

# Covariance alignment: from maximum likelihood estimation to Gromov–Wasserstein

Yanjun Han\*

Philippe Rigollet<sup>†‡</sup>George Stepaniants<sup>†‡</sup>

November 23, 2023

## Abstract

Feature alignment methods are used in many scientific disciplines for data pooling, annotation, and comparison. As an instance of a permutation learning problem, feature alignment presents significant statistical and computational challenges. In this work, we propose the *covariance alignment* model to study and compare various alignment methods and establish a minimax lower bound for covariance alignment that has a non-standard dimension scaling because of the presence of a nuisance parameter. This lower bound is in fact minimax optimal and is achieved by a natural quasi MLE. However, this estimator involves a search over all permutations which is computationally infeasible even when the problem has moderate size. To overcome this limitation, we show that the celebrated Gromov–Wasserstein algorithm from optimal transport which is more amenable to fast implementation even on large-scale problems is also minimax optimal. These results give the first statistical justification for the deployment of the Gromov–Wasserstein algorithm in practice.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Main results	4
1.2	Related work	5
1.3	Notation	7
<b>2</b>	<b>Covariance alignment estimators</b>	<b>8</b>
2.1	Estimators	8
2.1.1	Quasi Maximum Likelihood Estimator	8
2.1.2	Gromov–Wasserstein Estimator	9
2.2	Minimax optimality	10
2.3	Numerical experiments	10
<b>3</b>	<b>Proof of upper bounds</b>	<b>12</b>
3.1	Upper bound of the QMLE	13
3.1.1	Good events	13
3.1.2	Lower bound of the loss difference	14

---

\*Courant Institute of Mathematical Sciences and Center for Data Science, New York University, New York NY, United States

<sup>†</sup>Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge MA, United States

<sup>‡</sup>Department of Mathematics, Massachusetts Institute of Technology, Cambridge MA, United States

3.1.3	A trace-Frobenius inequality . . . . .	14
3.2	Upper bound of the GW estimator . . . . .	16
3.2.1	Good events . . . . .	16
3.2.2	Lower bound of the loss difference . . . . .	17
<b>4</b>	<b>Proof of lower bounds</b>	<b>18</b>
4.1	Construction of the prior . . . . .	18
4.2	Upper bounding the mutual information . . . . .	19
4.2.1	Reduction to the case $m = n$ . . . . .	20
4.2.2	Second moment method . . . . .	21
<b>5</b>	<b>Discussions</b>	<b>22</b>
5.1	Other notions of minimax risk . . . . .	22
5.2	Comparison with orthogonal statistical learning . . . . .	23
<b>A</b>	<b>Auxiliary lemmas</b>	<b>24</b>
<b>B</b>	<b>Deferred proofs in Section 3</b>	<b>25</b>
B.1	Proof of Lemma 2 . . . . .	25
B.2	Proof of Lemma 3 . . . . .	27
<b>C</b>	<b>Deferred proofs in Section 4</b>	<b>28</b>
C.1	Proof of Lemma 5 . . . . .	28
C.2	Proof of Lemma 6 . . . . .	29
C.3	Proof of Lemma 7 . . . . .	30
C.4	Proof of Lemma 8 . . . . .	30
C.5	Proof of Lemma 9 . . . . .	31
C.6	Proof of Lemma 10 . . . . .	31
<b>D</b>	<b>Deferred proofs in Section 5</b>	<b>32</b>
D.1	Proof of Theorem 5 . . . . .	32
D.1.1	Proof of (29) . . . . .	32
D.1.2	Proof of (30) . . . . .	33
D.2	Proof of Theorem 6 . . . . .	34
D.2.1	Proof of upper bound . . . . .	34
D.2.2	Proof of lower bound . . . . .	35

# 1 Introduction

Modern research is marked by an unprecedented access to vast computational, data collection, and storage resources. Standing in for hand-crafted and expensive statistical experiments, a deluge of data is available at a scale on which even straightforward statistical techniques have the potential to yield remarkably nontrivial insights. This paradigm shift has transformed countless domains in science, engineering, and management science, to name a few, and also brought a new set of challenges including the integration, or *alignment*, of several datasets collected from different sources. This work proposes a statistical framework to study the fundamental limits associated with this phenomenon and establishes the optimality of state-of-the-art alignment methods based on optimal transport.

Motivated by applications to biostatistics, we formulate the feature alignment problem as follows. Consider two datasets that consist of  $m$  and  $n$  observations respectively. Each dataset can be represented as a features-by-observations matrix:

$$X^m = [X_1, \dots, X_m] \in \mathbb{R}^{d \times m}, \quad Y^n = [Y_1, \dots, Y_n] \in \mathbb{R}^{d \times n}. \quad (1)$$

The goal of feature alignment is to find a one-to-one mapping (permutation)  $\pi : S \rightarrow [d]$  from a subset  $S \subseteq [d]$  of features in  $X^m$  to a subset  $\pi(S) \subseteq [d]$  of features in  $Y^n$ .

This setup arises for example in untargeted metabolomics, proteomics, and lipidomics studies which record the concentrations of  $d$  compounds (metabolites, proteins, and lipids) across a collection of  $n$  patients to find biomarker compounds that are associated with the health status of a patient. The term “untargeted” refers to the fact that each of the  $d$  features corresponds to the concentration level of an unnamed extracted compound measured across all patients in the study. Matching features (compounds) between two such studies is a necessary step for transferring annotated labels between studies, comparing studies to discover shared compounds, and merging multiple studies to increase their sample size and ultimately increase the statistical power of downstream inference tasks.

To investigate this question, we propose and analyze a novel statistical model where the observations are centered Gaussian vectors in the idealized case where there exists a matching of all  $d$  features ( $S = [d]$ ). More specifically, let  $\Sigma \succ 0$  be an unknown covariance matrix and let  $\pi^*$  be an unknown permutation of  $[d]$ . We assume that

$$X_1, \dots, X_m \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma) \quad \text{and} \quad Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma^{\pi^*})$$

where  $(\Sigma^{\pi^*})_{i,j} := \Sigma_{\pi^*(i), \pi^*(j)}$  and the two samples are assumed to be independent. It is not hard to see that in this context the sample covariance matrices  $\hat{\Sigma}_X = X^m(X^m)^\top / m$  and  $\hat{\Sigma}_Y = Y^n(Y^n)^\top / n$  associated to each of the two samples are sufficient statistics for  $\pi^*$  and the problem boils down to aligning these covariance matrices. We call this problem *covariance alignment*.

Treating the covariance matrix  $\Sigma$  as a nuisance parameter and  $\pi^*$  as the parameter of interest, one may write a profile log-likelihood for  $\pi$ —see (7)—which can be further simplified into a quasi log-likelihood that is quadratic in  $\pi$ —see (8). Our first estimator is obtained by maximizing this quasi log-likelihood and is thus called the *quasi maximum likelihood estimator* (QMLE in short):

$$\hat{\pi}^{\text{QMLE}} = \underset{\pi \in S_d}{\operatorname{argmin}} \langle P_\pi \hat{\Sigma}_X^{-1} P_\pi^\top, \hat{\Sigma}_Y \rangle, \quad (2)$$

where  $S_d$  denotes the set of all permutations and  $P_\pi$  is the permutation matrix associated to  $\pi$ . One contribution of the paper is that the QMLE is optimal in a minimax sense. Unfortunately, the statistical optimality of the QMLE is contrasted by its associated computational burden. Indeed, the QMLE is a quadratic assignment problem that requires optimization over the set  $S_d$  of  $d!$  permutations, which is not practically feasible, even for moderate values of  $d$ . To overcome this situation, we borrow from computational optimal transport [PC<sup>+</sup>17] and propose a Gromov–Wasserstein (GW) estimator which optimizes a similar objective over the Birkhoff polytope  $\text{BP}(d)$  defined in (11) as the convex hull of the set of permutation matrices:

$$\hat{P}^{\text{GW}} = \underset{P \in \text{BP}(d)}{\operatorname{argmax}} \langle P \hat{\Sigma}_X P^\top, \hat{\Sigma}_Y \rangle. \quad (3)$$

We show in Section 2 that  $\hat{P}^{\text{GW}} = P_{\hat{\pi}^{\text{GW}}}$  is in fact a permutation matrix corresponding to a permutation  $\hat{\pi}^{\text{GW}}$  so that this relaxation is tight. Despite the structural similarity between the objective

functions in the previous two displays, we note in Section 2 that the same relaxation for the QMLE is not tight; in other words, the solution to

$$\operatorname{argmin}_{P \in \text{BP}(d)} \langle P \hat{\Sigma}_X^{-1} P^\top, \hat{\Sigma}_Y \rangle,$$

is not achieved at a permutation in general, but rather in the interior of the Birkhoff polytope. While this observation does not preclude the existence of good rounding schemes, we do not pursue this route. Instead, we show that the GW estimator is also minimax optimal.

In light of its advantageous computational properties, this work provides theoretical support in favor of the Gromov–Wasserstein matcher recently proposed in a metabolomic study [BSKR<sup>+</sup>23].

**Remark 1.** *While the Birkhoff polytope is tractable in the sense that it is convex and described by  $O(n^2)$  inequalities, we still have to resort to a heuristic to solve (3) because the objective is nonconvex. In fact, it is a convex maximization problem and a classical maximum principle indicates that the solution is achieved at a vertex of the Birkhoff polytope, namely a permutation. Despite this lack of convexity, the objective function is smooth and lends itself to the gradient-based heuristic in [SPKS16] and is successfully deployed in a companion applied work [BSKR<sup>+</sup>23].*

## 1.1 Main results

We first recall the covariance alignment model. Let  $\Sigma \succ 0$  be an unknown covariance matrix and let  $\pi^*$  be an unknown permutation of  $[d]$ . Assume that we observe

$$X_1, \dots, X_m \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma) \quad \text{and} \quad Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma^{\pi^*})$$

where  $(\Sigma^{\pi^*})_{i,j} := \Sigma_{\pi^*(i), \pi^*(j)}$  and the two samples are assumed to be independent.

Our goal is to estimate the permutation  $\pi^*$ . While  $\Sigma$  is a nuisance parameter, it does play an important role in our ability to identify  $\pi^*$ . For example, if  $\Sigma$  is the identity or has two identical columns, then  $\pi^*$  is not identifiable. From this simple observation, it is clear that the performance of any estimator will be greatly influenced by the structure of  $\Sigma$ . Rather than imposing structural conditions on  $\Sigma$ , we propose to evaluate our estimation using a  $\Sigma$ -dependent (pseudo-)distance, following an approach similar to the one employed in [FMR19] for statistical seriation. More specifically, for any two permutations  $\pi, \pi' \in S_d$ , define

$$d_\Sigma(\pi, \pi') := \|\Sigma^\pi - \Sigma^{\pi'}\|_F.$$

It follows readily from the triangle inequality that  $d_\Sigma$  is a pseudo-distance for any  $\Sigma \succeq 0$ . It is less sensitive to errors that consist in permuting two rows/columns that are close in Euclidean norm. More discussions on the loss function can be found in Section 5.1.

Our main result is the following characterization of the optimal sample complexity for estimating the permutation  $\pi$  under the Frobenius loss.

**Theorem 1.** *For  $\|\Sigma\|_{\text{op}} \leq 1$  and  $\varepsilon \in (0, 1)$ , both the QMLE  $\hat{\pi}^{\text{QMLE}}$  (2) and the GW estimator  $\hat{\pi}^{\text{GW}}$  (3) achieve  $\mathbb{E}\|\Sigma^{\hat{\pi}} - \Sigma^{\pi^*}\|_F \leq \varepsilon$  as long as*

$$\min\{m, n\} \geq C \frac{d \log d}{\varepsilon^2} \quad \text{and} \quad mn \geq C \frac{d^3 \log d}{\varepsilon^4}.$$

Here  $C > 0$  is a numerical constant independent of  $(d, \varepsilon)$ . Up to multiplicative constants, the above conditions on the sample complexity are not improvable in general.

The results on the minimax rate of convergence, as well as the high probability upper bounds on the Frobenius loss, are deferred to Theorems 2-4. Some remarks of Theorem 1 are in order:

(i)  $\min\{m, n\} = \Omega((d \log d)/\varepsilon^2)$  ensures that based solely on  $X^m$  or  $Y^n$ , there is an estimator  $\hat{\pi}$  with  $\mathbb{E}\|\Sigma^{\hat{\pi}} - \Sigma^{\pi^*}\|_F \leq \varepsilon$  when  $\Sigma$  is known. In other words, this is the sample complexity of permutation estimation with a known covariance.

(ii)  $mn = \Omega((d^3 \log d)/\varepsilon^4)$  accounts for the effect of an unknown covariance  $\Sigma$ , and interpolates between two extreme scenarios. First, if  $m = \Omega(d^2/\varepsilon^2)$  and  $n = \Omega((d \log d)/\varepsilon^2)$ , one may estimate  $\Sigma$  within Frobenius loss  $\varepsilon$  solely based on  $X^m$ , and then estimate  $\pi^*$  based on  $Y^n$  as if  $\Sigma$  were known—see (i) above. A more illuminating scenario arises if  $m = n$ . In that case, the above condition reduces to  $m = n = \Omega(\sqrt{d^3 \log d}/\varepsilon^2)$ . This peculiar sample complexity is in sharp contrast with the  $\Theta(d^2/\varepsilon^2)$  sample size required to estimate the covariance matrix  $\Sigma$  within accuracy  $\varepsilon$ ; in other words, the nuisance is allowed to be estimated at a slower rate without hurting the estimation of target parameters. This phenomenon is reminiscent of high-dimensional semiparametric statistics [CCD<sup>+</sup>17].

(iii)  $\|\Sigma\|_{\text{op}} \leq 1$  ensures that the hard problem instance appears in the high-dimensional scenario. If it is replaced by the Frobenius norm, the hard instance will become low-dimensional and we no longer enjoy the  $\text{poly}(d)$  reduction in the sample complexity compared with covariance estimation. This issue could be further mitigated by considering a rescaled Frobenius loss; we discuss the details in Section 5.1.

## 1.2 Related work

The estimation of permutations has become a preponderant challenge in various problems. It brings singular challenges, both from a statistical and computational perspective. In the rest of this section, we review some work in this growing area and comment on its connection with our setup and contributions.

**Dataset alignment and biological applications** Dataset alignment is a primitive problem in statistics and machine learning, used broadly in computer vision for matching local descriptors in images [Low04, BTVG06, SPKS16], machine translation for converting words between languages [AGCJ18, AMJ18, AMJJ19, GJB19], and biostatistics for matching untargeted compounds, transferring labels, and merging unlabeled datasets [LHM13, HKB<sup>+</sup>21, CPKL<sup>+</sup>22, BSKR<sup>+</sup>23]. The alignment problem is to find a correspondence between two  $d$ -dimensional datasets  $X^m = [X_1, \dots, X_m] \in \mathbb{R}^{d \times m}$ ,  $Y^n = [Y_1, \dots, Y_n] \in \mathbb{R}^{d \times n}$  either by matching their features (rows), matching their observations (columns), or both.

There are several biological applications where the features (rows) of  $X^m, Y^n$  are unordered or misaligned, and hence need to be matched. These applications include untargeted metabolomics, proteomics, and lipidomics experiments where unnamed features (compounds such as metabolites, proteins, and lipids) are collected and their concentrations are measured in a set of patients (observations). Untargeted studies are common practice in biology as they can be simpler to perform than targeted studies with labeled features and can discover new biomarker compounds that enable drug discovery and tracking of disease progression [PML<sup>+</sup>16, Wis19, LSV<sup>+</sup>21]. Recent works in metabolomics and proteomics have explored the applications of matching algorithms through combinatorial search, nearest neighbor methods, and optimal transport to solve the untargeted feature alignment problem [LHM13, HKB<sup>+</sup>21, CPKL<sup>+</sup>22, BSKR<sup>+</sup>23].

**Feature alignment and linear assignment** In many machine learning applications, data features have a predefined ordering and the aim is to match the observations in  $X^m$  and  $Y^n$ . Examples include matching a set of patches in two images or word embeddings from two different languages. A classical approach is to form a similarity matrix  $A_{ij} = a(X_i, Y_j)$  and maximize the objective

$$\max_{\pi} \sum_{i=1}^m \sum_{j=1}^n A_{i,\pi(i)} \quad (4)$$

known as the *(linear) assignment problem*. This combinatorial optimization problem (4) can be solved exactly in polynomial time using linear programming and the Hungarian algorithm [Mun57], or through optimal transport by lifting the matchings to the space of probabilistic couplings [PC<sup>+</sup>17] where the Sinkhorn algorithm has allowed researchers to process this problem in near-linear time [Cut13, AWR17]. A statistical model for the assignment problem was proposed in [CD16, GMD22] where it is assumed that  $X^m$  and  $Y^n$  share a subset of observations corrupted with additive noise. When  $X^m$  and  $Y^n$  are the set of vertices in a bipartite random graph, the sharp statistical properties of (4) have been established in [DWXY23] for a planted matching problem.

**Graph matching and quadratic assignment** Another fundamental instantiation of dataset alignment is through *graph matching*, which is closest to our covariance alignment model. Based on  $X^m$  and  $Y^n$ , form two similarity matrices  $A, B \in \mathbb{R}^{d \times d}$  over the features, treat them as weighted undirected graphs, and aim to find a permutation  $\pi \in S_d$  such that  $A^\pi \approx B$ . This is usually done by solving the following combinatorial optimization problem

$$\max_{\pi \in S_d} \sum_{i=1}^d \sum_{j=1}^d A_{\pi(i)\pi(j)} B_{ij}, \quad (5)$$

This is an example of the *quadratic assignment problem* (QAP) [KB57], “one of the most difficult problems in the NP-hard class” [LdABN<sup>+</sup>07]. This tantalizing difficulty has driven researchers to explore specific instantiations of QAP that are potentially easier to solve. Perhaps the most famous one is graph isomorphism which is strongly believed to not be NP-complete. Another route to solve this problem in polynomial time is to consider random instances of QAP. Most noticeably, a recent line of research culminating with [MWXY23, MRT23] has investigated the alignment of correlated Erdős-Rényi random graphs from both the statistical and computational perspective.

Although the QAP in (5) has appeared in the graph matching literature [HM23, WXS22], our covariance alignment model differs from the above graph matching papers in two aspects. First, under our Gaussian covariance model, the matrices  $\hat{\Sigma}_X$  and  $\hat{\Sigma}_Y$  being aligned are Wishart random graphs, departing from the traditional Erdős-Rényi, Wigner, or random geometric graphs. Second, and more important, all above works assume a *known* distribution of the random matrices  $(A, B)$ . In contrast, our covariance alignment model treats the true covariance  $\Sigma$  as an additional nuisance parameter, and we need to estimate the marginal distributions of  $\hat{\Sigma}_X$  and  $\hat{\Sigma}_Y$ . To the best of our knowledge, the statistical problem of estimating  $\pi$  in this setting has not been explored.

**Gromov–Wasserstein distance** The Gromov-Hausdorff distance [GKPS99] from metric geometry gives us yet another perspective on the quadratic assignment objective (5). Consider two finite metric spaces given by their  $d \times d$  pairwise distance metrics,  $A$  and  $B$  respectively. The Gromov-Hausdorff distance between these two metric spaces can be defined as

$$d_{\text{GH}}(A, B) := \min_{\pi \in S_d} \max_{1 \leq i, j \leq d} |A_{ij} - B_{\pi(i)\pi(j)}|.$$

Using an idea from optimal transport, Mémoli [Mém11] proposed the Gromov–Wasserstein distance using a double<sup>1</sup> relaxation:

$$d_{\text{GW}}^2(A, B) := \min_{P \in \text{BP}(d)} \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \sum_{l=1}^d P_{ij} P_{kl} (B_{ik} - A_{jl})^2.$$

Note that the above optimization problem is equivalent to

$$\max_{P \in \text{BP}(d)} \langle PAP^\top, B \rangle. \quad (6)$$

The above formulation (6) is equivalent to the quadratic assignment (5) for PSD  $A, B \succeq 0$  and has received significant attention from the optimization community with various algorithms ranging from Frank-Wolfe [FPPA14, FCG<sup>+</sup>21] to alternating linear programs [Mém11] and alternating Sinkhorn updates with entropic regularization [SPKS16, SVP21]. Some progress has been made in studying the convergence of GW optimization [LTK<sup>+</sup>22, LTK<sup>+</sup>23] and exact recovery of matchings for Bernoulli graphs [LFF<sup>+</sup>15]. In parallel, the statistical community has also studied the statistical properties of the Gromov–Wasserstein distance; see [ZGMS22] and references therein. The setup of this line of work is quite different from the one pursued in this paper: it consists in assuming that  $d$  points used to form the distance matrices  $A$  and  $B$  are sampled i.i.d. from two metric measure spaces  $\mathcal{A}$  and  $\mathcal{B}$  respectively, and the goal is to study the convergence of  $d_{\text{GH}}(A, B)$  to  $d_{\text{GH}}(\mathcal{A}, \mathcal{B})$  as  $d \rightarrow \infty$ .

The precise objective (6) studied here has been applied to match feature covariance between untargeted metabolomic datasets in [BSKR<sup>+</sup>23], while its statistical properties are underexplored. Our work presents the first statistical rates of estimation for GW in learning the matching between two datasets by aligning their covariance matrices.

**Additional literature on permutation estimation** Finally, we remark that our work falls in the broader line of work associated to permutation estimation. While more directly related to the graph matching problem mentioned above, our technical derivations present similarities with other statistical problems parameterized by a permutation. These include statistical seriation [FMR19], regression with shuffled data [PWC17], isotonic regression [PS22], and noisy sorting [MWR18], to name a few.

### 1.3 Notation

We use the following notations throughout the paper.

**Scalars, vectors and matrices** For  $d \in \mathbb{N}$ , let  $\mathbb{1}_d$  denote the  $d$ -length vector of all ones. For  $n \in \mathbb{N}$ , let  $[n] = \{1, \dots, n\}$ . Let  $S_n$  be the set of all permutations over  $[n]$ . For  $\pi \in S_n$  and  $A \in \mathbb{R}^{n \times n}$ , let  $A^\pi$  be the  $n \times n$  matrix with  $(A^\pi)_{i,j} = A_{\pi(i), \pi(j)}$ . Occasionally we also write  $A^\pi$  in equivalent matrix forms as  $A^\pi = P_\pi A P_\pi^\top$ , where  $P_\pi$  is the permutation matrix associated with  $\pi$ , formally defined as  $(P_\pi)_{i,j} = \mathbb{1}(\pi(i) = j)$ . For matrices  $A$  and  $B$  of the same size, let  $\langle A, B \rangle = \text{Tr}(AB^\top)$  be their matrix inner product. For a matrix  $A$ , let  $\|A\|_F = \sqrt{\langle A, A \rangle}$  be its Frobenius norm, and  $\|A\|_{\text{op}} = \max_{\|v\|_2=1} \|Av\|_2$  be its operator norm. Finally for two real numbers  $a$  and  $b$ , we use the shorthand notation  $a \wedge b := \min(a, b)$ ,  $a \vee b = \max(a, b)$ .

<sup>1</sup>The sup norm is relaxed to the squared Euclidean norm and the set of permutations is relaxed to the Birkhoff polytope.



**Information-theoretic quantities** For probability measures  $P$  and  $Q$  over the same probability space, let  $D_{\text{KL}}(P\|Q) = \int dP \log(dP/dQ)$  and  $\chi^2(P\|Q) = \int (dP)^2/dQ - 1$  be the Kullback-Leibler (KL) divergence and  $\chi^2$  divergence between  $P$  and  $Q$ , respectively. For a random vector  $(X, Y, Z) \sim P_{XYZ}$ , let  $I(X; Y \mid Z) = \mathbb{E}_{P_Z}[D_{\text{KL}}(P_{XY|Z}\|P_{X|Z} \otimes P_{Y|Z})]$  be the conditional mutual information between  $X$  and  $Y$  given  $Z$ .

**Asymptotics** For non-negative sequences  $\{a_n\}$  and  $\{b_n\}$ , we write  $a_n = O(b_n)$  whenever there exists a numerical constant  $C < \infty$  such that  $a_n \leq Cb_n$  for all  $n$ . Similarly, we write  $a_n = \Omega(b_n)$  if  $b_n = O(a_n)$ , and  $a_n = \Theta(b_n)$  or  $a_n \asymp b_n$  if both  $a_n = O(b_n)$  and  $b_n = O(a_n)$ .

## 2 Covariance alignment estimators

In this section we investigate the statistical performance of two estimators: the quasi maximum likelihood estimator (QMLE) and the Gromov–Wasserstein (GW) estimator. In Section 2.1, we introduce both the QMLE and the GW estimator. In Section 2.2, we establish the minimax rate-optimality of both estimators. Despite a similar statistical performance, Section 2.3 discusses the computational benefits of the GW estimator over the QMLE, supported by numerical experiments.

### 2.1 Estimators

We introduce the QMLE and the GW estimator separately.

#### 2.1.1 Quasi Maximum Likelihood Estimator

The log-likelihood in the covariance alignment model is given, up to an additive constant, by

$$\ell(\pi, \Sigma) = \log \det(\Sigma^{-1}) - \frac{1}{m+n} \sum_{i=1}^m X_i^\top \Sigma^{-1} X_i - \frac{1}{m+n} \sum_{i=1}^n (P_{\pi^{-1}} Y_i)^\top \Sigma^{-1} (P_{\pi^{-1}} Y_i),$$

which is concave in  $\Sigma^{-1}$  and maximized at

$$\hat{\Sigma} = \frac{m}{m+n} \hat{\Sigma}_X + \frac{n}{m+n} (\hat{\Sigma}_Y)^{(\pi^{-1})},$$

leading to the (profile) log-likelihood

$$\ell(\pi) := \sup_{\Sigma} \ell(\pi, \Sigma) = -\log \det \left( \frac{m}{m+n} \hat{\Sigma}_X + \frac{n}{m+n} (\hat{\Sigma}_Y)^{(\pi^{-1})} \right) - 1. \quad (7)$$

Due to the nonlinearity of the log det function, properties of the maximum likelihood estimator obtained by maximizing (7) are difficult to establish. Instead, consider a quasi-log-likelihood obtained as follows. Note that for that  $\pi \approx \pi^*$ , we have

$$\frac{m}{m+n} \hat{\Sigma}_X + \frac{n}{m+n} (\hat{\Sigma}_Y)^{(\pi^{-1})} \approx \hat{\Sigma}_X.$$

A first-order Taylor expansion of  $A \mapsto \log \det(A)$  around  $\hat{\Sigma}_X$ , yields

$$\log \det(A) \approx \langle \hat{\Sigma}_X^{-1}, A - \hat{\Sigma}_X \rangle = \langle \hat{\Sigma}_X^{-1}, A \rangle + \text{const.}$$

Hence

$$\ell(\pi) \approx -\frac{n}{m+n} \langle \hat{\Sigma}_X^{-1}, (\hat{\Sigma}_Y)^{(\pi^{-1})} \rangle + \text{const.}$$



Maximizing this approximation is equivalent to maximizing the quasi log-likelihood  $\tilde{\ell}$  defined by

$$\tilde{\ell}(\pi) := \langle \hat{\Sigma}_Y, (\hat{\Sigma}_X^\pi)^{-1} \rangle. \quad (8)$$

Consequently, our final quasi MLE  $\hat{\pi}^{\text{QMLE}}$  is given by

$$\hat{\pi}^{\text{QMLE}} = \underset{\pi \in S_d}{\operatorname{argmin}} R_{m,n}^{\text{QMLE}}(\pi) := \langle \hat{\Sigma}_Y, (\hat{\Sigma}_X^\pi)^{-1} \rangle. \quad (9)$$

This quasi MLE can also be viewed as a two-step estimator where we first estimate the covariance as  $\Sigma \approx \hat{\Sigma}_X$  and then plug it into the log-likelihood for  $\pi$ .

### 2.1.2 Gromov–Wasserstein Estimator

The GW estimator aims to solve the quadratic assignment problem (5) in random graph matching, with similarity matrices  $A, B$  given by the sample covariance matrices  $\hat{\Sigma}_X, \hat{\Sigma}_Y$ . Formally, the GW estimator  $\hat{\pi}^{\text{GW}}$  is given by

$$\hat{\pi}^{\text{GW}} = \underset{\pi \in S_d}{\operatorname{argmax}} \langle \hat{\Sigma}_X^\pi, \hat{\Sigma}_Y \rangle = \underset{\pi \in S_d}{\operatorname{argmin}} R_{m,n}^{\text{GW}}(\pi) := \left\| \hat{\Sigma}_X^\pi - \hat{\Sigma}_Y \right\|_{\text{F}}^2. \quad (10)$$

The name GW comes from the mathematical theory of optimal transport which lifts the problem of permutation estimation to probabilistic couplings. Specifically, the GW optimization is performed as

$$\hat{P}^{\text{GW}} = \underset{P \in \text{BP}(d)}{\operatorname{argmax}} \langle P \hat{\Sigma}_X P^\top, \hat{\Sigma}_Y \rangle,$$

where the maximization is over the Birkhoff polytope given by

$$\text{BP}(d) = \{P \in \mathbb{R}_+^{d \times d} : P \mathbb{1}_d = P^\top \mathbb{1}_d = \mathbb{1}_d\}. \quad (11)$$

Since this is a convex cost in  $P$ , the maximizer  $\hat{P}^{\text{GW}}$  is one of the vertices of the Birkhoff polytope which are exactly the set of permutation matrices  $S_d$ . Therefore, the relaxation to the Birkhoff polytope is tight, while allowing for a broad diversity of continuous optimization algorithms unlike the combinatorial optimization in (10).

Unlike the GW estimator, the objective for the quasi MLE defined over the set of permutations cannot in general be relaxed to the space of coupling matrices. In fact, even for the limiting case  $\hat{\Sigma}_X = \hat{\Sigma}_Y = \Sigma$  with  $m, n = \infty$ , von Neumann's trace inequality gives

$$\min_{\pi \in S_d} \langle \Sigma, (\Sigma^\pi)^{-1} \rangle \geq d.$$

However, for the choice  $P = \frac{1}{d} \mathbb{1}_d \mathbb{1}_d^\top \in \text{BP}(d)$  of the coupling matrix, it holds that

$$\langle P \Sigma^{-1} P^\top, \Sigma \rangle = \frac{\mathbb{1}_d^\top \Sigma^{-1} \mathbb{1}_d}{d} \cdot \frac{\mathbb{1}_d^\top \Sigma \mathbb{1}_d}{d},$$

which could be smaller than  $d$  for many covariance matrices  $\Sigma$  (e.g. when  $\Sigma = \text{diag}(1, 1/2, \dots, 1/2)$ ). This means that the minimizing coupling matrix may lie in the interior of the Birkhoff polytope.

## 2.2 Minimax optimality

Here we state the main results showing the minimax optimality of the quasi MLE and GW estimators.

**Theorem 2** (Upper bound for QMLE). *Fix  $\delta \in (0, 1)$ . Assume that*

$$m \wedge n \geq c_0(d \log d + \log(1/\delta)) \quad \text{and} \quad mn \geq c_0 d(d \log d + \log(1/\delta))(d + \log(1/\delta))$$

*for a large enough numerical constant  $c_0 > 0$ . Then for every  $(\pi^*, \Sigma)$  with  $\|\Sigma\|_{\text{op}} \leq 1$ , with probability at least  $1 - \delta$  it holds that*

$$\left\| \Sigma^{\hat{\pi}^{\text{QMLE}}} - \Sigma^{\pi^*} \right\|_{\text{F}}^2 \lesssim \frac{d \log d + \log(1/\delta)}{m \wedge n} + \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}}.$$

**Theorem 3** (Upper bound for GW). *Fix  $\delta \in (0, 1)$ , and assume that*

$$m \wedge n \geq d \log d + \log(1/\delta).$$

*Then for every  $(\pi^*, \Sigma)$  with  $\|\Sigma\|_{\text{op}} \leq 1$ , with probability at least  $1 - \delta$  it holds that*

$$\left\| \Sigma^{\hat{\pi}^{\text{GW}}} - \Sigma^{\pi^*} \right\|_{\text{F}}^2 \lesssim \frac{d \log d + \log(1/\delta)}{m \wedge n} + \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}}.$$

Integrating the tails over  $\delta \in (0, 1)$  we readily obtain the sample complexity upper bounds in Section 1. Note that Theorem 2 holds under an additional condition on the product  $mn$  which does not appear in Theorem 3; we believe this extra condition to be superfluous. It comes from a technicality in our analysis that would be overcome using a concentration result for linear functionals of inverse-Wishart random matrices. In absence of such a result, we perform a local Taylor approximation instead. On the other hand, this condition remains necessary for either bound to be meaningful so we do not pursue this question here.

The following minimax lower bound shows that the rates of convergence in Theorem 2 and 3 are tight.

**Theorem 4** (Minimax lower bound). *Given observations  $X^m \sim \mathcal{N}(0, \Sigma)^{\otimes m}$  and  $Y^n \sim \mathcal{N}(0, \Sigma^{\pi^*})^{\otimes n}$ , the following minimax lower bound holds: if  $m \wedge n \geq \log d$  and  $m \vee n \geq d$ , there exists an absolute constant  $c_0 > 0$  independent of  $(m, n, d)$  such that*

$$\inf_{\hat{\pi}} \sup_{\pi^*, \Sigma: \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{\Sigma} \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_{\text{F}}^2 \geq c_0 \left( \frac{d \log d}{m \wedge n} + \sqrt{\frac{d^3 \log d}{mn}} \right).$$

We note that the conditions  $m \vee n \geq d$  and  $m \wedge n \geq \log d$  in Theorem 4 are not entirely superfluous: they ensure that the minimax lower bound is no larger than  $\Theta(d)$ , as  $\|\Sigma^{\hat{\pi}} - \Sigma^{\pi^*}\|_{\text{F}}^2 = O(d)$  trivially holds for every  $\hat{\pi}$ .

## 2.3 Numerical experiments

Although the GW estimator and the QMLE are both minimax optimal, the GW estimator presents significant computational advantages over the QMLE. The objective of the QMLE (9) is a general quadratic assignment problem over the set of permutation matrices, and it is NP-hard to find a constant approximation for QAPs in general [SG76]. In contrast, although the GW estimator (10) is also NP-hard, the lifting to probabilistic couplings offers more computational flexibility in the

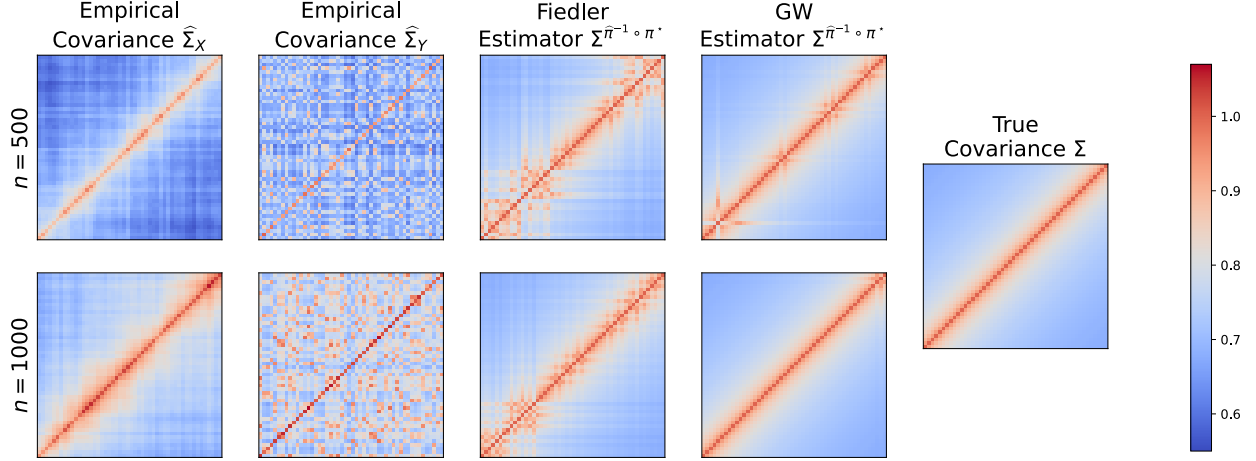


Figure 1: Covariance alignment for structured Robinson covariance matrices  $\Sigma \in \mathbb{R}^{d \times d}$  with  $\Sigma_{ij} = (1 + |i - j|)^{-\gamma}$  and decaying off-diagonals for dimension  $d = 50$ . Here we compare the Sinkhorn GW estimator to a classical Fiedler vector spectral estimator. Across all sample sizes  $n$  we find that GW outperforms the spectral estimator and almost perfectly reconstructs the structured covariance matrix at  $n = 1,000$ .

development of numerical heuristics as gradient information can now be used and potentially leads to better convergence results empirically. This approach has been successfully leveraged in a variety of contexts [SPKS16, PC<sup>+</sup>17]. As discussed above, the combinatorial optimization (9) for the QMLE cannot be lifted to probabilistic couplings, for (9) is a minimization of convex functions and gives a non-permutation after the lifting. In the experiments below, we optimize the GW objective (10) over the space of couplings by adding a small entropic penalty (with penalty parameter  $\varepsilon > 0$ ) to the cost and using Sinkhorn projections [SPKS16, PC<sup>+</sup>17, FCG<sup>+</sup>21]. At convergence, we round the final coupling to the closest permutation matrix by solving a linear assignment problem; we call it the Sinkhorn GW estimator. We also optimize both GW and QMLE objectives directly over the set of permutation matrices using the commercial LocalSolver optimizer, a state-of-the-art local search mixed-integer program solver [BEGN10].

We begin with a structured covariance alignment problem where  $\Sigma \in \mathbb{R}^{d \times d}$  with  $d = 50$  is a Robinson matrix with decaying off-diagonals  $\Sigma_{ij} = (1 + |i - j|)^{-\gamma}$  for  $\gamma = 0.1$ . In the experiment we set  $m = n$ , and compare two estimators of  $\pi^*$ : one is the Sinkhorn GW estimator with  $\varepsilon = 5 \times 10^{-4}$ , and the other is a classical spectral estimator [ABH98] which estimates the permutation  $\hat{\pi}$  by sorting the entries of the Fiedler vector (the eigenvector for the smallest nonzero eigenvalue) of the unnormalized Laplacian matrix  $L = \text{diag}(\hat{\Sigma}_Y \mathbb{1}_d) - \hat{\Sigma}_Y$ . We note that the second estimator crucially relies on the Robinson structure of  $\Sigma$  where the Fiedler vector is monotone. Figure 1 displays the visualizations for  $\hat{\Sigma}_X$ ,  $\hat{\Sigma}_Y$ , and the permuted covariances for both estimators under two sample sizes  $n \in \{500, 1000\}$ , and shows that the GW estimator achieves a smaller Frobenius loss than the Fiedler vector estimator. In particular, it is strikingly visible that when  $n = 1,000$ , the GW estimator almost perfectly recovers the structured covariance  $\Sigma$ .

We proceed with random covariance matrices  $\Sigma$ , and compare the Sinkhorn GW estimator with the QMLE and GW estimators computed via LocalSolver. We generate random Wishart distributed covariances  $\Sigma \sim \mathcal{W}_d(I_d, d)$  with  $d \in [10, 10^3]$ , choose various sample sizes  $m = n \in [10, 10^6]$ , and set  $\varepsilon = 1/d^2$  for the Sinkhorn GW estimator. Figure 2 displays the heatmaps of the mean squared Frobenius losses (averaged over 1,000 experiments for Sinkhorn GW and 100 experiments for the

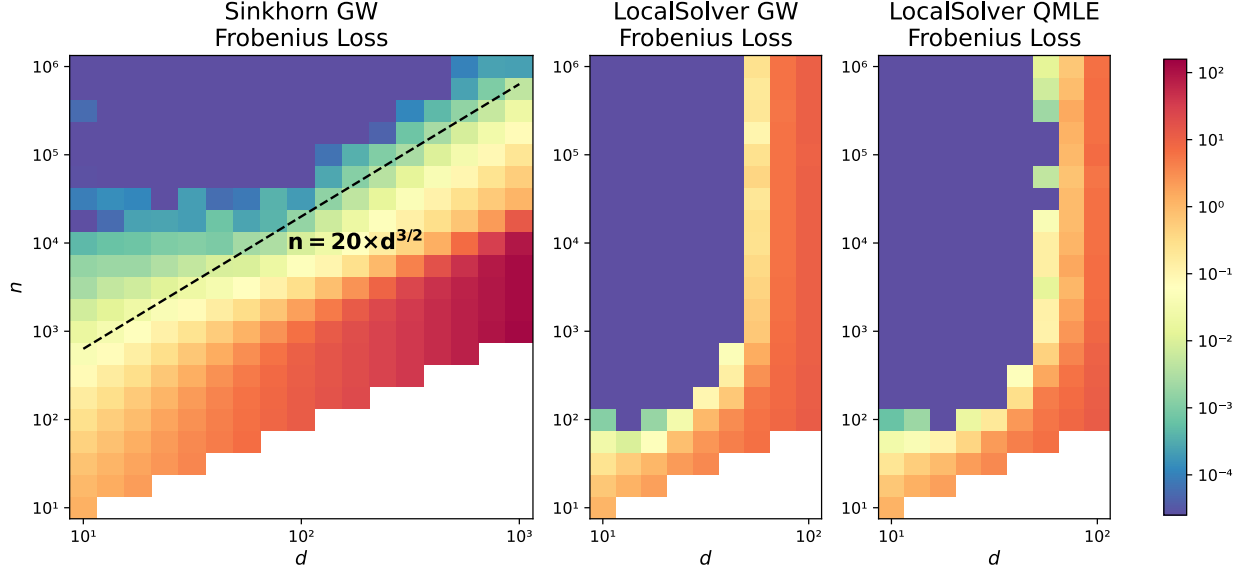


Figure 2: Comparison of Sinkhorn GW, LocalSolver GW, and LocalSolver QMLE for covariance alignment with random Wishart distributed covariances  $\Sigma \sim \mathcal{W}_d(I_d, d)$ . Here we vary the sample size  $m = n \in [10, 10^6]$ , and the dimension  $d \in [10, 10^3]$  of the covariance matrix. Three heatmaps show the mean squared Frobenius norm of these algorithms as a function of  $(d, n)$  averaged over 1,000 experiments for Sinkhorn GW and 100 experiments for the more time-consuming LocalSolver estimators. We observe that the LocalSolver combinatorial search often achieves exact recovery for small values of  $d$ , while only the Sinkhorn GW algorithm scales to high dimensions and attains the minimax optimal scaling  $n = \tilde{\Theta}(d^{3/2})$  for sufficiently large  $n$ .

more time-consuming LocalSolver estimators) achieved by all estimators for each  $(n, d)$  pair. We observe that for small values of  $d$ , the combinatorial search performed by LocalSolver can often achieve exact recovery (dark purple color) and result in smaller losses than the Sinkhorn GW estimator. However, starting from moderate values of  $d$  (say  $d = 100$ ), the LocalSolver combinatorial search fails and outputs random results even for very large  $n$ . In contrast, the Sinkhorn GW estimator still achieves vanishing Frobenius losses for large problems with  $d = 1,000$ , with a running time on the order of several seconds. We also find that for sufficiently large  $n$ , the minimax optimal scaling  $n = \tilde{\Theta}(d^{3/2})$  (indicated by the black dashed line in Figure 2) holds for the Sinkhorn GW estimator.

The above experimental results show that the Sinkhorn GW estimator is competitive for covariance alignment both statistically and computationally. However, we do not have a formal proof that the Sinkhorn GW estimator achieves the minimax rate in Theorems 2-4. We leave the statistical properties of the Sinkhorn GW estimator, or more generally the construction of a polynomial-time minimax optimal estimator, as outstanding open problems.

### 3 Proof of upper bounds

In this section we establish the upper bounds for both the QMLE and GW estimators in Theorem 2 and Theorem 3, respectively. Our high-level idea is to establish a high-probability lower bound of

$R_{m,n}(\hat{\pi}) - R_{m,n}(\pi^*)$  for both the QMLE and GW estimators: with probability exceeding  $1 - \delta$ ,

$$R_{m,n}(\hat{\pi}) - R_{m,n}(\pi^*) \geq \rho \left( \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F \right)$$

holds for some properly chosen  $\rho(\cdot)$ . Since  $R_{m,n}(\hat{\pi}) - R_{m,n}(\pi^*) \leq 0$  by definition of  $\hat{\pi}$ , solving the above inequality will give the target upper bound of  $\left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F$ .

### 3.1 Upper bound of the QMLE

We first establish the upper bound for the QMLE and prove Theorem 2. In Section 3.1.1, we define several good events which happen simultaneously with probability at least  $1 - \delta$ . In Section 3.1.2, we condition on these events and derive a deterministic lower bound of  $R_{m,n}^{\text{QMLE}}(\hat{\pi}^{\text{QMLE}}) - R_{m,n}^{\text{QMLE}}(\pi^*)$ . In Section 3.1.3, we prove a useful trace-Frobenius inequality which relates the above lower bound to the final Frobenius norm.

#### 3.1.1 Good events

Without loss of generality, throughout our proof we assume that  $\pi^* = \text{Id}$ ; by swapping the roles of  $X^m$  and  $Y^n$  we also assume that  $m \geq n$ . Denote  $\hat{\pi}^{\text{QMLE}}$  by  $\hat{\pi}$  for notational simplicity. The good event  $\mathcal{G}$  that we will condition on afterwards is defined to be the intersection of events  $\mathcal{G} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ , where

$$\begin{aligned} \mathcal{E}_1 := & \left\{ \langle \hat{\Sigma}_Y, (\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1} \rangle \geq \langle \Sigma, (\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1} \rangle \right. \\ & \left. - c \sqrt{\frac{d \log d + \log(1/\delta)}{n}} \left\| \Sigma^{1/2} ((\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1}) \Sigma^{1/2} \right\|_F \right\}, \end{aligned} \quad (12)$$

$$\begin{aligned} \mathcal{E}_2 := & \left\{ \langle \Sigma^{-1} - \Sigma^{-1} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1}, \hat{\Sigma}_X \rangle \geq \langle \Sigma^{-1} - \Sigma^{-1} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1}, \Sigma \rangle \right. \\ & \left. - c \sqrt{\frac{d \log d + \log(1/\delta)}{m}} \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_F \right\}, \end{aligned} \quad (13)$$

$$\mathcal{E}_3 := \left\{ \left\| \Sigma^{-1/2} \hat{\Sigma}_X \Sigma^{-1/2} - I \right\|_{\text{op}} \leq \min \left\{ c \sqrt{\frac{d + \log(1/\delta)}{m}}, \frac{1}{2} \right\} \right\}. \quad (14)$$

Here  $c > 0$  is a numerical constant chosen in the following lemma.

**Lemma 1.** *There exist absolute constants  $c, c_0 > 0$  independent of  $(n, m, d, \delta)$  such that, if  $m \geq n \geq c_0(d \log d + \log(1/\delta))$ , it holds that  $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$ .*

*Proof.* We show that  $\mathbb{P}(\mathcal{E}_i) \geq 1 - \delta/3$  for all  $i = 1, 2, 3$ . For  $\mathcal{E}_1$ , first consider the scenario where  $\hat{\pi}$  is replaced by any fixed  $\pi \in S_d$ ; we denote the resulting event by  $\mathcal{E}_1(\pi)$ . After conditioning on  $X^m$ , the Hanson–Wright inequality in Lemma 11 tells that  $\mathbb{P}(\mathcal{E}_1(\pi)) \geq 1 - \delta/(3 \cdot d!)$  for any fixed  $\pi$ , for a large numerical constant  $c > 0$ . Since  $\cap_{\pi \in S_d} \mathcal{E}_1(\pi) \subseteq \mathcal{E}_1$ , the union bound yields  $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta/3$ . The analysis of  $\mathcal{E}_2$  follows from a similar application of the Hanson–Wright inequality to  $\hat{\Sigma}_X$ .

Finally, the lower bound of  $\mathbb{P}(\mathcal{E}_3)$  follows from the concentration of the sample covariance matrix, cf. [RH15, Theorem 5.7]. The upper bound  $1/2$  could be ensured by  $m \geq c_0(d \log d + \log(1/\delta))$  and a sufficiently large constant  $c_0$ .  $\square$

The specific form of the good event  $\mathcal{G}$  is motivated by the calculation of  $R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id})$  in the following section.

### 3.1.2 Lower bound of the loss difference

Throughout this section we condition on the good event  $\mathcal{G}$  defined in (12)-(14) and prove a deterministic lower bound of  $R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id})$ . Specifically,

$$\begin{aligned} & R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id}) \\ & \stackrel{(a)}{=} \langle \hat{\Sigma}_Y, (\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1} \rangle \\ & \stackrel{(b)}{\geq} \langle \Sigma, (\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1} \rangle - c \sqrt{\frac{d \log d + \log(1/\delta)}{n}} \underbrace{\left\| \Sigma^{1/2} ((\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1}) \Sigma^{1/2} \right\|_{\text{F}}}_{=: R_1}. \end{aligned}$$

Here (a) follows from the definition of the empirical loss in (9), and (b) is due to the event  $\mathcal{E}_1$  in (12). Deferring the analysis of the remainder term  $R_1$  for a moment, we continue to lower bound the main term as

$$\begin{aligned} & \langle \Sigma, (\hat{\Sigma}_X^{-1})^{\hat{\pi}} - \hat{\Sigma}_X^{-1} \rangle = \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \hat{\Sigma}_X^{-1} \rangle \\ & \stackrel{(c)}{=} \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle + \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1}(\Sigma - \hat{\Sigma}_X)\Sigma^{-1} \rangle + \underbrace{\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \hat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \hat{\Sigma}_X)\Sigma^{-1} \rangle}_{=: R_2} \\ & \stackrel{(d)}{\geq} \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle - c \sqrt{\frac{d \log d + \log(1/\delta)}{m}} \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_{\text{F}} + R_2. \end{aligned}$$

Here (c) is motivated by the approximation  $\hat{\Sigma}_X^{-1} \approx \Sigma^{-1} + \Sigma^{-1}(\Sigma - \hat{\Sigma}_X)\Sigma^{-1}$  if  $\hat{\Sigma}_X \approx \Sigma$ ; this step linearizes the dependence on  $\hat{\Sigma}_X$ . As a result, (d) makes use of the event  $\mathcal{E}_2$  in (13).

The remainder terms  $R_1$  and  $R_2$  are bounded in the following lemma.

**Lemma 2.** *Under the event  $\mathcal{E}_3$  in (14), it holds that*

$$\begin{aligned} R_1 & \leq \left\| \Sigma^{1/2} (\Sigma^{-1})^{\hat{\pi}} \Sigma^{1/2} - I_d \right\|_{\text{F}} + 4c\sqrt{d} \cdot \sqrt{\frac{d + \log(1/\delta)}{m}}, \\ |R_2| & \leq 2c^2 \cdot \frac{\sqrt{d}(d + \log(1/\delta))}{m} \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_{\text{F}}. \end{aligned}$$

Plugging Lemma 2 into the previous lower bound yields

$$\begin{aligned} & R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id}) \\ & \geq \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle - C \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}} \\ & \quad - C \sqrt{\frac{d \log d + \log(1/\delta)}{n}} \left\| \Sigma^{1/2} (\Sigma^{-1})^{\hat{\pi}} \Sigma^{1/2} - I_d \right\|_{\text{F}} \\ & \quad - C \left( \sqrt{\frac{d \log d + \log(1/\delta)}{m}} + \frac{\sqrt{d}(d + \log(1/\delta))}{m} \right) \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_{\text{F}}, \end{aligned} \tag{15}$$

for some absolute constant  $C > 0$  independent of  $(n, m, d)$ .

### 3.1.3 A trace-Frobenius inequality

Based on (15), the remaining question is to relate the Frobenius norms in (15) to the inner product  $\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle$ . Roughly speaking, for the matrix  $A = \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d$ , the Frobenius norm

$\|A\|_F$  corresponds to the  $L_2$  norm of its eigenvalues, and the inner product  $\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle = \text{Tr}(A)$  corresponds to the sum of its eigenvalues. Although  $\|A\|_F \leq \text{Tr}(A)$  holds for PSD matrices  $A \succeq 0$ , this inequality may break down if some eigenvalues of  $A$  are negative.

To this end, we present a useful technical inequality in the following lemma.

**Lemma 3.** *Let  $x_1, \dots, x_d > 0$  with  $\prod_{i=1}^d x_i = 1$ . Then*

$$\sum_{i=1}^d (x_i - 1)^2 \leq 4 \left( \sum_{i=1}^d (x_i - 1) + \left( \sum_{i=1}^d (x_i - 1) \right)^2 \right).$$

The usefulness of Lemma 3 lies in the following application: let  $x_1, \dots, x_d$  be all eigenvalues of  $\Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2}$ , then  $x_1, \dots, x_d > 0$ , and  $\prod_{i=1}^d x_i = \det(\Sigma^{\hat{\pi}^{-1}}) \det(\Sigma^{-1}) = 1$ . Therefore, we arrive at the following corollary, which we term as a *trace-Frobenius inequality*.

**Corollary 1.** *For positive definite  $\Sigma$ , the following inequalities hold:*

$$\begin{aligned} \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_F^2 &\leq 4(\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle + \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle^2), \\ \left\| \Sigma^{1/2} (\Sigma^{-1})^{\hat{\pi}} \Sigma^{1/2} - I_d \right\|_F^2 &\leq 4(\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle + \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle^2). \end{aligned}$$

By (15) and Corollary 1, we conclude that

$$R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id}) \geq (1-u) \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle - u \sqrt{\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle} - v \quad (16)$$

where

$$\begin{aligned} u &= 2C \left( \sqrt{\frac{d \log d + \log(1/\delta)}{n}} + \sqrt{\frac{d \log d + \log(1/\delta)}{m}} + \frac{\sqrt{d}(d + \log(1/\delta))}{m} \right), \\ v &= C \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}}. \end{aligned}$$

By our assumption  $m \geq n \geq c_0(d \log d + \log(1/\delta))$  and  $mn \geq c_0 d(d \log d + \log(1/\delta))(d + \log(1/\delta))$ , choosing a large enough numerical constant  $c_0 > 0$  ensures that  $u, v \leq 1/2$ . Moreover, by definition of  $\hat{\pi} = \hat{\pi}^{\text{QMLE}}$  we have  $R_{m,n}^{\text{QMLE}}(\hat{\pi}) - R_{m,n}^{\text{QMLE}}(\text{Id}) \leq 0$ . Hence, by (16) and Lemma 13, we conclude that

$$\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle \leq \left( \frac{u}{1-u} + \sqrt{\frac{v}{1-u}} \right)^2 \leq 4(u + \sqrt{v})^2 \leq 8(u^2 + v). \quad (17)$$

From (17) we are ready to prove Theorem 2: conditioned on  $\mathcal{G}$ ,

$$\begin{aligned} \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_F^2 &= \left\| \Sigma^{\hat{\pi}^{-1}} - \Sigma \right\|_F^2 \stackrel{(a)}{\leq} \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_F^2 \\ &\stackrel{(b)}{\leq} 4(\langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle + \langle \Sigma^{\hat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle^2) \\ &\stackrel{(c)}{\leq} 288(u^2 + v) \\ &\stackrel{(d)}{=} O \left( \frac{d \log d + \log(1/\delta)}{n} + \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}} \right), \end{aligned}$$

where (a) is due to the assumption  $\|\Sigma\|_{\text{op}} \leq 1$ , (b) uses Corollary 1, (c) follows from (17) and that  $u^2 + v < 1$ , and (d) plugs in the definition of  $(u, v)$  and uses  $m \geq \sqrt{mn} \geq \sqrt{c_0 d}(d + \log(1/\delta))$ . Since  $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$  by Lemma 1, this proves the upper bound of Theorem 2.



## 3.2 Upper bound of the GW estimator

Now we establish the upper bound for the GW estimator and prove Theorem 3 using similar arguments as in our proof for the QMLE. Again, in Section 3.2.1 we define the good events, and in Section 3.2.2 we condition on the good events and derive a deterministic lower bound of the loss difference  $R_{m,n}^{\text{GW}}(\hat{\pi}^{\text{GW}}) - R_{m,n}^{\text{GW}}(\pi^*)$ .

### 3.2.1 Good events

As before, without loss of generality we assume that  $\pi^* = \text{Id}$  and refer to  $\hat{\pi}^{\text{GW}}$  as  $\hat{\pi}$  for notational convenience. We condition on a good event  $\mathcal{G} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$  defined as the following intersection:

$$\mathcal{E}_1 := \left\{ \langle \hat{\Sigma}_Y, \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \rangle \geq \langle \Sigma, \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \rangle - c \sqrt{\frac{d \log d + \log(1/\delta)}{n}} \left\| \Sigma^{1/2} (\hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}}) \Sigma^{1/2} \right\|_{\text{F}} \right\}, \quad (18)$$

$$\mathcal{E}_2 := \left\{ \langle \Sigma, \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \rangle \geq \langle \Sigma, \Sigma - \Sigma^{\hat{\pi}} \rangle - c \sqrt{\frac{d \log d + \log(1/\delta)}{m}} \left\| \Sigma^{1/2} (\Sigma - \Sigma^{\hat{\pi}}) \Sigma^{1/2} \right\|_{\text{F}} \right\}, \quad (19)$$

$$\mathcal{E}_3 := \left\{ \left\| \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \right\|_{\text{F}} \leq \left\| \Sigma - \Sigma^{\hat{\pi}} \right\|_{\text{F}} + c \sqrt{\frac{d(d + \log(1/\delta))}{m}} \right\}. \quad (20)$$

Here  $c > 0$  is a numerical constant chosen in the following lemma.

**Lemma 4.** *Let  $m \wedge n \geq d \log d + \log(1/\delta)$ . There exists absolute constant  $c > 0$  independent of  $(n, m, d, \delta)$  such that  $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$ .*

*Proof.* The high probabilities of events  $\mathcal{E}_1, \mathcal{E}_2$  follows from similar applications of the Hanson-Wright inequality as in the proof of Lemma 1. For  $\mathcal{E}_3$ , applying the triangle inequality twice yields

$$\left\| \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \right\|_{\text{F}} - \left\| \Sigma - \Sigma^{\hat{\pi}} \right\|_{\text{F}} \leq \left\| (\hat{\Sigma}_X - \Sigma)^{\hat{\pi}} - (\hat{\Sigma}_X - \Sigma) \right\|_{\text{F}} \leq 2 \left\| \hat{\Sigma}_X - \Sigma \right\|_{\text{F}},$$

where in the last step we have used the permutation invariance of the Frobenius norm. Since

$$\left\| \hat{\Sigma}_X - \Sigma \right\|_{\text{F}} \leq \left\| \Sigma \right\|_{\text{op}} \left\| \Sigma^{-1/2} \hat{\Sigma}_X \Sigma^{-1/2} - I_d \right\|_{\text{F}} \leq \sqrt{d} \left\| \Sigma^{-1/2} \hat{\Sigma}_X \Sigma^{-1/2} - I_d \right\|_{\text{op}},$$

the rest of the proof follows from (14) and Lemma 1.  $\square$

### 3.2.2 Lower bound of the loss difference

Again, we condition on the good event  $\mathcal{G}$  defined in (18)-(20) and prove a deterministic lower bound of  $R_{m,n}^{\text{GW}}(\hat{\pi}) - R_{m,n}^{\text{GW}}(\text{Id})$ . We begin by writing

$$\begin{aligned}
R_{m,n}^{\text{GW}}(\hat{\pi}) - R_{m,n}^{\text{GW}}(\text{Id}) &\stackrel{(a)}{=} 2\langle \hat{\Sigma}_Y, \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \rangle \\
&\stackrel{(b)}{\geq} 2\langle \Sigma, \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \rangle - 2c\sqrt{\frac{d \log d + \log(1/\delta)}{n}} \underbrace{\left\| \Sigma^{1/2} (\hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}}) \Sigma^{1/2} \right\|_{\text{F}}}_{=:R_1} \\
&\stackrel{(c)}{\geq} 2\langle \Sigma, \Sigma - \Sigma^{\hat{\pi}} \rangle - 2c\sqrt{\frac{d \log d + \log(1/\delta)}{n}} \left( R_1 + \underbrace{\left\| \Sigma^{1/2} (\Sigma - \Sigma^{\hat{\pi}^{-1}}) \Sigma^{1/2} \right\|_{\text{F}}}_{=:R_2} \right) \\
&\stackrel{(d)}{=} \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 - 2c\sqrt{\frac{d \log d + \log(1/\delta)}{n}} (R_1 + R_2).
\end{aligned}$$

Here (a) follows from the definition of the empirical loss in (10), (b) is due to the event  $\mathcal{E}_1$  in (18), (c) is due to the event  $\mathcal{E}_2$  in (19) and  $m \geq n$ , and (d) uses the identity

$$2\langle \Sigma, \Sigma - \Sigma^{\hat{\pi}} \rangle = \left\| \Sigma \right\|_{\text{F}}^2 + \left\| \Sigma^{\hat{\pi}} \right\|_{\text{F}}^2 - 2\langle \Sigma, \Sigma^{\hat{\pi}} \rangle = \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2.$$

Next we upper bound the remainder terms.

- The remainder term  $R_1$ : by the event  $\mathcal{E}_3$  in (20) and  $\|\Sigma\|_{\text{op}} \leq 1$ , it directly follows that

$$R_1 \leq \|\Sigma\|_{\text{op}} \left\| \hat{\Sigma}_X - (\hat{\Sigma}_X)^{\hat{\pi}} \right\|_{\text{F}} \leq \left\| \Sigma - \Sigma^{\hat{\pi}} \right\|_{\text{F}} + c\sqrt{\frac{d(d + \log(1/\delta))}{m}}.$$

- The remainder term  $R_2$ : by the permutation invariance of the Frobenius norm, and by  $\|\Sigma\|_{\text{op}} \leq 1$  again, we have

$$R_2 \leq \|\Sigma\|_{\text{op}} \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_{\text{F}} \leq \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_{\text{F}} = \left\| \Sigma - \Sigma^{\hat{\pi}} \right\|_{\text{F}}.$$

A combination of the above displays tells that

$$R_{m,n}^{\text{GW}}(\hat{\pi}) - R_{m,n}^{\text{GW}}(\text{Id}) \geq \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 - u \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}} - v, \quad (21)$$

where

$$u = C\sqrt{\frac{d \log d + \log(1/\delta)}{n}}, \quad v = C\sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}},$$

and  $C > 0$  is an absolute constant independent of  $(n, m, d)$ .

Finally, by definition of  $\hat{\pi} = \hat{\pi}^{\text{GW}}$  we have  $R_{m,n}^{\text{GW}}(\hat{\pi}) - R_{m,n}^{\text{GW}}(\text{Id}) \leq 0$ . Hence, by (21) and Lemma 13 we conclude that when conditioned on  $\mathcal{G}$ ,

$$\begin{aligned}
\left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 &\leq (u + \sqrt{v})^2 \leq 2(u^2 + v) \\
&= O\left(\frac{d \log d + \log(1/\delta)}{n} + \sqrt{\frac{d(d \log d + \log(1/\delta))(d + \log(1/\delta))}{mn}}\right).
\end{aligned}$$

The above inequality together with  $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$  in Lemma 4 completes the proof of Theorem 3.

## 4 Proof of lower bounds

In this section we prove the minimax lower bound in Theorem 4, and by symmetry we assume that  $m \geq n$ . We use an information-theoretic argument: after specifying the prior distribution on  $(\Sigma, \pi^*)$  in Section 4.1, we invoke Fano's inequality and upper bound the mutual information  $I(\pi^*; X^m, Y^n)$  in Section 4.2. The proof of the mutual information upper bound consists of three steps:

1. By the variational representation of the mutual information, we first reduce the weak recovery of  $\pi^*$  to the detection of  $\Sigma$ . This step is similar to [HM23, WXS22];
2. By the superadditivity of the KL divergence between a stationary process and an iid process, we reduce the mutual information to the case with an equal sample size, where we also have  $m$  observations from  $Y$ ;
3. By passing to the  $\chi^2$ -divergence, we carry out the second moment computation of the likelihood ratio and derive the final upper bound of  $I(\pi^*; X^m, Y^n)$ .

### 4.1 Construction of the prior

To prove a Bayes risk lower bound, we impose the following prior distribution on  $(\Sigma, \pi^*)$ . The choice of the prior on  $\pi^*$  is natural: we simply take  $\pi^* \sim \text{Unif}(S_d)$  to be a uniformly random permutation. As for the prior distribution over  $\Sigma$ , we set

$$\Sigma = \frac{1}{2}(I_d + \eta S),$$

with some parameter

$$\eta \in \left(0, \frac{1}{2c_1\sqrt{d}}\right) \tag{22}$$

to be chosen later (here  $c_1 > 0$  is an absolute constant appearing in Lemma 5), and a random matrix  $S \sim \text{Unif}(\mathcal{S}_0)$ . The set of matrices  $\mathcal{S}_0$  is constructed so that the following three properties hold:

1. It is a subset of symmetric Rademacher matrices, i.e.  $\mathcal{S}_0 \subseteq \mathcal{S} \triangleq \{S \in \{\pm 1\}^{d \times d} : S = S^\top\}$ ;
2. Every matrix in  $\mathcal{S}_0$  has a small operator norm:  $\|S\|_{\text{op}} \leq c_1\sqrt{d}$  for all  $S \in \mathcal{S}_0$ ;
3. Different permutations to a matrix in  $\mathcal{S}_0$  give different matrices: for all  $S \in \mathcal{S}_0$  and  $\pi_1, \pi_2 \in S_d$  with  $\sum_{i=1}^d \mathbb{1}(\pi_1(i) \neq \pi_2(i)) \geq d/10$ , it holds that  $\|S^{\pi_1} - S^{\pi_2}\|_{\text{F}}^2 \geq c_2 d^2$ .

Note that the choice of  $\eta$  in (22) and the second property of  $\mathcal{S}_0$  ensure that  $\|\Sigma\|_{\text{op}} \leq 1$  almost surely. The next lemma shows that  $\mathcal{S}_0$  can be constructed as a sufficiently large subset of  $\mathcal{S}$ .

**Lemma 5.** *There exist absolute constants  $c_1, c_2, d_0 > 0$  such that the following holds: if  $d \geq d_0$ , there exists some  $\mathcal{S}_0 \subseteq \mathcal{S}$  such that all the above properties hold, and*

$$\frac{|\mathcal{S}_0|}{|\mathcal{S}|} \geq \frac{1}{2}.$$

We proceed to lower bound the Bayes risk under the above prior, and to this end we introduce the (generalized) Fano's inequality. Consider a general scenario where  $\theta \in \Theta$  is an unknown parameter, the learner observes  $X \sim P_\theta$ , and  $L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}_+$  is a non-negative loss function with a generic action space  $\mathcal{A}$ . The following lemma presents a general lower bound of the Bayes risk.

**Lemma 6.** *Let  $\pi$  be any probability distribution over  $\Theta$ . For any  $\Delta > 0$ , define*

$$p_\Delta := \sup_a \pi\{\theta \in \Theta : L(\theta, a) \leq \Delta\}.$$

*Then for  $\theta \sim \pi$ , the following Bayes risk lower bound holds:*

$$\inf_{T(\cdot)} \mathbb{E}_\pi \mathbb{E}_\theta [L(\theta, T(X))] \geq \Delta \left(1 - \frac{I(\theta; X) + \log 2}{\log(1/p_\Delta)}\right).$$

Lemma 6 was essentially proved in [DW13, CGZ16], and for completeness we include the proof of the above form in Appendix C.2. To apply this lemma, we will choose  $\theta = \pi^*$ , so that  $P_{\pi^*}$  is the mixture distribution  $\mathbb{E}_\Sigma[\mathcal{N}(0, \Sigma)^{\otimes m} \otimes \mathcal{N}(0, \Sigma^{\pi^*})^{\otimes n}]$ . The loss function  $L(\pi^*, \hat{\pi})$  is defined as

$$L(\pi^*, \hat{\pi}) = \min_{\Sigma=(I_d+\eta S)/2: S \in \mathcal{S}_0} \left\| \Sigma^{\pi^*} - \Sigma^{\hat{\pi}} \right\|_F^2 = \frac{\eta^2}{4} \min_{S \in \mathcal{S}_0} \left\| S^{\pi^*} - S^{\hat{\pi}} \right\|_F^2. \quad (23)$$

The next lemma provides an upper bound on the quantity  $p_\Delta$  used in Lemma 6.

**Lemma 7.** *Let the loss  $L$  be given in (23), and  $p_\Delta$  be defined in Lemma 6. Then for  $\Delta = c_2 \eta^2 d^2 / 4$ , it holds that  $p_\Delta \leq d^{-c_3 d}$ , where  $c_3 > 0$  is an absolute numerical constant.*

By Lemma 6 and Lemma 7, as well as the definition of  $L$  in (23), we have

$$\mathbb{E} \left[ \left\| \Sigma^{\pi^*} - \Sigma^{\hat{\pi}} \right\|_F^2 \right] \geq \mathbb{E} [L(\pi^*, \hat{\pi})] \geq \frac{c_2 \eta^2 d^2}{4} \left( 1 - \frac{I(\pi^*; X^m, Y^n) + \log 2}{c_3 d \log d} \right). \quad (24)$$

It remains to upper bound the mutual information  $I(\pi^*; X^m, Y^n)$ , which is deferred to Section 4.2.

## 4.2 Upper bounding the mutual information

In this section we aim to prove the following upper bound of mutual information:

$$I(\pi^*; X^m, Y^n) \leq c_4 \min \{ nd^2 \eta^2, mnd^2 \eta^4 + 1 \}, \quad (25)$$

where  $c_4 > 0$  is an absolute constant independent of  $(n, m, d, \eta)$ . We first show that (25) implies the desired minimax lower bound in Theorem 4. To see it, we choose

$$\eta = c_5 \max \left\{ \sqrt{\frac{\log d}{nd}}, \left( \frac{\log d}{mnd} \right)^{1/4} \right\}$$

with a small absolute constant  $c_5 > 0$ , then Theorem 4 is a direct consequence of (24). Also note that since  $m \geq d$  and  $n \geq \log d$ , our choice of  $\eta$  satisfies the constraint in (22).

In the rest of this section we prove (25). The  $O(nd^2 \eta^2)$  upper bound essentially corresponds to the known covariance case where  $m = \infty$ , and is easily established in the following lemma.

**Lemma 8.** *Regardless of the value of  $m$ , it always holds that*

$$I(\pi^*; X^m, Y^n) \leq \frac{nd^2 \eta^2}{2}.$$

In the sequel we establish the  $O(mnd^2 \eta^4 + 1)$  upper bound.

### 4.2.1 Reduction to the case $m = n$

We first prove an upper bound of  $I(\pi^*; X^m, Y^n)$  which no longer involves  $\pi^*$ ; the high-level idea is to relate the recovery of  $\pi^*$  (under the loss  $L(\pi^*, \hat{\pi})$ ) to a detection problem of telling if  $X^m, Y^n$  are i.i.d. drawn from  $\mathcal{N}(0, I_d/2)$ . To this end, we write

$$\begin{aligned}
I(\pi^*; X^m, Y^n) &= I(\pi^*; X^m) + I(\pi^*; Y^n | X^m) \\
&\stackrel{(a)}{=} I(\pi^*; Y^n | X^m) \\
&\stackrel{(b)}{=} \mathbb{E}_{X^m, \pi^*} [D_{\text{KL}}(P_{Y^n | \pi^*, X^m} \| \mathcal{N}(0, I_d/2)^{\otimes n})] - \mathbb{E}_{X^m} [D_{\text{KL}}(P_{Y^n | X^m} \| \mathcal{N}(0, I_d/2)^{\otimes n})] \\
&\leq \mathbb{E}_{X^m, \pi^*} [D_{\text{KL}}(P_{Y^n | \pi^*, X^m} \| \mathcal{N}(0, I_d/2)^{\otimes n})] \\
&\stackrel{(c)}{=} \mathbb{E}_{X^m, \pi^*} [D_{\text{KL}}(\Phi_{\#}^{\pi^*} P_{Y^n | X^m} \| \Phi_{\#}^{\pi^*} \mathcal{N}(0, I_d/2)^{\otimes n})] \\
&\stackrel{(d)}{=} \mathbb{E}_{X^m} [D_{\text{KL}}(P_{Z^n | X^m} \| \mathcal{N}(0, I_d/2)^{\otimes n})].
\end{aligned}$$

Here (a) is due to the independence between  $\pi^*$  and  $X^m$ , (b) follows from simple algebra (this step is also known as the variational representation of the mutual information). In (c), we define a bijection  $\Phi^{\pi^*} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  with

$$\Phi^{\pi^*}(y_1, \dots, y_d) = (y_{(\pi^*)^{-1}(1)}, \dots, y_{(\pi^*)^{-1}(d)}).$$

This definition is also extended in the natural way to  $(\mathbb{R}^d)^n$  as  $\Phi^{\pi^*}(Y_1, \dots, Y_n) = (\Phi^{\pi^*} Y_1, \dots, \Phi^{\pi^*} Y_n)$ . Clearly  $\Phi^{\pi^*}$  is a bijection, so (c) holds. As for (d), we introduce an auxiliary sequence of random vectors  $Z_1, Z_2, \dots$ , which conditioned on  $\Sigma$  are i.i.d. drawn from  $\mathcal{N}(0, \Sigma)$ . Since  $\pi^*$  is independent of  $X^m$ , it is clear that  $\Phi_{\#}^{\pi^*} P_{Y^n | X^m} = P_{Z^n | X^m}$ . In addition,  $\Phi_{\#}^{\pi^*} \mathcal{N}(0, I_d/2)^{\otimes n} = \mathcal{N}(0, I_d/2)^{\otimes n}$  trivially holds, so we arrive at (d). Note that the final expression no longer depends on  $\pi^*$ .

In the second step, we note that  $P_{Z^n | X^m} = \mathbb{E}_{\Sigma | X^m} [P_{Z^n | \Sigma}]$  is a mixture of product distributions and therefore exchangeable. Meanwhile, the second argument  $\mathcal{N}(0, I_d/2)^{\otimes n}$  in the KL divergence is a product distribution. This structure enables us to prove the following lemma and reduce two parameters  $(m, n)$  to a single parameter  $m$ .

**Lemma 9.** *Let  $(Z_1, Z_2, \dots)$  be a stationary process and  $Q$  be an arbitrary distribution. Then the sequence  $a_n = D_{\text{KL}}(P_{Z^n} \| Q^{\otimes n})$  satisfies that  $\{a_{n+1} - a_n\}$  is non-decreasing, i.e.*

$$a_1 \leq a_2 - a_1 \leq a_3 - a_2 \leq \dots$$

*In particular, if  $m \geq n$ , then  $a_m/m \geq a_n/n$ .*

Consequently, we proceed the upper bound as

$$\begin{aligned}
I(\pi^*; X^m, Y^n) &\leq \mathbb{E}_{X^m} [D_{\text{KL}}(P_{Z^n | X^m} \| \mathcal{N}(0, I_d/2)^{\otimes n})] \\
&\stackrel{(e)}{\leq} \frac{n}{m} \mathbb{E}_{X^m} [D_{\text{KL}}(P_{Z^m | X^m} \| \mathcal{N}(0, I_d/2)^{\otimes m})] \\
&\stackrel{(f)}{=} \frac{n}{m} \left( D_{\text{KL}}(P_{X^m, Z^m} \| \mathcal{N}(0, I_d/2)^{\otimes (2m)}) - D_{\text{KL}}(P_{X^m} \| \mathcal{N}(0, I_d/2)^{\otimes m}) \right) \\
&\leq \frac{n}{m} D_{\text{KL}}(P_{X^m, Z^m} \| \mathcal{N}(0, I_d/2)^{\otimes (2m)}),
\end{aligned}$$

where (e) follows from Lemma 9 and  $m \geq n$ , and (f) is the chain rule of KL divergence. In summary, we have shown that

$$I(\pi^*; X^m, Y^n) \leq \frac{n}{m} D_{\text{KL}}(\mathbb{E}_{\Sigma} [\mathcal{N}(0, \Sigma)^{\otimes (2m)}] \| \mathcal{N}(0, I_d/2)^{\otimes (2m)}). \quad (26)$$

Note that in (26) we essentially reduce to the case with equal sample sizes. Without this reduction we would only have  $I(\pi^*; X^m, Y^n) \leq D_{\text{KL}}(\mathbb{E}_\Sigma[\mathcal{N}(0, \Sigma)^{\otimes(n+m)}] \|\mathcal{N}(0, I_d/2)^{\otimes(n+m)}\)$ , which could be shown to be loose compared with the target upper bound in (25).

#### 4.2.2 Second moment method

Based on (26) we further have

$$I(\pi^*; X^m, Y^n) \stackrel{(a)}{\leq} \frac{n}{m} \log(1 + \chi^2(\mathbb{E}_\Sigma[\mathcal{N}(0, \Sigma)^{\otimes(2m)}] \|\mathcal{N}(0, I_d/2)^{\otimes(2m)}))$$

which follows from  $D_{\text{KL}}(P\|Q) \leq \log(1 + \chi^2(P\|Q))$ . Now to bound the  $\chi^2$ -divergence we derive the following lemma based on the standard second moment computation (cf. [FRW15, Appendix A]) which we prove for completeness in Appendix C.6.

**Lemma 10.** *Let  $\Sigma_S$  be a symmetric matrix in  $\mathbb{R}^{d \times d}$  indexed by a random variable  $S \sim \mu$ . Further assume that  $0 \prec \Sigma_S \prec 2I_d$  almost surely. Then for  $p \in \mathbb{N}$ , we have the identity*

$$\chi^2(\mathbb{E}_S[\mathcal{N}(0, \Sigma_S)^{\otimes p}] \|\mathcal{N}(0, I_d)^{\otimes p}) + 1 = \mathbb{E}_{S,T} \left[ \det \left( I - (\Sigma_S - I)(\Sigma_T - I) \right)^{-\frac{p}{2}} \right],$$

where  $S, T \sim \mu$  are independent.

Using Lemma 10 we can write

$$\log(1 + \chi^2(\mathbb{E}_\Sigma[\mathcal{N}(0, \Sigma)^{\otimes(2m)}] \|\mathcal{N}(0, I_d/2)^{\otimes(2m)})) \stackrel{(b)}{=} \log \mathbb{E}_{S,T} [\det(I_d - \eta^2 ST)^{-m}],$$

with  $T$  being an independent copy of  $S$ . Here (b) applies Lemma 10 to  $\Sigma_S = I_d + \eta S$ , where we have also used the invariance of the  $\chi^2$ -divergence with respect to bijections. The condition of Lemma 10 is fulfilled thanks to  $\eta < 1/(2c_1\sqrt{d})$  from (22) and  $\|S\|_{\text{op}} \leq c_1\sqrt{d}$  from the second property of  $\mathcal{S}_0$ .

To proceed, let  $\lambda_1, \dots, \lambda_d$  be the (positive real) eigenvalues of  $ST$ , and we write

$$\begin{aligned} \mathbb{E}_{S,T} [\det(I_d - \eta^2 ST)^{-m}] &= \mathbb{E}_{S,T} \left[ \exp \left( -m \sum_{i=1}^d \log(1 - \eta^2 \lambda_i) \right) \right] \\ &\stackrel{(c)}{\leq} \mathbb{E}_{S,T} \left[ \exp \left( -m \sum_{i=1}^d (-\eta^2 \lambda_i - \eta^4 \lambda_i^2) \right) \right] \\ &= \mathbb{E}_{S,T} \left[ \exp \left( m\eta^2 \langle S, T \rangle + m\eta^4 \|ST\|_{\text{F}}^2 \right) \right] \\ &\stackrel{(d)}{\leq} \exp(c_1^2 m d^3 \eta^4) \cdot \mathbb{E}_{S,T} [\exp(m\eta^2 \langle S, T \rangle)] \end{aligned}$$

where in (c) we have used that

$$\eta^2 |\lambda_i| \leq \left( \frac{1}{2c_1\sqrt{d}} \right)^2 \|S\|_{\text{op}} \|T\|_{\text{op}} \leq \frac{1}{4}$$

from (22) and the second property for  $\mathcal{S}_0$ , as well as  $\log(1+x) \geq x - x^2$  for all  $|x| \leq 1/4$ ; (d) follows from  $\|ST\|_F \leq \|S\|_{\text{op}} \|T\|_F \leq c_1 d^{3/2}$  almost surely. For the remaining quantity, we have

$$\begin{aligned} \mathbb{E}_{S,T} [\exp(m\eta^2 \langle S, T \rangle)] &\leq \left( \frac{|\mathcal{S}|}{|\mathcal{S}_0|} \right)^2 \mathbb{E}_{S,T \sim \text{Unif}(\mathcal{S})} [\exp(m\eta^2 \langle S, T \rangle)] \\ &\stackrel{(e)}{\leq} 4 \mathbb{E}_{S,T \sim \text{Unif}(\mathcal{S})} \left[ \exp \left( m\eta^2 \left( \sum_{i=1}^d S_{ii} T_{ii} + 2 \sum_{1 \leq i < j \leq d} S_{ij} T_{ij} \right) \right) \right] \\ &\stackrel{(f)}{\leq} 4 \exp \left( \frac{1}{2} m^2 \eta^4 d + 2m^2 \eta^4 \binom{d}{2} \right) \leq 4 \exp(m^2 \eta^4 d^2). \end{aligned}$$

Here (e) follows from Lemma 5, and (f) makes use of the independence and 1-subGaussianity of the Rademacher random variables. Piecing everything together gives that

$$I(\pi^*; X^m, Y^n) = O(mnd^2\eta^4 + nd^3\eta^4 + 1) = O(mnd^2\eta^4 + 1),$$

where the last step is due to our assumption  $m \geq d$ . This proves the second upper bound in (25).

## 5 Discussions

### 5.1 Other notions of minimax risk

Instead of using the squared Frobenius norm  $\|\Sigma^{\hat{\pi}} - \Sigma^{\pi^*}\|_F^2$  over the parameter set  $\{\Sigma \succeq 0 : \|\Sigma\|_{\text{op}} \leq 1\}$ , one may also consider a normalized Frobenius norm

$$d_{\text{NF},\Sigma}(\pi^*, \hat{\pi}) := \left\| (\Sigma^{\pi^*})^{-1/2} (\Sigma^{\hat{\pi}} - \Sigma^{\pi^*}) (\Sigma^{\pi^*})^{-1/2} \right\|_F, \quad (27)$$

as well as another parameter set based on the trace:

$$\Sigma(R) := \{\Sigma \succeq 0 : \text{Tr}(\Sigma) \leq R\}, \quad R > 0. \quad (28)$$

Note that  $d_{\text{NF},\Sigma}$  in (27) is closely related to the total variation distance  $\text{TV}(\mathcal{N}(0, \Sigma^{\hat{\pi}}), \mathcal{N}(0, \Sigma^{\pi^*}))$  for density estimation (cf. [DMR18, Theorem 1.1]), and (28) uses an  $\ell_1$  ball (instead of  $\ell_\infty$ ) of the spectrum of the covariance  $\Sigma$  to promote sparsity. The following theorem summarizes two different minimax risks over  $\Sigma(R)$ .

**Theorem 5.** *Let  $R > 0$ ,  $d \geq 2$ , and  $m \geq n$  be fixed sample sizes. Then*

$$\inf_{\hat{\pi}} \sup_{\pi^* \in S_d, \Sigma \in \Sigma(R)} \mathbb{E}_{(\pi^*, \Sigma)} \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \asymp \frac{R^2}{n}, \quad (29)$$

*and the GW estimator achieves this upper bound. If in addition  $n \geq c_0 d \log d$  and  $mn \geq c_0 \sqrt{d^3 \log d}$  for a sufficiently large numerical constant  $c_0 > 0$ , it holds that*

$$\inf_{\hat{\pi}} \sup_{\pi^* \in S_d, \Sigma \in \Sigma(R)} \mathbb{E}_{(\pi^*, \Sigma)} [d_{\text{NF},\Sigma}(\pi^*, \hat{\pi})^2] \asymp \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}}, \quad (30)$$

*and the QMLE achieves this upper bound.*



Lemma 5 shows that if we have the constraint on the trace instead of the operator norm, the minimax Frobenius risk in (29) coincides with the minimax Frobenius risk  $\inf_{\hat{\Sigma}} \sup_{\Sigma} \|\hat{\Sigma} - \Sigma\|_F^2$  of estimating the covariance matrix  $\Sigma$  using  $n$  observations. In other words, under the trace constraint, the minimax optimal estimator essentially needs to estimate the nuisance parameter  $\Sigma$ . This is because the hard instance of  $\Sigma$  under the trace constraint has a very sparse spectrum: only  $O(1)$  eigenvalues of  $\Sigma$  are non-zero. This essentially reduces the dimension of the problem to  $d = O(1)$ , and therefore saving a  $\text{poly}(d)$  factor in the minimax risk becomes vacuous. We also note that the GW estimator remains minimax rate-optimal in this scenario.

We could circumvent the above problem by considering a different loss. By using the normalized Frobenius norm  $d_{\text{NF},\Sigma}$  in (27), the minimax rate in (30) of Theorem 5 coincides with the minimax rate in Theorem 2 – 4 again, even if  $\Sigma$  has a sparse spectrum. Therefore, our main result in this paper is not an artifact of considering a special family of covariance matrices.

## 5.2 Comparison with orthogonal statistical learning

One particular feature of covariance alignment is that the dimension ( $\asymp d^2$ ) of the nuisance parameter  $\Sigma$  is much larger than the counterpart ( $\asymp d$ ) of the target parameter  $\pi^*$ , and the minimax rate takes a semiparametric form that interpolates between the rates for covariance estimation and permutation estimation with known covariance. This feature shares some similarities with the classical semiparametric statistics [BKB<sup>+</sup>93] or more recently orthogonal statistical learning [CCD<sup>+</sup>17, FS23], where a typical result is that the parametric rate can be achieved for the target parameter even if the nuisance is estimated with a slower rate. The semiparametric analysis typically relies on the Neyman orthogonality condition [Ney59] or its higher-order counterparts [MSZ18]. Specifically, let  $r_T$  be an achievable rate of convergence for estimating the target *if the nuisance were known*, and  $r_N$  be an achievable rate of convergence for estimating the nuisance. Then under the Neyman orthogonality condition, the rate of convergence for estimating the target under an unknown nuisance is shown [FS23] to be  $O(r_T + r_N^2)$ , or even  $O(r_T + r_N^4)$  if a certain strong convexity condition is met.

We illustrate our main distinction from the above line of research. First, the Neyman orthogonality condition does not hold in our problem. For example, the GW estimator uses the loss function  $\ell(y; P, \Sigma) = -y^\top P \Sigma P^\top y$  in the sense that  $\hat{P}^{\text{GW}} = \arg\min_{P \in \text{BP}(d)} \frac{1}{n} \sum_{i=1}^n \ell(Y_i; P, \hat{\Sigma}_X)$ , while

$$\mathbb{E}[\nabla_P \nabla_\Sigma \ell(Y; P_{\pi^*}, \Sigma)] = \mathbb{E}[\nabla_P (-P^\top y y^\top P) |_{P=P_{\pi^*}}] = -\nabla_P (P^\top P_{\pi^*} \Sigma P_{\pi^*}^\top P) |_{P=P_{\pi^*}} \neq 0.$$

Second, instead of using a Taylor expansion to decouple the target estimation and nuisance estimation, we crucially make use of the specific form of the nuisance estimator when used for target estimation. This point can be rigorously established by comparing with the following oracle model.

Under an oracle model, we still observe the sample  $Y^n$ ; however, instead of observing the other sample  $X^m$ , now we have access to an oracle which outputs a covariance estimate  $\hat{\Sigma}_m$ . The only property we know about  $\hat{\Sigma}_m$  is that  $\|\hat{\Sigma}_m - \Sigma\|_{\text{op}} \leq \sqrt{d/m}$  almost surely. This would essentially be the case if  $\hat{\Sigma}_m$  were the empirical covariance computed from  $X^m$  in the original observation model, while under the oracle model it is unknown if  $\hat{\Sigma}_m$  is unbiased and/or Wishart distributed. We note that the statistical guarantees of orthogonal statistical learning [FS23] only rely on such a weaker oracle model for the nuisance estimation.

In covariance alignment, however, we prove the following theorem stating that the minimax rate of convergence for estimating  $\pi^*$  is strictly larger under the above oracle model.

**Theorem 6.** *For  $m \geq n \geq d \log d$ , under the oracle model it holds that*

$$\inf_{\hat{\pi}} \sup_{\pi^*, \Sigma: \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_\Sigma \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \asymp \frac{d \log d}{n} + \frac{d^2}{m}.$$

In addition, the GW estimator (with  $\widehat{\Sigma}_X$  replaced by  $\widehat{\Sigma}_m$ ) achieves this upper bound.

By a simple AM-GM inequality, the above rate is no smaller than the minimax rate in Theorems 2-4 under the original observation model. In particular, if  $m = n$ , the sample complexity for consistent estimation increases from  $n = \Theta(\sqrt{d^3 \log d})$  to  $n = \Theta(d^2)$  in the oracle model. Theorem 6 shows that, the nuisance estimation error  $\Theta(d^2/m)$  for  $\|\widehat{\Sigma}_m - \Sigma\|_F^2$  is an unavoidable price to pay in the oracle model. Therefore, one cannot hope to fit the covariance alignment model using only the orthogonal statistical learning framework; instead, our proof in Section 3 relies on the Wishart distribution of  $\widehat{\Sigma}_X$  to prove a *high-probability* curvature property of the empirical loss. We note that both the oracle model and our finding in Theorem 6 are similar to in spirit to [BKW23] for several nonparametric functional estimation problems.

**Funding** Yanjun Han was generously supported by the Norbert Wiener postdoctoral fellowship in statistics at MIT IDSS. Philippe Rigollet is supported by NSF grants IIS-1838071, DMS-2022448, and CCF-2106377. George Stepaniants is supported through a National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

## A Auxiliary lemmas

The following lemma follows from the well-known Hanson–Wright inequality [HW71, Wri73].

**Lemma 11.** *Let  $X_1, \dots, X_n \in \mathbb{R}^d$  be i.i.d. random vectors, with independent zero-mean 1-subGaussian coordinates. Then there exists an absolute constant  $C > 0$  independent of  $(n, d)$  such that, for every symmetric matrix  $A \in \mathbb{R}^{d \times d}$  and  $\delta \in (0, 1)$ ,*

$$\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n X_i^\top A X_i \geq \mathbb{E}[X^\top A X] - C \left( \sqrt{\frac{\log(1/\delta)}{n}} \|A\|_F + \frac{\log(1/\delta)}{n} \|A\|_{\text{op}} \right) \right) \geq 1 - \delta.$$

In particular, whenever  $n \geq \log(1/\delta)$ , it holds that

$$\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n X_i^\top A X_i \geq \mathbb{E}[X^\top A X] - 2C \sqrt{\frac{\log(1/\delta)}{n}} \|A\|_F \right) \geq 1 - \delta.$$

*Proof.* For the first statement, let  $X^{(n)} \in \mathbb{R}^{nd}$  be a long vector via the concatenation of  $X_1, \dots, X_n$ , and  $A^{(n)} \in \mathbb{R}^{nd \times nd}$  be the matrix with  $n$  repeated appearances of  $A/n$  in the diagonal. We have

$$\frac{1}{n} \sum_{i=1}^n X_i^\top A X_i = (X^{(n)})^\top A^{(n)} X^{(n)},$$

and  $\|A^{(n)}\|_F = \|A\|_F / \sqrt{n}$ ,  $\|A^{(n)}\|_{\text{op}} = \|A\|_{\text{op}} / n$ . Now the first statement follows from the standard Hanson–Wright inequality (cf. [RV13, Theorem 1.1]). Since  $\|A\|_{\text{op}} \leq \|A\|_F$  and  $\log(1/\delta)/n \leq 1$ , the second statement directly follows.  $\square$

The following inequality appears in [DMR18, Page 4], which was stated without proof. For completeness we include a proof here.

**Lemma 12.** *If  $\Sigma \in \mathbb{R}^{d \times d}$  is a positive definite matrix then*

$$\frac{1}{2} \leq \frac{\min(1, \|\Sigma^{-1} - I\|_F)}{\min(1, \|\Sigma - I\|_F)} \leq 2. \quad (31)$$

*Proof.* We only prove the upper bound, as the lower bound is entirely symmetric. Also note that the upper bound holds trivially when  $\|\Sigma - I\|_F \geq 1/2$ , so in the sequel we assume that  $\|\Sigma - I\|_F < 1/2$ . Let  $\lambda_1, \dots, \lambda_d > 0$  be the eigenvalues of  $\Sigma$  (with multiplicities). Since  $\|\Sigma - I\|_F \geq \max_{i \in [d]} |\lambda_i - 1|$ , we have  $\lambda_i > 1/2$  for all  $i \in [d]$ . Therefore,

$$\begin{aligned} \min(1, \|\Sigma^{-1} - I\|_F)^2 &\leq \|\Sigma^{-1} - I\|_F^2 = \sum_{i=1}^d \left( \frac{1}{\lambda_i} - 1 \right)^2 = \sum_{i=1}^d \frac{(\lambda_i - 1)^2}{\lambda_i^2} \\ &\leq 4 \sum_{i=1}^d (\lambda_i - 1)^2 = 4 \|\Sigma - I\|_F^2 = 4 \min(1, \|\Sigma - I\|_F)^2, \end{aligned}$$

which is the claimed upper bound.  $\square$

The following lemma is a simple quadratic inequality which will be useful in the proof.

**Lemma 13.** *If  $x \in \mathbb{R}_+$  satisfies  $ax^2 \leq bx + c$  for  $a, b, c > 0$ . Then  $x \leq \frac{b}{a} + \sqrt{\frac{c}{a}}$ .*

*Proof.* If  $ax > b + \sqrt{ac}$ , then

$$ax^2 - bx - c = (ax - b)x - c > \sqrt{ac} \cdot \sqrt{\frac{c}{a}} - c = 0,$$

a contradiction.  $\square$

## B Deferred proofs in Section 3

### B.1 Proof of Lemma 2

We first deal with the term  $R_1$ . By triangle inequality,

$$R_1 \leq \left\| \Sigma^{1/2} (\widehat{\Sigma}_X^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F + \left\| \Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d \right\|_F. \quad (32)$$

Under  $\mathcal{E}_3$ , the second term could be easily upper bounded by

$$\begin{aligned} \left\| \Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d \right\|_F &\leq \sqrt{d} \left\| \Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d \right\|_{\text{op}} \\ &\stackrel{(a)}{\leq} \sqrt{d} \cdot \frac{\left\| \Sigma^{-1/2} \widehat{\Sigma}_X \Sigma^{-1/2} - I_d \right\|_{\text{op}}}{1 - \left\| \Sigma^{-1/2} \widehat{\Sigma}_X \Sigma^{-1/2} - I_d \right\|_{\text{op}}} \\ &\stackrel{(b)}{\leq} 2c\sqrt{d} \cdot \sqrt{\frac{d + \log(1/\delta)}{m}}. \end{aligned} \quad (33)$$

Here (a) follows from  $\|A^{-1} - I\|_{\text{op}} \leq \|A - I\|_{\text{op}} / (1 - \|A - I\|_{\text{op}})$  as long as  $\|A - I\|_{\text{op}} < 1$ ; the last condition is fulfilled by the event  $\mathcal{E}_3$  in (14). Step (b) directly follows from the event  $\mathcal{E}_3$  in (14).

The upper bound of the first term is slightly more involved. By triangle inequality,

$$\begin{aligned}
\left\| \Sigma^{1/2}(\widehat{\Sigma}_X^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F &= \left\| (\Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2})(\Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2})(\Sigma^{-1/2} P_{\widehat{\pi}}^{\top} \Sigma^{1/2}) - I_d \right\|_F \\
&\leq \left\| (\Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2})(\Sigma^{-1/2} P_{\widehat{\pi}}^{\top} \Sigma^{1/2}) - I_d \right\|_F \\
&\quad + \left\| (\Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2})(\Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d)(\Sigma^{-1/2} P_{\widehat{\pi}}^{\top} \Sigma^{1/2}) \right\|_F \\
&\stackrel{(c)}{\leq} \left\| \Sigma^{1/2}(\Sigma^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F + \left\| \Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2} \right\|_{\text{op}}^2 \left\| \Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d \right\|_F \\
&\stackrel{(d)}{=} \left\| \Sigma^{1/2}(\Sigma^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F + \left\| \Sigma^{1/2} \widehat{\Sigma}_X^{-1} \Sigma^{1/2} - I_d \right\|_F. \tag{34}
\end{aligned}$$

Here (c) uses  $\|ABC\|_F \leq \|A\|_{\text{op}} \|BC\|_F \leq \|A\|_{\text{op}} \|B\|_F \|C\|_{\text{op}}$ , and (d) uses the similarity between  $\Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2}$  and  $P_{\widehat{\pi}}$  to conclude that  $\left\| \Sigma^{1/2} P_{\widehat{\pi}} \Sigma^{-1/2} \right\|_{\text{op}} = 1$ .

A combination of (32), (33), and (34) gives that

$$R_1 \leq \left\| \Sigma^{1/2}(\Sigma^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F + 4c\sqrt{d} \cdot \sqrt{\frac{d + \log(1/\delta)}{m}},$$

which is the first statement of the lemma.

As for  $R_2$ , we begin with an upper bound on the operator norm of

$$\Sigma^{1/2}(\widehat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \widehat{\Sigma}_X)\Sigma^{-1})\Sigma^{1/2} =: A^{-1} + A - 2I_d,$$

where  $A := \Sigma^{-1/2} \widehat{\Sigma}_X \Sigma^{-1/2}$ . By the definition of  $\mathcal{E}_3$  in (14), all eigenvalues of  $A$  lie in the interval  $[1 - \eta, 1 + \eta]$ , with

$$\eta = \min \left\{ c\sqrt{\frac{d + \log(1/\delta)}{m}}, \frac{1}{2} \right\}.$$

It is clear that all eigenvalues of  $A^{-1} + A - 2I_d$  take the form  $\lambda^{-1} + \lambda - 2$ , with  $\lambda$  being an eigenvalue of  $A$ . For  $\lambda \in [1 - \eta, 1 + \eta]$ , we have

$$\frac{1}{\lambda} + \lambda - 2 = \frac{(\lambda - 1)^2}{\lambda} \leq \frac{\eta^2}{1 - \eta} \leq 2c^2 \cdot \frac{d + \log(1/\delta)}{m}.$$

Therefore, we have established that

$$\left\| \Sigma^{1/2}(\widehat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \widehat{\Sigma}_X)\Sigma^{-1})\Sigma^{1/2} \right\|_{\text{op}} \leq 2c^2 \cdot \frac{d + \log(1/\delta)}{m}. \tag{35}$$

Consequently,

$$\begin{aligned}
|R_2| &= |\langle \Sigma^{-1/2}(\Sigma^{\widehat{\pi}^{-1}} - \Sigma)\Sigma^{-1/2}, \Sigma^{1/2}(\widehat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \widehat{\Sigma}_X)\Sigma^{-1})\Sigma^{1/2} \rangle| \\
&\leq \left\| \Sigma^{-1/2}(\Sigma^{\widehat{\pi}^{-1}} - \Sigma)\Sigma^{-1/2} \right\|_F \cdot \left\| \Sigma^{1/2}(\widehat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \widehat{\Sigma}_X)\Sigma^{-1})\Sigma^{1/2} \right\|_F \\
&\leq \left\| \Sigma^{-1/2}(\Sigma^{\widehat{\pi}^{-1}} - \Sigma)\Sigma^{-1/2} \right\|_F \cdot \sqrt{d} \left\| \Sigma^{1/2}(\widehat{\Sigma}_X^{-1} - \Sigma^{-1} - \Sigma^{-1}(\Sigma - \widehat{\Sigma}_X)\Sigma^{-1})\Sigma^{1/2} \right\|_{\text{op}} \\
&\stackrel{(35)}{\leq} 2c^2 \cdot \frac{\sqrt{d}(d + \log(1/\delta))}{m} \left\| \Sigma^{-1/2} \Sigma^{\widehat{\pi}^{-1}} \Sigma^{-1/2} - I_d \right\|_F.
\end{aligned}$$

This completes the proof of the upper bound on  $|R_2|$ .  $\square$

## B.2 Proof of Lemma 3

Let  $S \triangleq \sum_{i=1}^d (x_i - 1)$ . By AM-GM inequality we have  $S \geq 0$ .

We first prove an upper bound on  $\max_i x_i$  in terms of  $S$ . For each  $i \in [d]$ , it holds that

$$S + d = x_i + \sum_{j \neq i} x_j \geq x_i + (d-1) \left( \prod_{j \neq i} x_j \right)^{\frac{1}{d-1}} = x_i + \frac{d-1}{x_i^{1/(d-1)}}.$$

As  $(1/x_i)^{1/(d-1)} = e^{-\log(x_i)/(d-1)} \geq 1 - \log(x_i)/(d-1)$ , the above inequality gives

$$S + 1 \geq x_i - \log x_i.$$

As the function  $x \mapsto h(x) \triangleq x - \log x$  is increasing on  $[1, \infty)$ , and  $h(2S+2) = 2S+2 - \log(2) - \log(S+1) \geq 2S+2 - 1 - S = S+1$ , we conclude that  $x_i \leq 2(S+1)$ . This inequality holds for all  $i \in [d]$ , so  $\max_i x_i \leq 2(S+1)$ .

Next we introduce the following lemma, the proof of which is deferred to the end of this section.

**Lemma 14.** *For  $x \in (-1, r]$  we have that*

$$\log(1+x) \leq x - \frac{x^2}{2(r+1)}.$$

Applying Lemma 14 with  $r = 2S+1$  gives

$$0 = \sum_{i=1}^d \log(x_i) \leq \sum_{i=1}^d \left( (x_i - 1) - \frac{(x_i - 1)^2}{4(S+1)} \right) = S - \frac{1}{4(S+1)} \sum_{i=1}^d (x_i - 1)^2,$$

which implies that

$$\sum_{i=1}^d (x_i - 1)^2 \leq 4S(S+1),$$

as claimed. □

*Proof of Lemma 14.* Define

$$f(x) = x - \frac{x^2}{2(r+1)} - \log(1+x).$$

We have

$$f'(x) = 1 - \frac{x}{r+1} - \frac{1}{x+1} = \frac{x(r-x)}{(x+1)(r+1)},$$

which satisfies  $f'(x) \geq 0$  for  $x \in [0, r]$  and  $f'(x) \leq 0$  for  $x \in (-1, 0]$ . Hence,  $f(x) \geq f(0) = 0$  for all  $x \in (-1, r]$ , as claimed. □

## C Deferred proofs in Section 4

### C.1 Proof of Lemma 5

By the probabilistic argument, it suffices to show that by drawing a random matrix  $S \sim \text{Unif}(\mathcal{S})$ , the probability that  $S$  satisfies all three properties is at least  $1/2$ . The first property holds trivially. For the second property, it was shown in [Ver18, Corollary 4.4.8] that

$$\mathbb{P}(\|S\|_{\text{op}} \geq c_1 \sqrt{d}) \leq 4 \exp(-c_1 d)$$

holds for some absolute constant  $c_1 > 0$  independent of  $d$ . By choosing  $d_0 > 0$  large enough we may ensure that  $\mathbb{P}(\|S\|_{\text{op}} \geq c_1 \sqrt{d}) \leq 1/4$  for all  $d \geq d_0$ .

For the last property, note that for fixed  $\pi_1, \pi_2 \in S_d$ , we have

$$\|S^{\pi_1} - S^{\pi_2}\|_{\text{F}}^2 = \|P_{\pi_1} S P_{\pi_1}^\top - P_{\pi_2} S P_{\pi_2}^\top\|_{\text{F}}^2 = \|P_{\pi_2^{-1}} P_{\pi_1} S P_{\pi_1}^\top P_{\pi_2^{-1}}^\top - S\|_{\text{F}}^2 = \|P_{\pi_1 \circ \pi_2^{-1}} S P_{\pi_1 \circ \pi_2^{-1}}^\top - S\|_{\text{F}}^2,$$

where we have used that  $P_\pi^{-1} = P_{\pi^{-1}}$  and  $P_{\pi_1} P_{\pi_2} = P_{\pi_2 \circ \pi_1}$ . Let  $\pi = \pi_1 \circ \pi_2^{-1} \in S_d$ . Note that

$$\|S^\pi - S\|_{\text{F}}^2 = \sum_{i,j=1}^d (S_{\pi(i), \pi(j)} - S_{i,j})^2 =: X^\top A X,$$

where  $X$  is the  $\binom{d+1}{2}$ -dimensional column vector  $(S_{i,j})_{1 \leq i \leq j \leq d}$ , and the matrix  $A$  is given as follows: for  $i \leq j$  and  $k \leq \ell$ ,

$$A_{(i,j),(k,\ell)} = \begin{cases} 2 \cdot \mathbb{1}(i = k) - \mathbb{1}(\pi(i) = k) - \mathbb{1}(\pi(k) = i), & \text{if } i = j, k = \ell; \\ 0, & \text{if } i = j, k \neq \ell; \\ 4[1 - \mathbb{1}((\pi(i), \pi(j)) = (i, j) \text{ or } (j, i))], & \text{if } i \neq j, (k, \ell) = (i, j); \\ -2[\mathbb{1}((\pi(i), \pi(j)) = (k, \ell) \text{ or } (\ell, k)) \\ \quad + \mathbb{1}((\pi(k), \pi(\ell)) = (i, j) \text{ or } (j, i))], & \text{if } i \neq j, (k, \ell) \neq (i, j). \end{cases}$$

Clearly  $A$  is symmetric, and

$$\begin{aligned} & \max_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} |A_{(i,j),(k,\ell)}| \\ & \leq \max_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} 4 \cdot \mathbb{1}(\{k, \ell\} = \{i, j\}, \{\pi(i), \pi(j)\} \text{ or } \{\pi^{-1}(i), \pi^{-1}(j)\}) \\ & \leq 12. \end{aligned}$$

As a result, we have

$$\|A\|_{\text{op}} \leq \max_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} |A_{(i,j),(k,\ell)}| \leq 12, \quad (36)$$

$$\|A\|_{\text{F}}^2 \leq 4 \sum_{1 \leq i \leq j \leq d} \sum_{1 \leq k \leq \ell \leq d} |A_{(i,j),(k,\ell)}| \leq 48 \binom{d+1}{2} = 24d(d+1). \quad (37)$$

In addition, whenever  $\sum_{i=1}^d \mathbb{1}(\pi(i) \neq i) = \sum_{i=1}^d \mathbb{1}(\pi_1(i) \neq \pi_2(i)) \geq d/10$ , it holds that

$$\begin{aligned} \text{Tr}(A) &= \sum_{1 \leq i \leq j \leq d} A_{(i,j),(i,j)} = 2 \sum_{i,j=1}^d \mathbb{1}((\pi(i), \pi(j)) \neq (i,j), (j,i)) \\ &\geq 2 \sum_{i,j=1}^d \mathbb{1}(\pi(i) \neq i) \mathbb{1}(\pi(i) \neq j) \\ &= 2(d-1) \sum_{i=1}^d \mathbb{1}(\pi(i) \neq i) \geq \frac{d^2}{10}. \end{aligned} \quad (38)$$

With the help of (36)-(38), we are ready to prove a high-probability lower bound of  $\|S^{\pi_1} - S^{\pi_2}\|_{\text{F}}^2 = X^\top AX$ . Applying the Hanson–Wright inequality (cf. Lemma 11) with  $n = 1$  to  $X^\top AX$ , with probability exceeding  $1 - 1/[4(d!)^2]$  it holds that

$$\begin{aligned} \|S^{\pi_1} - S^{\pi_2}\|_{\text{F}}^2 &= X^\top AX \geq \mathbb{E}[X^\top AX] - C \left( \sqrt{d \log d} \|A\|_{\text{F}} + (d \log d) \|A\|_{\text{op}} \right) \\ &= \text{Tr}(A) - C \left( \sqrt{d \log d} \|A\|_{\text{F}} + (d \log d) \|A\|_{\text{op}} \right) \\ &\geq \frac{d^2}{10} - O \left( \sqrt{d \log d} \cdot d + d \log d \right) \geq c_2 d^2 \end{aligned}$$

as long as  $d \geq d_0$  for a sufficiently large  $d_0$  and a sufficiently small  $c_2 > 0$ . Now applying the union bound over at most  $(d!)^2$  choices of  $(\pi_1, \pi_2)$ , we conclude that  $S$  satisfies the third property with probability at least  $3/4$ .

Finally, by a union bound again, with probability at least  $1/2$ , the random matrix  $S$  satisfies all three properties. This completes the proof of the lemma.  $\square$

## C.2 Proof of Lemma 6

Fix any estimator  $T$ . Let  $\mathbb{P}$  be the joint distribution of  $(\theta, X) \sim \pi(\theta)P_\theta(x)$  under the Bayes setup, and  $\mathbb{Q}$  be another joint distribution of  $(\theta, X) \sim \pi(\theta)Q(x)$ , where  $\theta$  and  $X$  are independent. Here  $Q$  is an arbitrary probability distribution over  $\mathcal{X}$ . Consider a map  $\Phi : \Theta \times \mathcal{X} \rightarrow \{0, 1\}$  defined as  $\Phi(\theta, x) = \mathbb{1}(L(\theta, T(X)) \leq \Delta)$ , the data-processing inequality gives that

$$\begin{aligned} D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) &\geq D_{\text{KL}}(\Phi_{\#} \mathbb{P} \parallel \Phi_{\#} \mathbb{Q}) \\ &= \mathbb{P}(L(\theta, T(X)) \leq \Delta) \log \frac{\mathbb{P}(L(\theta, T(X)) \leq \Delta)}{\mathbb{Q}(L(\theta, T(X)) \leq \Delta)} \\ &\quad + (1 - \mathbb{P}(L(\theta, T(X)) \leq \Delta)) \log \frac{1 - \mathbb{P}(L(\theta, T(X)) \leq \Delta)}{1 - \mathbb{Q}(L(\theta, T(X)) \leq \Delta)} \\ &\geq \mathbb{P}(L(\theta, T(X)) \leq \Delta) \log \frac{1}{\mathbb{Q}(L(\theta, T(X)) \leq \Delta)} - \log 2, \end{aligned}$$

where the last step is due to  $x \log x + (1-x) \log(1-x) \geq -\log 2$  for all  $x \in [0, 1]$ . Since

$$\mathbb{Q}(L(\theta, T(X)) \leq \Delta) \leq \sup_a \pi\{\theta \in \Theta : L(\theta, a) \leq \Delta\} = p_\Delta,$$

the above inequality rearranges to

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[L(\theta, T(X))] &\geq \Delta \cdot \mathbb{P}(L(\theta, T(X)) > \Delta) \\ &\geq \Delta \left( 1 - \frac{D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) + \log 2}{\log(1/p_\Delta)} \right). \end{aligned}$$



Finally, since the above inequality holds for all  $Q$ , using the variational representation of the mutual information  $I(\theta; X) = \min_Q D_{\text{KL}}(\pi(\theta)P_\theta(x) \parallel \pi(\theta)Q(x))$  completes the proof.  $\square$

### C.3 Proof of Lemma 7

By the third property of  $\mathcal{S}_0$ , we have the inclusion

$$\left\{ \pi^* : L(\pi^*, \hat{\pi}) \leq \frac{c_2 \eta^2 d^2}{4} \right\} = \left\{ \pi^* : \|S^{\pi^*} - S^{\hat{\pi}}\|_{\text{F}}^2 \leq c_2 d^2 \right\} \subseteq \left\{ \pi^* : \sum_{i=1}^d \mathbb{1}(\pi^*(i) \neq \hat{\pi}(i)) \leq \frac{d}{10} \right\}.$$

As  $\pi^* \sim \text{Unif}(S_d)$ , for any fixed  $\hat{\pi}$ , the final probability is at most

$$\frac{1}{d!} \sum_{k=0}^{\lfloor d/10 \rfloor} \binom{d}{k} k! = \sum_{k=0}^{\lfloor d/10 \rfloor} \frac{1}{(d-k)!} \leq \left( \frac{d}{10} + 1 \right) \frac{1}{\lceil 9d/10 \rceil!} \leq \exp(-c_3 d \log d)$$

for some absolute constant  $c_3 > 0$ . This shows that  $p_\Delta \leq d^{-c_3 d}$ .  $\square$

### C.4 Proof of Lemma 8

The following upper bound holds for  $I(\pi^*; X^m, Y^n)$ :

$$\begin{aligned} I(\pi^*; X^m, Y^n) &= I(\pi^*; X^m) + I(\pi^*; Y^n \mid X^m) \\ &\stackrel{(a)}{=} I(\pi^*; Y^n \mid X^m) \\ &\stackrel{(b)}{\leq} \mathbb{E}_{X^m, \pi^*} [D_{\text{KL}}(P_{Y^n \mid \pi^*, X^m} \parallel \mathcal{N}(0, I_d/2)^{\otimes n})] \\ &\stackrel{(c)}{\leq} \mathbb{E}_{\Sigma, \pi^*} [D_{\text{KL}}(P_{Y^n \mid \pi^*, \Sigma} \parallel \mathcal{N}(0, I_d/2)^{\otimes n})] \\ &\stackrel{(d)}{=} \frac{n}{2} \mathbb{E}_S [\eta \text{Tr}(S) - \log \det(I_d + \eta S)]. \end{aligned}$$

Here (a) is due to the independence between  $\pi^*$  and  $X^m$ , (b) follows from the variational representation of the mutual information, (c) follows from the convexity of the KL divergence and that  $P_{Y^n \mid \pi^*, X^m} = \mathbb{E}_{\Sigma \mid X^m} [P_{Y^n \mid \pi^*, \Sigma}]$ , (d) is the KL divergence between two Gaussians as well as the simple facts  $\text{Tr}(S) = \text{Tr}(S^{\pi^*})$ ,  $\det(I_d + \eta S) = \det(I_d + \eta S^{\pi^*})$ . To proceed, let  $\lambda_1, \dots, \lambda_d$  be the eigenvalues of  $S$ , so that

$$\begin{aligned} \eta \text{Tr}(S) - \log \det(I_d + \eta S) &= \sum_{i=1}^d (\eta \lambda_i - \log(1 + \eta \lambda_i)) \\ &\stackrel{(e)}{\leq} \sum_{i=1}^d \eta^2 \lambda_i^2 = \eta^2 \|S\|_{\text{F}}^2 = \eta^2 d^2. \end{aligned}$$

Here (e) follows from  $x - \log(1 + x) \leq x^2$  whenever  $|x| \leq 1/2$ , and we recall that  $\eta \leq 1/(2c_1 \sqrt{d})$  from (22), and that  $|\lambda_i| \leq c_1 \sqrt{d}$  almost surely from the second property of  $\mathcal{S}_0$ . A combination of the above inequalities gives that  $I(\pi^*; X^m, Y^n) \leq n\eta^2 d^2/2$ , as desired.  $\square$

## C.5 Proof of Lemma 9

For the first statement, note that

$$\begin{aligned}
a_{n+1} - a_n &= D_{\text{KL}}(P_{Z^{n+1}} \| Q^{\otimes(n+1)}) - D_{\text{KL}}(P_{Z^n} \| Q^{\otimes n}) \\
&\stackrel{(a)}{=} \mathbb{E}_{P_{Z^n}} [D_{\text{KL}}(P_{Z_{n+1}|Z^n} \| Q)] \\
&\stackrel{(b)}{\geq} \mathbb{E}_{P_{Z_2^n}} [D_{\text{KL}}(\mathbb{E}_{P_{Z_1|Z_2^n}} [P_{Z_{n+1}|Z^n}] \| Q)] \\
&= \mathbb{E}_{P_{Z_2^n}} [D_{\text{KL}}(P_{Z_{n+1}|Z_2^n} \| Q)] \\
&\stackrel{(c)}{=} \mathbb{E}_{P_{Z^{n-1}}} [D_{\text{KL}}(P_{Z_n|Z^{n-1}} \| Q)] \\
&= a_n - a_{n-1},
\end{aligned}$$

where (a) is crucially thanks to the product structure on  $Q$  and could be verified by simple algebra, (b) follows from the convexity of the KL divergence, and (c) is due to the stationarity of  $(Z_1, Z_2, \dots)$ . The first statement is therefore proved. The second statement directly follows from

$$\frac{a_m}{m} = \frac{\sum_{k=1}^m (a_k - a_{k-1})}{m} \geq \frac{\sum_{k=1}^n (a_k - a_{k-1})}{n} = \frac{a_n}{n}.$$

□

## C.6 Proof of Lemma 10

As a shorthand we denote the centered Gaussian distribution  $\mathcal{N}(0, \Sigma_S)$  by  $\mathbb{P}_S$ , and  $\mathcal{N}(0, I_d)$  by  $\mathbb{Q}$ . The standard second moment computation gives

$$\begin{aligned}
\chi^2(\mathbb{E}_S[\mathcal{N}(0, \Sigma_S)^{\otimes p}] \| \mathcal{N}(0, I_d)^{\otimes p}) + 1 &= \mathbb{E}_{\mathbb{Q}^{\otimes p}} \left[ \left( \frac{d\mathbb{E}_S[\mathbb{P}_S^{\otimes p}]}{d\mathbb{Q}^{\otimes p}} \right)^2 \right] \\
&= \mathbb{E}_{S,T} \left[ \mathbb{E}_{\mathbb{Q}^{\otimes p}} \left[ \frac{d\mathbb{P}_S^{\otimes p}}{d\mathbb{Q}^{\otimes p}} \frac{d\mathbb{P}_T^{\otimes p}}{d\mathbb{Q}^{\otimes p}} \right] \right] \\
&= \mathbb{E}_{S,T} \left[ \left( \mathbb{E}_{\mathbb{Q}} \left[ \frac{d\mathbb{P}_S}{d\mathbb{Q}} \frac{d\mathbb{P}_T}{d\mathbb{Q}} \right] \right)^p \right].
\end{aligned}$$

By Cauchy-Schwarz,

$$\mathbb{E}_{\mathbb{Q}} \left[ \frac{d\mathbb{P}_S}{d\mathbb{Q}} \frac{d\mathbb{P}_T}{d\mathbb{Q}} \right] \leq \left( \mathbb{E}_{\mathbb{Q}} \left[ \left( \frac{d\mathbb{P}_S}{d\mathbb{Q}} \right)^2 \right] \right)^{\frac{1}{2}} \left( \mathbb{E}_{\mathbb{Q}} \left[ \left( \frac{d\mathbb{P}_T}{d\mathbb{Q}} \right)^2 \right] \right)^{\frac{1}{2}} = (\chi^2(\mathbb{P}_S \| \mathbb{Q}) + 1)^{1/2} (\chi^2(\mathbb{P}_T \| \mathbb{Q}) + 1)^{1/2},$$

and

$$\chi^2(\mathbb{P}_S \| \mathbb{Q}) + 1 = \frac{\det(2\Sigma_S^{-1} - I_d)}{\det(\Sigma_S)} < \infty$$

holds as long as  $\Sigma_S \succ 0$  and  $2\Sigma_S^{-1} - I_d \succ 0$ . This is ensured by our assumption  $0 \prec \Sigma_S \prec 2I_d$ , so the above expression is finite. Now note that

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}} \left[ \frac{d\mathbb{P}_S}{d\mathbb{Q}} \frac{d\mathbb{P}_T}{d\mathbb{Q}} \right] &= \frac{1}{\sqrt{\det(\Sigma_S) \det(\Sigma_T)}} \mathbb{E}_{X \sim \mathbb{Q}} \left[ \exp \left( -\frac{1}{2} X^\top (\Sigma_S^{-1} + \Sigma_T^{-1} - 2I_d) X \right) \right] \\
&= \frac{1}{\sqrt{\det(\Sigma_S) \det(\Sigma_T)}} \mathbb{E}_{Z \sim \mathbb{Q}} \left[ \exp \left( Z^\top A Z \right) \right]
\end{aligned}$$

where  $A := I_d - (\Sigma_S^{-1} + \Sigma_T^{-1})/2$ . Let  $\lambda_1, \dots, \lambda_d$  be the eigenvalues of  $A$ . Since  $0 \prec \Sigma_S, \Sigma_T \prec 2I_d$ , we have  $\lambda_k < 1/2$  for all  $k \in [d]$ . Therefore, by diagonalization,

$$\mathbb{E}_{Z \sim \mathbb{Q}} \left[ \exp \left( Z^\top A Z \right) \right] = \prod_{k=1}^d \mathbb{E}_{Z_k \sim \mathcal{N}(0,1)} \left[ \exp \left( \lambda_k Z_k^2 \right) \right] = \prod_{k=1}^d \frac{1}{(1 - 2\lambda_k)^{1/2}}.$$

Hence, we can write

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}} \left[ \frac{d\mathbb{P}_S}{d\mathbb{Q}} \frac{d\mathbb{P}_T}{d\mathbb{Q}} \right] &= \frac{\det(I_d - 2A)^{-\frac{1}{2}}}{\sqrt{\det(\Sigma_S) \det(\Sigma_T)}} = \frac{\det(\Sigma_S^{-1} + \Sigma_T^{-1} - I_d)^{-\frac{1}{2}}}{\sqrt{\det(\Sigma_S) \det(\Sigma_T)}} \\ &= \left( \det(\Sigma_S) \det(\Sigma_S^{-1} + \Sigma_T^{-1} - I_d) \det(\Sigma_T) \right)^{-1/2} \\ &= \det(\Sigma_S + \Sigma_T - \Sigma_S \Sigma_T)^{-1/2} \\ &= \det(I_d - (\Sigma_S - I_d)(\Sigma_T - I_d))^{-1/2}. \end{aligned}$$

Finally, this proves that

$$\chi^2(\mathbb{E}_S[\mathcal{N}(0, \Sigma_S)^{\otimes p}] \| \mathcal{N}(0, I_d)^{\otimes p}) + 1 = \mathbb{E}_{S,T} \left[ \det \left( I_d - (\Sigma_S - I_d)(\Sigma_T - I_d) \right)^{-\frac{p}{2}} \right],$$

which is the desired result.  $\square$

## D Deferred proofs in Section 5

### D.1 Proof of Theorem 5

We establish (29) and (30) separately.

#### D.1.1 Proof of (29)

For the upper bound, we first note that

$$\begin{aligned} \mathbb{E} \left[ \left\| \widehat{\Sigma}_X - \Sigma \right\|_F^2 \right] &= \sum_{i,j=1}^d \mathbb{E}[(\widehat{\Sigma}_X)_{i,j} - \Sigma_{i,j}]^2 \\ &= \frac{1}{m} \sum_{i,j=1}^d \mathbb{E}[(X_{1,i} X_{1,j} - \Sigma_{i,j})^2] \\ &\stackrel{(a)}{=} \frac{1}{m} \sum_{i,j=1}^d (\mathbb{E}[X_{1,i}^2 X_{1,j}^2] - \Sigma_{i,j}^2) \\ &\stackrel{(b)}{=} \frac{1}{m} \sum_{i,j=1}^d (\mathbb{E}[X_{1,i}^2] \mathbb{E}[X_{1,j}^2] + 2\mathbb{E}[X_{1,i} X_{1,j}]^2 - \Sigma_{i,j}^2) \\ &= \frac{1}{m} \sum_{i,j=1}^d (\Sigma_{i,i} \Sigma_{j,j} + \Sigma_{i,j}^2) \\ &\stackrel{(c)}{\leq} \frac{2}{m} \sum_{i,j=1}^d \Sigma_{i,i} \Sigma_{j,j} \\ &= \frac{2\text{Tr}(\Sigma)^2}{m} \leq \frac{2R^2}{m}. \end{aligned}$$

Here (a) is due to the unbiasedness  $\mathbb{E}[X_{1,i}X_{1,j}] = \Sigma_{i,j}$ , (b) follows from Isserlis' theorem [Iss18], and (c) uses  $\Sigma_{i,j}^2 \leq \Sigma_{i,i}\Sigma_{j,j}$  for any PSD matrix  $\Sigma$ . Similarly, we have

$$\mathbb{E} \left[ \left\| \widehat{\Sigma}_Y - \Sigma^{\pi^*} \right\|_F^2 \right] \leq \frac{2R^2}{n}.$$

For the estimator  $\widehat{\pi}$ , we simply choose  $\widehat{\pi} = \widehat{\pi}^{\text{GW}}$  to be the GW estimator defined in (10). Then

$$\begin{aligned} \left\| \Sigma^{\widehat{\pi}} - \Sigma^{\pi^*} \right\|_F &\leq \left\| \Sigma^{\widehat{\pi}} - \widehat{\Sigma}_X^{\widehat{\pi}} \right\|_F + \left\| \widehat{\Sigma}_X^{\widehat{\pi}} - \widehat{\Sigma}_Y \right\|_F + \left\| \widehat{\Sigma}_Y - \Sigma^{\pi^*} \right\|_F \\ &\stackrel{(d)}{\leq} \left\| \Sigma^{\widehat{\pi}} - \widehat{\Sigma}_X^{\widehat{\pi}} \right\|_F + \left\| \widehat{\Sigma}_X^{\pi^*} - \widehat{\Sigma}_Y \right\|_F + \left\| \widehat{\Sigma}_Y - \Sigma^{\pi^*} \right\|_F \\ &\leq \left\| \Sigma^{\widehat{\pi}} - \widehat{\Sigma}_X^{\widehat{\pi}} \right\|_F + \left\| \widehat{\Sigma}_X^{\pi^*} - \Sigma^{\pi^*} \right\|_F + 2 \left\| \widehat{\Sigma}_Y - \Sigma^{\pi^*} \right\|_F \\ &= 2 \left\| \widehat{\Sigma}_X - \Sigma \right\|_F + 2 \left\| \widehat{\Sigma}_Y - \Sigma^{\pi^*} \right\|_F, \end{aligned}$$

where (d) follows from the defining property of the GW estimator in (10). Combining the above displays shows the upper bound

$$\mathbb{E} \left[ \left\| \Sigma^{\widehat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \right] = O \left( \frac{R^2}{n} \right).$$

For the lower bound in (29), by restricting the support of  $\Sigma$  within its upper left  $2 \times 2$  corner, it is clear that the minimax risk is no smaller than its counterpart with  $d = 2$ . In the sequel we assume that  $d = 2$  and lower bound the minimax risk as

$$\begin{aligned} \inf_{\widehat{\pi}} \sup_{\pi^* \in S_d, \Sigma \in \Sigma(R)} \mathbb{E}_{(\pi^*, \Sigma)} \left\| \Sigma^{\widehat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 &\geq \inf_{\widehat{\pi}} \sup_{(\pi^*, \Sigma): \|\Sigma\|_{\text{op}} \leq R/2} \mathbb{E}_{(\pi^*, \Sigma)} \left\| \Sigma^{\widehat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \\ &\stackrel{(e)}{=} \frac{R^2}{4} \inf_{\widehat{\pi}} \sup_{(\pi^*, \Sigma): \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{(\pi^*, \Sigma)} \left\| \Sigma^{\widehat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \\ &\stackrel{(f)}{=} \Omega \left( \frac{R^2}{n} \right), \end{aligned}$$

where (e) follows from simple scaling, and (f) is the standard  $\Omega(1/n)$  parametric rate (alternatively, this also follows from the  $\Omega(d \log d/n)$  lower bound established in Section 4 applied to  $d = 2$ ). This completes the proof of (29).

### D.1.2 Proof of (30)

For the upper bound, note that the analysis of  $\widehat{\pi} = \widehat{\pi}^{\text{QMLE}}$  in Section 3.1 holds for any  $\Sigma \succeq 0$  until (17). Assuming  $\pi^* = \text{Id}$ , integrating the tail over  $\delta \in (0, 1)$  in (17) gives that

$$\mathbb{E}[\langle \Sigma^{\widehat{\pi}^{-1}} - \Sigma, \Sigma^{-1} \rangle] = O \left( \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}} \right).$$

By the second inequality in Corollary 1, we have

$$\left\| \Sigma^{1/2} (\Sigma^{-1})^{\widehat{\pi}} \Sigma^{1/2} - I_d \right\|_F^2 \leq C \left( \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}} \right), \quad (39)$$

with  $C > 0$  being an absolute constant independent of  $(n, m, d)$ . Since  $n \geq c_0 d \log d$  and  $mn \geq c_0 d^3 \log d$ , for a large enough numerical constant  $c_0 > 0$  depending only on  $C$ , the RHS of (39) will be smaller than  $1/2$ . By Lemma 12,

$$\begin{aligned} \min \left\{ 1, \left\| \Sigma^{-1/2} \Sigma^{\hat{\pi}} \Sigma^{-1/2} - I_d \right\|_F^2 \right\} &\leq 2 \min \left\{ 1, \left\| \Sigma^{1/2} (\Sigma^{-1})^{\hat{\pi}} \Sigma^{1/2} - I_d \right\|_F^2 \right\} \\ &\leq 2C \left( \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}} \right), \end{aligned}$$

where the final quantity is smaller than 1. This gives

$$\left\| \Sigma^{-1/2} (\Sigma^{\hat{\pi}} - \Sigma) \Sigma^{-1/2} \right\|_F^2 = O \left( \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}} \right),$$

which is the desired upper bound.

The minimax lower bound is an easy consequence of Theorem 4:

$$\begin{aligned} \inf_{\hat{\pi}} \sup_{\pi^* \in S_d, \Sigma \in \Sigma(R)} \mathbb{E}_{(\pi^*, \Sigma)} [d_{\text{NF}, \Sigma}(\pi^*, \hat{\pi})^2] &\geq \inf_{\hat{\pi}} \sup_{(\pi^*, \Sigma): \|\Sigma\|_{\text{op}} \leq R/d} \mathbb{E}_{(\pi^*, \Sigma)} [d_{\text{NF}, \Sigma}(\pi^*, \hat{\pi})^2] \\ &\stackrel{(a)}{=} \inf_{\hat{\pi}} \sup_{(\pi^*, \Sigma): \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{(\pi^*, \Sigma)} [d_{\text{NF}, \Sigma}(\pi^*, \hat{\pi})^2] \\ &\stackrel{(b)}{\geq} \inf_{\hat{\pi}} \sup_{(\pi^*, \Sigma): \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{(\pi^*, \Sigma)} \left[ \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F^2 \right] \\ &\stackrel{(c)}{=} \Omega \left( \frac{d \log d}{n} + \sqrt{\frac{d^3 \log d}{mn}} \right), \end{aligned}$$

where (a) is the scale-invariance of the loss  $d_{\text{NF}, \Sigma}$ , i.e.  $d_{\text{NF}, \Sigma}(\pi^*, \hat{\pi}) = d_{\text{NF}, \lambda \Sigma}(\pi^*, \hat{\pi})$  for every  $\lambda > 0$ ; (b) uses that  $\left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_F \leq d_{\text{NF}, \lambda \Sigma}(\pi^*, \hat{\pi})$  if  $\|\Sigma\|_{\text{op}} \leq 1$ ; and (c) follows from Theorem 4. The proof is complete.

## D.2 Proof of Theorem 6

### D.2.1 Proof of upper bound

Similar to the proof of the GW upper bound, we begin with the definitions of several good events. Let  $\mathcal{G} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$  be the good event, where

$$\begin{aligned} \mathcal{E}_1 := \left\{ \langle \hat{\Sigma}_Y, \hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}} \rangle \geq \langle \Sigma, \hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}} \rangle \right. \\ \left. - c \sqrt{\frac{d \log d + \log(1/\delta)}{n}} \left\| \Sigma^{1/2} (\hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}}) \Sigma^{1/2} \right\|_F \right\}, \end{aligned} \quad (40)$$

$$\mathcal{E}_2 := \left\{ \langle \Sigma, \hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}} \rangle \geq \langle \Sigma, \Sigma - \Sigma^{\hat{\pi}} \rangle - \frac{d}{\sqrt{m}} \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_F \right\}, \quad (41)$$

$$\mathcal{E}_3 := \left\{ \left\| \hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}} \right\|_F \leq \left\| \Sigma - \Sigma^{\hat{\pi}} \right\|_F + \frac{2d}{\sqrt{m}} \right\}, \quad (42)$$

with a large enough numerical constant  $c > 0$ . We show that  $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$  under the oracle model. Using Hanson–Wright and triangle inequalities, the statements for  $\mathcal{E}_1$  and  $\mathcal{E}_3$  follow from the same arguments in Lemma 4 for their counterparts in (18) and (20). As for  $\mathcal{E}_2$ , we have

$$\begin{aligned} \langle \Sigma, \hat{\Sigma}_m - \hat{\Sigma}_m^{\hat{\pi}} \rangle &= \langle \Sigma - \Sigma^{\hat{\pi}^{-1}}, \hat{\Sigma}_m \rangle \\ &\stackrel{(a)}{\geq} \langle \Sigma - \Sigma^{\hat{\pi}^{-1}}, \Sigma \rangle - \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_{\text{F}} \left\| \hat{\Sigma}_m - \Sigma \right\|_{\text{F}} \\ &\stackrel{(b)}{\geq} \langle \Sigma - \Sigma^{\hat{\pi}^{-1}}, \Sigma \rangle - \frac{d}{\sqrt{m}} \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_{\text{F}} \\ &= \langle \Sigma, \Sigma - \Sigma^{\hat{\pi}} \rangle - \frac{d}{\sqrt{m}} \left\| \Sigma - \Sigma^{\hat{\pi}^{-1}} \right\|_{\text{F}}, \end{aligned}$$

where (a) is the triangle inequality  $\langle A, B \rangle \leq \|A\|_{\text{F}} \|B\|_{\text{F}}$ , and (b) follows from the oracle property of  $\hat{\Sigma}_m$ . This shows that  $\mathcal{E}_2$  holds almost surely.

Now using exactly the same arguments in Section 3.2.2, with (40)–(42) in place of (18)–(20), for the GW estimator  $\hat{\pi}$  we have

$$0 \geq \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 - u \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}} - v$$

under  $\mathcal{G}$ , where

$$u = 2c \sqrt{\frac{d \log d + \log(1/\delta)}{n}} + \frac{2d}{\sqrt{m}}, \quad v = 2c \sqrt{\frac{d^2(d \log d + \log(1/\delta))}{mn}}.$$

By Lemma 13, this implies

$$\left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 = O(u^2 + v) = O\left(\frac{d \log d + \log(1/\delta)}{n} + \frac{d^2}{m} + \sqrt{\frac{d^2(d \log d + \log(1/\delta))}{mn}}\right)$$

holds with probability at least  $1 - \delta$ . Integrating the tail over  $\delta \in (0, 1)$  gives the expectation upper bound

$$\mathbb{E} \left\| \Sigma^{\hat{\pi}} - \Sigma \right\|_{\text{F}}^2 = O\left(\frac{d \log d}{n} + \frac{d^2}{m} + \sqrt{\frac{d^3 \log d}{mn}}\right) = O\left(\frac{d \log d}{n} + \frac{d^2}{m}\right),$$

where the last inequality is due to AM-GM. This proves the claimed upper bound in Theorem 6.

### D.2.2 Proof of lower bound

We use the same lower bound program in Section 4, with an additional specification for the oracle  $\hat{\Sigma}_m$ . Recall that Section 4 shows the minimax lower bound

$$\inf_{\hat{\pi}} \sup_{\pi^*, \Sigma: \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{\Sigma} \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_{\text{F}}^2 \gtrsim \frac{d \log d}{n}$$

even if  $m = \infty$  (i.e. known covariance  $\Sigma$ ), the same lower bound also holds under the oracle setting by taking  $\hat{\Sigma}_m = \Sigma$ . Therefore, it remains to prove that if  $m \geq d$  under the oracle setting, we have

$$\inf_{\hat{\pi}} \sup_{\pi^*, \Sigma: \|\Sigma\|_{\text{op}} \leq 1} \mathbb{E}_{\Sigma} \left\| \Sigma^{\hat{\pi}} - \Sigma^{\pi^*} \right\|_{\text{F}}^2 \gtrsim \frac{d^2}{m}.$$

We choose the following joint distribution of  $(\pi^*, \Sigma)$ : set  $\pi^* \sim \text{Unif}(S_d)$ , and conditioned on  $\pi^*$ , set  $\Sigma = (I_d + \eta S^{(\pi^*)^{-1}})/2$  for  $S \sim \text{Unif}(\mathcal{S}_0)$ , where the set of matrices  $\mathcal{S}_0$  is constructed in Lemma 5. We choose (recall that  $m \geq d$ )

$$\eta = \frac{1}{2c_1\sqrt{m}} \leq \frac{1}{2c_1\sqrt{d}}, \quad (43)$$

and note that the loss function in (23) becomes

$$\begin{aligned} L(\pi^*, \hat{\pi}) &= \min_{\Sigma = (I_d + \eta S^{(\pi^*)^{-1}})/2 : S \in \mathcal{S}_0} \left\| \Sigma^{\pi^*} - \Sigma^{\hat{\pi}} \right\|_F^2 \\ &= \frac{\eta^2}{4} \min_{S \in \mathcal{S}_0} \left\| S - S^{(\pi^*)^{-1} \circ \hat{\pi}} \right\|_F^2 = \frac{\eta^2}{4} \min_{S \in \mathcal{S}_0} \left\| S^{(\pi^*)^{-1}} - S^{\hat{\pi}^{-1}} \right\|_F^2. \end{aligned}$$

As  $\sum_{i=1}^d \mathbb{1}(\pi_1^{-1}(i) \neq \pi_2^{-1}(i)) = \sum_{i=1}^d \mathbb{1}(\pi_2 \circ \pi_1^{-1}(i) \neq i) = \sum_{j=1}^d \mathbb{1}(\pi_2(j) \neq \pi_1(j))$  for every  $\pi_1, \pi_2 \in S_d$ , the same argument in Lemma 7 gives  $p_\Delta \leq d^{-c_3 d}$  for the quantity  $p_\Delta$  in Lemma 6. Therefore,

$$\mathbb{E} \left[ \left\| \Sigma^{\pi^*} - \Sigma^{\hat{\pi}} \right\|_F^2 \right] \geq \frac{c_2 \eta^2 d^2}{4} \left( 1 - \frac{I(\pi^*; \hat{\Sigma}_m, Y^n) + \log 2}{c_3 d \log d} \right). \quad (44)$$

Next we specify the choice of  $\hat{\Sigma}_m$  and upper bound the mutual information. The choice of the oracle is simple: we take  $\hat{\Sigma}_m \equiv I_d/2$ . Note that

$$\left\| \hat{\Sigma}_m - \Sigma \right\|_{\text{op}} = \frac{\eta}{2} \|S\|_{\text{op}} \stackrel{(c)}{\leq} \frac{\eta}{2} \cdot c_1 \sqrt{d} \stackrel{(43)}{\leq} \sqrt{\frac{d}{m}},$$

where (c) follows from the second property of  $\mathcal{S}_0$ . This shows that the oracle property is satisfied.

Finally we show that  $I(\pi^*; \hat{\Sigma}_m, Y^n) = 0$ , therefore the lower bound of Theorem 6 follows from (43) and (44). To see it, note that  $\hat{\Sigma}_m$  is deterministic, and  $\Sigma^{\pi^*} = (I_d + \eta S)/2$  with  $S \sim \text{Unif}(\mathcal{S}_0)$ , for any realization of  $\pi^*$ . Since  $Y^n \sim \mathcal{N}(0, \Sigma^{\pi^*})^{\otimes n}$ , this shows that  $Y^n$  and  $\pi^*$  are independent. It is consequently clear that  $I(\pi^*; \hat{\Sigma}_m, Y^n) = 0$ .

## References

- [ABH98] Jonathan E Atkins, Erik G Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.
- [AGCJ18] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124, 2018.
- [AMJ18] David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [AMJJ19] David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.



- [AWR17] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.
- [BEGN10] Thierry Benoist, Bertrand Estellon, Frederic Gardi, and Karim Nouioua. Toward local search programming: Localsolver 1.0. In *CPAIOR workshop: open source tools for constraint programming and mathematical programming*, volume 49, 2010.
- [BKB<sup>+</sup>93] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya’acov Ritov, J Klaassen, Jon A Wellner, and YA’Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.
- [BKW23] Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023.
- [BSKR<sup>+</sup>23] Marie Breeur, George Stepaniants, Pekka Keski-Rahkonen, Philippe Rigollet, and Vivian Viallon. Optimal transport for automatic alignment of untargeted metabolomic data. *arXiv preprint arXiv:2306.03218*, 2023.
- [BTVG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006.
- [CCD<sup>+</sup>17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.
- [CD16] Olivier Collier and Arnak S Dalalyan. Minimax rates in permutation estimation for feature matching. *The Journal of Machine Learning Research*, 17(1):162–192, 2016.
- [CGZ16] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.
- [CPKL<sup>+</sup>22] Rui Climaco Pinto, Ibrahim Karaman, Matthew R Lewis, Jenny Hällqvist, Manuja Kaluarachchi, Gonçalo Graça, Elena Chekmeneva, Brenan Durainayagam, Mohsen Ghanbari, M Arfan Ikram, et al. Finding correspondence between metabolomic features in untargeted liquid chromatography–mass spectrometry metabolomics datasets. *Analytical Chemistry*, 94(14):5493–5503, 2022.
- [Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.
- [DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 6, 2018.
- [DW13] John C Duchi and Martin J Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

- [DWXY23] Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. The planted matching problem: Sharp threshold and infinite-order phase transition. *Probability Theory and Related Fields*, pages 1–71, 2023.
- [FCG<sup>+</sup>21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boissunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.
- [FMR19] Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1):623 – 653, 2019.
- [FPPA14] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.
- [FRW15] Jianqing Fan, Philippe Rigollet, and Weichen Wang. Estimation of functionals of sparse covariance matrices. *Annals of statistics*, 43(6):2706, 2015.
- [FS23] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.
- [GJB19] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890. PMLR, 16–18 Apr 2019.
- [GKPS99] Mikhael Gromov, Misha Katz, Pierre Pansu, and Stephen Semmes. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Springer, 1999.
- [GMD22] Tigran Galstyan, Arshak Minasyan, and Arnak S Dalalyan. Optimal detection of the feature matching map in presence of noise and outliers. *Electronic Journal of Statistics*, 16(2):5720–5750, 2022.
- [HKB<sup>+</sup>21] Hani Habra, Maureen Kachman, Kevin Bullock, Clary Clish, Charles R Evans, and Alla Karnovsky. metabcombiner: paired untargeted lc-hrms metabolomics feature matching and concatenation of disparately acquired data sets. *Analytical chemistry*, 93(12):5028–5036, 2021.
- [HM23] Georgina Hall and Laurent Massoulié. Partial recovery in the graph alignment problem. *Operations Research*, 71(1):259–272, 2023.
- [HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.
- [Iss18] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

- [KB57] Tjalling C Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76, 1957.
- [LdABN<sup>+</sup>07] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European Journal of Operational Research*, 176(2):657 – 690, 2007.
- [LFF<sup>+</sup>15] Vince Lyzinski, Donniell E Fishkind, Marcelo Fiori, Joshua T Vogelstein, Carey E Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):60–73, 2015.
- [LHM13] Hao Lin, Lin He, and Bin Ma. A combinatorial approach to the peptide feature matching problem for label-free quantification. *Bioinformatics*, 29(14):1768–1775, 2013.
- [Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.
- [LSV<sup>+</sup>21] Erikka Loftfield, Magdalena Stepień, Vivian Viallon, Laura Trijsburg, Joseph A Rothwell, Nivonirina Robinot, Carine Biessy, Ingvar A Bergdahl, Stina Bodén, Matthias B Schulze, et al. Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality. *JNCI: Journal of the National Cancer Institute*, 113(11):1542–1550, 2021.
- [LTK<sup>+</sup>22] Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. Fast and provably convergent algorithms for gromov-wasserstein in graph learning. *arXiv preprint arXiv:2205.08115*, 2022.
- [LTK<sup>+</sup>23] Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. A convergent single-loop algorithm for relaxation of gromov-wasserstein in graph data. *arXiv preprint arXiv:2303.06595*, 2023.
- [Mém11] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.
- [MRT23] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Exact matching of random graphs with constant correlation. *Probability Theory and Related Fields*, 186(1-2):327–389, 2023.
- [MSZ18] Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pages 3375–3383. PMLR, 2018.
- [Mun57] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.
- [MWR18] Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR, 2018.

- [MWXY23] Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. Random graph matching at otter’s threshold via counting chandeliers. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1345–1356, 2023.
- [Ney59] Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and statistics*, pages 213–234, 1959.
- [PC<sup>+</sup>17] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.
- [PML<sup>+</sup>16] Leila Pirhaji, Pamela Milani, Mathias Leidl, Timothy Curran, Julian Avila-Pacheco, Clary B Clish, Forest M White, Alan Saghatelian, and Ernest Fraenkel. Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature methods*, 13(9):770–776, 2016.
- [PS22] Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation and adaptation. *The Annals of Statistics*, 50(1):324–350, 2022.
- [PWC17] Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017.
- [RH15] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.
- [RV13] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic communications in probability*, 18:1–9, 2013.
- [SG76] Sartaj Sahni and Teofilo Gonzalez. P-complete approximation problems. *Journal of the ACM (JACM)*, 23(3):555–565, 1976.
- [SPKS16] Justin Solomon, Gabriel Peyré, Vladimir G. Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13, 2016.
- [SVP21] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gro-mov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.
- [Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [Wis19] David S Wishart. Metabolomics for investigating physiological and pathophysiological processes. *Physiological reviews*, 99(4):1819–1875, 2019.
- [Wri73] Farrol Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.
- [WXS22] Yihong Wu, Jiaming Xu, and H Yu Sophie. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory*, 68(8):5391–5417, 2022.

- [ZGMS22] Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K Sriperumbudur. Gromov-wasserstein distances: Entropic regularization, duality, and sample complexity. *arXiv preprint arXiv:2212.12848*, 2022.