Next, Alice writes $M_3 = \mathcal{E}_3(\mathbf{X}_1, M_1, M_2)$, and this exchange continues alternately. The sequence of messages written on the board is denoted by $\Pi = (M_1, M_2, M_3, \dots)$. Ultimately, Alice and Bob compute an estimate $\widehat{\mathbf{C}}_{12} = \mathcal{D}(\Pi)$ of the cross covariance matrix $\mathbf{C}_{12}$. The communication protocol is subject to a rate constraint given by:

$$H(\Pi) = H(M_1, M_2, M_3, \dots) \leq B. \tag{57}$$

We are interested in the expected distortion of the protocol, defined as:

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}_{12}, \mathbf{C}_{12})\right] = \mathbb{E}_{\{\mathbf{Z}^{(i)}\}_{i=1}^m \sim P_{\mathbf{Z}}^{\otimes m}}\left[\left\|\widehat{\mathbf{C}}_{12} - \mathbf{C}_{12}\right\|_{\mathsf{dist}}\right], \tag{58}$$

where $\mathsf{dist}$ denotes either the operator norm or the Frobenius norm. Our goal is to characterize the minimax expected distortion, defined by:

$$\mathcal{M}^{\mathsf{int}}_{\mathsf{dist}}(m, d_1, d_2, B) := \inf_{(\Pi, \mathcal{D}): H(\Pi) \leq B} \sup_{\|\mathbf{C}_{12}\|_{\mathsf{op}} \leq 1} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}_{12}, \mathbf{C}_{12})\right]. \tag{59}$$

## 7.1 Upper bound

We'll now explain the upper bound for the distance in terms of the operator norm. The approach for the Frobenius norm is similar, so we'll omit it for brevity.

### Scenario 1: Alice has More Dimensions ($d_1 \geq d_2$)

Imagine Alice has access to more data dimensions than Bob. In this case, the process unfolds as follows:

1. **Alice's First Round:** Alice doesn't write anything on the board.
2. **Bob's Action:** Bob writes the same message he would in a distributed setting.
3. **Alice as Server:** Alice effectively acts as a central server. She uses her data and Bob's message to estimate the cross–covariance matrix.

This situation is like a distributed cross–covariance estimation where Alice can provide all her samples without any restrictions (like having an infinite communication budget, $B_1 = \infty$). Based on Theorem 4.5, Alice can estimate the cross–covariance within an error of $\frac{\varepsilon}{2}$ if:

- The number of samples, $m$, is sufficient: $m = \mathcal{O}\left(\frac{d}{\varepsilon^2}\right)$
- Bob's communication budget, $B_2$, is sufficient: $B_2 = \widetilde{\mathcal{O}}\left(\frac{d_1 d_2}{\varepsilon^2}\right)$

To get an estimate $\widehat{\mathbf{C}}_{12}$ of the true cross–covariance $\mathbf{C}_{12}$, Alice uses a covering argument (explained in Appendix A.6) to get a quantized version, $\widetilde{\mathbf{C}}_{12}$. This quantized version satisfies the condition $\left\|\widehat{\mathbf{C}}_{12} - \widetilde{\mathbf{C}}_{12}\right\|_{\text{op}} \leq \frac{\varepsilon}{2}$.

This quantization requires $\mathcal{O}\left(d_1 d_2 \log \frac{8}{\varepsilon}\right)$, (which is less than $\mathcal{O}\left(\frac{d_1 d_2}{\varepsilon^2}\right)$) bits as detailed in Equation (86) of Appendix A.6. Alice then writes these bits on the board. This ensures that both Alice and Bob can compute $\widetilde{\mathbf{C}}_{12}$, which is within an $\varepsilon$ error of the actual cross–covariance matrix $\mathbf{C}_{12}$.

The total number of bits consumed in this entire process is:

$$B = \widetilde{\mathcal{O}}\left(\frac{d_1 d_2}{\varepsilon^2}\right). \tag{60}$$

### Scenario 2: Bob has More Dimensions ($d_2 > d_1$)

If Bob has more dimensions than Alice, their roles are simply reversed, and we arrive at the same total number of bits consumed.

### Frobenius Norm Considerations

A similar line of reasoning applies when considering the Frobenius norm distance. For an interactive approximation of the cross–covariance matrix within an $\varepsilon$ error in the Frobenius norm, the following constraints on the number of samples ($m$) and the total communication budget ($B$) are sufficient:

$$m = \mathcal{O}\left(\frac{d^2}{\varepsilon^2}\right), \qquad B = \widetilde{\mathcal{O}}\left(\frac{d_1 d_2 d_{\min}}{\varepsilon^2}\right). \tag{61}$$

**Remark 7.1.** In [ST21, Theorem 6], the scenario where $d_1 = d - 1$ and $d_2 = 1$ is analyzed, asserting a communication budget of $\Theta\left(\frac{d^2}{\varepsilon^2}\right)$ for approximating the cross–covariance matrix. Contrary to this, our preceding protocol demonstrates that this can be accomplished with a substantially reduced communication budget of $\widetilde{\mathcal{O}}\left(\frac{d}{\varepsilon^2}\right)$. The root of this difference lies in a misapplication of a certain generalization of SDPI, a point we will elaborate on in the subsequent subsection.

## 7.2 Lower bound

We now proceed to prove the tightness of the upper bound (60). The tightness of the upper bound (61) can be established using a similar argument and is therefore omitted for brevity. The proof relies on the concept of *"symmetric–SDPI"* introduced in [LCV17] and further investigated in [HLPS19].

### 7.2.1 Overview of symmetric–SDPI

For a pair of random variables $(X, Y) \sim P_{XY}$, the symmetric–SDPI coefficient is defined as the minimum number $s_\infty$ such that the following inequality holds for any integer number $T$:

$$I(X;Y) - I(X;Y|U_1, \cdots, U_T) \leq s_\infty I(U_1, \cdots, U_T; X, Y) \tag{62}$$

where $U_1, \cdots, U_T$ satisfies

$$\begin{aligned} U_i &\multimap (X, U^{i-1}) \multimap Y, \quad i \in \mathbb{N} \setminus 2\,\mathbb{N}, \\ U_i &\multimap (Y, U^{i-1}) \multimap X, \quad i \in \mathbb{N} \cap 2\,\mathbb{N}. \end{aligned} \tag{63}$$

**Lemma 7.2** (Tensorization of symmetric–SDPI [HLPS19, Lemma 9.3])**.** *Let* $(X^n, Y^n) \sim \otimes_{i=1}^n P_{X_i Y_i}$. *Then* $s_\infty(X^n, Y^n) = \max_{1 \leq i \leq n} s_\infty(X_i, Y_i)$.

**Lemma 7.3** (symmetric–SDPI for Gaussian,[HLPS19, Lemma 9.4])**.** *Let* $(X, Y) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$. *Then* $s_\infty(X, Y) = \rho^2$.

**Corollary 7.4** (symmetric–SDPI for vector Gaussian)**.** *Let* $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ *be a zero mean Gaussian vector with covariance matrix* $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^\top & \mathbf{C}_{22} \end{bmatrix}$, *such that* $\mathbf{C}_{11}$ *and* $\mathbf{C}_{22}$ *are non–singular. Then*

$$s_\infty(\mathbf{X}_1, \mathbf{X}_2) = \left\| \mathbf{C}_{11}^{-1/2} \mathbf{C}_{12} \mathbf{C}_{22}^{-1/2} \right\|_{\mathsf{op}}^2.$$

**Remark 7.5.** In [ST21, Lemma 16], the symmetric–SDPI for the pair $(\mathbf{X}, Y)$ with covariance matrix $\mathbf{C} = \begin{bmatrix} \mathbf{I}_d & \rho \\ \rho^\top & 1 \end{bmatrix}$, $\rho \in \mathbb{R}^d$, is used as an intermediate step in deriving a lower bound for the DCME. The authors claim (without rigorous proof) that the symmetric–SDPI is given by $\max_i \rho_i^2$, where $\rho_i$ denotes the $i$–th coordinate of $\rho$. However, the correct symmetric–SDPI for this pair is $\|\rho\|^2$, which can be significantly larger than the claimed bound.

*Proof.* Let $\mathbf{A} = \mathbf{C}_{22}^{-1/2} \mathbf{C}_{12}^\top \mathbf{C}_{11}^{-1/2}$ and consider its singular value decomposition as $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are unitaries and $\mathbf{D}$ is diagonal with diagonal entries $(D_1, \cdots, D_r)$, where $r \leq \min\{d_1, d_2\}$. Define $\overline{\mathbf{X}}_1 = \mathbf{V}^\top \mathbf{C}_{11}^{-1/2} \mathbf{X}_1$, $\overline{\mathbf{X}}_2 = \mathbf{U}^\top \mathbf{C}_{11}^{-1/2} \mathbf{X}_2$. Observe that $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$ are in one–to–one correspondence with $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Thus $s_\infty(\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2) = s_\infty(\mathbf{X}_1, \mathbf{X}_2)$. Also it can be readily verified that

$$\begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_{d_1} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{I}_{d_2} \end{bmatrix}\right), \tag{64}$$

Consider the components of $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$ as $\overline{\mathbf{X}}_1 = (\bar{X}_{11}, \cdots, \bar{X}_{1d_1})$ and $\overline{\mathbf{X}}_2 = (\bar{X}_{21}, \cdots, \bar{X}_{2d_2})$. Then we observe that $(\bar{X}_{11}, \bar{X}_{21}), \cdots, (\bar{X}_{1r}, \bar{X}_{2r}), \{\bar{X}_{1i}\}_{i=r+1}^{d_1}, \{\bar{X}_{2i}\}_{i=r+1}^{d_2}$ are mutually independent. Thus by Lemma 7.2, we obtain:

$$s_\infty(\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2) = \max_{1 \leq i \leq r} s_\infty(\bar{X}_{1i}, \bar{X}_{2i}). \tag{65}$$

Further $(\bar{X}_{1i}, \bar{X}_{2i}) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & D_i \\ D_i & 1 \end{bmatrix}\right)$. Therefore $s_\infty(\bar{X}_{1i}, \bar{X}_{2i}) = D_i^2$. Subsequently, we have $s_\infty(\mathbf{X}_1, \mathbf{X}_2) = \max_{1\le i\le r} D_i^2 = \|\mathbf{A}\|_{\mathsf{op}}^2$. □

### 7.3 Minimax Lower bound

The argument is again based on Fano's method. Similar to the proof of Theorem 4.2, we consider the family of normal distributions $\mathcal{N}(\mathbf{0}, \mathbf{C}_v)$ with the covariance matrices $\mathbf{C}_v$ structured as:

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

where $\mathbf{D}_v$ is a matrix in $\mathbb{R}^{d_2 \times d_1}$ with $\left\|\mathbf{D}_v\right\|_{\mathsf{op}} \le 1$, and $\delta \le 1$ is a parameter which will determined subsequently. Now, we have:

$$\begin{aligned} \rho_{\mathsf{op}} &= \inf_{(v,v')\in\mathcal{V}^2, v\neq v'} \left\{ \left\| \mathbf{C}_v - \mathbf{C}_{v'} \right\|_{\mathsf{dist}} \right\} \\ &= \frac{\delta}{2} \inf_{(v,v')\in\mathcal{V}^2, v\neq v'} \left\{ \left\| \mathbf{D}_v - \mathbf{D}_{v'} \right\|_{\mathsf{dist}} \right\}, \end{aligned} \tag{66}$$

We consider the family $\{\mathbf{D}_v\}_{v\in\mathcal{V}}$ as the $\frac{1}{4}$–packing of $\mathcal{B}^{(d_1 d_2)}_{\|\|.\|\|_{\mathsf{op}}}(1)$ (see Equation (85)), under the $\|\|.\|\|_{\mathsf{op}}$ norm. Thus, $\inf_{v,v':v\neq v'} \|\mathbf{D}_v - \mathbf{D}_{v'}\|_{\mathsf{dist}} \ge \frac{1}{4}$, $\max_{v\in\mathcal{V}} \{\|\mathbf{D}_v\|_{\mathsf{op}}^2\} \le 1$, and from (86) and (92), we have $\log_2(|\mathcal{V}|) \ge 2d_1 d_2$. Invoking Fano's method yields:

$$\inf_{\Pi,\mathcal{D}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right] \ge \frac{\rho_{\mathsf{dist}}}{2} \left(1 - \frac{I(V;\Pi) + \log 2}{2d_1 d_2}\right) \tag{67}$$

Thus we need to upper bound $I(V;\Pi)$. We do this using symmetric–SDPI. Defining $\widetilde{P}_{\Pi,\mathbf{X}_1,\mathbf{X}_1} = P_{\mathbf{X}_1} P_{\mathbf{X}_2} P_{\Pi|\mathbf{X}_1,\mathbf{X}_2}$, where $P_{\mathbf{X}_i} = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_i})^{\otimes m}$ for $i = 1, 2$, we have:

$$\begin{aligned} I(V;\Pi) &\overset{(a)}{\le} \mathbb{E}_V \left[ D_{\mathrm{KL}} \left( P_\Pi^{(V)} \| \widetilde{P}_\Pi \right) \right] \\ &\overset{(b)}{\le} I(\mathbf{X}_1; \mathbf{X}_2 | V) - I(\mathbf{X}_1; \mathbf{X}_2 \mid \Pi, V), \end{aligned} \tag{68}$$

where $P_\Pi^(v)$ is the marginal distribution of $\Pi$ when $(\mathbf{X}_1, \mathbf{X}_2)$ is resulted from the pair with covariance matrix $\mathbf{C}_v$, (a) follows from [PW25, Corollary 4.2], and (b) follows from [HLPS19, Theorem 7.1]. Now, observe that $\Pi = (M_1, M_2, \cdots)$ satisfies the Markov chains in the definition of symmetric–SDPI (63)(with $U_i$ is replaced with $M_i$), thus we can further upper bound the r.h.s. of (68) as,

$$I(V;\boldsymbol{M}) \le \left( \max_v s_{\infty,v}(\mathbf{X}_1, \mathbf{X}_2) \right) I(\mathbf{X}_1, \mathbf{X}_2; \Pi | V) \le \delta^2 H(\Pi) \le \delta^2 B, \tag{69}$$

where we have used Corollary 7.4 to get $s_{\infty,v}(\mathbf{X}_1, \mathbf{X}_2) = \left\| \mathbf{C}_{11}^{-1/2} \mathbf{C}_{12} \mathbf{C}_{22}^{-1/2} \right\|_{\mathsf{op}}^2 = \delta^2 \|\mathbf{D}_v\|_{\mathsf{op}}^2 \le \delta^2$.

Substituting (69) in (67) with the choice $\delta = \sqrt{\frac{d_1 d_2}{2B}}$, we conclude:

$$\inf_{\boldsymbol{M},\mathcal{D}} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\mathcal{L}_{\mathsf{dist}}(\widehat{\mathbf{C}}, \mathbf{C})\right] = \Omega\left(\sqrt{\frac{d_1 d_2}{B}}\right). \tag{70}$$

### 7.4 Interaction Reduces the communication budget.

We now compare the total communication budgets required for cross–covariance estimation in the non–interactive and interactive settings. By Corollary 4.3 and Theorem 4.5, the total communication budget needed to estimate the cross–covariance matrix $\mathbf{C}_{12}$ up to an error of $\varepsilon$ in the non–interactive setting is $B_1 + B_2 = \widetilde{\Theta}\big(\frac{d^2}{\varepsilon^2}\big)$. In contrast, allowing interaction reduces this budget to $\widetilde{\Theta}\big(\frac{d_1 d_2}{\varepsilon^2}\big)$, which can be significantly smaller than $\widetilde{\Theta}\big(\frac{d^2}{\varepsilon^2}\big)$ when the dimensions $d_1$ and $d_2$ are imbalanced—for example, when $d_1 = 1$ and $d_2 = d-1$. Thus, interaction can significantly reduce the total communication budget for the DCCME task. This phenomenon has also been observed previously in the context of distributed nonparametric estimation [Liu23].

## 8 Conclusion

This paper rigorously investigated the fundamental limits and achievable performance for distributed covariance matrix estimation (DCME) in a feature–split setting under communication constraints. Our core contribution is the development of the Conditional Strong Data Processing Inequality (C–SDPI), a novel theoretical framework that enabled us to derive near–optimal minimax lower bounds for the DCME problem. These bounds precisely quantify the trade–offs between sample complexity, communication budgets, data dimensionality, and estimation error, highlighting the inherent constraints on accuracy. We also designed and analyzed an explicit estimation scheme that achieves these theoretical limits up to logarithmic factors, confirming the near–optimality of our approach.

Furthermore, our analysis extended to interactive settings, revealing that interaction can significantly reduce the total communication budget for cross–covariance estimation, particularly in scenarios with imbalanced agent dimensions. This work provides a comprehensive information–theoretic foundation for distributed covariance matrix estimation, offering both theoretical insights into its limits and practical, near–optimal solutions for this challenging problem.

#### Acknowledgements

We sincerely thank Mohammad Ali Maddah-Ali and Amin Gohari for their valuable comments and insightful discussions.

## References

[ABD+20] Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

[ACST23] Jayadev Acharya, Clément L Canonne, Ziteng Sun, and Himanshu Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 51133–51165. Curran Associates, Inc., 2023.

[ACT20a] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

[ACT20b] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints i: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020.

[AG76] Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, pages 925–939, 1976.

[AGKN13] Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *arXiv preprint arXiv:1304.6133*, 2013.

[AJN22] Venkat Anantharam, Varun Jog, and Chandra Nair. Unifying the brascamp-lieb inequality and the entropy power inequality. *IEEE Transactions on Information Theory*, 68(12):7665–7684, 2022.

[AKB+22] Corinne G Allaart, Björn Keyser, Henri Bal, and Aart Van Halteren. Vertical split learning-an exploration of predictive performance in medical and other use cases. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[AM05] Karim M Abadir and Jan R Magnus. *Matrix algebra*, volume 1. Cambridge University Press, 2005.

[BCL05] Zheng-Jian Bai, Raymond H Chan, and Franklin T Luk. Principal component analysis for distributed data sets with updating. In *International Workshop on Advanced Parallel Processing Technologies*, pages 471–483. Springer, 2005.

[BCÖ20] Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020.

[BGM+16] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. 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*, pages 1011–1020, 2016.

[BHO20] Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower bounds for learning distributions under communication constraints via fisher information. *Journal of Machine Learning Research*, 21(236):1–30, 2020.

[Bil99] Patrick Billingsley. *Convergence of probability measures; 2nd ed.* Wiley series in probability and statistics. Wiley, Hoboken, NJ, 1999.

[BK13] Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. *arXiv preprint arXiv:1301.7362*, 2013.

[BKLW14] Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David Woodruff. Improved distributed principal component analysis. *arXiv preprint arXiv:1408.5823*, 2014.

[BL08a] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[BL+08b] Peter J Bickel, Elizaveta Levina, et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[Bun34] L. N. H. Bunt. *Bijdrage tot de theorie der convexe puntverzamelingen*. PhD thesis, Rijksuniversiteit Groningen, Amsterdam, 1934. Doctoral dissertation.

[Car11] Constantin Carathéodory. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

[CGT12] Richard Y Chen, Alex Gittens, and Joel A Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

[CIR+93] Joel E Cohen, Yoh Iwasa, Gh Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179:211–235, 1993.

[Cou13] Thomas A Courtade. Outer bounds for multiterminal source coding via a strong data processing inequality. In *2013 IEEE International Symposium on Information Theory*, pages 559–563. IEEE, 2013.

[Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[CPW15] Flavio P Calmon, Yury Polyanskiy, and Yihong Wu. Strong data processing inequalities in power-constrained gaussian channels. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2558–2562. IEEE, 2015.

[CRS94] Man-Duen Choi, Mary Beth Ruskai, and Eugene Seneta. Equivalence of certain entropy contraction coefficients. *Linear algebra and its applications*, 208:29–36, 1994.

[CRZ13] T Tony Cai, Zhao Ren, and Harrison H Zhou. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143, 2013.

[CW24] T Tony Cai and Hongji Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *Journal of Machine Learning Research*, 25(37):1–63, 2024.

[CZZ+10] T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE, 2013.

[DKPN00] Jörg Dahmen, Daniel Keysers, Michael Pitz, and Hermann Ney. Structured covariance matrices for statistical image object recognition. In *Mustererkennung 2000*, pages 99–106. Springer, 2000.

[DMLM03] Pierre Del Moral, Michel Ledoux, and Laurent Miclo. On contraction properties of markov kernels. *Probability theory and related fields*, 126(3):395–420, 2003.

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and ising undirected graphical models. *Electronic Journal of Statistics*, 14:2338–2361, 2020.

[Dob56a] Roland L Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[Dob56b] Roland L Dobrushin. Central limit theorem for nonstationary markov chains. ii. *Theory of Probability & Its Applications*, 1(4):329–383, 1956.

[DV83] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[EK08] Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6), December 2008.

[ES99] William S Evans and Leonard J Schulman. Signal propagation and noisy circuits. *IEEE Transactions on Information Theory*, 45(7):2367–2373, 1999.

[FB07] Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

[FCG10] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. Consensus-based distributed linear support vector machines. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 35–46, 2010.

[GMN14] Ankit Garg, Tengyu Ma, and Huy L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. 27, 2014.

[GN14] Yanlin Geng and Chandra Nair. The capacity region of the two-receiver gaussian vector broadcast channel with private and common messages. *IEEE Transactions on Information Theory*, 60(4):2087–2104, 2014.

[GO62] SG Ghurye and Ingram Olkin. A characterization of the multivariate normal distribution. *The Annals of Mathematical Statistics*, 33(2):533–541, 1962.

[HLPL06] Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[HLPS19] Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. Communication complexity of estimating correlations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 792–803, 2019.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[HÖW18] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pages 3163–3188. PMLR, 2018.

[HS19] Uri Hadar and Ofer Shayevitz. Distributed estimation of gaussian correlations. *IEEE Transactions on Information Theory*, 65(9):5323–5338, 2019.

[JH85] Charles R Johnson and Roger A Horn. *Matrix analysis*. Cambridge university press Cambridge, 1985.

[KA12] Sudeep Kamath and Venkat Anantharam. Non-interactive simulation of joint distributions: The hirschfeld-gebelein-rényi maximal correlation and the hypercontractivity ribbon. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1057–1064. IEEE, 2012.

[KGK+17] Hyeji Kim, Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Discovering potential correlations via hypercontractivity. *Advances in Neural Information Processing Systems*, 30, 2017.

[KVW14] Ravi Kannan, Santosh Vempala, and David Woodruff. Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory*, pages 1040–1057. PMLR, 2014.

[LCCV18] Jingbo Liu, Thomas A Courtade, Paul W Cuff, and Sergio Verdú. A forward-reverse brascamp-lieb inequality: Entropic duality and gaussian optimality. *Entropy*, 20(6):418, 2018.

[LCV14] Jingbo Liu, Paul Cuff, and Sergio Verdú. Key capacity with limited one-way communication for product sources. In *2014 IEEE International Symposium on Information Theory*, pages 1146–1150. IEEE, 2014.

[LCV17] Jingbo Liu, Paul Cuff, and Sergio Verdú. Secret key generation with limited interaction. *IEEE Transactions on Information Theory*, 63(11):7358–7381, 2017.

[Liu23] Jingbo Liu. A few interactions improve distributed nonparametric estimation, optimally. *IEEE Transactions on Information Theory*, 69(12):7867–7886, 2023.

[LLZ09] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(3), 2009.

[LRV08] Yumao Lu, Vwani Roychowdhury, and Lieven Vandenberghe. Distributed parallel support vector machines in strongly connected networks. *IEEE Transactions on Neural Networks*, 19(7):1167–1178, 2008.

[LW03] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

[Mic97] Laurent Miclo. *Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies*. Springer, 1997.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[MZ15] Anuran Makur and Lizhong Zheng. Bounds between contraction coefficients. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1422–1429. IEEE, 2015.

[NRRW11] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *arXiv preprint arXiv:1106.5730*, 2011.

[NVG+06] Angel Navia-Vázquez, D Gutierrez-Gonzalez, Emilio Parrado-Hernández, and JJ Navarro-Abellan. Distributed support vector machines. *IEEE Transactions on Neural Networks*, 17(4):1091, 2006.

[PW16] Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

[PW17] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

[PW25] Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

[QOSG02] Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist. Principal component analysis for dimension reduction in massive distributed data sets. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, volume 1318, page 1788, 2002.

[Rag13] Maxim Raginsky. Logarithmic sobolev inequalities and strong data processing theorems for discrete channels. In *2013 IEEE International Symposium on Information Theory*, pages 419–423. IEEE, 2013.

[Rag16] Maxim Raginsky. Strong data processing inequalities and $\phi$-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

[Rud87] W. Rudin. *Real and Complex Analysis*. Mathematics series. McGraw-Hill, 1987.

[SFKM17] Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *International conference on machine learning*, pages 3329–3337. PMLR, 2017.

[ST21] KR Sahasranand and Himanshu Tyagi. Communication complexity of distributed high dimensional correlation testing. *IEEE Transactions on Information Theory*, 67(9):6082–6095, 2021.

[SVL13] Ramanan Subramanian, Badri N Vellambi, and Ingmar Land. An improved bound on information loss due to finite block length in a gaussian line network. In *2013 IEEE International Symposium on Information Theory*, pages 1864–1868. IEEE, 2013.

[SVVZ23] Botond Szabó, Lasse Vuursteen, and Harry Van Zanten. Optimal high-dimensional and nonparametric distributed testing under communication constraints. *The Annals of Statistics*, 51(3):909–934, 2023.

[SZZ+19] Meng Shen, Jie Zhang, Liehuang Zhu, Ke Xu, and Xiangyun Tang. Secure svm training over vertically-partitioned datasets using consortium blockchain for vehicular social networks. *IEEE Transactions on Vehicular Technology*, 69(6):5773–5783, 2019.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VN37] John Von Neumann. *Some matrix-inequalities and metrization of matric space*. 1937.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[WCX+20] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. Privacy preserving vertical federated learning for tree-based models. *arXiv preprint arXiv:2008.06170*, 2020.

[Wit75] Hans S Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

[WP09] Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768, 2009.

[XR15] Aolin Xu and Maxim Raginsky. Converses for distributed estimation via strong data processing inequalities. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2376–2380. IEEE, 2015.

[YFC+19] Kai Yang, Tao Fan, Tianjian Chen, Yuanming Shi, and Qiang Yang. A quasi-newton method based vertical federated learning framework for logistic regression. *arXiv preprint arXiv:1912.00513*, 2019.

[ZDJW13] Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

[ZWB+07] Kaihua Zhu, Hao Wang, Hongjie Bai, Jian Li, Zhihuan Qiu, Hang Cui, and Edward Chang. Parallelizing support vector machines on distributed computers. *Advances in neural information processing systems*, 20, 2007.

[ZWSL10] Martin Zinkevich, Markus Weimer, Alexander J Smola, and Lihong Li. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 4, page 4. Citeseer, 2010.

# A Some Preliminary Lemmas, Corollaries, and Propositions

## A.1 Proof of Lemma 2.2

*Proof.* Let $\mathbf{D}$ denote $\mathbb{E}\left[\mathbf{A}^\top\mathbf{B}\mathbf{A}\right]$. Consider an arbitrary signed permutation matrix $\tilde{\mathbf{A}} \in \mathcal{P}_d$. Since $\mathcal{P}_d$ forms a group under matrix multiplication, the random matrix $\mathbf{A}\tilde{\mathbf{A}}$ is also uniformly distributed over $\mathcal{P}_d$. Consequently, for any $\tilde{\mathbf{A}} \in \mathcal{P}_d$, we have:

$$\mathbf{D} = \mathbb{E}\left[\mathbf{A}^\top\mathbf{B}\mathbf{A}\right] = \mathbb{E}\left[\tilde{\mathbf{A}}^\top\mathbf{A}^\top\mathbf{B}\mathbf{A}\tilde{\mathbf{A}}\right] = \tilde{\mathbf{A}}^\top\mathbf{D}\tilde{\mathbf{A}} \tag{71}$$

We now demonstrate that $\mathbf{D}$ must be a scalar multiple of the identity matrix. For any distinct pair of indices $(i, j)$, define the matrices $\mathbf{A}_+^{(ij)} = \mathbf{e}_i\mathbf{e}_j^\top + \mathbf{e}_j\mathbf{e}_i^\top + \sum_{k\neq i,j}\mathbf{e}_k\mathbf{e}_k^\top$ and $\mathbf{A}_-^{(ij)} = \mathbf{e}_i\mathbf{e}_j^\top - \mathbf{e}_j\mathbf{e}_i^\top + \sum_{k\neq i,j}\mathbf{e}_k\mathbf{e}_k^\top$. Both $\mathbf{A}_+^{(ij)}$ and $\mathbf{A}_-^{(ij)}$ are signed permutation matrices, and thus $(\mathbf{A}_\pm^{(ij)})^\top\mathbf{D}\mathbf{A}_\pm^{(ij)} = \mathbf{D}$. This implies the following relationships:

$$\begin{aligned} D_{ii} &= \mathbf{e}_i^\top\mathbf{D}\mathbf{e}_i = \mathbf{e}_i^\top(\mathbf{A}_+^{(ij)})^\top\mathbf{D}\mathbf{A}_+^{(ij)}\mathbf{e}_i = \mathbf{e}_j^\top\mathbf{D}\mathbf{e}_j = D_{jj} \\ D_{ij} &= \mathbf{e}_i^\top\mathbf{D}\mathbf{e}_j = \mathbf{e}_i^\top(\mathbf{A}_+^{(ij)})^\top\mathbf{D}\mathbf{A}_+^{(ij)}\mathbf{e}_j = \mathbf{e}_j^\top\mathbf{D}\mathbf{e}_i = D_{ji} \\ D_{ij} &= \mathbf{e}_i^\top\mathbf{D}\mathbf{e}_j = \mathbf{e}_i^\top(\mathbf{A}_-^{(ij)})^\top\mathbf{D}\mathbf{A}_-^{(ij)}\mathbf{e}_j = -\mathbf{e}_j^\top\mathbf{D}\mathbf{e}_i = -D_{ji} \end{aligned} \tag{72}$$

From these relations, it follows that the off–diagonal entry $D_{ij}$ must be zero, and all diagonal entries are equal. Therefore, $\mathbf{D}$ is a scalar multiple of the identity matrix; that is, $\mathbf{D} = \alpha \mathbf{I}_d$ for some scalar $\alpha$. Recalling that $\alpha \mathbf{I}_d = \mathbf{D} = \mathbb{E}[\mathbf{A}^\top \mathbf{B} \mathbf{A}]$, taking the trace of both sides of this identity yields:

$$\alpha d = \mathsf{Tr}\left\{ \mathbb{E}[\mathbf{A}^\top \mathbf{B} \mathbf{A}] \right\} = \mathbb{E}\left[ \mathsf{Tr}\left\{ \mathbf{A}\mathbf{A}^\top \mathbf{B} \right\} \right] = \mathsf{Tr}\left\{ \mathbf{B} \right\}, \tag{73}$$

where we have utilized the orthogonality property of signed permutation matrices (i.e., $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$). Consequently, $\alpha = \frac{\mathsf{Tr}\{\mathbf{B}\}}{d}$, completing the proof. □

## A.2 A Lemma from Linear Algebra

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

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \qquad \mathbf{A}^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i. \tag{74}$$

We write:

$$\begin{aligned} \frac{1}{\sqrt{2}} \mathbf{B} \begin{bmatrix} \pm \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} &= \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \pm \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{A}\mathbf{v}_i \\ \pm \mathbf{A}^\top \mathbf{u}_i \end{bmatrix} \\ &= \frac{\sigma_i}{\sqrt{2}} \begin{bmatrix} \mathbf{u}_i \\ \pm \mathbf{v}_i \end{bmatrix} \\ &= \frac{\pm \sigma_i}{\sqrt{2}} \begin{bmatrix} \pm \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}. \end{aligned} \tag{75}$$

This completes the proof. □

## A.3 Some Properties of Sub–Gaussian Random Variables

To study some properties of sub–Gaussian random variables, familiarity with another family of random variables is necessary. This family extends the class of sub–Gaussian random variables and is called sub–Gamma random variables.

**Definition A.2** ([BLM13, Chapter 2.4])**.** A random variable $X$ is called $(\sigma, \alpha)$–sub–Gamma, if:

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

for all $\lambda$ such that: $|\lambda| < \frac{1}{\alpha}$.

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

**Lemma A.3** ([BLM13])**.** *Consider an independent sequence $\{X_i\}_{i=1}^m$ of random variables,*

- *if $X_i$, $i \in [m]$ is a $\sigma_i$–sub–Gaussian random variable, then $\sum_{i=1}^n X_i$ is $\sqrt{\sum_{i=1}^n \sigma_i^2}$–sub–Gaussian.*
- *if $X_i$, $i \in [m]$ is a $(\sigma_i, \alpha_i)$–sub–Gamma random variable, then $\sum_{i=1}^n X_i$ is $\Big(\sqrt{\sum_{i=1}^n \sigma_i^2}, \max_i\{\alpha_i\}\Big)$–sub–Gamma.*

**Lemma A.4.** *Any $(\sigma, \alpha)$–sub–Gamma random variable $X$ satisfies the following inequality:*

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

*Proof.* Some variations of this lemma are presented in different papers. For completeness, we provide a proof here. We write:

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

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 condition $0 < \lambda < \frac{1}{\alpha}$. Thus:

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

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 [BLM13, Corollary 2.6])**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of centered sub–Gamma random variables with the same parameters $(\sigma, \alpha)$. Then:*

$$\mathbb{E}\Big[\max_{i\in[n]} X_i\Big] \leq \sigma\sqrt{2\ln(n)} + \alpha\ln(n). \tag{76}$$

**Lemma A.6.** *Assume that $X$ and $Y$ are centered sub–Gaussian random variables with parameters $\sigma_1$ and $\sigma_2$, 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:

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(XY-\mathbb{E}[XY])}\right] &= 1 + \lambda\,\mathbb{E}\left[(XY - \mathbb{E}[XY])\right] + \sum_{k=2}^{+\infty} \frac{\lambda^k\,\mathbb{E}\left[(XY - \mathbb{E}[XY])^k\right]}{k!} \\
&= 1 + \sum_{k=2}^{+\infty} \frac{\lambda^k}{k!}\,\mathbb{E}\left[(XY - \mathbb{E}[XY])^k\right] \\
&\leq 1 + \sum_{k=2}^{+\infty} \frac{|\lambda|^k}{k!}\,\mathbb{E}\left[|XY - \mathbb{E}[XY]|^k\right] \\
&= 1 + \sum_{k=2}^{+\infty} \frac{|\lambda|^k}{k!}\|XY - \mathbb{E}[XY]\|_k^k && (77) \\
&\leq 1 + \sum_{k=2}^{+\infty} \frac{|\lambda|^k}{k!}\left(\|XY\|_k + \|\,\mathbb{E}[XY]\|_k\right)^k && (78) \\
&= 1 + \sum_{k=2}^{+\infty} \frac{|\lambda|^k}{k!}\left(\left(\mathbb{E}\left[|XY|^k\right]\right)^{\frac{1}{k}} + |\,\mathbb{E}[XY]|\right)^k \\
&\leq 1 + \sum_{k=2}^{+\infty} \frac{|\lambda|^k}{k!}\left(\left(\mathbb{E}\left[X^{2k}\right]\,\mathbb{E}\left[Y^{2k}\right]\right)^{\frac{1}{2k}} + \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}\right)^k && (79) \\
&\leq 1 + \sum_{k=2}^{+\infty} (|\lambda|\sigma_1\sigma_2)^k \frac{\left(\left(2^{k+1}k!\right)^{\frac{1}{k}} + 1\right)^k}{k!} && (80) \\
&\leq 1 + \sum_{k=2}^{+\infty} (|\lambda|\sigma_1\sigma_2)^k.2.(2.5)^k && (81) \\
&= 1 + \frac{25(\lambda\sigma_1\sigma_2)^2}{2(1 - 2.5|\lambda|\sigma_1\sigma_2)} && (82) \\
&\leq \exp\left(\frac{25(\lambda\sigma_1\sigma_2)^2}{2(1 - 2.5|\lambda|\sigma_1\sigma_2)}\right), && (83)
\end{aligned}$$

where:

- in (77), for a random variable $Z$, $\|Z\|_k := \mathbb{E}^{1/k}[|Z|^k]$ is the $L_k$ norm of the random variable $Z$,
- (78) follows directly from the application of Minkowski's inequality (also known as the triangle inequality) to the $L_k$ norm.
- (79) follows from Cauchy–Schwarz inequality,

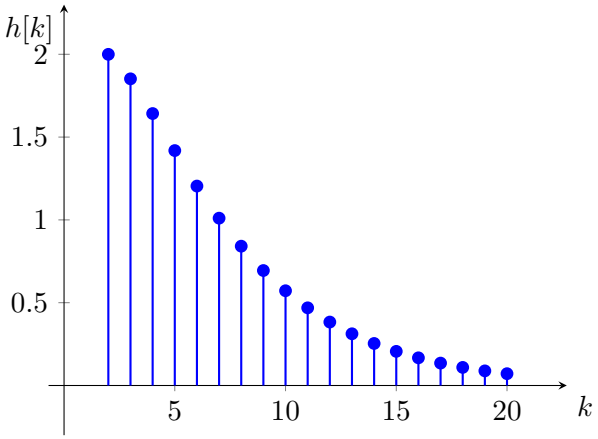


Figure 2: Diagram of the function $h[k] := \frac{\left(\left(2^{k+1}k!\right)^{\frac{1}{k}}+1\right)^k}{(2.5)^k k!}$.

- in (80), we use the following upper bound for the $2k$–th moment of a $\sigma$–sub–Gaussian random variable $Z$ (see [BLM13, Theorem 2.1]):

$$\mathbb{E}[Z^{2k}] \leq 2(2\sigma^2)^k k!,$$

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

**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 $\sigma_1$–sub–Gaussian and $\sigma_2$–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.* From Lemma A.6, we know 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:

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m (X_iY_i - \mathbb{E}[X_iY_i]) \geq 10\sigma_1\sigma_2 t\right] &= \mathbb{P}\left[\sum_{i=1}^m Z_i \geq 10m\sigma_1\sigma_2 t\right] \\
&\leq \exp\left(\frac{-100m^2\sigma_1^2\sigma_2^2t^2}{2(25\sigma_1^2\sigma_2^2m + 25\sigma_1^2\sigma_2^2mt)}\right) \\
&= \exp\left(\frac{-2mt^2}{1+t}\right) \\
&\leq \exp\left(-m.\min\{t, t^2\}\right).
\end{aligned} \tag{84}$$

□

## A.4 An Important Relation Between the Packing and the Covering Numbers of a Set

The packing and covering numbers are defined in Section 2.3. There is an important relationship between the packing and covering numbers of a set, stated in the following lemma:

**Lemma A.8** ([Wai19, 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.5 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** ([Ver18, Exercise 4.4.3]). *Let $\mathbf{A}$ be a $m \times n$ matrix. We define the sets $\mathcal{S}^{m-1} = \{\mathbf{u} \in \mathbb{R}^{m-1} : \|\mathbf{u}\| = 1\}$ and $\mathcal{S}^{n-1} = \{\mathbf{v} \in \mathbb{R}^{n-1} : \|\mathbf{v}\| = 1\}$. We fix an arbitrary $\epsilon > 0$ and denote an $\epsilon$–covering set of $\mathcal{S}^{m-1}$ by $\mathcal{N}_\epsilon^{(m)}$ and an $\epsilon$–covering set of $\mathcal{S}^{n-1}$ by $\mathcal{N}_\epsilon^{(n)}$. We have:*

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

## A.6 Packing and Covering in Matrix Spaces

Consider the family of matrices defined as follows:

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

We vectorize each member of this family as:

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

We convert the $\mathsf{dist}$ 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:

$$\mathcal{B}^{(mn)}_{\|\!|.|\!\|_{\mathsf{op}}}(r) = \left\{ \mathbf{x} \in \mathbb{R}^{mn} : \|\!|\mathbf{x}|\!\|_{\mathsf{op}} \leq r \right\} = \left\{ \mathrm{vec}(\mathbf{A}) : \mathbf{A} \in \mathbb{R}^{m \times n}, \|\mathbf{A}\|_{\mathsf{op}} \leq r \right\}. \tag{85}$$

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

- Consider the case $\mathsf{dist} = \mathsf{op}$, in this case, from [Wai19, Lemma 5.7], we have:

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

  From Lemma A.8 we conclude:

$$\left(\frac{r}{\epsilon}\right)^{mn} \leq \mathcal{N}(\mathcal{B}^{(mn)}_{\|\!|.|\!\|_{\mathsf{op}}}(r), \|\!|.|\!\|_{\mathsf{op}}, \epsilon) \leq \mathcal{M}(\mathcal{B}^{(mn)}_{\|\!|.|\!\|_{\mathsf{op}}}(r), \|\!|.|\!\|_{\mathsf{op}}, \epsilon) \leq \mathcal{N}(\mathcal{B}^{(mn)}_{\|\!|.|\!\|_{\mathsf{op}}}(r), \|\!|.|\!\|_{\mathsf{op}}, \frac{\epsilon}{2}) \leq \left(1 + \frac{4r}{\epsilon}\right)^{mn} \tag{86}$$

### A.6.1 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 $\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{op}}}(r)$ . 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:

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

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

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

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

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

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

where $\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}; \epsilon) = \{\mathrm{vec}(\mathbf{B}) : \left\|\mathbf{B} - \mathbf{A}\right\|_{\mathsf{F}} \leq \epsilon\}$ is the Frobenius ball with center $\mathbf{A}$ and radius $\epsilon$. 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 [Ver18, Theorem 7.3.1] that $\mathbb{E}\left[\|\boldsymbol{G}\|_{\mathsf{op}}\right] \leq \frac{1}{2}$. Thus Markov inequality yields:

$$\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{op}}}(1)\right] = \mathbb{P}\left[\|\boldsymbol{G}\|_{\mathsf{op}} \leq 1\right] = 1 - \mathbb{P}\left[\|\boldsymbol{G}\|_{\mathsf{op}} > 1\right] \geq 1 - \frac{1}{2} = \frac{1}{2}. \tag{87}$$

On the other side, union bound gives:

$$\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{op}}}(1)\right] \leq \sum_{i=1}^{M} \mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}_i; \epsilon)\right]. \tag{88}$$

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

$$\begin{aligned}
\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}_i; \epsilon)\right] &= \int_{\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)} \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} \exp\left(-2(\sqrt{m}+\sqrt{n})^2 \left\|\mathbf{G} - \mathbf{A}_i\right\|_{\mathsf{F}}^2\right) d\mathbf{G} \\
&\leq \int_{\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}} \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} d\mathbf{G} \\
&= \left(\frac{4(\sqrt{m}+\sqrt{n})^2}{2\pi}\right)^{\frac{mn}{2}} \mathsf{Vol}\left(\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}_i; \epsilon)\right) \\
&= \left(\frac{2\epsilon^2(\sqrt{m}+\sqrt{n})^2}{\pi}\right)^{\frac{mn}{2}} \mathsf{Vol}\left(\mathcal{B}^{(mn)}_{\|\|\|.\|\|\|_{\mathsf{F}}}(\mathbf{A}_i; 1)\right),
\end{aligned} \tag{89}$$

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

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

Using the bound $\Gamma(1+x) >> (\frac{x}{e})^x$, in (89), we obtain:

$$\begin{aligned}\mathbb{P}\left[\mathrm{vec}(\boldsymbol{G}) \in \mathcal{B}^{(mn)}_{|||.|||_{\mathsf{F}}}(\mathbf{A}_i;\epsilon)\right] &\leq \left(\frac{4e\epsilon^2(\sqrt{m}+\sqrt{n})^2}{mn}\right)^{\frac{mn}{2}} \\ &\leq \left(\frac{16e\epsilon^2\max\{m,n\}}{mn}\right)^{\frac{mn}{2}} \\ &= \left(\frac{16e\epsilon^2}{\min\{m,n\}}\right)^{\frac{mn}{2}}.\end{aligned} \tag{90}$$

Putting (87), (88) and (90) together yields:

$$M \geq \frac{1}{2}\left(\frac{\min\{m,n\}}{16e\epsilon^2}\right)^{\frac{mn}{2}} \geq \left(\frac{\sqrt{\min\{m,n\}}}{8\sqrt{e}\epsilon}\right)^{mn} \geq \left(\frac{\sqrt{\min\{m,n\}}}{14\epsilon}\right)^{mn}. \tag{91}$$

This concludes the proof. □

From Lemma A.10, we conclude that:

$$\mathcal{M}(\mathcal{B}^{(mn)}_{|||.|||_{\mathsf{op}}}(r), |||.|||_{\mathsf{F}}, \epsilon) \geq \left(\frac{r\cdot\sqrt{\min\{m,n\}}}{14\epsilon}\right)^{mn}. \tag{92}$$

# B Proof of Lemma 3.7

Here, we state the proof of Lemma 3.7.

*Proof.* Applying the chain rule for KL–divergence yields the following sequence of equalities:

$$\begin{aligned}D_{\mathrm{KL}}(Q_{Y_1,Y_2|V}\|P_{Y_1|V}P_{Y_2|V}|P_V) &\overset{(a)}{=} D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2|Y_1,V}\|P_{Y_2|V}|Q_{Y_1,V}) \\ &\overset{(b)}{=} D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2,X_1|Y_1,V}\|Q_{X_1|Y_1,V}P_{Y_2|V}|Q_{Y_1,V}) \\ &\quad - D_{\mathrm{KL}}(Q_{X_1|Y_1,Y_2,V}\|Q_{X_1|Y_1,V}|Q_{Y_1,Y_2,V}) \\ &\overset{(c)}{=} D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2,X_1|Y_1,V}\|Q_{X_1|Y_1,V}P_{Y_2|V}|Q_{Y_1,V}) \\ &\quad - I_Q(X_1;Y_2|Y_1,V) \\ &\overset{(d)}{=} D_{\mathrm{KL}}(Q_{Y_1|V}\|P_{Y_1|V}|P_V) + D_{\mathrm{KL}}(Q_{Y_2|X_1,Y_1,V}\|P_{Y_2|V}|Q_{X_1,Y_1,V}) \\ &\quad - I_Q(X_1;Y_2|Y_1,V),\end{aligned} \tag{93}$$

where (a) follows from the chain rule for KL–divergence applied to $D_{\mathrm{KL}}(Q_{Y_1,Y_2|V}\|P_{Y_1|V}P_{Y_2|V}|P_V)$, (b) follows from the chain rule for KL–divergence applied to $D_{\mathrm{KL}}(Q_{Y_2,X_1|Y_1,V}\|Q_{X_1|Y_1,V}P_{Y_2|V}|Q_{Y_1,V})$, (c) follows from the definition of conditional mutual information, and (d) follows from the chain rule for KL–divergence applied to $D_{\mathrm{KL}}(Q_{Y_2,X_1|Y_1,V}\|Q_{X_1|Y_1,V}P_{Y_2|V}|Q_{Y_1,V})$ and the fact that $D_{\mathrm{KL}}(Q_{X_1|Y_1,V}\|Q_{X_1|Y_1,V}|Q_{Y_1,V}) = 0$.

The final identity follows from the fact that the product channel $T_{Y_1|X_1,V}T_{Y_2|X_2,V}$ induces the Markov chain $Y_1 \multimap (X_1, V) \multimap Y_2$. This implies:

$$D_{\mathrm{KL}}(Q_{Y_2|X_1,Y_1,V}\|P_{Y_2|V}|Q_{X_1,Y_1,V}) = D_{\mathrm{KL}}(Q_{Y_2|X_1,V}\|P_{Y_2|V}|Q_{X_1,V}).$$

□

# C Gaussian Optimality

The goal of this section is to prove Lemma 3.10, which asserts the Gaussian optimality for the optimization problem (18). For technical reason, we first consider a smoothed version of the optimization problem (18) and show the Gaussian optimality for the smoothed version. Then we argue that the desired result can be deduced from the result for the smoothed version.

For a positive value $\delta$, define

$$\mathsf{g}_\delta(Q_{\mathbf{X}}) := \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}}) - D_{\mathrm{KL}}(Q_{\mathbf{Y}|V} * \mathcal{N}(0, \delta\mathbf{I})\|P_{\mathbf{Y}|V}|P_V). \tag{94}$$

where $P_{\mathbf{X}} = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{X}}}), P_{\mathbf{Y}} = P_{\mathbf{Y}|V} = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{Y}}})$ and $*$ represent the *convolution* operator. That is the output $\mathbf{Y}$ is smoothed by adding a noise $\sqrt{\delta}\mathbf{W}$ (where $\mathbf{W} \sim \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{Y}}})$) which is independent from all other random variables.

**Lemma C.1.** *Let $\mathcal{F}_{\mathsf{G}}$ be the set of centered normal distribution $Q_{\mathbf{X}}$ with the property $D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})$ is finite. Then:*

$$\inf_{\substack{Q_{\mathbf{X}}:\mathbb{E}_Q[\|\mathbf{X}\|^2]\le t\\ D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty}} \mathsf{g}_\delta(Q_{\mathbf{X}}) = \inf_{\substack{Q_{\mathbf{X}}:\mathbb{E}_Q[\|\mathbf{X}\|^2]\le t\\ Q_{\mathbf{X}}\in\mathcal{F}_{\mathsf{G}}}} \mathsf{g}_\delta(Q_{\mathbf{X}}). \tag{95}$$

The proof of Lemma 3.10 is divided to three steps:

1. **Existence of an optimizer**: We first consider a constrained variation of the optimization (95) and show that there exists a minimizer for which infimum is attained.
2. **Gaussian optimality for the optimizer**: In the second step, we prove that any minimizer from the previous step should be a centered normal distribution. From here, we deduce Lemma C.1. The argument closely follows the doubling–trick argument in [GN14], in which we will show that the minimizer should be rotationally invariant, thus it should be Gaussian distribution.
3. We argue that Lemma C.1 yields Lemma 3.10.

## C.1 Existence of optimizer

Consider the following constrained version of the optimization (95),

$$\mathsf{v}_\delta(t) := \inf_{\substack{Q_{\mathbf{X}}:\mathbb{E}_Q[\|\mathbf{X}\|^2]\le t,\\ D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty}} \mathsf{g}_\delta(Q_{\mathbf{X}}). \tag{96}$$

Further, let $\mathsf{G}_\delta(Q_{\mathbf{X}})$ be the convex envelop of the function $\mathsf{g}_\delta(Q_{\mathbf{X}})$. Then it can be computed using the following expression

$$\begin{aligned}\mathsf{G}_\delta(Q_{\mathbf{X}}) &:= \inf_{n\ge 1} \inf_{\substack{\{p_i\}_{i=1}^n, \{Q_{\mathbf{X}}^{(i)}\}_{i=1}^n: p_i\ge 0, \sum_i p_i = 1\\ \sum_i p_i Q_{\mathbf{X}}^{(i)} = Q_{\mathbf{X}}}} \sum_{i=1}^n p_i \mathsf{g}_\delta(Q_{\mathbf{X}}^{(i)})\\ &= \inf_{\widetilde{Q}_{U\mathbf{X}}:\widetilde{Q}_{\mathbf{X}}=Q_{\mathbf{X}}} \mathbb{E}_U\left[\mathsf{g}_\delta(\widetilde{Q}_{\mathbf{X}|U}(.|U))\right],\end{aligned} \tag{97}$$

where $U$ is an auxiliary random variable.

The main statement of this section is the following:

**Proposition C.2.** *Let:*

$$\begin{aligned}\mathsf{V}_\delta(t) &:= \inf_{\substack{Q_{\mathbf{X}}:\mathbb{E}_Q[\|\mathbf{X}\|^2]\le t,\\ D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty}} \mathsf{G}_\delta(Q_{\mathbf{X}}) \\ &= \inf_{\substack{Q_{U\mathbf{X}}:\mathbb{E}_Q[\|\mathbf{X}\|^2]\le t,\\ D_{\mathrm{KL}}(Q_{\mathbf{X}}\|P_{\mathbf{X}})<\infty}} \mathbb{E}_U\left[\mathsf{g}_\delta\big(Q_{\mathbf{X}|U}(.|U)\big)\right].\end{aligned} \tag{98}$$

*Then there exists a* binary *random variable* $U \in \{1,2\}$ *and a probability measure* $Q^*_{U\mathbf{X}}$ *such that:*

$$\begin{aligned}\mathbb{E}_{U\sim Q^*_U}\left[\mathsf{g}_\delta(Q^*_{\mathbf{X}|U}(.|U))\right] &= \mathsf{V}_\delta(t),\\ \mathbb{E}_{Q^*}[\mathbf{X}|U] &= 0,\\ \mathbb{E}_{Q^*}[\|\mathbf{X}\|^2] &\le t,\\ D_{\mathrm{KL}}(Q^*_{\mathbf{X}}\|P_{\mathbf{X}}) &< \infty.\end{aligned} \tag{99}$$

*In other words,* $Q^*_{U\mathbf{X}}$ *is a minimizer. Further,* $\mathbf{X}$ *is a centered random variable under* $Q^*_{\mathbf{X}}$.

We defer the proof of this proposition to the end of this section. We now investigate some properties of the functions $\mathsf{g}_\delta$ and $\mathsf{v}_\delta$ that will be used in the proof of Proposition C.2.

The function $\mathsf{g}_\delta(Q_{\mathbf{X}})$ can be represented using differential entropy as follows:

$$\begin{aligned}\mathsf{g}_\delta(Q_{\mathbf{X}}) &= h_Q(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h_Q(\mathbf{X}) + \frac{1}{2}\left(\mathbb{E}\left[\mathsf{s}^\star\|\mathbf{X}\|^2 - \|\mathbf{Y}+\sqrt{\delta}\mathbf{W}\|^2\right]\right) + C_1\\ &= h_Q(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h_Q(\mathbf{X}) + \frac{1}{2}\left(\mathbb{E}\left[\mathbf{X}^\top(\mathsf{s}^\star\mathbf{I} - \mathbf{A}_V\mathbf{A}_V^\top)\mathbf{X}\right] - \delta d_{\mathbf{Y}}\right) + C_1\\ &= h_Q(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h_Q(\mathbf{X}) + \frac{1}{2}\left(\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top(\mathsf{s}^\star\mathbf{I} - \mathbb{E}_V[\mathbf{A}_V\mathbf{A}_V^\top])\mathbf{X}\right]\right) + C\\ &= h_Q(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h_Q(\mathbf{X}) + \frac{1}{2}\,\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] + C,\end{aligned} \tag{100}$$

where $\mathbf{X}\sim Q_{\mathbf{X}}$, $\mathbf{Y} = \mathbf{A}_V\mathbf{X} + \mathbf{Z}_V$, $C_1 = \frac{1}{2}(\mathsf{s}^\star d_{\mathbf{X}} - d_{\mathbf{Y}})\log(2\pi)$, $C = C_1 - \frac{\delta d_{\mathbf{Y}}}{2}$ and $\mathbf{B} := \mathsf{s}^\star\mathbf{I} - \mathbb{E}_V[\mathbf{A}_V\mathbf{A}_V^\top]$. Observe that $\mathbf{B}\succeq 0$ by the definition of $\mathsf{s}^\star$.

**Lemma C.3.** *Suppose* $Q_X^{(i)} \Rightarrow Q_X^*$ *be a weakly convergent sequence of probability measure satisfying the moment constraint* $\mathbb{E}_{Q^{(i)}}[\|X\|^2] < b$ *for all* $i$ *and some positive value* $b$*, then:*

$$\liminf_{n\to\infty} \mathsf{g}_\delta(Q_X^{(i)}) \ge \mathsf{g}_\delta(Q_X^*).$$

The proof is based on the following fact about weaky converegen sequence of probability measures [Bil99],

*Fact.* For any weakly convergent sequence $Q_{\mathbf{X}}^{(i)}$ and any lower semi–continuous and bounded from below function $\varphi(\mathbf{X})$, we have:

$$\liminf_{n\to\infty} \mathbb{E}_{Q_{\mathbf{X}}^{(i)}}\left[\varphi(\mathbf{X})\right] \ge \mathbb{E}_{Q^*_{\mathbf{X}}}\left[\varphi(\mathbf{X})\right].$$

*Proof of Lemma C.3.* We show that each term in the expression (100) satisfies the desired limit behavior, separately.

1. Using [GN14, Proposition 18] (see also [PW16, Corollary 4]), we have $\lim_{n\to\infty} h_{Q^{(i)}}(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V = v) = h_{Q^*}(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V=v)$. Further, $h_Q(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V=v) \ge h(\sqrt{\delta}\mathbf{W})$. Thus by Fatou's lemma, we have:

$$\begin{aligned}\liminf_{n\to\infty} h_{Q^{(i)}}(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V) &= \liminf_{n\to\infty} \mathbb{E}_{P_V}[h_{Q^{(i)}}(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V=v)]\\ &\ge \mathbb{E}_{P_V}[\lim_{n\to\infty} h_{Q^{(i)}}(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V=v)]\\ &= h_{Q^*}(\mathbf{Y}+\sqrt{\delta}\mathbf{W}|V).\end{aligned} \tag{101}$$

2. It is shown in [LCCV18, Lemma A2], that any weakly convergent sequence $Q_{\mathbf{X}}^{(i)} \Rightarrow Q_{\mathbf{X}}^*$ with moment constraint $\mathbb{E}_Q^{(i)}[\mathbf{X}\mathbf{X}^\top] \preceq \mathbf{C}$ satisfies the following inequality:

$$\limsup_{n\to\infty} h_{Q^{(i)}}(\mathbf{X}) \leq h_{Q^*}(\mathbf{X}). \tag{102}$$

However, inspecting the proof in [LCCV18] shows that (102) also holds under weaker constraint $\mathbb{E}_Q^{(i)}[\|\mathbf{X}\|^2] \leq t$.

3. Since $\mathbf{B} \succeq 0$, the function $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{B}\mathbf{x}$ is bounded from below. Thus utilizing Fact C.1 yields:

$$\liminf_{n\to\infty} \mathbb{E}_{Q^{(i)}}[\mathbf{X}^\top \mathbf{B}\mathbf{X}] \geq \mathbb{E}_{Q^*}[\mathbf{X}^\top \mathbf{B}\mathbf{X}].$$

□

**Proposition C.4.** *The optimization problem (96) has a minimizer. More precisely, there exists $Q_{\mathbf{X}}^*$ satisfying the constraints $\mathbb{E}_{Q^*}[\|\mathbf{X}\|^2] \leq t$, and $D_{\mathrm{KL}}(Q_{\mathbf{X}}^* \| P_{\mathbf{X}}) < \infty$ such that $\mathsf{v}_\delta(t) = \mathsf{g}_\delta(Q_{\mathbf{X}}^*)$.*

*Proof.* Take a sequence of probability measure $\{Q_{\mathbf{X}}^{(i)}\}_{i=1}^n$ along which $\mathsf{g}_\delta(Q_{\mathbf{X}}^{(i)})$ approaches $\mathsf{v}_\delta(t)$. The moment constraint $\mathbb{E}_{Q^{(i)}}[\|\mathbf{X}\|^2] \leq t$ ensures that $\{Q_{\mathbf{X}}^{(i)}\}_{i=1}^n$ is a *tight* sequence [Bil99]. Thus we can extract a subsequence of it that weakly converges to a probability measure $Q_{\mathbf{X}}^*$. Without loss of generality, assume that the sequence $\{Q_{\mathbf{X}}^{(i)}\}_{i=1}^n$ is itself convergent to $Q_{\mathbf{X}}^*$. We show that $Q_{\mathbf{X}}^*$ is a minimizer.

- Utilizing Fact C.1 with $\varphi(\mathbf{x}) = \|\mathbf{x}\|^2$ implies that $Q_{\mathbf{X}}^*$ satisfies the constraint $\mathbb{E}_{Q^*}[\|\mathbf{X}\|^2] \leq t$. Also from the semi–lower continuity of KL–divergence [PW25], we have $D_{\mathrm{KL}}(Q_{\mathbf{X}}^* \| P_{\mathbf{X}}) \leq \liminf_{n\to\infty} D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(i)} \| P_{\mathbf{X}}) < \infty$. Thus $Q_{\mathbf{X}}^*$ satisfies the constraint of (96), so $\mathsf{g}_\delta(Q_{\mathbf{X}}^*) \geq \mathsf{v}_\delta(t)$.
- Conversely, Lemma C.3 says $\mathsf{v}_\delta(t) = \liminf_{n\to\infty} \mathsf{g}_\delta(Q_{\mathbf{X}}^{(i)}) \geq \mathsf{g}_\delta(Q_{\mathbf{X}}^*)$. Therefore $Q_{\mathbf{X}}^*$ must be the minimizer.

□

*Proof of Proposition C.2.* Utilizing Proposition C.4, we can express $\mathsf{V}_\delta(t)$ in the following way:

$$\mathsf{V}_\delta(t) = \inf_{n\geq 1} \inf_{\substack{\{p_i\}_{i=1}^n, t_i : p_i \geq 0, \sum_i p_i = 1, \\ \sum_i p_i t_i \leq t}} \sum_{i=1}^n p_i \mathsf{v}_\delta(t). \tag{103}$$

By Bunt–Carathéodorý theorem [Bun34, Car11], for any point on the boundary of convex envelop of the set $\mathcal{M} := \Big\{\big(t_i, \mathsf{v}_\delta(t_i)\big) : 1 \leq i \leq n\Big\}$ can be expressed as the convex combination of at most two points of $\mathcal{M}$. Thus (103) can be rewritten as:

$$\mathsf{V}_\delta(t) = \inf_{\substack{\alpha, t_1, t_2 : 0 \leq \alpha \leq \frac{1}{2}, \\ \alpha t_1 + \bar{\alpha} t_2 \leq t}} \alpha \mathsf{v}_\delta(t_1) + \bar{\alpha} \mathsf{v}_\delta(t_2), \tag{104}$$

where $\bar{\alpha} = 1 - \alpha$. Let $\{(\alpha_i, t_{1i}, t_{2i})\}_{i\geq 1}$ be a sequence satisfying $\alpha_i t_{1i} + \bar{\alpha}_i t_{2i} \leq t$ such that $\alpha_i \mathsf{v}_\delta(t_{1i}) + \bar{\alpha}_i \mathsf{v}_\delta(t_{2i})$ converges to $\mathsf{V}_\delta(t)$. Also let $Q_1^{(i)}$ and $Q_2^{(i)}$ be probability measures on $\mathcal{X}$ such that $\mathsf{v}_\delta(t_{ki}) = \mathsf{g}_\delta(Q_k^{(i)})$ for $k = 1, 2$. Existence of such $Q_k^{(i)}$ with the property $\mathbb{E}_{Q_k^{(i)}}[\|\mathbf{X}\|^2] \leq t_{ki}$ is guaranteed by Proposition C.4. Since there exists a convergent subsequence of $\{\alpha_i\}_{i\geq 1}$, we can assume that the sequence $\{\alpha_i\}_{i\geq 1}$ is itself convergent to some $\alpha^* \in [0, \frac{1}{2}]$. We investigate the cases $\alpha^* > 0$ and $\alpha^* = 0$, separately.

- **Case I:** $\alpha^* > 0$.
  After discarding some initial terms in the sequence $\{\alpha_i\}_{i\geq 1}$, we can assume that for all $i > 0$, we have: $\alpha_i > \frac{\alpha^*}{2}$. This implies that for all $i > 0$, $t_{1i} \leq \frac{2t}{\alpha^*}$ and $t_{2i} \leq 2t$. Thus $Q_1^{(i)}$ and $Q_2^{(i)}$ have common finite second moments, thus there exist a subsequence $\{i_j\}_{j=1}^{\infty}$ for which $Q_1^{(i_j)} \Rightarrow Q_1^*$ and $Q_2^{(i_j)} \Rightarrow Q_2^*$. Rename the sequence $\{i_j :\}_{j\geq 1}$ to $\{1, 2, \dots\}$. Let $U \in \{1, 2\}$ be a binary random variable and define $Q_U^*(U = 1) = \alpha^*$ and $Q^*_{\mathbf{X}|U=k} = Q_k^*$ for $k = 1, 2$. Also let $Q_U^{(i)}(U = 1) = \alpha_i$ and $Q^{(i)}_{\mathbf{X}|U=k} = Q_k^{(i)}$ for $k = 1, 2$. We have $Q_{\mathbf{X}}^{(i)} \Rightarrow Q_{\mathbf{X}}^*$. Now consider:
  $$\mathbb{E}_{Q_{\mathbf{X}}^*}\left[\|\mathbf{X}\|^2\right] \leq \liminf_{i\to\infty} \mathbb{E}_{Q_{\mathbf{X}}^{(i)}}\left[\|\mathbf{X}\|^2\right] = \liminf_{i\to\infty} \alpha_i \,\mathbb{E}_{Q_1^{(i)}}\left[\|\mathbf{X}\|^2\right] + \bar{\alpha_i}\,\mathbb{E}_{Q_2^{(i)}}\left[\|\mathbf{X}\|^2\right] \leq t. \tag{105}$$
  Also $Q_{\mathbf{X}}^*$ satisfies the constraint $D_{\mathrm{KL}}(Q_{\mathbf{X}}^*\|P_{\mathbf{X}}) < \infty$, by the semi–lower continuity of KL–divergence. Thus:
  $$\mathsf{V}_\delta(t) \leq \mathsf{G}_\delta(Q_{\mathbf{X}}^*) \leq \mathbb{E}\left[\mathsf{g}_\delta\big(Q^*_{\mathbf{X}|U}(.|U)\big)\right] \tag{106}$$
  However, Lemma C.3 asserts:
  $$\begin{aligned}\mathbb{E}\left[\mathsf{g}_\delta\big(Q^*_{\mathbf{X}|U}(.|U)\big)\right] &= \alpha^*\mathsf{g}_\delta(Q_1^*) + \bar{\alpha^*}\mathsf{g}_\delta(Q_2^*)\\ &\leq \liminf_{i\to\infty} \alpha_i \mathsf{g}_\delta(Q_1^{(i)}) + \liminf_{i\to\infty} \bar{\alpha_i}\mathsf{g}_\delta(Q_2^{(i)})\\ &\leq \mathsf{V}_\delta(t).\end{aligned} \tag{107}$$
  Thus $Q^*_{U\mathbf{X}}$ is the desired minimizer.

- **Case II:** $\alpha^* = 0$.
  We have again $t_{2i} \leq 2t$. Thus we can assume that $Q_2^{(i)} \Rightarrow Q_2^*$. The assumption $\alpha_i \to 0$ implies $\alpha_i Q_1^{(i)} + \bar{\alpha_i} Q_2^{(i)} \Rightarrow Q_2^* = Q_{\mathbf{X}}^*$. In this case, the inequality (105) is still holds. Further we have:
  $$\mathsf{V}_\delta(t) \leq \mathsf{G}_\delta(Q_{\mathbf{X}}^*) \leq \mathsf{g}_\delta(Q_{\mathbf{X}}^*). \tag{108}$$
  In the other side, we show that $\mathsf{V}_\delta(t) \geq \mathsf{g}_\delta(Q_{\mathbf{X}}^*)$. The function $\mathsf{g}_\delta(Q_1^{(i)})$ can be bounded from below as:
  $$\begin{aligned}\mathsf{g}_\delta(Q_1^{(i)}) &= h_{Q_1^{(i)}}(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h_{Q_1^{(i)}}(\mathbf{X}) + \frac{1}{2}\,\mathbb{E}_{\mathbf{X}\sim Q_1^{(i)}}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] + C\\ &\geq h(\sqrt{\delta}\mathbf{W}) - \mathsf{s}^\star h_{\mathbf{X}\sim\mathcal{N}(0,\frac{t_{1i}}{d})}(\mathbf{X}) + C\\ &= D - \frac{\mathsf{s}^\star}{2}\log(t_{1i}),\end{aligned} \tag{109}$$
  where we used the facts that $h(\mathbf{Y}+\mathbf{Z}) \geq h(\mathbf{Z})$ for any independent random variable $\mathbf{Y}$ and $\mathbf{Z}$, and the normal distribution maximizes the entropy under second moment constraint. We also used the positive–definiteness of $\mathbf{B}$. In the last line, $D$ is a constant which is not depending on $t$.
  Since $\mathsf{G}_\delta$ is a convex envelop of $\mathsf{g}_\delta$ and the right hand side of (109) is a convex function, the inequality (109) continues to hold for it. Now observe $t_{1i} \leq \frac{t}{\alpha_i}$. Inequality (109) implies $\liminf_{i\to\infty} \alpha_i \mathsf{g}_\delta(Q_1^{(i)}) \geq 0$. Thus by Lemma C.3 we have:
  $$\begin{aligned}\mathsf{g}_\delta(Q_{\mathbf{X}}^*) &= \mathsf{g}_\delta(Q_2^*)\\ &\leq \liminf_{i\to\infty} \bar{\alpha_i}\mathsf{g}_\delta(Q_2^{(i)})\\ &\leq \liminf_{i\to\infty} \alpha_i \mathsf{g}_\delta(Q_1^{(i)}) + \liminf_{i\to\infty} \bar{\alpha_i}\mathsf{g}_\delta(Q_2^{(i)})\\ &\leq \mathsf{V}_\delta(t).\end{aligned} \tag{110}$$

Therefore $\mathsf{V}_\delta(t) = \mathsf{g}_\delta(Q^*_{\mathbf{X}})$. Hence in this case, we can assume that $U$ is a constant random variable and $Q^*_{\mathbf{X}}$ is the desired minimizer.

*Proof of* $\mathbb{E}_{Q^*}[\mathbf{X}|U] = 0$.

Let $\boldsymbol{\mu}_u = \mathbb{E}_{Q^*}[\mathbf{X}|U = u]$ for $u \in \{1, 2\}$. We show that for an optimal $Q^*_{U\mathbf{X}}$, $\mathbf{X}$ must be a conditionally centered random variable. Take $\tilde{\mathbf{X}} := \mathbf{X} - \boldsymbol{\mu}_U$ and let $\tilde{\mathbf{Y}}$ be the output of the Markov kernel (channel) $T_{\mathbf{Y}|\mathbf{X},V}$ given the input $\tilde{\mathbf{X}}$, that is, given $V = v$, $\tilde{\mathbf{Y}} = \mathbf{A}_v\tilde{\mathbf{X}} + \mathbf{Z}_v$. Suppose $(\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top) \neq (\mathbf{0}^\top, \mathbf{0}^\top)$. We show that $Q^*_{U\tilde{\mathbf{X}}}$ achieves smaller $\mathbb{E}[\mathsf{g}(Q_{\mathbf{X}|U})]$ than $Q^*_{U\mathbf{X}}$, contradicting the assumption that $Q^*_{U\mathbf{X}}$ is the minimizer. The definition of $\tilde{\mathbf{X}}$ implies:

$$\mathbb{E}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] = \mathbb{E}\left[\tilde{\mathbf{X}}^\top\mathbf{B}\tilde{\mathbf{X}}\right] + \mathbb{E}_U\left[\boldsymbol{\mu}_U^\top\mathbf{B}\boldsymbol{\mu}_U\right] \geq \mathbb{E}\left[\tilde{\mathbf{X}}^\top\mathbf{B}\tilde{\mathbf{X}}\right],$$

where we used the fact that $\mathbf{B}$ is a positive semi–definite matrix. Using this and (100), we obtain:

$$\begin{aligned}
\mathbb{E}[\mathsf{g}_\delta(Q^*_{\tilde{\mathbf{X}}|U})] &= \mathbb{E}_{Q^*_U}\left[h(\tilde{\mathbf{Y}} + \sqrt{\delta}\mathbf{W}|V, U = u) - \mathsf{s}^\star h(\tilde{\mathbf{X}}|U = u)\right] + \frac{1}{2}\mathbb{E}\left[\tilde{\mathbf{X}}^\top\mathbf{B}\tilde{\mathbf{X}}\right] + C \\
&\overset{\text{(a)}}{=} \mathbb{E}_{Q^*_U}\left[h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V, U = u) - \mathsf{s}^\star h(\mathbf{X}|U = u)\right] + \frac{1}{2}\mathbb{E}\left[\tilde{\mathbf{X}}^\top\mathbf{B}\tilde{\mathbf{X}}\right] + C \\
&\leq \mathbb{E}_{Q^*_U}\left[h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V, U = u) - \mathsf{s}^\star h(\mathbf{X}|U = u)\right] + \frac{1}{2}\mathbb{E}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] + C \\
&= \mathbb{E}\left[\mathsf{g}_\delta(Q^*_{\mathbf{X}|U})\right],
\end{aligned}$$

where (a) follows from the fact that differential entropy is invariant under the shift. □

## C.2 Gaussian Optimality–Proof of Lemma C.1

For the sake of brevity, let $\mathsf{g}_\delta(Q_{U\mathbf{X}}) := \mathbb{E}_U[\mathsf{g}_\delta(Q_{\mathbf{X}|U})]$, $\overline{\mathbf{Y}} = \mathbf{Y} + \sqrt{\delta}\mathbf{W}$ and $T_{\overline{\mathbf{Y}}|V,\mathbf{X}} = T_{\mathbf{Y}|V,\mathbf{X}} * \mathcal{N}(0, \delta\mathbf{I}_{d_\mathbf{Y}})$.

Let $Q^*_{U\mathbf{X}}$ be the minimizer for which $U \in \{1, 2\}$, $\mathsf{g}_\delta(Q^*_{U\mathbf{X}}) = \mathsf{V}_\delta(t)$ and $\mathbb{E}_{Q^*}[\mathbf{X}|U] = 0$. Take $(U_1, \mathbf{X}_1), (U_2, \mathbf{X}_2)$ be two i.i.d. samples drawn from $Q^*_{U\mathbf{X}}$. Let $Q_{U_1U_2\mathbf{X}_1\mathbf{X}_2}(u_1, u_2, \mathbf{x}_1, \mathbf{x}_2) = Q^*_{U\mathbf{X}}(u_1, \mathbf{x}_1)Q^*_{U\mathbf{X}}(u_2, \mathbf{x}_2)$ be the joint distribution of $(U_1, U_2, \mathbf{X}_1, \mathbf{X}_2)$. Set $\mathbf{U} = [U_1, U_2]^\top$ and define:

$$\begin{aligned}
\mathbf{X}_+ &= \frac{\mathbf{X}_1 + \mathbf{X}_2}{\sqrt{2}}, \quad \mathbf{X}_- = \frac{\mathbf{X}_1 - \mathbf{X}_2}{\sqrt{2}}, \\
\overline{\mathbf{Y}}_+ &= \frac{\overline{\mathbf{Y}}_1 + \overline{\mathbf{Y}}_2}{\sqrt{2}}, \quad \overline{\mathbf{Y}}_- = \frac{\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2}{\sqrt{2}}.
\end{aligned} \tag{111}$$

The random variables $(\mathbf{U}, V, \mathbf{X}_+, \mathbf{X}_-, \mathbf{Y}_+, \mathbf{Y}_-)$ has the following properties:

1. The joint distribution of $(\mathbf{U}, V, \mathbf{X}_+, \mathbf{X}_-, \mathbf{Y}_+, \mathbf{Y}_-)$ factors as:

$$P_V Q_{\mathbf{U},\mathbf{X}_+,\mathbf{X}_-} T_{\mathbf{Y}_+|V,\mathbf{X}_+} T_{\mathbf{Y}_-|V,\mathbf{X}_-}. \tag{112}$$

2. The rotational invariance of the Gaussian noises $\mathbf{Z}_v$ and $\mathbf{W}$ implies $T_{\overline{\mathbf{Y}}_+|V,\mathbf{X}_+}$ and $T_{\overline{\mathbf{Y}}_-|V,\mathbf{X}_-}$ are two *independent* copies of the channel $T_{\overline{\mathbf{Y}}|V,\mathbf{X}}$, that is:

$$T_{\overline{\mathbf{Y}}_+|V,\mathbf{X}_+}(\mathbf{y}|v, \mathbf{x}) = T_{\overline{\mathbf{Y}}_-|V,\mathbf{X}_-}(\mathbf{y}|v, \mathbf{x}) = T_{\overline{\mathbf{Y}}|V,\mathbf{X}}(\mathbf{y}|v, \mathbf{x}). \tag{113}$$

3. The following moment constraint holds:

$$\mathbb{E}[\|\mathbf{X}_+\|^2] = \mathbb{E}[\|\mathbf{X}_-\|^2] = \mathbb{E}[\|\mathbf{X}_1\|^2] \leq t.$$

We now proceed to show $\mathbf{X}_+ \multimap (\overline{\mathbf{Y}}_+, V) \multimap \overline{\mathbf{Y}}_-$. To accomplish this, consider:

$$\begin{aligned}
2\mathsf{V}_\delta(t) &= \mathsf{g}_\delta(Q^*_{U_1\mathbf{X}_1}) + \mathsf{g}_\delta(Q^*_{U_2\mathbf{X}_2}) \\
&= \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}_1\mathbf{X}_2|U_1,U_2}\|P_{\mathbf{X}}^{\otimes 2}|Q_{U_1U_2}) - D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_1\overline{\mathbf{Y}}_2|U_1,U_2,V}\|P_{\mathbf{Y}}^{\otimes 2}|Q_{U_1U_2V}) && (114)\\
&= \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}_+\mathbf{X}_-|U_1,U_2}\|P_{\mathbf{X}}^{\otimes 2}|Q_{U_1U_2}) - D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_+\overline{\mathbf{Y}}_-|U_1,U_2,V}\|P_{\mathbf{Y}}^{\otimes 2}|Q_{U_1U_2V}) && (115)\\
&= \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}_+|\mathbf{U}}\|P_{\mathbf{X}}|Q_{\mathbf{U}}) - D(Q_{\overline{\mathbf{Y}}_+|\mathbf{U}V}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V}) \\
&\quad + \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}_-|\mathbf{U}\mathbf{X}_+}\|P_{\mathbf{X}}|Q_{\mathbf{U}\mathbf{X}_+}) - D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_-|\mathbf{U}V\mathbf{X}_+}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V\mathbf{X}_+}) \\
&\quad + I_Q(\mathbf{X}_+;\overline{\mathbf{Y}}_-|\mathbf{U}V\overline{\mathbf{Y}}_+) && (116)\\
&\geq \mathbb{E}_{\mathbf{U}}\left[\mathsf{g}_\delta(Q_{\mathbf{X}_+|\mathbf{U}})\right] + \mathbb{E}_{\mathbf{U},\mathbf{X}_+}\left[\mathsf{g}_\delta(Q_{\mathbf{X}_-|\mathbf{U}\mathbf{X}_+})\right] && (117)\\
&\geq 2\mathsf{V}_\delta(t), && (118)
\end{aligned}$$

where (116) follows from the identity in Lemma 3.7, (117) follows because mutual information term is non–negative, and (118) is due to the definition of $\mathsf{V}_\delta(t)$.

The chain of inequalities (114)–(118) shows:

- $Q_{\mathbf{U}\mathbf{X}_+}$ is also a minimizer attaining:

$$\mathsf{g}_\delta(Q_{\mathbf{U}\mathbf{X}_+}) = \mathsf{V}_\delta(t). \tag{119}$$

  Similarly $Q_{\mathbf{U}\mathbf{X}_-}$ is a minimizer.

- More importantly, $I_Q(\mathbf{X}_+;\mathbf{Y}_-|\mathbf{U}V\mathbf{Y}_+) = 0$ which is equivalent to $\mathbf{X}_+ \multimap (\overline{\mathbf{Y}}_+, V, \mathbf{U}) \multimap \overline{\mathbf{Y}}_-$. However, we also have the trivial Markov chain $\overline{\mathbf{Y}}_+ \multimap (\mathbf{X}_+, V) - (\overline{\mathbf{Y}}_-, \mathbf{U})$. It is easy to show that if $A \multimap B \multimap C$ and $B \multimap A \multimap C$ simultaneously hold and $P_{AB} \ll\gg P_AP_B$, then $C$ is independent of $(A, B)$ (see e.g. [AJN22])[1]. Observe that for each $V = v$, $Q_{\overline{\mathbf{Y}}_+,\mathbf{X}_+|V=v} \ll\gg Q_{\overline{\mathbf{Y}}_+|V=v}Q_{\mathbf{X}_+}$ (due to the smoothing term $\mathbf{W}$ in the definition of $\overline{\mathbf{Y}}$), hence we can deduce:

$$I_Q(\mathbf{X}_+\overline{\mathbf{Y}}_+;\overline{\mathbf{Y}}_-|\mathbf{U},V) = I_Q(X_-\overline{\mathbf{Y}}_-;\overline{\mathbf{Y}}_+|\mathbf{U},V) = 0, \tag{120}$$

  where the second equality comes from swapping the role of $+$ and $-$.

We now show that (120) yields $I_Q(\mathbf{X}_+;\mathbf{X}_-|\mathbf{U}) = 0$. To accomplish this, we invoke the following identities to expand (115) in a different way:

$$\begin{aligned}
D_{\mathrm{KL}}(Q_{\mathbf{X}_+\mathbf{X}_-|\mathbf{U}}\|P_{\mathbf{X}}^{\otimes 2}|Q_{\mathbf{U}}) \quad &= D_{\mathrm{KL}}(Q_{\mathbf{X}_+|\mathbf{U}}\|P_{\mathbf{X}}|Q_{\mathbf{U}}) + D_{\mathrm{KL}}(Q_{\mathbf{X}_-|\mathbf{U}}\|P_{\mathbf{X}}|Q_{\mathbf{U}}) \\
&\quad + I_Q(\mathbf{X}_+;\mathbf{X}_-|\mathbf{U})
\end{aligned} \tag{121}$$

$$\begin{aligned}
D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_+\overline{\mathbf{Y}}_-|\mathbf{U}V}\|P_{\mathbf{Y}}^{\otimes 2}|Q_{\mathbf{U}V}) &= D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_+|\mathbf{U}V}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V}) + D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_-|\mathbf{U}V}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V}) \\
&\quad + I_Q(\mathbf{Y}_+;\mathbf{Y}_-|\mathbf{U}V) \\
&\overset{(120)}{=} D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_+|\mathbf{U}V}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V}) + D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_-|\mathbf{U}V}\|P_{\mathbf{Y}}|Q_{\mathbf{U}V}).
\end{aligned} \tag{122}$$

By substituting this in (115), we obtain:

$$\begin{aligned}
2\mathsf{V}_\delta(t) &= \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}_+\mathbf{X}_-|\mathbf{U}}\|P_{\mathbf{X}}^{\otimes 2}|Q_{\mathbf{U}}) - D_{\mathrm{KL}}(Q_{\overline{\mathbf{Y}}_+\overline{\mathbf{Y}}_-|\mathbf{U},V}\|P_{\mathbf{Y}}^{\otimes 2}|Q_{\mathbf{U}V}) \\
&= \mathsf{g}_\delta(Q_{\mathbf{U}\mathbf{X}_+}) + \mathsf{g}_\delta(Q_{\mathbf{U}\mathbf{X}_-}) + \mathsf{s}^\star I_Q(\mathbf{X}_+;\mathbf{X}_-|\mathbf{U}) \geq 2\mathsf{V}_\delta(t),
\end{aligned}$$

where we have used (119) in the last inequality. Hence we should have $I_Q(\mathbf{X}_+;\mathbf{X}_-|\mathbf{U}) = 0$.

In summary, $(U_1, U_2, \mathbf{X}_+, \mathbf{X}_-, \mathbf{X}_1, \mathbf{X}_2)$ has the following properties:

[1] A simple information theoretic argument is as follows: The two Markov chains imply $D_{\mathrm{KL}}(P_{C|A=a}\|P_{C|B=b}) = 0$, a.s. $P_{AB}$. The condition $P_{AB} \gg P_AP_B$ ensures that this also holds a.s. $P_AP_B$. Thus by Jensen inequality $I(A;C) = D_{\mathrm{KL}}(P_{C|A}\|P_C) \leq \mathbb{E}_B[D_{\mathrm{KL}}(P_{C|A}\|P_{C|B})] = 0$. This and the Markov chain $B \multimap A \multimap C$ imply $(A, B)\perp C$

- Given $\mathbf{U} = [u_1, u_2]^\top$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent.
- Given $\mathbf{U} = [u_1, u_2]^\top$, $\mathbf{X}_+$ and $\mathbf{X}_-$ are independent.
- $U_i \in \{1, 2\}, i = 1, 2$. Moreover, $(U_1, \mathbf{X}_1)$ and $(U_2, \mathbf{X}_2)$ are independent and have common distribution $Q^*_{U\mathbf{X}}$.

The first two items and the multivariate extension of Darmois–Skitovich theorem [GO62] imply $\mathbf{X}_1$ and $\mathbf{X}_2$ have to be Gaussian with the same covariance matrix. In particular for $u_1 = 1, u_2 = 2$, $Q_{\mathbf{X}_1|U_1=1,U_2=2} = Q_{\mathbf{X}_2|U_1=1,U_2=2} = \mathcal{N}(0, \mathbf{C})$. This and the third item yield $Q^*_{\mathbf{X}|U=1} = Q^*_{\mathbf{X}|U=2} = \mathcal{N}(0, \mathbf{C})$, thus $\mathbf{X}$ and $U$ are independent under $Q^*$. Hence $\mathsf{V}_\delta(t) = \mathsf{g}_\delta(Q^*_{\mathbf{X}})$, where $Q^*_{\mathbf{X}} = \mathcal{N}(0, \mathbf{C})$. Therefore the infimum in (95) is attained by a Gaussian distribution. This concludes the proof of Lemma C.1.

### C.3 Proof of Lemma 3.10

Let $B_{\mathsf{G}} = \inf_{\substack{Q_{\mathbf{X}}: Q_{\mathbf{X}} \in \mathcal{F}_{\mathsf{G}}, \\ D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) < \infty}} \mathsf{g}_0(Q_{\mathbf{X}})$, where $\mathsf{g}_0(Q_X) = \mathsf{s}^\star D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) - D_{\mathrm{KL}}(Q_{\mathbf{Y}|V} \| P_{\mathbf{Y}} | P_V)$. It suffices to show that $\mathsf{g}_0(Q_{\mathbf{X}}) \geq B_{\mathsf{G}}$, for any $Q_{\mathbf{X}}$ with $D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) < \infty$. We need the following observation about the finiteness of the second moment of $\mathbf{X} \sim Q_{\mathbf{X}}$.

**Claim 1.** *Every distribution $Q_{\mathbf{X}}$ with $D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) < \infty$, has finite second moments in the sense that $\mathbb{E}_Q\left[\|\mathbf{X}\|^2\right] < \infty$.*

*Proof.* The claim follows from the Donsker–Varadhan variational representation of KL divergence [DV83]:

$$D_{\mathrm{KL}}(Q_{\mathbf{X}} \| P_{\mathbf{X}}) = \sup_{f: \mathcal{X} \mapsto \mathbb{R}} \mathbb{E}_Q\left[f(\mathbf{X})\right] - \log \mathbb{E}_P\left[\exp(f(\mathbf{X}))\right]. \tag{123}$$

Thus the finiteness assumption of KL divergence yields that the function inside the suprimum is always finite. In particular, for $f(\mathbf{x}) = \frac{\|\mathbf{x}\|^2}{4}$ we have:

$$\frac{1}{4}\mathbb{E}_Q\left[\|\mathbf{X}\|^2\right] - \log \mathbb{E}_P\left[\exp(\frac{\|\mathbf{X}\|^2}{4})\right] < \infty. \tag{124}$$

However, for $P = \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{X}}})$ we have $\mathbb{E}_P\left[\exp(\frac{\|\mathbf{X}\|^2}{4})\right] < \infty$, therefore the second moment $\mathbb{E}_Q[\|\mathbf{X}\|^2]$ should be finite. □

Assume $\mathbb{E}_Q[\|\mathbf{X}\|^2] = t < \infty$. Lemma C.1 shows that for any $\delta > 0$, there exists a normal distribution $Q^{(\delta)}_{\mathbf{X}}$ with $\mathbb{E}_{Q^{(\delta)}}\left[\|\mathbf{X}\|^2\right] \leq t$ such that $\mathsf{g}_\delta(Q^{(\delta)}_{\mathbf{X}}) \leq \mathsf{g}_\delta(Q_{\mathbf{X}})$. Invoking (100), we can relate the function $\mathsf{g}_\delta(Q_{\mathbf{X}})$ to $\mathsf{g}_0(Q_{\mathbf{X}})$ as follows:

$$\begin{aligned}
\mathsf{g}_\delta(Q_{\mathbf{X}}) &= h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h(\mathbf{X}) + \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top \mathbf{B}\mathbf{X}\right] + C_1 - \frac{\delta d_{\mathbf{Y}}}{2} \\
&\geq h(\mathbf{Y}|V) - \mathsf{s}^\star h(\mathbf{X}) + \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top \mathbf{B}\mathbf{X}\right] + C_1 - \frac{\delta d_{\mathbf{Y}}}{2} \\
&= \mathsf{g}_0(Q_{\mathbf{X}}) - \frac{\delta d_{\mathbf{Y}}}{2},
\end{aligned} \tag{125}$$

where we used the fact that adding independent noise $\sqrt{\delta}\mathbf{W}$ increases the differential entropy. This yields:

$$\mathsf{g}_\delta(Q_{\mathbf{X}}) \geq \mathsf{g}_\delta(Q^{(\delta)}_{\mathbf{X}}) \geq \mathsf{g}_0(Q^{(\delta)}_{\mathbf{X}}) - \frac{\delta d_{\mathbf{Y}}}{2} \geq B_{\mathsf{G}} - \frac{\delta d_{\mathbf{Y}}}{2}, \tag{126}$$

where we used the definition of $B_{\mathsf{G}}$ and the assumption that $Q^{(\delta)}_{\mathbf{X}}$ is a normal distribution.

**Claim 2** (Continuity of the function $\mathsf{g}_\delta$.)**.** *For any $Q_{\mathbf{X}}$ with finite second moments, the following continuity property holds:*

$$\lim_{\delta \downarrow 0} \mathsf{g}_\delta(Q_{\mathbf{X}}) = \mathsf{s}_0(Q_{\mathbf{X}}). \tag{127}$$

We defer the proof of this claim to the end of this section. Putting (126) and the continuity of $\mathsf{g}_\delta$ at $\delta = 0$ together implies:

$$\mathsf{g}_0(Q_{\mathbf{X}}) \geq B_{\mathsf{G}} - \lim_{\delta \downarrow 0} \frac{\delta d_{\mathbf{Y}}}{2} = B_{\mathsf{G}}. \tag{128}$$

This concludes the proof of Lemma 3.10.

*Proof of Claim 2.* The proof is based on the following continuity lemma.

**Lemma C.5** (Lemma 12 in [AJN22])**.** *For any $\mathbb{R}^n$–valued random variable $\mathbf{U}$ with finite second moment (i.e. $\mathbb{E}[\|\mathbf{U}\|^2]$) and random variable $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I}_n)$, the following continuity property hold:*

$$\lim_{\delta \downarrow 0} h(\mathbf{U} + \sqrt{\delta}\mathbf{N}) = h(\mathbf{U}). \tag{129}$$

Define $g(v;\delta) = h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V = v)$. For each $v$, the function $g(v;\delta)$ is an increasing function of $\delta$. Further, for $\delta < 1$ and conditioned on any $V = v$, $\mathbf{Y} + \sqrt{\delta}\mathbf{W}$ has a universal upper bound on the second moment. To see this, recall that $\mathbf{Y} = \mathbf{A}_v\mathbf{X} + \mathbf{Z}_v$, where $\|\mathbf{A}_v\|_{\mathsf{op}} \leq 1$ and $\mathbf{Z}_v \sim \mathcal{N}(0, \mathbf{I}_{d_{\mathbf{Y}}} - \mathbf{A}_v\mathbf{A}_v^\top)$. Thus for $\delta < 1$:

$$\begin{aligned} \mathbb{E}_Q\left[\|\mathbf{Y} + \sqrt{\delta}\mathbf{W}\|^2 | V = v\right] &= \mathbb{E}_Q\left[\|\mathbf{A}_v\mathbf{X}\|^2\right] + \mathbb{E}_Q[\|\mathbf{Z}_v\|^2\Big] + \delta d_{\mathbf{Y}} \\ &\leq \mathbb{E}_Q\left[\|\mathbf{X}\|^2\right] + (1+\delta)d_{\mathbf{Y}} < K, \end{aligned} \tag{130}$$

where $K = \mathbb{E}_Q\left[\|\mathbf{X}\|^2\right] + 2d_{\mathbf{Y}} < \infty$ does not depend on $v$. This implies $g(v;\delta) \leq h(\mathbf{A}) = \frac{d_{\mathbf{Y}}}{2}\log\frac{2\pi e K}{d_{\mathbf{Y}}} < \infty$, where $\mathbf{A} \sim \mathcal{N}(0, \frac{K}{d_{\mathbf{Y}}}\mathbf{I}_{d_{\mathbf{Y}}})$.

Also, by Lemma C.5, we have $\lim_{\delta \downarrow 0} g(v;\delta) = g(v,0)$. Now the monotone convergence theorem [Rud87] implies:

$$\lim_{\delta \downarrow 0} h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V) = \lim_{\delta \downarrow 0} \mathbb{E}_V\left[g(V;\delta)\right] = \mathbb{E}_V\left[g(V;0)\right] = h(\mathbf{Y}|V). \tag{131}$$

Now, we are ready to complete the proof of the Claim. Invoking (100), we have:

$$\begin{aligned} \lim_{\delta \downarrow 0} \mathsf{g}_\delta(Q_{\mathbf{X}}) &= \lim_{\delta \downarrow 0} h(\mathbf{Y} + \sqrt{\delta}\mathbf{W}|V) - \mathsf{s}^\star h(\mathbf{X}) + \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] + C_1 - \frac{\delta d_{\mathbf{Y}}}{2} \\ &= h(\mathbf{Y}|V) - \mathsf{s}^\star h(\mathbf{X}) + \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}^\top\mathbf{B}\mathbf{X}\right] + C_1 \\ &= \mathsf{g}_0(Q_{\mathbf{X}}), \end{aligned} \tag{132}$$

which is the statement of Claim 2. □

# D SDPI Coefficient of The Gaussian Mixture Channel

We state the proof of Proposition 3.12 here.

*Proof of Proposition 3.12.* The direction $s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) \leq \|\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}}$ has been established in Subsection 3.3. Therefore, our primary objective is to prove the converse: $s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) \geq \|\mathbb{E}[\mathbf{A}_V^\top\mathbf{A}_V]\|_{\mathsf{op}}$.

For a scalar $\beta \geq 1$ and a positive semi–definite matrix $\mathbf{B}$ such that $\mathbf{I} \preceq \mathbf{B}$, let $Q_{\mathbf{X}}^{(\beta,\mathbf{B})}$ be a Gaussian distribution defined as $Q_{\mathbf{X}}^{(\beta,\mathbf{B})} = \mathcal{N}(\mathbf{0}, \mathbf{I} + \beta\mathbf{B})$. The SDPI coefficient can then be bounded as:

$$s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) \geq \sup_{\beta,\mathbf{B}} \frac{D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})} \| P_{\mathbf{Y}})}{D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}})} \geq \sup_{\mathbf{B}} \lim_{\beta\to\infty} \frac{D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})} \| P_{\mathbf{Y}})}{D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}})},$$

where $Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}$ is the output distribution of the channel $\widetilde{T}_{\mathbf{Y}|\mathbf{X}}$ when the input distribution is $Q_{\mathbf{X}}^{(\beta,\mathbf{B})}$, and $P_{\mathbf{X}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\mathbf{X}}})$, $P_{\mathbf{Y}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\mathbf{Y}}})$ are the reference input and output distributions (standard normal distributions in $d_{\mathbf{X}}$ and $d_{\mathbf{Y}}$ dimensions, respectively). We first focus on evaluating the limit $\lim_{\beta\to\infty} \frac{D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})} \| P_{\mathbf{Y}})}{D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}})}$.

**Analyzing $D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}})$:** The KL divergence between $Q_{\mathbf{X}}^{(\beta,\mathbf{B})} = \mathcal{N}(\mathbf{0}, \mathbf{I} + \beta\mathbf{B})$ and $P_{\mathbf{X}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is given by:

$$\begin{aligned} D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}}) &= \frac{1}{2} \mathsf{Tr}\left\{\beta\mathbf{B} - \log(\mathbf{I} + \beta\mathbf{B})\right\} && (133) \\ &= \frac{1}{2}\left(\beta \mathsf{Tr}\left\{\mathbf{B}\right\} - \sum_{k=1}^{d_{\mathbf{X}}} \log(1 + \beta\lambda_i(\mathbf{B}))\right) && (134) \\ &= \frac{1}{2}\left(\beta \mathsf{Tr}\left\{\mathbf{B}\right\} - \mathcal{O}\left(\log\beta\right)\right) && (135) \end{aligned}$$

where $\lambda_i(\mathbf{B})$ are the eigenvalues of $\mathbf{B}$. Since $\mathbf{I} \preceq \mathbf{B}$, all eigenvalues $\lambda_i(\mathbf{B}) \geq 1$. For large $\beta$, $\log(1 + \beta\lambda_i(\mathbf{B})) \approx \log(\beta\lambda_i(\mathbf{B})) = \log\beta + \log\lambda_i(\mathbf{B})$. Thus, the sum can be approximated as $\sum_{i=1}^{d_{\mathbf{X}}}(\log\beta + \log\lambda_i(\mathbf{B})) = d_{\mathbf{X}}\log\beta + \sum_{i=1}^{d_X}\log\lambda_i(\mathbf{B})$. Therefore, for large $\beta$:

$$D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})} \| P_{\mathbf{X}}) = \frac{1}{2}\left(\beta \mathsf{Tr}\left\{\mathbf{B}\right\} - \mathcal{O}(\log\beta)\right)$$

The $\mathcal{O}(\log\beta)$ term encompasses constants depending on $d_{\mathbf{X}}$ and $\mathbf{B}$.

**Analyzing $D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})} \| P_{\mathbf{Y}})$:** The KL divergence for the output distribution is given by:

$$\begin{aligned} D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})} \| P_{\mathbf{Y}}) &= -h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) + \mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\log\frac{1}{P_{\mathbf{Y}}(\mathbf{Y})}\right] \\ &= -h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) + \frac{d_{\mathbf{Y}}}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\|\mathbf{Y}\|^2\right] \\ &= -h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) + \frac{d_{\mathbf{Y}}}{2}\log(2\pi) + \frac{1}{2}\mathsf{Tr}\left\{\mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right]\right\}, \end{aligned}$$

where $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})})$ is the differential entropy of $Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}$.

Let's simplify each term. First, consider the expected outer product $\mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right]$. We can use the law of total expectation:

$$\begin{aligned} \mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right] &= \mathbb{E}_V\left[\mathbb{E}_{\mathbf{Y}|V,\mathbf{X}\sim Q_{\mathbf{X}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top|V\right]\right] \\ &= \mathbb{E}_V\left[\mathrm{Cov}(\mathbf{Y}|V,\mathbf{X}) + \mathbb{E}[\mathbf{Y}|V,\mathbf{X}]\,\mathbb{E}[\mathbf{Y}|V,\mathbf{X}]^\top\right] \\ &= \mathbb{E}_V\left[(\mathbf{I} - \mathbf{A}_V\mathbf{A}_V^\top) + \mathbf{A}_V\,\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{A}_V^\top\right]. \end{aligned}$$

Since $\mathbf{X} \sim Q_{\mathbf{X}}^{(\beta,\mathbf{B})} = \mathcal{N}(\mathbf{0}, \mathbf{I} + \beta\mathbf{B})$, we have $\mathbb{E}[\mathbf{X}\mathbf{X}^\top] = \mathbf{I} + \beta\mathbf{B}$. Substituting this:

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right] &= \mathbb{E}_V\left[\mathbf{I} - \mathbf{A}_V\mathbf{A}_V^\top + \mathbf{A}_V(\mathbf{I}+\beta\mathbf{B})\mathbf{A}_V^\top\right] \\ &= \mathbf{I} + \beta\,\mathbb{E}_V\left[\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top\right].\end{aligned}$$

Taking the trace:

$$\mathsf{Tr}\left\{\mathbb{E}_{\mathbf{Y}\sim Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right]\right\} = \mathsf{Tr}\left\{\mathbf{I} + \beta\,\mathbb{E}_V\left[\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top\right]\right\} = d_{\mathbf{Y}} + \beta\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\}.$$

Next, we need to analyze the differential entropy $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})})$. We know that the differential entropy of a distribution is bounded by the entropy of a Gaussian distribution with the same covariance matrix [Cov99, Theorem 8.6.5]. The covariance matrix of $Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}$ is $\mathrm{Cov}_Q(\mathbf{Y}) = \mathbb{E}_{Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}}\left[\mathbf{Y}\mathbf{Y}^\top\right] = \mathbf{I} + \beta\,\mathbb{E}_V\left[\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top\right]$. The trace of this covariance matrix is $\mathsf{Tr}\left\{\mathrm{Cov}_Q(\mathbf{Y})\right\} = d_{\mathbf{Y}} + \beta\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\}$.

We can establish bounds for $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})})$:

1. Lower Bound: By the conditioning inequality for entropy, $h(\mathbf{Y}) \geq h(\mathbf{Y}|V)$. Since $\mathbf{Y}|\{V, \mathbf{X} = \mathbf{x}\} \sim \mathcal{N}(\mathbf{A}_V\mathbf{x}, \mathbf{I} - \mathbf{A}_V\mathbf{A}_V^\top)$ and $\mathbf{X} \sim Q_{\mathbf{X}}^{(\beta,\mathbf{B})}$, we have $\mathbf{Y}|\{V = v\} \sim \mathcal{N}(\mathbf{0}, \mathbf{I} + \beta\mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top)$. Thus:

$$\begin{aligned}h(\mathbf{Y}|V=v) &= \frac{1}{2}\log\det\left\{(2\pi e(\mathbf{I} + \beta\mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top))\right\} \\ &\geq \frac{1}{2}\log\det\left\{(2\pi e\mathbf{I})\right\} = \frac{d_{\mathbf{Y}}}{2}\log(2\pi e) = \mathcal{O}(1),\end{aligned}$$

   where we have used the fact that $\mathbf{A}_v\mathbf{B}\mathbf{A}_v^\top \succeq \mathbf{0}$, since $\mathbf{B} \succeq \mathbf{0}$. This implies $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) \geq \mathbb{E}_V[h(\mathbf{Y}|V)] = \mathcal{O}(1)$.

2. Upper Bound: Let $G_{\mathbf{Y}} = \mathcal{N}(\mathbf{0}, \mathrm{Cov}_Q(\mathbf{Y}))$ be a Gaussian distribution with the same covariance as $Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}$. Then $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) \leq h(G_{\mathbf{Y}})$. The entropy of $G_{\mathbf{Y}}$ is:

$$h(G_{\mathbf{Y}}) = \frac{1}{2}\log\det\left\{(2\pi e\,\underset{Q}{\mathrm{Cov}}(\mathbf{Y}))\right\}.$$

   Since $\mathrm{Cov}_Q(\mathbf{Y}) = \mathbf{I} + \beta\,\mathbb{E}_V\left[\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top\right]$, and $\mathbf{A}_V\mathbf{B}\mathbf{A}_V^\top$ is positive semi–definite (as $\mathbf{B}$ is positive semi–definite), the eigenvalues of $\mathrm{Cov}_Q(\mathbf{Y})$ will grow linearly with $\beta$. Therefore, $\log\det\left\{(\mathrm{Cov}_Q(\mathbf{Y}))\right\} = \mathcal{O}(\log\beta)$. Thus, $h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}) \leq h(G_{\mathbf{Y}}) = \mathcal{O}(\log\beta)$.

Combining the lower and upper bounds, we conclude that $|h(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})})| = \mathcal{O}(\log\beta)$.

Now, substituting these findings back into the expression for $D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}\|P_{\mathbf{Y}})$:

$$D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}\|P_{\mathbf{Y}}) = \frac{1}{2}\beta\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\} + \mathcal{O}(\log\beta).$$

Finally, we take the limit of the ratio:

$$\lim_{\beta\to\infty}\frac{D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}\|P_{\mathbf{Y}})}{D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})}\|P_{\mathbf{X}})} = \lim_{\beta\to\infty}\frac{\frac{1}{2}\beta\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\} + \mathcal{O}(\log\beta)}{\frac{1}{2}\left(\beta\mathsf{Tr}\left\{\mathbf{B}\right\} - \mathcal{O}(\log\beta)\right)}.$$

Dividing both numerator and denominator by $\beta$ and taking the limit as $\beta \to \infty$:

$$\lim_{\beta\to\infty} \frac{D_{\mathrm{KL}}(Q_{\mathbf{Y}}^{(\beta,\mathbf{B})}\|P_{\mathbf{Y}})}{D_{\mathrm{KL}}(Q_{\mathbf{X}}^{(\beta,\mathbf{B})}\|P_{\mathbf{X}})} = \frac{\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\}}{\mathsf{Tr}\left\{\mathbf{B}\right\}}.$$

To complete the proof, we need to maximize this expression over all valid matrices $\mathbf{B}$:

$$\sup_{\mathbf{B}:\mathbf{I}\preceq\mathbf{B}} \frac{\mathsf{Tr}\left\{\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\mathbf{B}\right\}}{\mathsf{Tr}\left\{\mathbf{B}\right\}}$$

Let $\mathbf{M} = \mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]$. We are maximizing $\frac{\mathsf{Tr}\{\mathbf{MB}\}}{\mathsf{Tr}\{\mathbf{B}\}}$. It is a known result in matrix analysis (e.g., related to Rayleigh quotients and Courant–Fischer theorem for matrices) that for a positive semi–definite matrix $\mathbf{M}$ and a positive definite matrix $\mathbf{B}$, this ratio is maximized when $\mathbf{B}$ is chosen to align with the dominant eigenvector of $\mathbf{M}$. Specifically, the supremum of this ratio is equal to the largest eigenvalue of $\mathbf{M}$, which is its operator norm.

$$\sup_{\mathbf{B}:\mathbf{I}\preceq\mathbf{B}} \frac{\mathsf{Tr}\left\{\mathbf{MB}\right\}}{\mathsf{Tr}\left\{\mathbf{B}\right\}} = \|\mathbf{M}\|_{\mathsf{op}} = \left\|\mathbb{E}_V\left[\mathbf{A}_V^\top\mathbf{A}_V\right]\right\|_{\mathsf{op}}.$$

Thus, we have shown that $s(P_{\mathbf{X}}, \widetilde{T}_{\mathbf{Y}|\mathbf{X}}) \geq \left\|\mathbb{E}_V[\mathbf{A}_V^\top\mathbf{A}_V]\right\|_{\mathsf{op}}$. Combining this with the previously established upper bound, the equality holds.

□

# E Proof Completion for Theorems 4.1 and 4.2

In this section, we complete the proof of Theorems 4.1 and 4.2. Note that the main part of proof is stated in Section 5 and we complete the proof here.

## E.1 Lower Bounds Related to Sample Complexity

Up to this point, we have established bounds related to the constraints on the communication budget. We now turn our attention to the bounds related to the sample size $m$. Here, we employ the standard Fano method. We analyze the cross–covariance case and the full covariance case separately.

### E.1.1 Cross Covariance

We utilize the same family of distributions employed in Subsections 5.3 and 5.6. The sole difference is the omission of the random variable $W$, which was used in the averaged Fano's method. More precisely, consider a set $\mathcal{V} = [1 : |\mathcal{V}|]$ and a corresponding family of distributions $\mathcal{P}_{\mathcal{V}} = \{P_v\}_{v\in\mathcal{V}}$, where $P_v = \mathcal{N}(\mathbf{0}, \mathbf{C}_v)$, and:

$$\mathbf{C}_v = \frac{\sigma^2}{2}\begin{bmatrix}\mathbf{I}_{d_1} & \delta\mathbf{D}_v^\top \\ \delta\mathbf{D}_v & \mathbf{I}_{d_2}\end{bmatrix}, \tag{136}$$

where $\mathbf{D}_v$ is some matrix in $\mathbb{R}^{d_1\times d_2}$ with $\|\mathbf{D}_v\|_{\mathsf{op}} \leq 1$, and $\delta \leq 1$ is a parameter to be determined subsequently. Analogous to (45), we have $I(V; M_1, M_2) \leq I(M_1; M_2|V)$. Furthermore, for a fixed $V = v$, we have the following Markov chain: $M_1 \multimap \mathbf{X}_1 \multimap \mathbf{X}_2 \multimap M_2$. The data processing inequality then implies $I(M_1; M_2|V) \leq I(\mathbf{X}_1; \mathbf{X}_2|V)$. In summary, we have:

$$I(V; M_1, M_2) \leq I(\mathbf{X}_1; \mathbf{X}_2|V). \tag{137}$$

We now proceed to derive an upper bound on $I(\mathbf{X}_1;\mathbf{X}_2|V)$:

$$\begin{aligned}
I(\mathbf{X}_1;\mathbf{X}_2|\,V=v) &= m\,I\left(\mathbf{X}_1;\mathbf{X}_2|V=v\right)\\
&= m\left[h(\mathbf{X}_1|V=v)+h(\mathbf{X}_2|V=v)-h(\mathbf{X}_1,\mathbf{X}_2|V=v)\right]\\
&\overset{\text{(a)}}{=} \frac{m}{2}\log_2\left(\frac{\det\left\{\frac{\sigma^2}{2}\mathbf{I}_{d1}\right\}\det\left\{\frac{\sigma^2}{2}\mathbf{I}_{d2}\right\}}{\det\{\mathbf{C}_v\}}\right)\\
&= \frac{m}{2}\left(2r_v\log_2(\frac{\sigma^2}{2})-\sum_{i=1}^{r_v}\log_2\left(\frac{\sigma^4}{4}\left(1-\delta^2\sigma_i^2(\mathbf{D}_v\right)\right)\right)\\
&= -\frac{m}{2}\left(\sum_{i=1}^{r_v}\log_2\left(1-\delta^2\sigma_i^2(\mathbf{D}_v\right)\right)\\
&\overset{\text{(b)}}{\leq} \frac{-mr_v}{2}\log_2\left(1-\delta^2\right)\\
&\overset{\text{(c)}}{\leq} mr_v\delta^2\\
&\overset{\text{(d)}}{\leq} m(d_1\wedge d_2)\delta^2,
\end{aligned} \tag{138}$$

where (a) follows from [Cov99, Theorem 8.4.1] and $r_v = \mathrm{rank}(\mathbf{D}_v)$. Furthermore, (b) holds because $g(x) = -\log_2(1-x) = \log_2(\frac{1}{1-x})$ is an increasing function, and $\|\mathbf{D}\|_{\mathsf{op}} \leq 1$. (c) follows from the fact that for all $x \in [0,1/2]$, we have $\log_2(\frac{1}{1-x}) \leq 2x$, and we assume that $\delta \leq \frac{1}{\sqrt{2}}$.

Then, Lemma 5.1 yields:

$$\begin{aligned}
\mathcal{M}_{\mathsf{dist}} &\geq \frac{\rho_{\mathsf{dist}}^{(\mathsf{cross})}}{2}\left[1-\frac{I(V;M_1,M_2)+1}{\log_2(|\mathcal{V}|)}\right]\\
&\geq \frac{\delta\sigma^2\nu_{\mathsf{dist}}^{(d_1,d_2)}}{8}\left[1-\frac{1+m(d_1\wedge d_2)\delta^2}{2d_1d_2}\right],
\end{aligned} \tag{139}$$

where $\nu_{\mathsf{dist}}^{(d_1,d_2)}$ was defined earlier, and we have used (53). Setting $\delta = \sqrt{\frac{d_1\vee d_2}{2m}\wedge\frac{1}{2}}$ implies:

$$\begin{aligned}
\mathcal{M}_{\mathsf{op}}^{(\mathsf{cross})} &\geq \frac{\sigma^2}{32}\left(\alpha_{\mathsf{op}}^{(\mathsf{sc})}\bigwedge 2\right)\\
\mathcal{M}_{\mathsf{F}}^{(\mathsf{cross})} &\geq \frac{\sigma^2}{32}\left(\alpha_{\mathsf{F}}^{(\mathsf{sc},\mathsf{cross})}\bigwedge\frac{\sqrt{d_1\wedge d_2}}{7}\right)
\end{aligned} \tag{140}$$

### E.1.2 Full Covariance

As mentioned in Remark 4.4, any lower bound in the centralized setting—where all data are aggregated on a central server—also applies to the distributed setting. It is folklore that the operator norm distortion in the centralized setting has a lower bound of $\Omega\left(\sqrt{\frac{d}{m}}\wedge 1\right)$, while the Frobenius norm distortion admits a lower bound of $\Omega\left(\sqrt{\frac{d^2}{m}}\wedge\sqrt{d}\right)$ [ABD+20, DMR20]. The arguments in [ABD+20] and [DMR20] are for $d \geq 9$ and $d \geq 5$, respectively. However, the argument used for the distributed setting also applies to the centralized case for $d \geq 2$.

To establish this, we split the vector $\mathbf{Z} = \{\mathbf{Z}^{(i)}\}_{i=1}^m$ into two equal–length subvectors: $\mathbf{X}_1' = \{\mathbf{Z}_{[1:d/2]}^{(i)}\}_{i=1}^m$ and $\mathbf{X}_2' = \{\mathbf{Z}_{[d/2+1:d]}^{(i)}\}_{i=1}^m$. We then use the same distributed argument with a new

family of distributions: $P_v = \mathcal{N}(\mathbf{0}, \mathbf{C}_v)$, where the covariance matrix has the form

$$\mathbf{C}_v = \frac{\sigma^2}{2}\begin{bmatrix} \mathbf{I}_{d/2} & \delta\mathbf{D}_v^\top \\ \delta\mathbf{D}_v & \mathbf{I}_{d/2} \end{bmatrix}. \tag{141}$$

Note that within the normal distributions characterized by (141), $\mathbf{X}_1'$ and $V$ are independent. Similarly, $\mathbf{X}_2'$ and $V$ are independent. By the Markov chain $V \multimap \mathbf{Z} = (\mathbf{X}_1', \mathbf{X}_2') \multimap (M_1, M_2)$ and the data processing inequality (DPI), we obtain:

$$\begin{aligned} I(V; M_1, M_2) &\le I(V; \mathbf{X}_1', \mathbf{X}_2') \\ &= I(V; \mathbf{X}_1') + I(V; \mathbf{X}_2') + I(\mathbf{X}_1'; \mathbf{X}_2'|V) - I(\mathbf{X}_1'; \mathbf{X}_2') \\ &\le I(\mathbf{X}'; \mathbf{Y}' \mid V). \end{aligned} \tag{142}$$

The rest of the proof follows similarly to the distributed case. Specifically, one can show

$$I(\mathbf{X}_1'; \mathbf{X}_2' \mid V) \le \frac{md\delta^2}{2},$$

as in (138). Then, applying the same inequalities used in (139) and (140), we obtain:

$$\begin{aligned} \mathfrak{M}_{\mathsf{op}}^{(\mathsf{cross})} &\ge \frac{\sigma^2}{32}\left(\alpha_{\mathsf{op}}^{(\mathsf{sc})} \bigwedge 2\right) \\ \mathfrak{M}_{\mathsf{F}}^{(\mathsf{cross})} &\ge \frac{\sigma^2}{32}\left(\alpha_{\mathsf{F}}^{(\mathsf{sc})} \bigwedge \frac{\sqrt{d_1 \wedge d_2}}{7}\right) \end{aligned} \tag{143}$$

## E.2 Lower bounds related to communication budget for self–covariance estimation.

In this section, we utilize the second family of Gaussian distributions. For the set $\mathcal{U}$, consider the set of distributions $\{P_u\}_{u\in\mathcal{U}}$, where $P_u = \mathcal{N}(\mathbf{0}, \mathbf{C}_u)$ and:

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

where $\mathbf{D}_u$ is a matrix in $\mathbb{R}^{d_1/2\times d_1/2}$.

As in the previous case, we must have $\mathbf{C}_u \succeq \mathbf{0}$ and $\|\mathbf{C}_u\|_{\mathsf{op}} \le \sigma^2$. Note that the eigenvalues of $\mathbf{C}_u$ are $\left\{\frac{\sigma^2}{2}\left(1 \pm \delta\sigma_i(\mathbf{D}_u)\right)\right\}_{i=1}^{\mathrm{rank}(\mathbf{D}_u)}$. Thus, if we assume that $\|\mathbf{C}_u\|_{\mathsf{op}} \le 1$ and $\delta \le 1$, the conditions $\mathbf{C}_u \succeq \mathbf{0}$ and $\|\mathbf{C}_u\|_{\mathsf{op}} \le \sigma^2$ are satisfied.

We then write:

$$\begin{aligned} \rho_{\mathsf{dist}} &= \inf_{u,u':u\neq u'} \|\mathbf{C}_u - \mathbf{C}_{u'}\|_{\mathsf{dist}} \\ &= \frac{\sigma^2}{2}\inf_{u,u':u\neq u'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_u - \mathbf{D}_{u'}) & \mathbf{0} & \mathbf{0} \\ \delta(\mathbf{D}_u - \mathbf{D}_{u'})^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\ &= \frac{\sigma^2}{2}\inf_{u,u':u\neq u'} \left\| \begin{bmatrix} \mathbf{0} & \delta(\mathbf{D}_u - \mathbf{D}_{u'}) \\ \delta(\mathbf{D}_u - \mathbf{D}_{u'})^\top & \mathbf{0} \end{bmatrix} \right\|_{\mathsf{dist}} \\ &\overset{(a)}{=} \sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\frac{\sigma^2\delta}{2}\inf_{u,u':u\neq u'} \|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\mathsf{dist}}, \end{aligned} \tag{144}$$

where (a) follows from Lemma A.1.

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

$$\begin{aligned} I(U; M_1, M_2) &= I(U; M_1) + I(U; M_2|M_1) \\ &\leq I(U; M_1) + I(U; M_2|M_1) + I(M_1; M_2) \\ &= I(U; M_1) + I(U, M_1; M_2) \\ &= I(U; M_1) + I(U; M_2) + I(M_1; M_2|U) \\ &\leq I(U; M_1) + I(U; \mathbf{Y}) + I(\mathbf{X}; \mathbf{Y}|U) \\ &= I(U; M_1) \\ &\leq B_1. \end{aligned} \tag{145}$$

Next, if we define the set $\{\mathbf{D}_u\}_{u\in\mathcal{U}}$ as the $\epsilon$–packing points of $\mathcal{B}^{(d_1^2/4)}_{|||.|||_{\mathsf{op}}}(1)$, under the $|||.|||_{\mathsf{dist}}$ norm, we have $\inf_{u,u':u\neq u'} \|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\mathsf{dist}} \geq \epsilon$, $\max_{u\in\mathcal{U}} \left\{ \left\|\mathbf{D}_u\right\|_{\mathsf{op}} \right\} \leq 1$, and $\log_2(|\mathcal{U}|) \geq \frac{d_1^2}{4} \log_2 \left( \frac{\nu^{(d_1/2)}_{\mathsf{dist}}}{\epsilon} \right)$, where:

$$\nu^{(d_1/2)}_{\mathsf{dist}} = \begin{cases} 1 & \text{if } \mathsf{dist} = \mathsf{op} \\ \frac{\sqrt{d_1}}{14\sqrt{2}} & \text{if } \mathsf{dist} = \mathsf{F} \end{cases}. \tag{146}$$

Now, if we set $\epsilon = \nu^{(d_1/2)}_{\mathsf{dist}} \cdot 2^{\frac{-16B_1}{d_1^2}}$ and $\delta = 1$, from Lemma 5.1 we have the following minimax lower bound:

$$\begin{aligned} \mathcal{M}_{\mathsf{dist}}(\sigma, B_1, B_2, d_1, d_2, m) &\geq \frac{\rho_{\mathsf{dist}}}{2} \left[ 1 - \frac{I(U; M_1, M_2) + 1}{\log_2(|\mathcal{U}|)} \right] \\ &\geq \frac{\sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\sigma^2}{2} \left[ 1 - \frac{B_1 + 1}{\log_2(|\mathcal{U}|)} \right] \inf_{u,u':u\neq u'} \|\mathbf{D}_u - \mathbf{D}_{u'}\|_{\mathsf{op}} \\ &\geq \frac{\sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}}\sigma^2}{2} \left[ 1 - \frac{B_1 + 1}{\frac{d_1^2}{4} \log_2 \left( \frac{\nu^{(d_1/2)}_{\mathsf{dist}}}{\epsilon} \right)} \right] \epsilon \\ &= \frac{\sigma^2}{4} \sqrt{1 + \mathbb{1}_{\{\mathsf{dist}=\mathsf{F}\}}} \nu^{(d_1/2)}_{\mathsf{dist}} \cdot 2^{\frac{-16B_1}{d_1^2}}. \end{aligned} \tag{147}$$

Therefore, we have:

$$\begin{aligned} \mathcal{M}_{\mathsf{op}}(\sigma, B_1, B_2, d_1, d_2, m) &\geq \frac{\sigma^2}{4} \cdot \left( 2^{\frac{-16B_1}{d_1^2}} \bigvee 2^{\frac{-16B_2}{d_2^2}} \right), \\ \mathcal{M}_{\mathsf{F}}(\sigma, B_1, B_2, d_1, d_2, m) &\geq \frac{\sigma^2}{56} \left( \sqrt{d_1} \cdot 2^{\frac{-16B_1}{d_1^2}} \bigvee \sqrt{d_2} \cdot 2^{\frac{-16B_2}{d_2^2}} \right). \end{aligned} \tag{148}$$

# F Some Concentration Inequalities for Random Matrices

In this section, we obtain two lemmas and one proposition that are useful in proving Theorem 4.5.

**Lemma F.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* $\mathbf{C}_{\mathbf{XY}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top]$ *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:*

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

*and:*

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

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

$$\begin{aligned}\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq t\Big] &\leq \mathbb{P}\Big[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}} \mathbf{u}^\top(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v} \geq \frac{t}{2}\Big] \\ &\leq \sum_{j=1}^{|\mathcal{N}_{1/4}^{(d_1)}|}\sum_{k=1}^{|\mathcal{N}_{1/4}^{(d_2)}|} \mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big],\end{aligned} \tag{149}$$

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 [Ver18, Corollary 4.2.13] that $\mathcal{N}_{1/4}^{(d)} \leq 9^d$.

We have:

$$\begin{aligned}&\mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] = \mathbb{P}\Big[\mathbf{u}^{(j)\top}(\frac{1}{m}\sum_{i=1}^{m}\mathbf{X}^{(i)}\mathbf{Y}^{(i)\top} - \mathbb{E}[\mathbf{X}\mathbf{Y}^\top])\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] \\ &= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^{m}(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}) - \mathbb{E}\Big[(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)})\Big] \geq \frac{t}{2}\Big].\end{aligned} \tag{150}$$

We know that $\mathbf{X}^{(i)}$ is a $\sigma_1$–sub–Gaussian vector, therefore, from Definition 2.4, we conclude that $U_i = \mathbf{u}^{(j)\top}\mathbf{X}^{(i)}$ is a $\sigma_1$–sub–Gaussian random variable. Similarly we conclude that $V_i = \mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}$ is a $\sigma_2$–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:

$$\begin{aligned}&\mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq \frac{t}{2}\Big] \\ &= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^{m}(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)}) - \mathbb{E}\Big[(\mathbf{u}^{(j)\top}\mathbf{X}^{(i)})(\mathbf{v}^{(k)\top}\mathbf{Y}^{(i)})\Big] \geq \frac{t}{2}\Big] \\ &= \mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^{m}(U_iV_i - \mathbb{E}[U_iV_i]) \geq \frac{t}{2}\Big] \\ &\leq \exp\Big(-m.\min\Big\{\frac{t}{10\sigma_1\sigma_2}, (\frac{t}{10\sigma_1\sigma_2})^2\Big\}\Big).\end{aligned} \tag{151}$$

Therefore we combine (149), (150), and (151) and write:

$$\begin{aligned}\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 10\sigma_1\sigma_2 t\Big] &\leq \sum_{j=1}^{|\mathcal{N}_{1/4}^{(d_1)}|}\sum_{k=1}^{|\mathcal{N}_{1/4}^{(d_2)}|} \mathbb{P}\Big[\mathbf{u}^{(j)\top}(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}^{(k)} \geq 5\sigma_1\sigma_2 t\Big] \\ &\leq (9)^{d_1+d_2}\exp\Big(-m.\min\Big\{t, t^2\Big\}\Big).\end{aligned} \tag{152}$$

Thus:

$$\begin{aligned}\mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 11\sigma_1\sigma_2\Big] &\leq \mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} + \big\|\mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 11\sigma_1\sigma_2\Big] \\ &\leq \mathbb{P}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}} \geq 10\sigma_1\sigma_2\Big] \\ &\leq \min\Big\{1, \exp\big((d_1+d_2)\ln(9) - m\big)\Big\} \\ &\leq \min\Big\{1, \exp\big(3(d_1+d_2) - m\big)\Big\}.\end{aligned} \tag{153}$$

□

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

**Proposition F.2.** *Assume that $\mathbf{X} \in \mathbb{R}^{d_1}$ is a zero mean, $\sigma_1^2$–sub–Gaussian random 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, $\sigma_2^2$–sub–Gaussian random 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 $\mathbf{C}_{\mathbf{XY}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top]$ 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:*

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

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

$$\begin{aligned}\mathbb{E}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}}\Big] &\leq 2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}} \mathbf{u}^\top(\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}})\mathbf{v}\right] \\ &= 2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}} \frac{1}{m}\sum_{i=1}^m \Big\{(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)}) - \mathbb{E}\Big[(\mathbf{u}^\top\mathbf{X}^{(i)})(\mathbf{v}^\top\mathbf{Y}^{(i)})\Big]\Big\}\right].\end{aligned} \tag{154}$$

Let:

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

Using similar reasoning to the one used in establishing (151), 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:

$$\begin{aligned}\mathbb{E}\Big[\big\|\widetilde{\mathbf{C}}_{\mathbf{XY}} - \mathbf{C}_{\mathbf{XY}}\big\|_{\mathsf{op}}\Big] &\leq 2\,\mathbb{E}\left[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d_1)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(d_2)}} Z_{\mathbf{u},\mathbf{v}}\right] \\ &\leq 10\sigma_1\sigma_2\sqrt{\frac{2\ln\big(|\mathcal{N}_{1/4}^{(d_1)}|.|\mathcal{N}_{1/4}^{(d_2)}|\big)}{m}} + 5\sigma_1\sigma_2\frac{\ln\big(|\mathcal{N}_{1/4}^{(d_1)}|.|\mathcal{N}_{1/4}^{(d_2)}|\big)}{m} \\ &\leq 10\sigma_1\sigma_2\sqrt{\frac{2(d_1+d_2)\ln(9)}{m}} + 5\sigma_1\sigma_2\frac{(d_1+d_2)\ln(9)}{m} \\ &\leq 32\sigma_1\sigma_2\max\left\{\sqrt{\frac{d_1+d_2}{m}}, \frac{d_1+d_2}{m}\right\}\end{aligned} \tag{155}$$

□

**Lemma F.3.** *Let* $\mathbf{A}$ *be a* $d \times n$ *random matrix whose columns* $\mathbf{A}_i$ *are independent, mean zero,* $\sigma$*–sub–Gaussian random vectors, then we have:*

$$\mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d+n}\Big] \leq \exp\big(-2(d+n)\big).$$

*Moreover, for* $q \in \{1,2\}$ *the following inequality holds:*

$$\mathbb{E}\Big[\|\mathbf{A}\|_{\mathsf{op}}^q\Big] \leq C_q \sigma^q (d+n)^{q/2}, \tag{156}$$

*for some universal constant* $C_q$ *depending only on* $q$*.*

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

$$\begin{aligned}\mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq t\Big] &\leq \mathbb{P}\Big[\max_{\mathbf{u}\in\mathcal{N}_{1/4}^{(d)}, \mathbf{v}\in\mathcal{N}_{1/4}^{(n)}} \mathbf{u}^\top \mathbf{A}\mathbf{v} \geq \frac{t}{2}\Big] \\ &\leq \sum_{i=1}^{|\mathcal{N}_{1/4}^{(d)}|}\sum_{j=1}^{|\mathcal{N}_{1/4}^{(n)}|} \mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big],\end{aligned} \tag{157}$$

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:

$$\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} = \sum_{k=1}^{n} v_k^{(j)} \mathbf{u}^{(i)\top}\mathbf{A}_k, \tag{158}$$

where $v_k^{(j)}$ is the $k$–th element of $\mathbf{v}^{(j)}$. Therefore, from Definition 2.4 and Lemma A.3, $\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)}$ is a sub–Gaussian random variable with parameter $\sigma$. Therefore we write:

$$\mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big] \leq \exp\Big(\frac{-t^2}{8\sigma^2}\Big). \tag{159}$$

We know from [Wai19, Lemma 5.7] that:

$$\mathcal{N}_{1/4}^{(d)} \leq 9^d, \qquad \mathcal{N}_{1/4}^{(n)} \leq 9^n. \tag{160}$$

Then from (157), we have:

$$\begin{aligned}\mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq t\Big] &\leq \sum_{i=1}^{|\mathcal{N}_{1/4}^{(d)}|}\sum_{j=1}^{|\mathcal{N}_{1/4}^{(n)}|} \mathbb{P}\Big[\mathbf{u}^{(i)\top}\mathbf{A}\mathbf{v}^{(j)} \geq \frac{t}{2}\Big] \\ &\leq 9^{n+d}\exp\Big(\frac{-t^2}{8\sigma^2}\Big).\end{aligned} \tag{161}$$

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

$$\begin{aligned}\mathbb{P}\Big[\|\mathbf{A}\|_{\mathsf{op}} \geq 6\sigma\sqrt{d+n}\Big] &\leq 9^{n+d}\exp\Big(\frac{-9(d+n)}{2}\Big) \\ &\leq \exp\big(-2(n+d)\big).\end{aligned} \tag{162}$$

The inequality (156) is a straightforward consequence of (161) and the integral representation of expectation. First we consider the case $q = 1$:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{A}\|_{\mathsf{op}}\right] &= \int_0^{+\infty} \mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \geq t\right] dt \\
&\leq \int_0^{+\infty} \min\left\{1, 9^{n+d} \exp\left(\frac{-t^2}{8\sigma^2}\right)\right\} dt \\
&= \int_0^{2\sigma\sqrt{2\ln(9)(n+d)}} dt + \int_{2\sigma\sqrt{2\ln(9)(n+d)}}^{+\infty} 9^{n+d} \exp\left(\frac{-t^2}{8\sigma^2}\right) dt \\
&\leq 2\sqrt{2\ln(9)}\sigma(n+d)^{1/2} + 1 \\
&\leq 9\sigma(n+d)^{1/2}.
\end{aligned} \tag{163}$$

Now we consider the case $q = 2$:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{A}\|_{\mathsf{op}}^2\right] &= \int_0^{+\infty} \mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}}^2 \geq u\right] du \\
&= 2\int_0^{+\infty} t\,\mathbb{P}\left[\|\mathbf{A}\|_{\mathsf{op}} \geq t\right] dt \\
&\leq 2\int_0^{+\infty} t \min\left\{1, 9^{n+d} \exp\left(\frac{-t^2}{8\sigma^2}\right)\right\} dt \\
&= 2\int_0^{2\sigma\sqrt{2\ln(9)(n+d)}} t dt + 2\int_{2\sigma\sqrt{2\ln(9)(n+d)}}^{+\infty} 9^{n+d} t \exp\left(\frac{-t^2}{8\sigma^2}\right) dt \\
&= 8\ln(9)\sigma^2(n+d) + 8\sigma^2 9^{n+d} \int_{\ln(9)(n+d)}^{+\infty} e^{-t'} dt' \\
&= 8\ln(9)\sigma^2(n+d) + 8\sigma^2 \\
&\leq 36\sigma^2(n+d).
\end{aligned} \tag{164}$$

□

# G Detailed Proof of Theorems 4.5 and 4.6

## G.1 Proof of Theorem 4.5

*Proof.* We prove the existence of a scheme having distortion error less than $\varepsilon$ (under operator norm), with the following choices for the sample number and communication budgets:

$$m \geq 2^{19} \frac{d}{\tilde{\varepsilon}^2}, \tag{165}$$

$$B_k \geq \frac{2^{18}\beta d_k d}{\tilde{\varepsilon}^2}, \tag{166}$$

$$n = \frac{\frac{B_1}{d_1} \wedge \frac{B_2}{d_2}}{\beta} \bigwedge m, \tag{167}$$

where for brevity we set $\tilde{\varepsilon} = \frac{\varepsilon}{\sigma^2} \leq 1$ and $\beta$ will be determined in the sequel. Also observe that (165)–(166) imply:

$$n \geq \frac{2^{18} d}{\tilde{\varepsilon}^2}. \tag{168}$$

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 $\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} > 11\sigma^2$.", "Receiving error from Agent 2, when $\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} > 11\sigma^2$.", "Receiving error from Agent

1, when $\|\mathbf{X}_1\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}$.", and "Receiving error from Agent 2, when $\|\mathbf{X}_2\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}$.", 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:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] \mathbb{P}\left[\mathcal{E}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right]. \tag{169}$$

We find an upper bound for every term of (169). First, notice that when an error is received in central server, the central server returns $\widehat{\mathbf{C}} = \mathbf{0}$, therefore:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] &= \mathbb{E}\left[\left\|\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] \\ &\leq \sigma^2.\end{aligned} \tag{170}$$

From Lemma F.1, we have:

$$\begin{aligned}\mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] &\leq \min\left\{1, \exp\left(6d_1 - m\right)\right\}, \\ \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] &\leq \min\left\{1, \exp\left(6d_2 - m\right)\right\}.\end{aligned}$$

Also Lemma F.3 yields:

$$\begin{aligned}\mathbb{P}\left[\left\|\mathbf{X}_1\right\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}\right] &\leq \exp\left(-2(d_1+n)\right), \\ \mathbb{P}\left[\left\|\mathbf{X}_2\right\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}\right] &\leq \exp\left(-2(d_2+n)\right)\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}\right] &\leq \mathbb{P}[\mathcal{E}_{1,1}] + \mathbb{P}[\mathcal{E}_{2,1}] + \mathbb{P}[\mathcal{E}_{1,2}] + \mathbb{P}[\mathcal{E}_{2,2}] \\ &= \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] + \mathbb{P}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \geq 11\sigma^2\right] \\ &\quad + \mathbb{P}\left[\left\|\mathbf{X}_1\right\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_1+n}\right] \\ &\quad + \mathbb{P}\left[\left\|\mathbf{X}_2\right\|_{\mathsf{op}} \geq 6\sigma\sqrt{d_2+n}\right] \\ &\leq \exp\left(6d_1 - m\right) + \exp\left(6d_2 - m\right) \\ &\quad + \exp\left(-2(d_1+n)\right) + \exp\left(-2(d_2+n)\right) \\ &\leq 2\exp\left(6d - m\right) + 2\exp(-2(n+1)) \\ &< \frac{\tilde{\varepsilon}}{10000}\end{aligned} \tag{171}$$

where the last equation follows from the inequalities $\exp(6d - m) \leq \exp(d(6 - 2^{19}\tilde{\varepsilon}^{-2})) \leq \exp(-2^{18})\tilde{\varepsilon}$ and $\exp(-2n) \leq \exp(-2^{19}d\tilde{\varepsilon}^{-2}) \leq \exp(-2^{19})\tilde{\varepsilon}$. Since $\tilde{\varepsilon} \leq 1$, we have:

$$\mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right] \geq 0.9999. \tag{172}$$

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

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+ - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*_+ - \widehat{\mathbf{C}}^*\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &= \mathbb{E}\left[\left|\lambda_{\min}(\widehat{\mathbf{C}}^*)\right| \mathbb{1}_{\{\lambda_{\min}(\widehat{\mathbf{C}}^*) < 0\}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\leq \mathbb{E}\left[\left|\lambda_{\min}(\widehat{\mathbf{C}}^*) - \lambda_{\min}(\mathbf{C})\right| \mathbb{1}_{\{\lambda_{\min}(\widehat{\mathbf{C}}^*) < 0\}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\quad + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\overset{(a)}{\leq} 2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right],\end{aligned} \tag{173}$$

where (a) is a consequence of Weyl's inequality [JH85, Section 4.3]. Also we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &= \mathbb{E}\left[\left\|\begin{bmatrix} \widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} & \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top \\ \frac{1}{n}\widehat{\mathbf{X}}_2\widehat{\mathbf{X}}_1^\top & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2} \end{bmatrix}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&= \mathbb{E}\left[\left\|\begin{bmatrix} \widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ \frac{1}{n}\widehat{\mathbf{X}}_2\widehat{\mathbf{X}}_1^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2} \end{bmatrix}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right].
\end{aligned} \tag{174}$$

We use matrix quantization scheme defined in Appendix A.6 to quantize matrices $\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}$, $\widetilde{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}$, $\mathbf{X}_1$, and $\mathbf{X}_2$. Therefore, we can use the relation between communication load and the resolution of this quantization, which is stated in Appendix A.6.

- Quantization of $\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} \in \mathbb{R}^{d_1\times d_1}$: $r = 11\sigma^2$, therefore:

$$d_1^2 \log_2\left(\frac{33\sigma^2}{\epsilon_1'}\right) = B_1' = \frac{B_1}{2} \Rightarrow \epsilon_1' = 33\sigma^2 \cdot 2^{\frac{-B_1}{2d_1^2}}. \tag{175}$$

- Quantization of $\widetilde{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} \in \mathbb{R}^{d_2\times d_2}$: $r = 11\sigma^2$, therefore:

$$d_2^2 \log_2\left(\frac{33\sigma^2}{\epsilon_2'}\right) = B_2' = \frac{B_2}{2} \Rightarrow \epsilon_2' = 33\sigma^2 \cdot 2^{\frac{-B_2}{2d_2^2}}. \tag{176}$$

- Quantization of $\mathbf{X}_1 \in \mathbb{R}^{d_1\times n}$: $r = 6\sigma\sqrt{d_1+n}$, therefore:

$$nd_1 \log_2\left(\frac{18\sigma\sqrt{d_1+n}}{\epsilon_1''}\right) = B_1'' = \frac{B_1}{2} \Rightarrow \epsilon_1'' = 18\sigma\sqrt{d_1+n} \cdot 2^{\frac{-B_1}{2nd_1}}. \tag{177}$$

- Quantization of $\mathbf{X}_2 \in \mathbb{R}^{d_2\times n}$: $r = 6\sigma\sqrt{d_2+n}$, therefore:

$$nd_2 \log_2\left(\frac{18\sigma\sqrt{d_2+n}}{\epsilon_2''}\right) = B_2'' = \frac{B_2}{2} \Rightarrow \epsilon_2'' = 18\sigma\sqrt{d_2+n} \cdot 2^{\frac{-B_2}{2nd_2}}. \tag{178}$$

From Proposition F.2 and the choice of $m$ in (165), we have (for $k = 1, 2$):

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k} - \mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \leq \frac{\mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k} - \mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}}\right]}{\mathbb{P}[\mathcal{E}^c]} \leq 32\sigma^2\sqrt{\frac{2d_k}{m}} \tag{179}$$

We also have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}_1\mathbf{X}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^c\right] &\leq \frac{\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}_1\mathbf{X}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}\right]}{\mathbb{P}[\mathcal{E}^c]} \\
&\leq 32\sigma^2 \max\left\{\sqrt{\frac{d_1+d_2}{n}}, \frac{d_1+d_2}{n}\right\} \\
&= 32\sigma^2\sqrt{\frac{d}{n}}.
\end{aligned} \tag{180}$$

From (179) and (175) we write:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}-\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \epsilon_1' + 32\sigma^2\sqrt{\frac{2d_1}{m}} \\
&\overset{\text{(a)}}{<} 33\sigma^2 \cdot 2^{\frac{-B_1}{2d_1^2}} + 32\sigma^2\sqrt{\frac{2d}{m}} \\
&\overset{\text{(b)}}{\leq} \sigma^2\left(33\cdot 2^{\frac{-2^{17}\beta d}{d_1\tilde{\varepsilon}^2}} + \frac{\tilde{\varepsilon}}{16}\right) \\
&\overset{\text{(c)}}{\leq} \sigma^2\left(33\cdot 2^{-2^{17}\beta} + \frac{\tilde{\varepsilon}}{16}\right)
\end{aligned} \tag{181}$$

where (a) and (b) follow from (165) and (166) and (c) follows because $d_1 < d$ and $\tilde{\varepsilon} \leq 1$.

Similarly we have:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}-\mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \leq \sigma^2\left(33\cdot 2^{-2^{17}\beta} + \frac{\tilde{\varepsilon}}{16}\right). \tag{182}$$

Next, we consider the estimation error of the *cross–covariance* matrix. From (180), (177), and (178) we write:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq \mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1(\widehat{\mathbf{X}}_2-\mathbf{X}_2)^\top\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\frac{1}{n}(\widehat{\mathbf{X}}_1-\mathbf{X}_1)\mathbf{X}_2^\top\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}_1\mathbf{X}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\left\|\widehat{\mathbf{X}}_1\right\|_{\mathsf{op}}\left\|\widehat{\mathbf{X}}_2-\mathbf{X}_2\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \frac{1}{n}\mathbb{E}\left[\left\|\mathbf{X}_2\right\|_{\mathsf{op}}\left\|\widehat{\mathbf{X}}_1-\mathbf{X}_1\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}_1\mathbf{X}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \frac{6\sigma\sqrt{d_1+n}}{n}\epsilon_2'' + \frac{6\sigma\sqrt{d_2+n}}{n}\epsilon_1'' + 32\sigma^2\sqrt{\frac{d}{n}} \\
&= \frac{108\sigma^2\sqrt{(d_1+n)(d_2+n)}}{n}\left(2^{\frac{-B_1}{2nd_1}} + 2^{\frac{-B_2}{2nd_2}}\right) + 32\sigma^2\sqrt{\frac{d}{n}} \\
&\leq \frac{108\sigma^2\sqrt{(d_1+n)(d_2+n)}}{n}\left(2^{\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}}\right) + 32\sigma^2\sqrt{\frac{d}{n}}.
\end{aligned} \tag{183}$$

The choice of $n$ in (167) implies:

$$2^{\frac{-1}{2n}\min\{\frac{B_1}{d_1},\frac{B_2}{d_2}\}} \leq 2^{-\frac{\beta}{2}}. \tag{184}$$

Also (168) implies:

$$\frac{d_k+n}{n} < 1 + \frac{d}{n} < 1 + \tilde{\varepsilon}^2 \leq 2. \tag{185}$$

Thus:

$$\frac{\sqrt{(d_1+n)(d_2+n)}}{n} = \sqrt{\frac{d_1+n}{n}.\frac{d_2+n}{n}} < 2. \tag{186}$$

Next consider:

$$\begin{aligned}\sqrt{\frac{d}{n}} &= \sqrt{\frac{d\beta}{\frac{B_1}{d_1}\wedge\frac{B_2}{d_2}}\bigvee\frac{d}{m}}\\ &= \sqrt{\frac{\beta d d_1}{B_1}\vee\frac{\beta d d_2}{B_2}\vee\frac{d}{m}}\\ &< \frac{\tilde{\varepsilon}}{512}.\end{aligned} \tag{187}$$

In summary, we have:

$$\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}\mid\mathcal{E}^c\right]\le\sigma^2\left(432\cdot 2^{-\frac{\beta}{2}}+\frac{\tilde{\varepsilon}}{16}\right). \tag{188}$$

Choosing $\beta=2\log_2\frac{6912}{\tilde{\varepsilon}}$ yields:

$$\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}\mid\mathcal{E}^c\right]\le\frac{\sigma^2\tilde{\varepsilon}}{8}=\frac{\varepsilon}{8}. \tag{189}$$

Also substituting the value of $\beta$ in (181) and (182) implies:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k}-\mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}}\mid\mathcal{E}^c\right]\le\frac{\varepsilon}{8},\qquad k=1,2. \tag{190}$$

Putting (174), (189) and (190) together, gives:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\right]\le\frac{3\varepsilon}{8}. \tag{191}$$

Finally from (173) we can write:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}}\mid\mathcal{E}\right]\mathbb{P}\left[\mathcal{E}\right]+\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{op}}\mid\mathcal{E}^{\mathsf{c}}\right]\mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right]\\ &\le\sigma^2\,\mathbb{P}\left[\mathcal{E}\right]+2\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^*-\mathbf{C}\right\|_{\mathsf{op}}\mid\mathcal{E}^c\right]\\ &<\frac{\varepsilon}{10}+\frac{3\varepsilon}{4}<\varepsilon.\end{aligned} \tag{192}$$

The proof of Theorem 4.5 is completed. □

## G.2 Proof of Theorem 4.6

*Proof.* The proof of Theorem 4.6 is approximately the same as the proof of Theorem 4.5 with some minor modifications.

We prove the existence of a scheme having distortion error less than $\varepsilon$ (under Frobenius norm), with the following choices for the sample number and communication budgets:

$$m\ge 2^{19}\frac{d}{\tilde{\varepsilon}^2}, \tag{193}$$

$$B_k\ge\frac{2^{18}\beta d_k d_{\min}}{\tilde{\varepsilon}^2}\bigvee 2d_k^2\log_2\left(\frac{528}{\tilde{\varepsilon}}\right), \tag{194}$$

$$n=\frac{\frac{B_1}{d_1}\wedge\frac{B_2}{d_2}}{\beta}\bigwedge m, \tag{195}$$

where for brevity we set $\tilde{\varepsilon}=\frac{\varepsilon}{\sigma^2\sqrt{d}}\le 1$, and $\beta$ will be determined in the sequel. Also observe that (193)–(194) imply:

$$n\ge\frac{2^{18}d_{\min}}{\tilde{\varepsilon}^2}. \tag{196}$$

We use the same error events $(\mathcal{E}_{i,j} : i = 1, 2, j = 1, 2)$ as in the proof of Theorem 4.5. Then similar calculations to (171) shows the inequality $\mathbb{P}[\mathcal{E}] < \frac{\tilde{\varepsilon}}{10000}$ still holds for new assignment of $m, n, B_k$. We consider two regimes for the distortion error $\varepsilon$ and specify the covariance matrix estimator $\widehat{\mathbf{C}}$ for each regime attaining the distortion error $\varepsilon$, separately.

**Case I: Reasonable distortion error**. In this regime, the distortion error satisfies $\varepsilon < 512\sigma^2\sqrt{d_{\min}}$. Here, we take the estimator $\widehat{\mathbf{C}}$ as in (56). Further we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}\right]\mathbb{P}[\mathcal{E}] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right]\mathbb{P}[\mathcal{E}^{\mathsf{c}}] \\ &\leq \sigma^2\sqrt{d}\frac{\tilde{\varepsilon}}{10000} + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right]\mathbb{P}[\mathcal{E}^{\mathsf{c}}] \\ &= \frac{\varepsilon}{10000} + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right]\mathbb{P}[\mathcal{E}^{\mathsf{c}}]\,.\end{aligned} \tag{197}$$

Next consider:

$$\begin{aligned}\mathbb{P}[\mathcal{E}^{\mathsf{c}}]\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_+^* - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\overset{\text{(a)}}{\leq} \mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right],\end{aligned} \tag{198}$$

where (a) follows from the following inequalities:

$$\begin{aligned}\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{F}}^2 &= \mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}^* - \mathbf{C}\right)^2\right\} \\ &= \mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}^* - \widehat{\mathbf{C}}_+^*\right)^2\right\} + \mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}_+^* - \mathbf{C}\right)^2\right\} + 2\mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}^* - \widehat{\mathbf{C}}_+^*\right)\left(\widehat{\mathbf{C}}_+^* - \mathbf{C}\right)\right\} \\ &\geq \left\|\widehat{\mathbf{C}}_+^* - \mathbf{C}\right\|_{\mathsf{F}}^2 + 2\mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}^* - \widehat{\mathbf{C}}_+^*\right)\left(\widehat{\mathbf{C}}_+^* - \mathbf{C}\right)\right\} \\ &= \left\|\widehat{\mathbf{C}}_+^* - \mathbf{C}\right\|_{\mathsf{F}}^2 + 2\mathsf{Tr}\left\{\left(\widehat{\mathbf{C}}_+^* - \widehat{\mathbf{C}}^*\right)\mathbf{C}\right\} \qquad (199) \\ &\geq \left\|\widehat{\mathbf{C}}_+^* - \mathbf{C}\right\|_{\mathsf{F}}^2, \qquad (200)\end{aligned}$$

in which (199) is true because the positive part $\widehat{\mathbf{C}}_+^*$ of $\widehat{\mathbf{C}}^*$ is orthogonal to the negative part $\widehat{\mathbf{C}}_-^* = \widehat{\mathbf{C}}^* - \widehat{\mathbf{C}}_+^*$ of it, and (200) is due to the fact that the trace of the multiplication of two positive semi–definite matrices is non–negative [AM05, Exercise 12.14].

Now, we use the following counterpart of (174):

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^* - \mathbf{C}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] &= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} & \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top \\ \frac{1}{n}\widehat{\mathbf{X}}_2\widehat{\mathbf{X}}_1^\top & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix} - \begin{bmatrix}\mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & \frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ \frac{1}{n}\widehat{\mathbf{X}}_2\widehat{\mathbf{X}}_1^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \qquad (201) \\ &\quad + \sqrt{2}\,\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\leq \sqrt{d_1}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \sqrt{d_2}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\ &\quad + \sqrt{2d_{\min}}\,\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^\top - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right].\end{aligned}$$

Using the same inequalities (175)–(178) and (181) with the new assignments (193)–(194) imply:

$$\begin{aligned}\sqrt{d_k}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k}-\mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}}\mid\mathcal{E}^{\mathsf{c}}\right]&\leq\sqrt{d_k}\sigma^2\left(33\cdot 2^{\frac{-B_k}{2d_k^2}}+32\sqrt{\frac{2d}{m}}\right)\\&\leq\sqrt{d_k}\sigma^2\left(33\cdot 2^{\log_2\left(\frac{\tilde{\varepsilon}}{528}\right)}+\frac{\tilde{\varepsilon}}{16}\right)\\&\leq\frac{\sqrt{d}\sigma^2\tilde{\varepsilon}}{8}<\frac{\varepsilon}{8}.\end{aligned}\tag{202}$$

Further (180) still holds in the reasonable distortion error regime $\varepsilon<512\sigma^2\sqrt{d_{\min}}$, because in this regime $d<n$, which is a consequence of (196). Moreover the inequalities (183)–(186) are still valid. The assignment (194) guaranties:

$$\begin{aligned}\sqrt{\frac{d}{n}}&=\sqrt{\frac{d\beta}{\frac{B_1}{d_1}\wedge\frac{B_2}{d_2}}\bigvee\frac{d}{m}}\\&=\sqrt{\frac{\beta dd_1}{B_1}\vee\frac{\beta dd_2}{B_2}\vee\frac{d}{m}}\\&<\frac{\tilde{\varepsilon}}{512}\sqrt{\frac{d}{d_{\min}}}.\end{aligned}\tag{203}$$

In summary, we have:

$$\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^{\top}-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}\mid\mathcal{E}^{\mathsf{c}}\right]\leq\sigma^2\left(432\cdot 2^{-\frac{\beta}{2}}+\frac{\tilde{\varepsilon}}{16}\sqrt{\frac{d}{d_{\min}}}\right).\tag{204}$$

Choosing $\beta=2\log_2\frac{6912\sigma^2\sqrt{d_{\min}}}{\varepsilon}$ yields:

$$\sqrt{d_{\min}}\,\mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^{\top}-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{F}}\mid\mathcal{E}^{\mathsf{c}}\right]\leq\frac{\varepsilon}{8}.\tag{205}$$

Putting (201), (202) and (205) together, gives:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^{*}-\mathbf{C}\right\|_{\mathsf{F}}\mid\mathcal{E}^{\mathsf{c}}\right]\leq\frac{\varepsilon}{2}.\tag{206}$$

Finally from (197) we write:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{F}}\right]&=\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{F}}\mid\mathcal{E}\right]\mathbb{P}\left[\mathcal{E}\right]+\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{F}}\mid\mathcal{E}^{\mathsf{c}}\right]\mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right]\\&\leq\sigma^2\sqrt{d}\,\mathbb{P}\left[\mathcal{E}\right]+\mathbb{E}\left[\left\|\widehat{\mathbf{C}}^{*}-\mathbf{C}\right\|_{\mathsf{F}}\mid\mathcal{E}^{\mathsf{c}}\right]\\&<\varepsilon.\end{aligned}\tag{207}$$

This concludes the proof of Theorem 4.6 in the reasonable distortion error regime.

**Case II: High distortion error**. In this regime, the distortion error satisfies $\varepsilon\geq 512\sigma^2\sqrt{d_{\min}}$. Here, we slightly modify the estimator $\widehat{\mathbf{C}}$ given in (56). In this regime, we do not quantize the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Instead, agent $k$ devotes all the communication budget $B_k$ for transmitting its self–covariance matrix estimator $\widehat{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k}$. The central server simply returns:

$$\widehat{\mathbf{C}}=\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} & \mathbf{0}_{d_1\times d_2}\\ \mathbf{0}_{d_2\times d_1} & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix}.\tag{208}$$

In this case, error events $\mathcal{E}_{1,2}$ and $\mathcal{E}_{2,2}$ will never occur, because we don't aim to quantize matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Note that if we define $\widetilde{\mathcal{E}}$ as $\widetilde{\mathcal{E}} = \mathcal{E}_{1,1} \vee \mathcal{E}_{2,1}$, then $\mathbb{P}[\widetilde{\mathcal{E}}] \leq \mathbb{P}[\mathcal{E}] < \frac{\tilde{\varepsilon}}{10000}$ and we can write:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}}\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}\right] \mathbb{P}\left[\widetilde{\mathcal{E}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}^{\mathsf{c}}\right] \mathbb{P}\left[\widetilde{\mathcal{E}}^{\mathsf{c}}\right] \\
&\leq \sigma^2\sqrt{d}\frac{\tilde{\varepsilon}}{10000} + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}^{\mathsf{c}}\right] \mathbb{P}\left[\widetilde{\mathcal{E}}^{\mathsf{c}}\right] \\
&= \frac{\varepsilon}{10000} + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}^{\mathsf{c}}\right] \mathbb{P}\left[\widetilde{\mathcal{E}}^{\mathsf{c}}\right].
\end{aligned} \tag{209}$$

Now, we use the following counterpart of (174):

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}^{\mathsf{c}}\right] &= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} & \mathbf{0}_{d_1\times d_2} \\ \mathbf{0}_{d_2\times d_1} & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix} - \begin{bmatrix}\mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ \mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&= \mathbb{E}\left[\left\|\begin{bmatrix}\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1} & -\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2} \\ -\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}^\top & \widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\end{bmatrix}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{F}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \sqrt{2}\left\|\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{F}} \\
&\leq \sqrt{d_1}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1} - \mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \sqrt{d_2}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2} - \mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \sqrt{2d_{\min}}\left\|\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}.
\end{aligned} \tag{210}$$

Using (175)– (176) with $B_1' = B_1, B_2' = B_2$, values of $m, n, B_k$ in (193)–(195), and $\beta = 2\log_2(\frac{6912}{\tilde{\varepsilon}})$, we have for $k = 1, 2$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k} - \mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq \mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k} - \widetilde{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\widetilde{\mathbf{C}}_{\mathbf{X}_k\mathbf{X}_k} - \mathbf{C}_{\mathbf{X}_k\mathbf{X}_k}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \epsilon_k' + 32\sigma^2\sqrt{\frac{2d_k}{m}} \\
&< 33\sigma^2 \cdot 2^{\frac{-B_k}{d_k^2}} + 32\sigma^2\sqrt{\frac{2d}{m}} \\
&\leq \sigma^2\left(33 \cdot 2^{\log_2(\frac{\tilde{\varepsilon}}{528})} + \frac{\tilde{\varepsilon}}{16}\right) \\
&\leq \frac{\sigma^2\tilde{\varepsilon}}{8}.
\end{aligned} \tag{211}$$

We can use this upper bound for $\left\|\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}}$:

$$\begin{aligned}
\left\|\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} &= \sup_{\substack{\mathbf{u}\in\mathbb{R}^{d_1},\mathbf{v}\in\mathbb{R}^{d_2} \\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}} \left\{\mathbf{u}^\top\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\mathbf{v}\right\} \\
&= \sup_{\substack{\mathbf{u}\in\mathbb{R}^{d_1},\mathbf{v}\in\mathbb{R}^{d_2} \\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}} \left\{\mathbb{E}\left[\mathbf{u}^\top\mathbf{X}_1\mathbf{X}_2^\top\mathbf{v}\right]\right\} \\
&\leq \sup_{\substack{\mathbf{u}\in\mathbb{R}^{d_1},\mathbf{v}\in\mathbb{R}^{d_2} \\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}} \left\{\sqrt{\mathbb{E}\left[(\mathbf{u}^\top\mathbf{X}_1)^2\right]\,\mathbb{E}\left[(\mathbf{v}^\top\mathbf{X}_2)^2\right]}\right\} \\
&\leq \sigma^2.
\end{aligned} \tag{212}$$

Now from (209), (210), (211), and (212), we conclude:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{F}}\right] &\leq \frac{\varepsilon}{10000}+\mathbb{E}\left[\left\|\widehat{\mathbf{C}}-\mathbf{C}\right\|_{\mathsf{F}} \mid \widetilde{\mathcal{E}}^{\mathsf{c}}\right]\mathbb{P}\left[\widetilde{\mathcal{E}}^{\mathsf{c}}\right] \\
&\leq \frac{\varepsilon}{10000}+\sqrt{d_1}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_1\mathbf{X}_1}-\mathbf{C}_{\mathbf{X}_1\mathbf{X}_1}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right]+\sqrt{d_2}\,\mathbb{E}\left[\left\|\widehat{\mathbf{C}}_{\mathbf{X}_2\mathbf{X}_2}-\mathbf{C}_{\mathbf{X}_2\mathbf{X}_2}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad+\sqrt{2d_{\min}}\left\|\mathbf{C}_{\mathbf{X}_1\mathbf{X}_2}\right\|_{\mathsf{op}} \\
&\leq \frac{\varepsilon}{10000}+\frac{\sigma^2\sqrt{d_1}\tilde{\varepsilon}}{8}+\frac{\sigma^2\sqrt{d_2}\tilde{\varepsilon}}{8}+\sqrt{2d_{\min}}\sigma^2 \qquad (213)\\
&\overset{(a)}{\leq} \frac{\varepsilon}{10000}+\frac{\varepsilon}{4}+\frac{\varepsilon}{256\sqrt{2}} \\
&\leq \varepsilon,
\end{aligned}$$

where (a) follows from the condition $\varepsilon \geq 512\sigma^2\sqrt{d_{\min}}$. □

# H Achievable Scheme for Multi–Agent Scenario: Proof of Theorem 4.10

*Proof.* We prove Theorem 4.10 by first focusing on a simplified case, specifically the fully distributed scenario where there is a distinct agent for each dimension ($K = d$) and each agent controls only one dimension ($d_k = 1$).

This proof holds for the more general case as well. Any agent that possesses $d_k > 1$ dimensions can be treated as $d_k$ "virtual" agents, each responsible for a single dimension. By applying our coding scheme to these virtual agents, we can reduce the general problem to the fully distributed case.

Let's make the following choices for the number of samples and the parameter $\beta$:

$$\begin{aligned}
m \geq n &:= \frac{d}{\tilde{\varepsilon}^2}, \\
\beta &= \frac{\varepsilon}{2\sigma^2}.
\end{aligned} \qquad (214)$$

The exact value of $\tilde{\varepsilon}$ will be determined in the sequel. We can proof Theorem 4.10 in 3 steps:

**Randomized quantization at user $k$** Let's represent the first $n$ samples from agent $k$ as the vector $\mathbf{X}_k = \left[X_k^{(1)}, X_k^{(2)}, \ldots, X_k^{(n)}\right] \in \mathbb{R}^n$. It is well–known that, with high probability, the magnitude of every sample is bounded. Specifically, we have $\left|X_k^{(i)}\right| \leq L$ for all agents $k$ and all samples $i$, with probability at least $1-\beta$, where $L := \sigma\sqrt{2\log(dn/\beta)}$. This means the entire collection of $nd$ samples is guaranteed to lie within the $nd$–dimensional hypercube $[-L, L]^{nd}$. If $\|\mathbf{X}_k\|_\infty > L$, Agent $k$ transmits an error signal. Otherwise, Agent $k$ quantizes its observation vector coordinate–wise as follows. For simplicity, we assume $\frac{L}{\sigma\tilde{\varepsilon}}$ is a positive integer, and define $N = \frac{L}{\sigma\tilde{\varepsilon}}$. The quantization is performed by partitioning the interval $[-L, L]$ into $2N$ contiguous intervals of length $\sigma\tilde{\varepsilon}$, denoted by $\left\{\left[j\sigma\tilde{\varepsilon}, (j+1)\sigma\tilde{\varepsilon}\right)\right\}_{j=-N}^{N-2}$ and $\left[(N-1)\sigma\tilde{\varepsilon}, N\sigma\tilde{\varepsilon}\right]$. Now for each $k \in [d]$ and $i \in [n]$, if $X_k^{(i)} \in \left[j\sigma\tilde{\varepsilon}, (j+1)\sigma\tilde{\varepsilon}\right)$ for some integer $j \in \{-N, -N+1, \ldots, N-1\}$, it is randomly and independently quantized to $\hat{X}_k^{(i)} \in \left\{j\sigma\tilde{\varepsilon}, (j+1)\sigma\tilde{\varepsilon}\right\}$ such that $\mathbb{E}\left[\hat{X}_k^{(i)} \mid X_k^{(i)}\right] = X_k^{(i)}$.
This quantization requires $B_k^{(i)} := \log_2(2N+1)$ bits per coordinate. Thus, $\mathbf{X}_k$ is quantized to $\widehat{\mathbf{X}}_k := \left[\hat{X}_k^{(1)}, \ldots, \hat{X}_k^{(n)}\right]$ using $B_k := \sum_{i=1}^n B_k^{(i)} = n\log_2(2N+1)$ bits.

**Covariance Matrix Estimation at the Server** If the central server receives an error signal from any agent, it immediately sets $\widehat{\mathbf{C}} = \mathbf{0}$. Otherwise, after receiving the quantized vectors $\widehat{\mathbf{X}}_k$ from all agents, the server constructs a single aggregated matrix by vertically stacking the received vectors:

$$\widehat{\mathbf{X}} := \begin{bmatrix} \widehat{\mathbf{X}}_1 \\ \vdots \\ \widehat{\mathbf{X}}_d \end{bmatrix}. \tag{215}$$

The server then computes the final covariance matrix estimate using the sample covariance estimator:

$$\widehat{\mathbf{C}} = \frac{1}{n}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top. \tag{216}$$

**Analysis:** Let $\mathcal{E}$ be the event of receiving the error signal from at least one agent. Thus $\mathcal{E}$ occurs if $|X_k^{(i)}| > L$, for some $(k, i)$. We have:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] = \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] \mathbb{P}\left[\mathcal{E}\right] + \mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right]. \tag{217}$$

We find an upper bound for every term of (217). First, notice that when an error is received in central server, the central server returns $\widehat{\mathbf{C}} = \mathbf{0}$, therefore:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] = \mathbb{E}\left[\left\|\mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}\right] \leq \sigma^2. \tag{218}$$

Now we can find an upper bound for $\mathbb{P}[\mathcal{E}]$ and $\mathbb{P}[\mathcal{E}^{\mathsf{c}}]$:

$$\mathbb{P}[\mathcal{E}] \leq \beta, \qquad \mathbb{P}[\mathcal{E}^{\mathsf{c}}] \leq 1. \tag{219}$$

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

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\left((\widehat{\mathbf{X}} - \mathbf{X})(\widehat{\mathbf{X}} - \mathbf{X})^\top + (\widehat{\mathbf{X}} - \mathbf{X})\mathbf{X}^\top + \mathbf{X}(\widehat{\mathbf{X}} - \mathbf{X})^\top\right) + \frac{1}{n}\mathbf{X}\mathbf{X}^\top - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\left\|\left(\widehat{\mathbf{X}} - \mathbf{X}\right)\left(\widehat{\mathbf{X}} - \mathbf{X}\right)^\top\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \frac{2}{n}\mathbb{E}\left[\left\|\left(\widehat{\mathbf{X}} - \mathbf{X}\right)\mathbf{X}^\top\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}\mathbf{X}^\top - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}}^2 \mid \mathcal{E}^{\mathsf{c}}\right] + \frac{2}{n}\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}}\left\|\mathbf{X}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] + \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}\mathbf{X}^\top - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right].
\end{aligned} \tag{220}$$

We observe that when $\mathbf{X}$ satisfies the event $\mathcal{E}^{\mathsf{c}}$, the entries of the random matrix $\widehat{\mathbf{X}} - \mathbf{X}$ are conditionally independent. Furthermore, all the entries are bounded, $|\widehat{X}_k^{(i)} - X_k^{(i)}| < \sigma\tilde{\varepsilon}$, and are zero mean. Thus each entry is $\sigma\tilde{\varepsilon}$–sub–Gaussian. Hence given $(\mathbf{X}, \mathcal{E}^{\mathsf{c}})$, $\widehat{\mathbf{X}} - \mathbf{X}$ satisfies the conditions of Lemma F.3, and we can invoke this lemma to obtain:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}}^2 \mid (\mathbf{X}, \mathcal{E}^{\mathsf{c}})\right] &\leq 36\sigma^2\tilde{\varepsilon}^2(d+n), \\
\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}} \mid (\mathbf{X}, \mathcal{E}^{\mathsf{c}})\right] &\leq 9\sigma\tilde{\varepsilon}\sqrt{d+n}.
\end{aligned} \tag{221}$$

These yield:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}}^2 \mid \mathcal{E}^{\mathsf{c}}\right] &\leq 36\sigma^2\tilde{\varepsilon}^2(d+n), \\
\mathbb{E}\left[\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_{\mathsf{op}}\left\|\mathbf{X}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq 9\sigma\tilde{\varepsilon}\sqrt{d+n}\,\mathbb{E}\left[\left\|\mathbf{X}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right].
\end{aligned} \tag{222}$$

Furthermore $\mathbf{X}$ satisfies the conditions of Lemma F.3, and we have:

$$\mathbb{E}\left[\left\|\mathbf{X}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \leq \frac{\mathbb{E}\left[\left\|\mathbf{X}\right\|_{\mathsf{op}}\right]}{\mathbb{P}\left[\mathcal{E}^{\mathsf{c}}\right]} \leq \frac{9\sigma\sqrt{d+n}}{1-\beta}. \tag{223}$$

We also have:

$$\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{X}\mathbf{X}^{\top} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] \leq 32\sigma^2\sqrt{\frac{d}{n}}. \tag{224}$$

Putting (222), (223), and (224) in (220) yields:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}} \mid \mathcal{E}^{\mathsf{c}}\right] &\leq 36\sigma^2\tilde{\varepsilon}^2\frac{d+n}{n} + \frac{162}{1-\beta}\sigma^2\tilde{\varepsilon}\frac{d+n}{n} + 32\sigma^2\sqrt{\frac{d}{n}} \\ &\leq 360\sigma^2\frac{d+n}{n}\tilde{\varepsilon} + 32\sigma^2\tilde{\varepsilon} \\ &\leq 760\sigma^2\tilde{\varepsilon}.\end{aligned} \tag{225}$$

By setting $\tilde{\varepsilon} = \frac{\varepsilon}{1520\sigma^2}$ and substituting (225) in the expression (217), we conclude that the achievable scheme has the desired distortion error less than $\varepsilon$:

$$\mathbb{E}\left[\left\|\widehat{\mathbf{C}} - \mathbf{C}\right\|_{\mathsf{op}}\right] \leq \beta\sigma^2 + \frac{\varepsilon}{2} \leq \varepsilon. \tag{226}$$

Finally the communication budget of each agent in this scheme satisfies:

$$B_k = n\log_2(2N+1) \leq \tau'\frac{\sigma^4 d}{\varepsilon^2}\log_2\left(\tau''\frac{\sigma^4}{\varepsilon^2}\log(d\sigma^2/\varepsilon)\right), \tag{227}$$

for some constant $\tau', \tau''$. This concludes the proof.

□