Submitted to the Annals of Statistics

NEAR OPTIMAL SAMPLE COMPLEXITY FOR MATRIX AND TENSOR NORMAL MODELS VIA GEODESIC CONVEXITY

BY COLE FRANKS[1], RAFAEL OLIVEIRA[2], AKSHAY RAMACHANDRAN[2] AND MICHAEL WALTER[3]

[1]*Massachusetts Institute of Technology*
[2]*University of Waterloo*
[3]*University of Amsterdam*

The matrix normal model, the family of Gaussian matrix-variate distributions whose covariance matrix is the Kronecker product of two lower dimensional factors, is frequently used to model matrix-variate data. The tensor normal model generalizes this family to Kronecker products of three or more factors. We study the estimation of the Kronecker factors of the covariance matrix in the matrix and tensor models. We show nonasymptotic bounds for the error achieved by the maximum likelihood estimator (MLE) in several natural metrics. In contrast to existing bounds, our results do not rely on the factors being well-conditioned or sparse. For the matrix normal model, all our bounds are minimax optimal up to logarithmic factors, and for the tensor normal model our bound for the largest factor and overall covariance matrix are minimax optimal provided there are enough samples for any estimator to obtain better than constant Frobenius error. In the same regimes as our sample complexity bounds, we show that an iterative procedure to compute the MLE known as the flip-flop algorithm converges linearly with high probability. Our main tool is geodesic convexity in the geometry on positive-definite matrices induced by the Fisher information metric. We also provide numerical evidence that combining the flip-flop algorithm with a simple shrinkage estimator can improve performance in the undersampled regime.

CONTENTS

MSC2020 subject classifications: Primary 62F12; secondary 62F30.
Keywords and phrases: Covariance estimation, matrix normal model, maximum likelihood estimation, geodesic convexity, operator scaling.

1. Introduction. Covariance matrix estimation is an important task in statistics, machine learning, and the empirical sciences. We consider covariance estimation for matrix-variate and tensor-variate Gaussian data, that is, when individual data points are matrices or tensors. Matrix-variate data arises naturally in numerous applications like gene microarrays, spatio-temporal data, and brain imaging. A significant challenge is that the dimensionality of these problems is frequently much higher than the number of samples, making estimation information-theoretically impossible without structural assumptions.

To remedy this issue, matrix-variate data is commonly assumed to follow the *matrix normal distribution* (Dutilleul, 1999; Werner, Jansson and Stoica, 2008). Here the matrix follows a multivariate Gaussian distribution and the covariance between any two entries in the matrix is a product of an inter-row factor and an inter-column factor. In spatio-temporal statistics this is referred to as a separable covariance structure. Formally, if a matrix normal random variable X takes values in the $d_1 \times d_2$ matrices, then its covariance matrix Σ is a $d_1 d_2 \times d_1 d_2$ matrix that is the Kronecker product $\Sigma_1 \otimes \Sigma_2$ of two positive-semidefinite matrices Σ_1 and Σ_2 of dimension $d_1 \times d_1$ and $d_2 \times d_2$, respectively. This naturally extends to the *tensor normal model*, where X is a k-dimensional array, with covariance matrix equal to the Kronecker product of k many positive semidefinite matrices $\Sigma_1, \ldots, \Sigma_k$. In this paper we consider the estimation of $\Sigma_1, \ldots, \Sigma_k$ from n samples of a matrix or tensor normal random variable X.'

A great deal of research has been devoted to estimating the covariance matrix for the matrix and tensor normal models, but gaps in rigorous understanding remain. In unstructured covariance matrix estimation, i.e. $k = 1$, it is well-known that the maximum likelihood estimator exists whenever $d \geq n$ and achieves mean-squared Frobenius norm error $O(d^2/n)$ and mean-squared operator norm error $O(d/n)$, which are both minimax optimal. This fact is the starting point for a vibrant area of research attempting to estimate the covariance or precision matrix with fewer samples under structural assumptions. Particularly important is the study of graphical models, which seeks to better estimate the precision matrix under the assumption that it is *sparse* (has few nonzero entries).

For the matrix and tensor normal models, much of the work (apart from an initial flurry of work on the asymptotic properties of the maximum likelihood estimator) has approached the

sparse case directly. In contrast to the unstructured problem above, the fundamental problem of determining the optimal rates of estimation *without* sparsity assumptions is still largely open. We study this basic question in order to provide a firmer foundation for the large body of work studying its many variants, including the sparse case. We begin by discussing the related work in detail.

In the asymptotic arena, Dutilleul (1999) and later Werner, Jansson and Stoica (2008) proposed an iterative algorithm, known as the *flip-flop algorithm*, to compute the maximum likelihood estimator (MLE). In the latter work, the authors also showed that the MLE is consistent and asymptotically normal, and showed the same for the estimator obtained by terminating the flip-flop after three steps. For the tensor normal model, a natural generalization of the flip-flop algorithm was proposed to compute the MLE (Mardia and Goodall, 1993; Manceur and Dutilleul, 2013), but its convergence was not proven. Here we will be interested in non-asymptotic rates.

For the matrix normal model, treating the covariance matrix Σ as unstructured and estimating it by the sample covariance matrix (the MLE in the unstructured case) yields a mean-squared Frobenius norm error of $O(d_1^2 d_2^2/n)$ assuming $n \geq C d_1 d_2$ for a large enough constant C. The matrix normal model has only $\Theta(d_1^2 + d_2^2)$ parameters, so it should be possible to do much better. The state of the art towards optimal rates for the matrix normal model without sparsity assumptions is the work of Tsiligkaridis, Hero and Zhou (2013), which showed that a three-step flip-flop estimator has mean-squared Frobenius error of $O((d_1^2 + d_2^2)/n)$ for the full matrix Σ. However, their result requires the individual factors have constant condition number and that n is at least $\tilde{\Omega}(\max\{d_1, d_2\})$. They also did not state a bound for the individual factors Σ_1, Σ_2, and did not state bounds for estimation in the operator norm. For the tensor normal model, Sun et al. (2015) present an estimator with tight rates assuming bounded constant condition number of the true covariances and foreknowledge of initializers within constant Frobenius distance of the true precision matrices. In both the matrix and tensor case, no estimator for the Kronecker factors has been proven to have tight rates without additional assumptions on the factors' structure.

Regarding the sparse case, simply setting $\Sigma_2 = I_{d_2}$ or $\Sigma_1 = I_{d_1}$, in which case the matrix normal model reduces to standard covariance estimation with $d_1 n$ (resp. $d_2 n$) samples, shows the necessity of additional assumptions like sparsity or well-conditionedness if $n < \max\{d_1/d_2, d_2/d_1\}$. Tsiligkaridis, Hero and Zhou (2013) also propose a penalized estimator which obtains tighter rates that hold even for $n \ll d_i$ under the additional assumption that the precision matrices Σ_i^{-1} are sparse. In the extremely undersampled regime, Zhou (2014) demonstrated a single-step penalized estimator that converges even for a single matrix $(n = 1)$ when the precision matrices have constant condition number, are highly sparse, and have bounded ℓ_1 norm off the diagonal. Allen and Tibshirani (2010) also considered penalized estimators for the purpose of missing data imputation.

Even characterizing the existence of the MLE for the matrix and tensor normal model has remained elusive until recently, in contrast to the unstructured case $(k = 1)$. Améndola et al. (2020) recently noted that the matrix normal and tensor MLEs are equivalent to algebraic problems about a group action called the *left-right action* and the *tensor action*, respectively. In the computer science literature these two problems are called *operator* and *tensor scaling*, respectively. Independently from Améndola et al. (2020), it was pointed out by Franks and Moitra (2020) that the Tyler's M estimator for elliptical distributions (which is the MLE for the matrix normal model under the additional promise that Σ_2 is diagonal) is a special case of operator scaling. Using the connection to group actions, exact sample size thresholds for the existence of the MLE were recently determined in Derksen and Makam (2020) for the matrix normal model and subsequently for the tensor normal model in Derksen, Makam and Walter (2020). In the context of operator scaling, Gurvits (2004) showed much earlier that the flip-flop

algorithm converges to the matrix normal MLE whenever it exists. Recently it was shown that the number of flip-flop steps to obtain a gradient of magnitude ε in the log-likelihood function for the tensor and matrix normal model is polynomial in the input size and $1/\varepsilon$ (Garg et al., 2019; Bürgisser et al., 2018, 2019).

1.1. *Our contributions.* We take a geodesically convex optimization approach to provide rigorous nonasymptotic guarantees for the estimation of the precision matrices, without any assumptions on their structure. For the matrix normal model we provide high probability bounds on the estimator that are tight up to logarithmic factors. For the tensor normal model, our bounds are tight up to factors of k whenever it is information-theoretically possible to recover the factors to constant Frobenius error.

In the current literature on matrix normal and tensor models, typically the estimators are assessed using Frobenius or spectral norm between the estimated parameter and the truth. However, neither of these metrics bound statistical dissimilarity measures of interest such as the total variation or Kullback-Leibler divergence between the true distribution and that corresponding to the estimated parameter, or the Fisher-Rao distance. Such statistical measures enjoy an invariance property for multivariate normals - namely, they are preserved under acting on the both random variable by the same invertible linear transformation. Such transformations only change the basis in which the data is represented; ideally the performance of estimators should not depend on this basis and hence should not require the covariance matrix to be well-conditioned.

Here we consider the *relative Frobenius error* $D_F(A\|B) = \|I - B^{-1/2}AB^{-1/2}\|_F$ of the precision matrices. It is natural to consider the relative Frobenius error rather because it is invariant under the linear transformations discussed above, whereas the the Frobenius error is not. Moreover, the dissimilarity measures $D_F(\Theta_1||\Theta_2)$, total variation distance $D_{TV}(\mathcal{N}(0,\Theta_1^{-1}),\mathcal{N}(0,\Theta_2^{-1}))$, the square of the KL-divergence $D_{KL}(\mathcal{N}(0,\Theta_1^{-1}),\mathcal{N}(0,\Theta_2^{-1}))^2$, and the Fisher-Rao distance between Θ_1,Θ_2 all coincide "locally." That is, if any of them is at most a small constant then they are all on the same order. The estimation of precision and covariance matrices under D_{KL} was suggested by in (James and Stein, 1992) due to its natural invariance properties, and has been studied extensively (e.g. Ledoit and Wolf (2012)). To obtain the sharpest possible results, we also consider the *relative spectral error* $D_{\text{op}}(A\|B) = \|I - B^{-1/2}AB^{-1/2}\|_{\text{op}}$, which has been studied in the context of spectral graph sparsification. The dissimilarity measure $D_F(A||B)$ (resp. $D_{op}(A||B)$) can be related to the usual norm $\|A-B\|_F$, (resp. $\|A-B\|_{op}$) by a constant factor assuming the spectral norms of B, B^{-1} are bounded by a constant. Though we caution that D_F and D_{op} are not truly metrics, we will call them distances because they approximately (or "locally") obey symmetry and the triangle inequality. See Appendix E for a discussion of these properties.

Informally, our contributions are as follows:

1. Consider the matrix normal model for $d_1 \le d_2$. We show that for some $n_0 = \widetilde{O}(d_2/d_1)$, if $n \ge n_0$ then the MLE for the precision matrices Θ_1, Θ_2 has error $\widetilde{O}(\sqrt{d_1/nd_2})$ for Θ_1 and $O(\varepsilon\sqrt{d_2/nd_1})$ for Θ_2 in D_{op} with probability $1 - O(e^{-\Omega(d_1)})$. For estimating Θ_1 alone, we obtain the error $\widetilde{O}(\sqrt{d_1/n\min\{d_2, nd_1\}})$ for any n.
2. In the tensor normal model, for k fixed we show that for some $n_0 = O(\max\{d_i^3\}/\prod_i d_i)$, if $n \ge n_0$ then the MLE for the precision matrix Θ has error $O(\frac{d_{\max}}{\sqrt{n}})$ in D_F with probability $1 - (nD/\max\{d_i\})^{-\Omega(\min\{d_i\})}$. We also give bounds for growing k and a tight bound for the error of the largest Kronecker factor Θ_i.
3. Under the same sample requirements as above in each case, the flip-flop algorithm of (Mardia and Goodall, 1993; Manceur and Dutilleul, 2013) converges exponentially quickly to the MLE with high probability. As a consequence, the output of the flip-flop algorithm

with $O\left(d_{\max} + \log n\right)$ iterations is an efficiently computable estimator that enjoys statistical guarantees at least as strong (up to constant factors) as those we show for the MLE.

4. To handle the undersampled case, we introduce a shrinkage-based estimator that is much simpler to compute than the LASSO-type estimators of Tsiligkaridis, Hero and Zhou (2013); Sun et al. (2015); Zhou (2014) and give empirical evidence that it improves on them in a generative model for *dense* precision matrices.

We now discuss the tightness of our results. Our first result is tight up to logarithmic factors in the sense that it is information-theoretically impossible to obtain an error bound that is smaller by a polylogarithmic factor and holds with constant probability. For n a polylogarithmic factor smaller than n_0, it is information-theoretically impossible to obtain any finite bound independent of Θ with high probability. We also show that our results for estimating Θ_1 alone are tight up to logarithmic factors; i.e. that it is impossible to obtain a rate better than $O(\sqrt{d_1/n \min\{nd_1, d_2\}})$. Similarly, for the second result, provided $n \geq n_0$ it is impossible to obtain an error bound that is smaller than ours by a factor tending to infinity that holds with constant probability. For $n \ll n_0$, no constant error bound on the D_F error of the largest Kronecker factor can hold with constant probability. Apart from the lower bound for estimating Θ_1 alone, which to our knowledge is novel, these results follow by reduction to known results on the Frobenius and operator error for covariance estimation; see Section 6.

For interesting cases of the tensor normal model such as $d \times d \times d$ tensors we just require that n is at least a large constant. For the matrix normal model, our first result removes the added constraint $n \geq C \max\{d_1, d_2\}$ in Tsiligkaridis, Hero and Zhou (2013). We leave extending the D_{op} bounds for the matrix normal model to the tensor normal model as an open problem.

We now briefly discuss our estimator for the undersampled case, which is described in detail in Section 7. The MLE is a function of the sample covariance matrix, but in the undersampled case the MLE need not exist. To remedy this, we replace the sample covariance matrix by a shrinkage estimator for it (in particular, by taking a convex combination with the identity matrix) and then compute the MLE for the "shrunk" covariance matrix. Though our estimator is empirically outperformed by Zhou (2014) for sparse precision matrices, it emprically outperforms Zhou (2014) in a natural *dense* generative model of random approximately low-rank Kronecker factors which we refer to as the "spiked" model. Moreover, we found that on average our estimator is faster to compute than the estimator of Zhou (2014) by a factor of 500 for the matrix model with $d_1 = 100, d_2 = 200$. Given this empirical evidence, we view our shrinkage-based estimator as a potentially useful tool for the undersampled tensor normal model which merits further theoretical attention.

1.2. *Outline.* In the next section, Section 2, we precisely describe the model and formally state our results. In Section 3, we prove our main sample complexity bound for the tensor normal model, Theorem 2.1. In Section 4 we prove our improved bound for the matrix normal model, Theorem 2.3. Our results on the flip-flop algorithm for tensor and matrix normal models (Theorems 2.5 and 2.6, respectively) are proven in Section 5. Section 6 contains the proofs of our lower bound for the matrix normal model, and Section 7 contains empirical observations about the performance of our regularized estimator.

1.3. *Notation.* We write $\text{Mat}(d)$ for the space of $d \times d$ matrices with real entries, $\text{PD}(d)$ for the convex cone of $d \times d$ symmetric positive definite matrices; $\text{GL}(d)$ denotes the group of invertible $d \times d$ matrices with real entries. For a matrix A, $\|A\|_{\text{op}}$ denotes the operator norm, $\|A\|_F = (\text{Tr}\, A^T A)^{\frac{1}{2}}$ the Frobenius norm, and $\langle A, B \rangle = \text{Tr}\, A^T B$ for the Hilbert-Schmidt inner product. We extend these definitions to tuples $A = (A_0; A_1, \ldots, A_k)$, where $A_0 \in \mathbb{R}$

and the A_a for $a \in [k]$ are matrices and denote them by the same symbol, i.e., $\|A\|_F = (|A_0|^2 + \sum_{a=1}^k \|A_a\|_F^2)^{1/2}$ and similarly for the inner product. For functions $f, g : S \to \mathbb{R}$ for any set S, we say $f = O(g)$ if there is a constant $C > 0$ independent of x such that $f(x) \leq Cg(x)$ for all $x \in S$, and similarly $f = \Omega(g)$ if there is a constant $c > 0$ independent of x such that $f(x) \geq cg(x)$ for all $x \in S$. If $f = O(g)$ and $g = O(f)$ we write $f \asymp g$.

2. Model and main results. In this section we define the matrix and tensor normal models and we state our main technical results.

2.1. *Matrix and tensor normal model.* The tensor normal model, of which the matrix normal model is a particular case, is formally defined as follows.

DEFINITION. For positive definite matrices $\Sigma_1, \ldots, \Sigma_k$, we define the *tensor normal model* as the centered multivariate Gaussian distribution with covariance matrix given by the Kronecker product $\Sigma = \Sigma_1 \otimes \ldots \otimes \Sigma_k$. For $k = 2$, this is known as the *matrix normal model*.

Note that if each Σ_a is a $d_a \times d_a$ matrix then Σ is a $D \times D$-matrix, where $D = d_1 \cdots d_k$. Our goal is to estimate the k Kronecker factors $\Sigma_1, \ldots, \Sigma_k$ given access to n i.i.d. random samples $x_1, \ldots, x_n \in \mathbb{R}^D$ drawn from the model.

One may also think of each random sample x_j as taking values in the set of $d_1 \times \cdots \times d_k$ arrays of real numbers. There are k natural ways to "flatten" x_j to a matrix: for example, we may think of it as a $d_1 \times d_2 d_3 \cdots d_k$ matrix whose column indexed by $(i_2, \ldots, i_k)$ is the vector in $\mathbb{R}^{d_1}$ with i_1^{th} entry equal to $(x_j)_{i_1,\ldots,i_k}$. In an analogous way we may flatten it to a $d_2 \times d_1 d_3 \cdots d_k$ matrix, and so on. In the tensor normal model, the $d_2 d_3 \cdots d_k$ many columns are each distributed as a Gaussian random vector with covariance proportional to Σ_1. Similarly the columns of the $d_2 \times d_1 d_3 \cdots d_k$ flattening have covariance proportional to Σ_2, and so on. As such, the columns of the a^{th} flattening can be used to estimate Σ_a up to a scalar. However, doing so naïvely (e.g. using the sample covariance matrix of the columns) can result in an estimator with very high variance. This is because the columns of the flattenings are not independent. In fact they may be so highly correlated that they effectively constitute only one random sample rather than $d_2 \ldots d_k$ many. The MLE decorrelates the columns to obtain rates like those one would obtain if the columns were independent.

The MLE is easier to describe in terms of the precision matrices, the inverses of the covariance matrices. Let Θ denote the *precision matrix*, i.e., $\Theta = \bigotimes_{a=1}^k \Theta_a$, where $\Theta_a = \Sigma_a^{-1}$. Let $\mathbb{P}$ denote the manifold of all such Θ, i.e.

$$\mathbb{P} = \{\Theta_1 \otimes \ldots \otimes \Theta_k \in \mathrm{PD}(d_1) \times \cdots \times \mathrm{PD}(d_k)\}.$$

Given a tuple x of samples $x_1, \ldots, x_n \in \mathbb{R}^D$, the following function is proportional to the negative log-likelihood:

$$f_x(\Theta) = \frac{1}{nD} \sum_{i=1}^n x_i^T \Theta x_i - \frac{1}{D} \log\det \Theta.$$

Though Θ_a are not identifiable, the above expression is nonetheless well-defined. The *maximum likelihood estimator (MLE)* for Θ is then

$$\widehat{\Theta} := \operatorname*{arg\,min}_{\Theta \in \mathbb{P}} f_x(\Theta) \tag{2.1}$$

whenever the minimizer exists and is unique. We write $\widehat{\Theta} = \widehat{\Theta}(x)$ when we want to emphasize the dependence of the MLE on the samples x, and we say $(\widehat{\Theta}_1, \ldots, \widehat{\Theta}_k)$ is *an* MLE for $(\Theta_1, \ldots, \Theta_k)$ if $\otimes_{a=1}^k \widehat{\Theta}_a = \widehat{\Theta}$. Note that $\mathbb{P}$ is not a convex domain under the Euclidean geometry on the $D \times D$ matrices. This is in similarity with the fact that the set of rank-1 matrices is not convex in the space of all matrices.

2.2. *Results on the MLE.* We may now state our result for the tensor normal models precisely. As mentioned in the introduction, we use the following natural distance measures.

DEFINITION. For positive definite matrices A, B, define their *relative Frobenius error* (or *Mahalanobis distance*) as

$$D_F(A\|B) = \|I - B^{-1/2}AB^{-1/2}\|_F.$$

Similarly, define the *relative spectral error* as

$$D_{\mathrm{op}}(A\|B) = \|I - B^{-1/2}AB^{-1/2}\|_{\mathrm{op}}.$$

To state our results, and throughout this paper, we write $d_{\min} = \min_a d_a$, $d_{\max} = \max_a d_a$. Recall also that $D = \prod_{i=1}^k d_a$.

THEOREM 2.1 (Tensor normal Frobenius error). *There is a universal constant $C > 0$ such that the following holds. Suppose*

$$\max\{1, \varepsilon^2\} \le \frac{nD}{Ck^2 d_{\max}^2 \max\{k, d_{\max}\}}. \tag{2.2}$$

Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \cdots \otimes \widehat{\Theta}_k$ for n independent samples of the tensor normal model with precision matrix $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k$ satisfies

$$D_F(\widehat{\Theta}_a\|\Theta_a) = O\left(k^{1/2}\frac{\sqrt{d_a}\, d_{\max}}{\sqrt{nD}}\varepsilon\right) \quad \text{for all } a \in [k],$$

$$D_F(\widehat{\Theta}\|\Theta) = O\left(k^{3/2}\frac{d_{\max}}{\sqrt{n}}\varepsilon\right),$$

where the Kronecker factors of $\widehat{\Theta}$ and Θ are chosen such that $\det\widehat{\Theta}_1 = \cdots = \det\widehat{\Theta}_k$ and $\det\Theta_1 = \cdots = \det\Theta_k$, with probability at least

$$1 - ke^{-\Omega\left(\varepsilon^2 d_{\max}\right)} - k^2\left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})},$$

The error for the precision matrix Θ_a with $d_a = d_{\max}$ matches that of the MLE for the precision matrix of a single Gaussian with $D/d_{\max}$ samples, which is the special case when all the other Kronecker factors are the identity.

REMARK 2.2 (Fisher-Rao distance). *Our bounds on D_F in the above theorem follow from a stronger bound on the* Fisher-Rao distance $d(\widehat{\Theta}, \Theta) := \|\log\Theta^{-1/2}\widehat{\Theta}\Theta^{-1/2}\|_F$. *As we will discuss, the Fisher-Rao distance arises from the Fisher information metric for centered Gaussians parameterized by their covariance matrices (Skovgaard, 1984).*[1] *With the same hypotheses and failure probability as Theorem 2.1, we have $d(\widehat{\Theta}, \Theta) = O(\sqrt{k}\, d_{\max}\varepsilon/\sqrt{nD})$.*

For the matrix normal model ($k = 2$), we obtain a stronger result. In the following theorem we identify Θ_1, Θ_2 from Θ using the convention $\det\Theta_1 = 1$, and define the MLE's $\widehat{\Theta}_1, \widehat{\Theta}_2$ to be the minimizers of f restricted to the subset $\{P \in \mathrm{PD}(d_1) : \det P = 1\} \times \mathrm{PD}(d_2)$.

[1] We omit the $1/\sqrt{2}$ factor for notational convenience.

THEOREM 2.3 (Matrix normal spectral error). *There is a universal constant $C > 0$ with the following property. Suppose $d_1 \le d_2$ and $n \ge C\frac{d_2}{d_1}\max\{\log\frac{d_2}{d_1}, \frac{\log^2 d_1}{\varepsilon^2}, \varepsilon^2\}$. Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \widehat{\Theta}_2$ for n independent samples from the matrix normal model with precision matrix $\Theta = \Theta_1 \otimes \Theta_2$ satisfies*

$$D_{\mathrm{op}}(\widehat{\Theta}_1 \| \Theta_1) = O\left(\varepsilon\sqrt{\frac{d_1}{nd_2}}\log d_1\right) \quad \textit{and} \quad D_{\mathrm{op}}(\widehat{\Theta}_2 \| \Theta_2) = O\left(\varepsilon\sqrt{\frac{d_2}{nd_1}}\right),$$

with probability at least $1 - O(e^{-\Omega(d_1\varepsilon^2)})$.

In applications such as brain fMRI, one is interested only in Θ_1, and Θ_2 is treated as a nuisance parameter. Assuming the nuisance parameter Θ_2 is known, we can compute $(I \otimes \Theta_2^{1/2})X$, which is distributed as nd_2 independent samples from a Gaussian with precision matrix Θ_1. In this case, one can estimate Θ_1 in operator norm with an RMSE rate of $O(\sqrt{d_1/nd_2})$ no matter how large d_2 is. One could hope that this rate holds for Θ_1 even when Θ_2 is not known. In Section 6 we show that, to the contrary, the rate for Θ_1 cannot be better than $O(\sqrt{d_1/n\min(nd_1, d_2)})$. Thus, for $d_2 > nd_1$, it is impossible to estimate Θ_1 as well as one could if Θ_2 were known. Note that in this regime there is no hope of recovering Θ_2 even if Θ_1 is known. As the random variable Y_i obtained by ignoring all but $d_2' \approx nd_1$ columns of each X_i is also distributed according to the matrix normal model with covariance matrix $\Sigma_1 \otimes \Sigma_2'$ for $\Sigma_2 \in \mathrm{PD}(d_2')$, the MLE for Y obtains a matching upper bound.

COROLLARY 2.4 (Estimating only Θ_1.). *Let X be distributed according to the matrix normal model with precision matrix $\Theta = \Theta_1 \otimes \Theta_2$ and suppose $d_1 \le d_2$. Let $Y = (Y_1, \ldots, Y_n)$ be the random variable obtained by removing all but $d_2' = \min\{d_2, nd_1/C\max\{\log n, \varepsilon^{-2}\log^2 d_1\}\}$ columns of X_i for each $i \in [n]$. Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \widehat{\Theta}_2$ for Y satisfies*

$$D_{\mathrm{op}}(\widehat{\Theta}_1 \| \Theta_1) = O\left(\varepsilon\sqrt{\frac{d_1}{nd_2'}}\log d_1\right).$$

with probability at least $1 - O(e^{-\Omega(d_1\varepsilon^2)})$. Moreover, this rate is tight up to logarithmic factors in d_1, n, ε.

2.3. *Flip-flop algorithm.* The MLE can be computed by a natural iterative procedure known as the *flip-flop algorithm* (Dutilleul, 1999; Gurvits, 2004).

For simplicity, we describe it for the matrix normal model ($k = 2$), so that the samples x_i can be viewed as $d_1 \times d_2$ matrices which we denote by X_i. Initialize $\overline{\Theta}_1 = I_{d_1}$, $\overline{\Theta}_2 = I_{d_2}$, and choose a distance measure, say D_F in the case below, and a tolerance $\varepsilon > 0$.

1. Set $\overline{\Theta}_1 \leftarrow (\frac{1}{nd_2}\sum_{i=1}^n X_i\overline{\Theta}_2 X_i^T)^{-1}$.
2. Set $\Upsilon = \frac{1}{nd_1}\sum_{i=1}^n X_i^T\overline{\Theta}_1 X_i$. If $D_F(\Upsilon^{-1}||\overline{\Theta}_2) > \varepsilon$, set $\overline{\Theta}_2 \leftarrow \Upsilon^{-1}$ and return to Step 1.
3. Output $\overline{\Theta}_1, \overline{\Theta}_2$.

We can motivate this procedure by noting that if in the first step we already have $\overline{\Theta}_2 = \Theta_2$, then $\frac{1}{nd_2}\sum_{i=1}^n X_i\overline{\Theta}_2 X_i^T$ is simply a sum of outer products of nd_2 many independent random vectors with covariance $\Sigma_1 = \Theta_1^{-1}$; as such the inverse is a good estimator for Θ_1. As we don't know Θ_2, the flip-flop algorithm instead uses $\overline{\Theta}_2$ as our current best guess.

For the general tensor normal model, in each step the flip flop algorithm chooses one of the dimensions $a \in [k]$ and uses the a^{th} flattening of the samples x_i (which are just X_i and X_i^T in the matrix case) to update $\overline{\Theta}_a$.

2.4. *Results on the flip-flop algorithm.* Our next results show that the flip-flop algorithm can efficiently compute the MLE with high probability when the hypotheses of Theorem 2.1 or Theorem 2.3 hold. We first state our result for the general tensor normal model and then give an improved version for the matrix normal model.

THEOREM 2.5 (Tensor normal flip-flop). *If $\widehat{\Theta}$ denotes the MLE estimator for Θ, then provided $n = \Omega(k^2 \cdot d_{\max}^3/D)$, the flip-flop algorithm computes $\overline{\Theta}$ with*

$$D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon$$

in $O(k \log(1/\varepsilon))$ iterations with probability at least

$$1 - k^2 \cdot \left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})} - 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}.$$

THEOREM 2.6 (Matrix normal flip-flop). *Let $(\widehat{\Theta}_1, \widehat{\Theta}_2)$ denote the MLE for (Θ_1, Θ_2). There exists a universal constant $\Gamma > 0$ such that when given*

$$n \geq \Gamma \cdot \frac{d_{\max}}{d_{\min}} \cdot \max\left\{\log\left(\frac{d_{\max}}{d_{\min}}\right), \frac{\log^2 d_{\min}}{\varepsilon^2}\right\}$$

samples in the matrix normal model, the flip-flop algorithm computes $(\overline{\Theta}_1, \overline{\Theta}_2)$ with

$$D_{op}(\overline{\Theta}_a, \widehat{\Theta}_a) \leq \varepsilon$$

for $a \in \{1, 2\}$ in $O\left(d_{\max} + \log(1/\varepsilon)\right)$ iterations with probability at least $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$.

As a corollary of Theorems 2.5 and 2.6, the output of the flip-flop algorithm with $O\left(d_{\max} + \log n\right)$ iterations is an efficiently computable estimator with the same statistical guarantees as shown for the MLE in in Theorems 2.1 and 2.3 for the tensor and matrix normal models, respectively.

One may wonder why in the above theorems we consider the distances for the individual tensor factors and not the covariance matrix itself, but tight bounds for the covariance matrix itself follow from the above bounds (apart from the logarithmic factor in the matrix normal case and a constant factor in general).

3. Sample complexity for the tensor normal model. It was observed by Wiesel (2012) that the negative log-likelihood exhibits a certain variant of convexity known as *geodesic convexity*. In this section, we use geodesic convexity, following a strategy similar to Franks and Moitra (2020), to prove Theorem 2.1. Our improved result for the matrix normal model, Theorem 2.3, requires additional tools and will be proved later in Section 4.

3.1. *Geometry and geodesic convexity.* We now discuss the geodesic convexity used here and outline the strategy for our proof. We start by introducing a Riemannian metric on the manifold $\mathrm{PD}(D)$ of positive-definite $D \times D$ matrices. Rather than simply considering the metric induced by the Euclidean metric on the symmetric matrices, we consider the metric whose geodesics starting at a point $\Theta \in \mathrm{PD}(D)$ are of the form $t \mapsto \Theta^{1/2} e^{Ht} \Theta^{1/2}$ for $t \in \mathbb{R}$ and a symmetric matrix H. This metric arises from the Hessian of the log-determinant (Bhatia, 2009) and also as the Fisher information metric on centered Gaussians parametrized by their covariance matrices (Skovgaard, 1984). If Θ is positive definite and A an invertible matrix then $A\Theta A^T$ is again positive definite. The transformation $\Theta \mapsto A\Theta A^T$ is an isometry, i.e., it

preserves the geodesic distance. Importantly, the statistical distances we use are also *invariant* under such transformations:

$$D_F(A\Theta A^T \| A\Theta' A^T) = D_F(\Theta \| \Theta')$$

and likewise for the distance D_{op}. This invariance is natural because changing a pair of precision matrices in this way does not change the statistical relationship between the corresponding Gaussians; in particular the total variation distance, Fisher-Rao, and Kullback-Leibler divergence are unchanged.

As observed by Wiesel (2012), the negative log-likelihood is convex as the precision matrix moves along the geodesics of the Fisher information metric, and in particular for the tensor normal model it is convex along geodesics in $\mathbb{P} = \{\Theta_1 \otimes \ldots \otimes \Theta_k \in \mathrm{PD}(d_1) \times \cdots \times \mathrm{PD}(d_k)\}$. This is because the geodesics in $\mathrm{PD}(D)$ between elements of the manifold $\mathbb{P} = \{\Theta_1 \otimes \ldots \otimes \Theta_k \in \mathrm{PD}(d_1) \times \cdots \times \mathrm{PD}(d_k)\}$ remain in $\mathbb{P}$. That is, $\mathbb{P}$ is a *totally geodesic submanifold* of $\mathrm{PD}(D)$. The tangent space of $\mathbb{P}$ can be identified with the real vector space

$$\mathbb{H} = \{(H_0, H_1, \ldots, H_k) \,:\, H_0 \in \mathbb{R} \text{ and } H_a \text{ a symmetric traceless } d_a \times d_a \text{ matrix } \forall a \in [k]\}.$$

The direction $(1, 0, \ldots, 0)$ changes Θ by an overall scalar, and tangent directions supported only in the a^{th} component for $a \in [k]$ only change Θ_a, subject to its determinant staying fixed. The geodesics on $\mathbb{P}$ are simply the geodesics of the Fisher-information metric on $\mathrm{PD}(D)$, but we may define them more precisely in terms of the tangent space $\mathbb{H}$ as follows.

DEFINITION (Geodesics and balls). Let $P \in \mathbb{P}$. The *exponential map* $\exp_\Theta \colon \mathbb{H} \to \mathbb{P}$ at Θ is defined by

$$\exp_\Theta(H) = e^{H_0} \cdot (\Theta_1^{1/2} e^{\sqrt{d_1} H_1} \Theta_1^{1/2}) \otimes \cdots \otimes (\Theta_k^{1/2} e^{\sqrt{d_k} H_k} \Theta_k^{1/2}).$$

The *geodesics* through Θ are the curves $t \mapsto \exp_\Theta(tH)$ for $t \in \mathbb{R}$ and $H \in \mathbb{H}$.

Up to reparameterization, there is a unique geodesic between any two points of $\mathbb{P}$. We take the convention that the geodesics have unit speed if $\|H\|_F^2 = |H_0|^2 + \sum_{a=1}^k \|H_a\|_F^2 = 1$. The geodesic distance $d(\Theta, \Theta')$ between two points Θ and $\Theta' = \exp_\Theta(H)$ is then equal to $\|H\|_F$, which is equal to the Fisher-Rao distance between Θ and Θ'. The closed *(geodesic) ball* of radius $r > 0$ about Θ is given by

$$B_r(\Theta) = \{\exp_\Theta(H) : \|H\|_F \leq r\},$$

The manifold $\mathrm{PD}(D)$, and hence $\mathbb{P}$, is a *Hadamard manifold*, i.e. a complete, simply connected Riemannian manifold of non-positive sectional curvature (Bacák, 2014). Thus geodesic balls are *geodesically convex* subsets of $\mathbb{P}$, that is, if $\gamma(t)$ is a geodesic such that $\gamma(0), \gamma(1) \in B_r(\Theta)$ then $\gamma(t) \in B_r(\Theta)$ for all $t \in [0, 1]$.

Using our definition of geodesics, we obtain the following notion of geodesic convexity of functions.

DEFINITION (Geodesic convexity). A twice differentiable function $f \colon \mathbb{P} \to \mathbb{R}$ is said to be *geodesically convex* at $\Theta \in \mathbb{P}$ if $\partial^2_{t=0} f(\exp_\Theta(tH)) \geq 0$ for all $H \in \mathbb{H}$. It is called λ*-strongly geodesically convex* at Θ for some $\lambda > 0$ if $\partial^2_{t=0} f(\exp_P(tH)) \geq \lambda \|H\|_F^2$ for all $H \in \mathbb{H}$.

We say that for a geodesically convex domain $D \subseteq \mathbb{P}$, a function f is (strongly) geodesically convex on D if, and only if, the function $t \mapsto f(\gamma(t))$ is (strongly) convex on $[0, 1]$ for any (unit-speed) geodesic $\gamma(t)$ with $\gamma(0), \gamma(1) \in D$. In other words, geodesic convexity simply means convexity in the ordinary Euclidean sense when restricted to geodesics.

The invariance properties described above for $\mathrm{PD}(D)$ are directly inherited to $\mathbb{P}$. The manifold $\mathbb{P}$ carries a natural action by the group

$$\mathbb{G} = \{G_1 \otimes \ldots \otimes G_1 : \mathrm{GL}(d_1) \otimes \times \cdots \times \otimes \mathrm{GL}(d_k)\}$$

Namely, if $\Theta \in \mathbb{P}$ and $A \in \mathbb{G}$ then the $A\Theta A^T$ is in $\mathbb{P}$. Moreover, the mapping $\Theta \mapsto A\Theta A^T$ is an isometry of the Riemannian manifold P. As discussed above, it preserves the statistical distances D_F and D_{op}.

3.2. *Sketch of proof.* With these definitions in place, we are able to state a proof plan for Theorem 2.1. The proof is a Riemannian version of the standard approach using strong convexity.

1. **Reduce to identity:** We can obtain n independent samples from $\mathcal{N}(0, \Theta^{-1})$ as $x_i' = \Theta^{-1/2} x_i$, where $x_1, \ldots, x_n$ are distributed as n independent samples from a standard Gaussian. The MLE $\widehat{\Theta}(x')$ for the former is exactly $\Theta^{1/2}\widehat{\Theta}(x)\Theta^{1/2}$. By invariance of the relative Frobenius error, $D_F(\widehat{\Theta}(x')\|\Theta) = D_F(\widehat{\Theta}(x)\|I_D)$; the same is true for D_{op}. This shows that to prove Theorem 2.1 it is enough to consider the case that $\Theta = I_D$, i.e., the standard Gaussian.
2. **Bound the gradient:** Show that the gradient $\nabla f_x(I_D)$ (defined below) is small with high probability.
3. **Show strong convexity:** Show that, with high probability, f_x is $\Omega(1)$-strongly geodesically convex near I.

These together imply the desired sample complexity bounds – as in the Euclidean case, strong convexity in a suitably large ball about a point implies the optimizer cannot be far. Moreover, it happens that under alternating minimization f_x obeys a descent lemma (similar to what is shown in Bürgisser et al. (2018)); as such the flip-flop algorithm must converge exponentially quickly by the strong geodesic convexity of f_x.

To make this discussion more concrete, we now define the gradient and Hessian formally, and state the lemma that we will use to relate the gradient and strong convexity to the distance to the optimizer as in the plan above.

DEFINITION (Gradient and Hessian). Let $f\colon \mathbb{P} \to \mathbb{R}$ be a once or twice differentiable function and $\Theta \in \mathbb{P}$. The *(Riemannian) gradient* $\nabla f(\Theta)$ is the unique element in $\mathbb{H}$ such that

$$\langle \nabla f(\Theta), H\rangle = \partial_{t=0} f(\exp_\Theta(tH)) \qquad \forall H \in \mathbb{H}.$$

Similarly, the *(Riemannian) Hessian* $\nabla^2 f(\Theta)$ is the unique linear operator on $\mathbb{H}$ such that

$$\langle H, \nabla^2 f(\Theta) K\rangle = \partial_{s=0}\partial_{t=0} f(\exp_\Theta(sH + tK)) \qquad \forall H, K \in \mathbb{H}.$$

We abbreviate $\nabla f = \nabla f(I_D)$ and $\nabla^2 f = \nabla^2 f(I_D)$ for the gradient and Hessian, respectively, at the identity matrix, and we write $\nabla_a f$ and $\nabla^2_{ab} f$ for the components. As block matrices,

$$\nabla f = \begin{bmatrix} \nabla_0 f \\ \hline \nabla_1 f \\ \vdots \\ \nabla_k f \end{bmatrix}, \qquad \nabla^2 f = \left[\begin{array}{c|ccc} \nabla^2_{00} f & \nabla^2_{01} f & \ldots & \nabla^2_{0k} f \\ \hline \nabla^2_{10} f & \nabla^2_{11} f & \ldots & \nabla^2_{1k} f \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2_{k0} f & \nabla^2_{k1} f & \ldots & \nabla^2_{kk} f \end{array}\right].$$

Here, $\nabla_0 f \in \mathbb{R}$ and each $\nabla_a f(\Theta)$ is a $d_a \times d_a$ traceless symmetric matrix. Similarly, for $a, b \in [k]$ (i.e., for the blocks of the submatrix to the lower-right of the lines) the components $\nabla^2_{ab} f(\Theta)$ of the Hessian are linear operators from the space of traceless symmetric $d_b \times d_b$ matrices to the space of traceless symmetric $d_a \times d_a$ matrices. Moreover, $\nabla_{a0} f$ is a linear operator from

$\mathbb{R}$ to the space of traceless symmetric $d_a \times d_a$ matrices (hence can itself be viewed as such a matrix), $\nabla_{0a} f$ is the adjoint of this linear operator (which using the Hilbert-Schmidt inner product can be identified with the same matrix), and $\nabla^2_{00} f(\Theta)$ is a real number.

We note that the Hessian is symmetric with respect to the inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{H}$. Just like in the Euclidean case, the Hessian is convenient to characterize strong convexity. Indeed, $\langle H, \nabla^2 f(\Theta) H \rangle = \partial^2_{t=0} f(\exp_\Theta(tH))$ for all $H \in \mathbb{H}$. Thus, f is geodesically convex if and only if the Hessian is positive semidefinite, that is, $\nabla^2 f(\Theta) \succeq 0$. Similarly, f is λ-strongly geodesically convex if and only if $\nabla^2 f(\Theta) \succeq \lambda I$, i.e., the Hessian is positive definite with eigenvalues larger than or equal to λ.

We can now state and prove the following lemma, which shows that strong convexity in a ball about a point where the gradient is sufficiently small implies the optimizer cannot be far.

LEMMA 3.1. *Let $f : \mathbb{P} \to \mathbb{R}$ be geodesically convex and twice differentiable. Assume the gradient at some $\Theta \in \mathbb{P}$ is bounded as $\|\nabla f(\Theta)\|_F \leq \varepsilon$, and that f is λ-strongly geodesically convex in a ball $B_r(\Theta)$ of radius $r > \frac{2\varepsilon}{\lambda}$. Then the sublevel set $\{\Upsilon \in \mathbb{P} : f(\Upsilon) \leq f(\Theta)\}$ is contained in the ball $B_{2\varepsilon/\lambda}(\Theta)$, f has a unique minimizer $\widehat{\Theta}$, this minimizer is contained in $B_{\varepsilon/\lambda}(\Theta)$, and*

$$f(\widehat{\Theta}) \geq f(\Theta) - \frac{\varepsilon^2}{2\lambda}. \tag{3.1}$$

PROOF. We first show that the sublevel set of $f(\Theta)$ is contained in the ball of radius $\frac{2\varepsilon}{\lambda}$. Consider $g(t) := f(\exp_\Theta(tH))$, where $H \in \mathbb{H}$ is an arbitrary vector of unit norm $\|H\|_F = 1$. Then, using the assumption on the gradient,

$$g'(0) = \partial_{t=0} f(\exp_\Theta(tH)) = \langle \nabla f(\Theta), H \rangle \geq -\|\nabla f(\Theta)\|_F \|H\|_F \geq -\varepsilon. \tag{3.2}$$

Since f is λ-strongly geodesically convex on $B_r(\Theta)$, we have $g''(t) \geq \lambda$ for all $|t| \leq r$. It follows that for all $0 \leq t \leq r$ we have

$$g(t) \geq g(0) - \varepsilon t + \frac{1}{2}\lambda t^2. \tag{3.3}$$

Plugging in $t = r$ yields $g(r) \geq g(0) + \left(\frac{\lambda r}{2} - \varepsilon\right) r > g(0)$. Since g is convex due to the geodesic convexity of f, it follows that, for any $t \geq r$,

$$g(0) < g(r) \leq \left(1 - \frac{r}{t}\right) g(0) + \frac{r}{t} g(t),$$

hence

$$f(\Theta) = g(0) < g(t) = f(\exp_\Theta(tH)).$$

Thus the sublevel set of $f(\Theta)$ is contained in the ball of radius r about Θ. By replacing r with any smaller $r' > \frac{2\varepsilon}{\lambda}$, we see that the sublevel set is in fact contained in the ball of radius $\frac{2\varepsilon}{\lambda}$. In particular, the minimum of f is attained and any minimizer $\widehat{\Theta}$ is contained in this ball. Moreover, as the right-hand side of Eq. (3.3) takes a minimum at $t = \frac{\varepsilon}{\lambda}$, we have $g(t) \geq g(0) - \frac{\varepsilon^2}{2\lambda}$ for all $0 \leq t \leq r$. By definition of g, this implies Eq. (3.1).

Next, we prove that any minimizer of f is necessarily contained in the ball of radius $\frac{\varepsilon}{\lambda}$. To see this, take an arbitrary minimizer $\widehat{\Theta}$ and write it in the form $\widehat{\Theta} = \exp_\Theta(TH)$, where $H \in \mathbb{H}$ is a unit vector and $T > 0$.

As before, we consider the function $g(t) = f(\exp_\Theta(tH))$. Then, using Eq. (3.2), the convexity of $g(t)$ for all $t \in \mathbb{R}$ and the λ-strong convexity of $g(t)$ for $|t| \leq r$, we have

$$0 = g'(T) = g'(0) + \int_0^T g''(t)\, dt \geq \lambda \min(T, r) - \varepsilon.$$

If $T > r$ then we have a contradiction as $\lambda r - \varepsilon > \lambda r/2 - \varepsilon > 0$. Therefore we must have $T \leq r$ and hence $\lambda T - \varepsilon \leq 0$, so $T \leq \frac{\varepsilon}{\lambda}$. Thus we have proved that any minimizer of f is contained in the ball of radius $\frac{\varepsilon}{\lambda}$.

We still need to show that the minimizer is unique; that this follows from strong convexity is convex optimization "folklore," but we include a proof nonetheless. Indeed, suppose that $\widehat{\Theta}$ is a minimizer and let $H \in \mathbb{H}$ be arbitrary. Consider $h(t) := f(\exp_{\widehat{\Theta}}(tH))$. Then the function $h(t)$ is convex, has a minimum at $t = 0$, and satisfies $h''(0) > 0$, since f is λ-strongly geodesically convex near $\widehat{\Theta}$, as $\widehat{\Theta} \in B_r(\Theta)$ by what we showed above. It follows that $h(t) > h(0)$ for any $t \neq 0$. Since H was arbitrary, this shows that $f(\Upsilon) > f(\widehat{\Theta})$ for any $\Upsilon \neq \widehat{\Theta}$. □

Using the geodesic distance bounds from the previous lemma allows us to obtain bounds on the individual Kronecker factors under strong convexity.

COROLLARY 3.2 (From geodesic distance to D_F, D_{op}). *Let $f\colon \mathbb{P} \to \mathbb{R}$ be geodesically convex and twice differentiable. Suppose the gradient at some $\Theta \in \mathbb{P}$ is bounded as $\|\nabla f(\Theta)\|_F \leq \delta$. Suppose further that f is λ-strongly geodesically convex in a ball $B_r(\Theta)$ of radius $r > \frac{2\delta}{\lambda}$, and that $2\sqrt{d_{\max}}\,\delta/\lambda \leq 1$. Then f has a unique minimizer $\widehat{\Theta} \in \mathbb{P}$ and for $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \ldots \otimes \widehat{\Theta}_k$ and $\Theta = \Theta_1 \otimes \ldots \otimes \Theta_k$ where $|\widehat{\Theta}_1| = \ldots = |\widehat{\Theta}_k|$ and $|\Theta_1| = \ldots = |\Theta_k|$ we have*

$$D_{op}(\widehat{\Theta}_a \,||\, \Theta_a) \leq D_F(\widehat{\Theta}_a \,||\, \Theta_a) \leq 4\sqrt{d_a}\delta/\lambda$$

and

$$D_F(\widehat{\Theta}||\Theta) \leq 4ke^{4k\delta/\lambda}\sqrt{D}\delta/\lambda.$$

PROOF. By invariance of D_F, we may assume $\Theta_a = I_a$ for all a. By Lemma 3.1, we may write $\widehat{\Theta} = \exp_{I_D}(H)$ for $H \in \mathbb{H}$ and $\|H\|_F \leq 2\delta/\lambda$. Then the equal-determinant Kronecker factors are given by $\widehat{\Theta}_a = e^{\frac{H_0}{d_a k} I_{d_a} + \sqrt{d_a} H_a}$ for all $a \in [k]$. Since H_a is traceless, note that

$$\begin{aligned}\|\frac{H_0}{d_a k} I_{d_a} + \sqrt{d_a} H_a\|_F^2 &= \frac{|H_0|^2}{d_a k^2} + \|\sqrt{d_a} H_a\|_F^2 \\ &\leq d_a \left(|H_0|^2 + \|H_a\|_F^2\right) \leq d_a \|H\|_F^2 \leq 4 d_a \delta^2/\lambda^2.\end{aligned}$$

By assumption the above is at most one, so we obtain

$$\begin{aligned}D_F(\widehat{\Theta}_a \| I_a) &= \|I_a - e^{\frac{H_0}{d_a k} I_{d_a} + \sqrt{d_a} H_a}\|_F \\ &\leq 2\|\frac{H_0}{d_a k} I_{d_a} + \sqrt{d_a} H_a\|_F \leq 4\sqrt{d_a}\delta/\lambda,\end{aligned}$$

which establishes the first bound. The second now follows easily by a telescoping sum:

$$\begin{aligned}D_F(\widehat{\Theta}\|I_D) &\leq \sum_{a=1}^{k} \|\widehat{\Theta}_1\|_F \cdots \|\widehat{\Theta}_{a-1}\|_F \|I_{d_a} - \widehat{\Theta}_a\|_F \|I_{d_{a+1}}\|_F \cdots \|I_{d_k}\|_F \\ &\leq 4\sqrt{D} \cdot \frac{\delta}{\lambda} \cdot \sum_{a=1}^{k} \left(1 + \frac{4\delta}{\lambda}\right)^{a-1} \leq 4ke^{4k\delta/\lambda}\sqrt{D}\delta/\lambda.\end{aligned}$$

□

3.3. *Bounding the gradient.* Proceeding according to step 2 of the plan outlined in Section 3.2, we now compute the gradient of the objective function and bound it using basic matrix concentration results.

To calculate the gradient, we need a definition from linear algebra. Recall that our data comes as an n-tuple $x = (x_1, \dots, x_n)$ of k-tensors. Let $\rho := \frac{1}{nD} \sum_i x_i x_i^T$ denote the matrix of "second sample moments" of the data. Then we can rewrite the objective function as

$$f_x(\Theta) = \operatorname{Tr} \rho \Theta - \frac{1}{D} \log \det \Theta. \tag{3.4}$$

We may also consider the "second sample moments" of a subset of the coordinates $J \subseteq [k]$. For this the following definition is useful.

DEFINITION (Partial trace). Let ρ be an operator on $\mathbb{R}^{d_1} \otimes \ldots \otimes \mathbb{R}^{d_k}$, and $J \subseteq [k]$. Define the *partial trace* $\rho^{(J)}$ as the $d_J \times d_J$-matrix, $d_J = \prod_{a \in J} d_a$, that satisfies the property that

$$\operatorname{Tr} \rho^{(J)} H = \operatorname{Tr} \rho H_{(J)} \tag{3.5}$$

for any $d_J \times d_J$ matrix H, where $H_{(J)}$ denotes the operator on $\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k}$ that acts as H on the tensor factors labeled by J and as the identity on the rest. This property uniquely determines $\rho^{(J)}$. We write $\rho^{(a)}$ and $\rho^{(ab)}$ when $J = \{a\}$ and $J = \{a, b\}$, respectively.

If ρ is positive definite then so is $\rho^{(J)}$. Moreover, $\operatorname{Tr} \rho = \operatorname{Tr} \rho^{(J)}$ and $(\rho^{(J)})^{(K)} = \rho^{(K)}$ for $K \subseteq J$.

Concretely, the partial trace $\rho^{(I)}$ can be calculated as follows: Analogously to the discussion in Section 2.1, "flatten" the data x by regarding it as a $d_I \times N_I$ matrix $x^{(I)}$, where $N_I = \frac{nD}{d_I}$; then $\rho^{(I)} := \frac{1}{nD} x^{(I)} (x^{(I)})^T$.

The components of the gradient can be readily computed in terms of partial traces.

LEMMA 3.3 (Gradient). *Let $\rho = \frac{1}{nD} \sum_{i=1}^n x_i x_i^T$. Then the components of the gradient ∇f_x at the identity are given by*

$$\nabla_a f_x = \sqrt{d_a} \left(\rho^{(a)} - \frac{\operatorname{Tr} \rho}{d_a} I_{d_a} \right) \qquad \text{for } a \in [k], \tag{3.6}$$

$$\nabla_0 f_x = \operatorname{Tr} \rho - 1. \tag{3.7}$$

PROOF. For all $a \in [k]$ and any traceless symmetric $d_a \times d_a$ matrix H, we have

$$\begin{aligned} \langle \nabla_a f_x(I_D), H \rangle &= \partial_{t=0} f_x(e^{t\sqrt{d_a} H_{(a)}}) = \partial_{t=0} \operatorname{Tr} \rho e^{t\sqrt{d_a} H_{(a)}} - \frac{1}{D} \log \det(e^{t\sqrt{d_a} H_{(a)}}) \\ &= \sqrt{d_a} \operatorname{Tr} \rho H_{(a)} = \sqrt{d_a} \operatorname{Tr} \rho^{(a)} H \end{aligned}$$

using Eqs. (3.4) and (3.5) and that $\operatorname{Tr} H_{(a)} = \operatorname{Tr} H = 0$. Since $\nabla_a f_x$ is traceless and symmetric by definition, while $\rho^{(a)}$ is symmetric, this implies

$$\nabla_a f_x = \sqrt{d_a} \left(\rho^{(a)} - \frac{\operatorname{Tr} \rho^{(a)}}{d_a} I_{d_a} \right) = \sqrt{d_a} \left(\rho^{(a)} - \frac{\operatorname{Tr} \rho}{d_a} I_{d_a} \right).$$

Finally,

$$\nabla_0 f_x = \partial_{t=0} \left(\operatorname{Tr} \rho e^t - \frac{1}{D} \log \det(e^t I_D) \right) = \partial_{t=0} \left(\operatorname{Tr} \rho e^t - t \right) = \operatorname{Tr} \rho - 1.$$

□

REMARK 3.4 (Gradient at other points). *In the previous lemma we only computed the gradient at the identity. However, this is without loss of generality, since from the calculations above one easily obtains* $\nabla f_x(\Theta) = \nabla f_{\Theta^{1/2}x}(I)$. *That is, the gradient* $\nabla f_x(\Theta)$ *is given by Eqs.* (3.6) *and* (3.7) *with* ρ *replaced by* $\Theta^{1/2}\rho\,\Theta^{1/2}$.

Having calculated the gradient of the objective function, we are ready to state our bound:

PROPOSITION 3.5 (Gradient bound). *Let* $x = (x_1, \ldots, x_n)$ *consist of independent standard Gaussian random variables in* $\mathbb{R}^D$, *and let* $0 < \varepsilon < 1$. *Suppose* $n \geq \frac{d_{\max}^2}{D\varepsilon^2}$. *Then, the following occurs with probability at least* $1 - 2(k+1)e^{-\varepsilon^2 \frac{nD}{8d_{\max}}}$:

$$\|\nabla_a f_x\|_{\mathrm{op}} \leq \frac{9\varepsilon\sqrt{d_a}}{d_{max}} \qquad \text{for all } a \in [k],$$
$$|\nabla_0 f_x| \leq \varepsilon.$$

As a consequence, $\|\nabla_a f_x\|_{\mathrm{op}} \leq \frac{9\varepsilon}{\sqrt{d_{max}}} \leq \frac{9\varepsilon}{\sqrt{d_a}}$ *and hence*

$$\|\nabla f_x\|_F^2 \leq (1 + 81k)\varepsilon^2 \leq 82k\varepsilon^2.$$

To prove this we will need a standard result in matrix concentration. When the samples $x = (x_1, \ldots, x_n)$ are independent standard Gaussians in $\mathbb{R}^D$, then $\rho^{(a)}$ is distributed as $\frac{1}{nD}YY^T$, where Y is a random $d_a \times N_a$ matrix with independent standard Gaussian entries, where $N_a = \frac{nD}{d_a}$. The following result bounds the singular values of such random matrices.

THEOREM 3.6 (Corollary 5.35 of Vershynin (2010)). *Let* $Y \in \mathbb{R}^{d \times N}$ *have independent standard Gaussian entries where* $N \geq d$. *Then, for every* $t > 0$, *the following occurs with probability at least* $1 - 2e^{-t^2/2}$:

$$\sqrt{N} - \sqrt{d} - t \leq \sigma_d(Y) \leq \sigma_1(Y) \leq \sqrt{N} + \sqrt{d} + t,$$

where σ_j *denotes the* j*-th largest singular value.*

We will also need to bound $\operatorname{Tr}\rho = \frac{1}{nD}\|x\|_2^2$. Because $\|x\|_2^2$ is simply a sum of nD many χ-squared random variables, the next proposition follows from standard concentration bounds.

PROPOSITION 3.7 (Example 2.11 of Wainwright (2019)). *Let* $x = (x_1, \ldots, x_n)$ *consist of independent standard Gaussian random variables in* $\mathbb{R}^D$. *Then, for* $0 < t < 1$, *the following occurs with probability at least* $1 - 2e^{-t^2 nD/8}$:

$$(1-t)nD \leq \|x\|_2^2 \leq (1+t)nD.$$

Equipped with these results we now prove our gradient bound.

PROOF OF PROPOSITION 3.5. For any fixed $a \in [k]$, recall that $\rho^{(a)}$ has the same distribution as $\frac{1}{nD}YY^T$, where Y is an $d_a \times N_a$-matrix with standard Gaussian entries where $N_a := \frac{nD}{d_a}$. By Theorem 3.6, we have the following bound with failure probability at most $2e^{-t^2/2}$:

$$\sqrt{N_a} - \sqrt{d_a} - t \leq \sigma_d(Y) \leq \sigma_1(Y) \leq \sqrt{N_a} + \sqrt{d_a} + t.$$

This tells us that the eigenvalues of $d_a\rho^{(a)}$ are in the range $((1 - \frac{\sqrt{d_a}+t}{\sqrt{N_a}})^2, (1 + \frac{\sqrt{d_a}+t}{\sqrt{N_a}})^2)$ with failure probability at most $2e^{-t^2/2}$. Let $t = \varepsilon\sqrt{nDd_a/d_{\max}^2}$. Because $nD \geq d_{\max}^2/\varepsilon^2$

and $0 < \varepsilon \leq 1$, we have $\sqrt{d_a} \leq t \leq \sqrt{N_a}$. Hence, the eigenvalues of $d_a \rho^{(a)}$ are contained in $(1 - 4\frac{t}{\sqrt{N_a}}, 1 + 8\frac{t}{\sqrt{N_a}})$, and so the eigenvalues of $d_a \rho_a - I_{d_a}$ are bounded in absolute value by

$$\frac{8t}{\sqrt{N_a}} = 8\sqrt{\frac{\varepsilon^2 n D d_a / d_{\max}^2}{nD/d_a}} = \frac{8\varepsilon d_a}{d_{\max}}$$

with failure probability at most $2e^{-\varepsilon^2 n D d_a / d_{\max}^2}$. Moreover, by Proposition 3.7, we have that $|\operatorname{Tr} \rho - 1| \leq \varepsilon$ with failure probability at most $2e^{-\varepsilon^2 nD/8} \leq 2e^{-\varepsilon^2 n D d_a / d_{\max}^2}$. The formulae in Lemma 3.3 and the union bound imply

$$\|\nabla_a f_x\|_{\mathrm{op}} \leq \frac{1}{\sqrt{d_a}} \left\| d_a \rho^{(a)} - I_{d_a} \right\|_{\mathrm{op}} + \frac{|\operatorname{Tr} \rho - 1|}{\sqrt{d_a}} \leq \frac{9\varepsilon \sqrt{d_a}}{d_{\max}} \qquad \text{for all } a \in [k].$$

with failure probability $2(k+1)e^{-\varepsilon^2 n D d_a / d_{\max}^2}$. □

3.4. *Strong convexity.* In this section, we prove our strong convexity result, Theorem 3.14, in order to carry out step 3 of the plan from Section 3.2. The theorem states that, with high probability, f_x is strongly convex near the identity. We will prove it by first establishing strong convexity *at* the identity, Theorem 3.12, using quantum expansion techniques, and then giving a bound on how the Hessian changes away from the origin, Lemma 3.13. We first assemble these results and then prove Theorem 3.14 at the end of this subsection.

Similarly as for the gradient, we can compute the components of the Hessian in terms of partial traces, but now we also need to consider two coordinates at a time.

LEMMA 3.8 (Hessian). *Let $\rho = \frac{1}{nD} \sum_{i=1}^n x_i x_i^T$. Then the components of the Hessian $\nabla^2 f_x$ at the identity are given by*

$$\langle H, (\nabla^2_{aa} f_x) H \rangle = d_a \operatorname{Tr} \rho^{(a)} H^2$$
$$\langle H, (\nabla^2_{ab} f_x) K \rangle = \sqrt{d_a d_b} \operatorname{Tr} \rho^{(ab)} (H \otimes K)$$

for all $a \neq b \in [k]$ and traceless symmetric $d_a \times d_a$ matrices H, $d_b \times d_b$ matrices K, and

$$\nabla^2_{0a} f_x = \nabla^2_{a0} f_x = \sqrt{d_a} \left(\rho^{(a)} - \frac{\operatorname{Tr} \rho}{d_a} I_{d_a} \right)$$
$$\nabla^2_{00} f_x = \operatorname{Tr} \rho.$$

for all $a \in [k]$.

Again we caution the reader that $\nabla^2_{a0} f_x$ is a linear operator from the real numbers to the traceless symmetric matrices, which we identify with a traceless symmetric matrix. We do the same for its adjoint $\nabla^2_{0a} f_x$.

REMARK 3.9 (Hessian at other points). *Analogously to Remark 3.4, we can compute the Hessian at other points using $\nabla^2 f_x(\Theta) = \nabla^2 f_{\Theta^{1/2} x}$. That is, the Hessian $\nabla^2 f_x(\Theta)$ is given by Lemma 3.8 with ρ replaced by $\Theta^{1/2} \rho \, \Theta^{1/2}$.*

PROOF. Note that the Hessian of f_x coincides with the one of $\operatorname{Tr} \rho \Theta$. This follows from Eq. (3.4), since the Hessian of $\log \det \Theta$ vanishes identically. Accordingly, we will compute the Hessian of $\operatorname{Tr} \rho \Theta$. For $a \in [k]$ and any traceless symmetric $d_a \times d_a$ matrix H, we have

$$\langle H, (\nabla^2_{aa} f_x) H \rangle = \partial_{s=0} \partial_{t=0} \operatorname{Tr} \rho \, e^{(s+t)\sqrt{d_a} H_{(a)}} = d_a \operatorname{Tr} \rho H^2_{(a)} = d_a \operatorname{Tr} \rho^{(a)} H^2$$

using Eq. (3.5). Similarly, for $a \neq b \in [k]$, any traceless symmetric $d_a \times d_a$ matrix H, and any traceless symmetric $d_b \times d_b$ matrix K, we find that

$$\begin{aligned}\langle H, (\nabla^2_{ab} f_x) K\rangle &= \partial_{s=0}\partial_{t=0} \operatorname{Tr} \rho\, e^{s\sqrt{d_a} H_{(a)} + t\sqrt{d_b} K_{(b)}} \\ &= \sqrt{d_a d_b} \operatorname{Tr} \rho\, H_{(a)} K_{(b)} = \sqrt{d_a d_b} \operatorname{Tr} \rho^{(ab)} (H \otimes K)\end{aligned}$$

using Eq. (3.5). Next, for $a \in [k]$ and any traceless symmetric $d_a \times d_a$ matrix H, we have

$$\langle H, \nabla^2_{a0} f_x\rangle = \partial_{s=0}\partial_{t=0} \operatorname{Tr} \rho\, e^{s\sqrt{d_a} H_{(a)} + t} = \sqrt{d_a} \operatorname{Tr} \rho\, H_{(a)} = \sqrt{d_a} \operatorname{Tr} \rho^{(a)} H,$$

recalling that we identify $\nabla^2_{a0} f_x$ with a traceless symmetric $d_a \times d_a$ matrix; this shows that

$$\nabla^2_{a0} f_x = \sqrt{d_a} \left(\rho^{(a)} - \frac{\operatorname{Tr} \rho}{d_a} I_{d_a} \right),$$

and similarly for the transpose. Finally,

$$\nabla^2_{00} f_x = \partial_{s=0}\partial_{t=0} \operatorname{Tr} \rho\, e^{s+t} = \operatorname{Tr} \rho.$$

□

The most interesting part of the Hessian are the off-diagonal components for $a \neq b \in [k]$, which up to an overall factor $\sqrt{d_a d_b}$ can be seen as the restrictions of the linear maps

$$\Phi^{(ab)} \colon \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_a), \quad \langle H, \Phi^{(ab)}(K)\rangle = \operatorname{Tr} \rho^{(ab)} (H \otimes K) \tag{3.8}$$

to the traceless symmetric matrices. This is a special case of a *completely positive map*, which is a linear map of the form

$$\Phi_A \colon \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_a), \quad \Phi_A(X) = \sum_{i=1}^N A_i X A_i^T \tag{3.9}$$

for $d_a \times d_b$ matrices $A_1, \ldots, A_N$. To see the connection, note that since $\rho^{(ab)}$ is positive semidefinite, it can be written in the form $\sum_{i=1}^N \operatorname{vec}(A_i) \operatorname{vec}(A_i)^T$; then $\Phi^{(ab)} = \Phi_A$ follows. The matrices $A_1, \ldots, A_N$ are known as *Kraus operators*. Equation (3.9) can also be written as

$$\operatorname{vec}(\Phi_A(X)) = \sum_{i=1}^N (A_i \otimes A_i) \operatorname{vec}(X). \tag{3.10}$$

We denote by $\Phi^* \colon \operatorname{Mat}(d_a) \to \operatorname{Mat}(d_b)$ the adjoint of a completely positive map Φ with respect to the Hilbert-Schmidt inner product; this is again a completely positive map, with Kraus operators $A_1^T, \ldots, A_N^T$. In our proof of strong convexity, we will show that strong convexity follows if the completely positive maps $\Phi^{(ab)}$ are good *quantum expanders*.

DEFINITION (Quantum expansion). Let $\Phi \colon \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_a)$ be a completely positive map. Say Φ is ε-*doubly balanced* if

$$\left\| \frac{\Phi(I_{d_b})}{\operatorname{Tr} \Phi(I_{d_b})} - \frac{I_{d_a}}{d_a} \right\|_{\mathrm{op}} \leq \frac{\varepsilon}{d_a} \quad \text{and} \quad \left\| \frac{\Phi^*(I_{d_a})}{\operatorname{Tr} \Phi^*(I_{d_a})} - \frac{I_{d_b}}{d_b} \right\|_{\mathrm{op}} \leq \frac{\varepsilon}{d_b}. \tag{3.11}$$

Map Φ is an (ε, λ)-*quantum expander* if Φ is ε-doubly balanced and

$$\|\Phi\|_0 := \max_{\substack{H \text{ traceless symmetric} \\ \|H\|_F = 1}} \|\Phi(H)\|_F \leq \lambda \frac{\operatorname{Tr} \Phi(I_{d_b})}{\sqrt{d_a d_b}} \tag{3.12}$$

A $(0, \lambda)$-quantum expander is called a λ-*quantum expander*. We note that all these definitions are invariant under rescaling $\Phi \mapsto c\Phi$ for $c > 0$.

Quantum expanders play an important role in quantum information theory and quantum computation . There one typically takes $d_a = d_b$, so that Eq. (3.12) simplifies to $\|\Phi\|_0 \leq \lambda$. For us, their purpose will be the following lemma allowing us to translate quantum expansion properties into strong convexity.

LEMMA 3.10 (Strong convexity from expansion). *If the completely positive maps $\Phi^{(ab)}$ defined in Eq. (3.8) are (ε, λ)-quantum expanders for every $a \neq b \in [k]$, then*

$$\left\| \frac{\nabla^2 f_x}{\operatorname{Tr} \rho} - I_{\mathbb{H}} \right\|_{\mathrm{op}} \leq (k-1)\lambda + (1+\sqrt{k})\varepsilon.$$

Assuming $k \geq 2$, the right-hand side is at most $k(\lambda + \varepsilon)$.

It suffices to verify the hypothesis for $a < b$. Indeed, since $\operatorname{Tr} \Phi^*(I_{d_a}) = \operatorname{Tr} \Phi(I_{d_b})$, any Φ is an (ε, λ)-quantum expander if and only if this is the case for the adjoint Φ^*, but note that the adjoint of $\Phi^{(ab)}$ is $\Phi^{(ba)}$.

PROOF. We wish to bound the operator norm of $M = \frac{\nabla^2 f_x}{\operatorname{Tr} \rho} - I_{\mathbb{H}}$, which we consider as a block matrix as in Section 3.2. For this, we use the following basic estimate of the norm of a block matrix in terms of the norm of the matrix of block norms, i.e.,

$$\|M\|_{\mathrm{op}} \leq \|m\|_{\mathrm{op}}, \quad \text{where } m = (\|M_{ab}\|_{\mathrm{op}})_{a,b \in \{0,1,\ldots,k\}}. \tag{3.13}$$

We first bound the individual block norms. Recall from Lemma 3.8 that the off-diagonal blocks of the Hessian for $a \neq b \in [k]$ are given by $\nabla^2_{ab} f_x = \sqrt{d_a d_b}\Phi^{(ab)}$. Since $\Phi^{(ab)}$ is an (ε, λ)-quantum expander, we have

$$\|M_{ab}\|_{\mathrm{op}} = \frac{\|\nabla^2_{ab} f_x\|_{\mathrm{op}}}{\operatorname{Tr} \rho} = \frac{\sqrt{d_a d_b}}{\operatorname{Tr} \Phi^{(ab)}(I_{d_b})} \|\Phi^{(ab)}\|_0 \leq \lambda,$$

using that $\operatorname{Tr} \Phi^{(ab)}(I_{d_b}) = \operatorname{Tr} \rho$. The remaining off-diagonal blocks can be bounded as

$$\begin{aligned} \|M_{a0}\| = \frac{\|\nabla^2_{a0} f_x\|_{\mathrm{op}}}{\operatorname{Tr} \rho} &= \left\| \sqrt{d_a} \left(\frac{\rho^{(a)}}{\operatorname{Tr} \rho} - \frac{I_{d_a}}{d_a} \right) \right\|_F = \sqrt{d_a} \left\| \frac{\Phi^{(ab)}(I_{d_b})}{\operatorname{Tr} \Phi^{(ab)}(I_{d_b})} - \frac{I_{d_a}}{d_a} \right\|_F \\ &\leq d_a \left\| \frac{\Phi^{(ab)}(I_{d_b})}{\operatorname{Tr} \Phi^{(ab)}(I_{d_b})} - \frac{I_{d_a}}{d_a} \right\|_{\mathrm{op}} \leq \varepsilon, \end{aligned}$$

using that $\Phi^{(ab)}(I_{d_b}) = \rho^{(a)}$. On the other hand, the diagonal blocks for $a \in [k]$ can be bounded by observing that, for any traceless Hermitian H,

$$\begin{aligned} |\langle H, M_{aa} H \rangle| = \left| \langle H, \left(\frac{\nabla^2_{aa} f_x}{\operatorname{Tr} \rho} - I_{\mathbb{H}} \right) H \rangle \right| &= d_a \left| \operatorname{Tr} \left(\frac{\rho^{(a)}}{\operatorname{Tr} \rho} - \frac{I_{d_a}}{d_a} \right) H^2 \right| \\ &\leq d_a \left\| \frac{\rho^{(a)}}{\operatorname{Tr} \rho} - \frac{I_{d_a}}{d_a} \right\|_{\mathrm{op}} \|H\|_F^2 \leq \varepsilon \|H\|_F^2, \end{aligned}$$

hence $\|M_{aa}\|_{\mathrm{op}} \leq \varepsilon$, while $|M_{00}| = |\frac{\nabla^2_{00} f_x}{\operatorname{Tr} \rho} - 1| = 0$. To conclude the proof, decompose

$$m = \left[\begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 0 & m_{12} & \cdots & m_{1k} \\ 0 & m_{21} & 0 & & m_{2k} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & m_{k1} & m_{k2} & \cdots & 0 \end{array}\right] + \left[\begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & m_{11} & 0 & \cdots & 0 \\ 0 & 0 & m_{22} & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m_{kk} \end{array}\right] + \left[\begin{array}{c|cccc} 0 & m_{01} & m_{02} & \cdots & m_{0k} \\ \hline m_{10} & 0 & 0 & \cdots & 0 \\ m_{20} & 0 & 0 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ m_{k0} & 0 & 0 & \cdots & 0 \end{array}\right].$$

The nonzero entries of the first matrix are bounded by λ, hence its operator norm is at most $(k-1)\lambda$. The second matrix is diagonal with diagonal entries bounded by ε, hence its operator norm is at most ε. The third matrix has nonzero entries bounded by ε, hence its operator norm is bounded by $\sqrt{k}\varepsilon$. Using Eq. (3.13) we obtain the desired bound. □

We are concerned with $\Phi^{(ab)}$ that arise from random Gaussians. Just like random graphs can give rise to good expanders, it is known that random completely positive maps, namely Φ constructed by choosing Kraus operators at random from various well-behaved distributions, yield good quantum expanders. When the Kraus operators are chosen to be standard Gaussian we have the following result:

THEOREM 3.11 (Pisier (2012, 2014)). *Let $A_1, \dots, A_N$ be independent $d_a \times d_b$ random matrices with independent standard Gaussian entries. Then, for every $t \geq 2$, with probability at least $1 - t^{-\Omega(d_a+d_b)}$, the completely positive map Φ_A, defined as in Eq. (3.9), satisfies*

$$\|\Phi_A\|_0 \leq O\left(t^2\sqrt{N}\,(d_a+d_b)\right).$$

Pisier's actual result is slightly different. As stated, Theorem 3.11 is an easy consequence of Theorem 16.6 in Pisier (2012), together with a standard symmetrization trick (see, e.g., the proof of Lemma 4.1 in Pisier (2014)). We present the details in Appendix A.

When the samples $x = (x_1, \dots, x_n)$ are independent standard Gaussians in $\mathbb{R}^D$, the random completely positive maps $\Phi^{(ab)}$ we are interested in have the same distribution as $\frac{1}{nD}\Phi_A$, where the Kraus operators $A_1, \dots, A_N$ are $d_a \times d_b$ matrices with independent standard Gaussian entries and $N = \frac{nD}{d_a d_b}$. Accordingly, strong convexity at the identity follows quite easily from Theorem 3.11 once double balancedness can be controlled. For the latter, observe that

$$\left\| \frac{\Phi^{(ab)}(I_{d_b})}{\operatorname{Tr}\Phi^{(ab)}(I_{d_b})} - \frac{I_{d_a}}{d_a} \right\|_{\mathrm{op}} = \frac{1}{\operatorname{Tr}\rho}\left\| \rho^{(a)} - \frac{\operatorname{Tr}\rho}{d_a} I_{d_a} \right\|_{\mathrm{op}} = \frac{1}{1+\nabla_0 f_x}\frac{1}{\sqrt{d_a}}\|\nabla_a f_x\|_{\mathrm{op}},$$

by Lemma 3.3, and similarly for the adjoint. Therefore, the completely positive maps $\Phi^{(ab)}$ are ε-doubly balanced if and only if, for all $a \in [k]$,

$$\sqrt{d_a}\|\nabla_a f_x\|_{\mathrm{op}} \leq \varepsilon \operatorname{Tr}\rho = (1+\nabla_0 f_x)\,\varepsilon, \tag{3.14}$$

hence double balancedness can be controlled using the gradient bounds in Proposition 3.5.

We now state and prove our strong convexity result at the identity:

THEOREM 3.12 (Strong convexity at identity). *There is a universal constant $C > 0$ such that the following holds. Let $x = (x_1, \dots, x_n)$ be independent standard Gaussian random variables in $\mathbb{R}^D$, where $n \geq Ck\frac{d_{\max}^2}{D}$. Then, with probability at least $1 - k^2(\frac{\sqrt{nD}}{kd_{\max}})^{-\Omega(d_{\min})}$,*

$$\|\nabla^2 f_x - I_{\mathbb{H}}\|_{\mathrm{op}} \leq \frac{1}{4};$$

in particular, f_x is $\frac{3}{4}$-strongly convex at the identity.

PROOF. By Lemma 3.10, it is enough to prove that with the desired probability all $\Phi^{(ab)}$ are $(\varepsilon, \lambda) := (\frac{1}{40k^{1/2}}, \frac{1}{20k})$-quantum expanders for $a \neq b \in [k]$ and $\operatorname{Tr}\rho \in (\frac{7}{8}, \frac{9}{8})$. If that is the case, then

$$\begin{aligned}\left\|\nabla^2 f_x - I_{\mathbb{H}}\right\|_{\mathrm{op}} &\leq \operatorname{Tr}\rho \cdot \left\| \frac{\nabla^2 f_x}{\operatorname{Tr}\rho} - I_{\mathbb{H}} \right\|_{\mathrm{op}} + |1 - \operatorname{Tr}\rho| \\ &\leq \left((k-1)\lambda + (1+\sqrt{k})\varepsilon\right)\operatorname{Tr}\rho + |1-\operatorname{Tr}\rho| \leq \frac{1}{4}.\end{aligned}$$

Firstly, $\operatorname{Tr}\rho = \frac{1}{nD}\|X\|^2$ is in $(\frac{7}{8}, \frac{9}{8})$ with failure probability $e^{-\Omega(nD)}$ by Proposition 3.7.

Next, we describe an event that implies the $\Phi^{(ab)}$ are all ε-balanced for $\varepsilon = \frac{1}{40k^{1/2}}$. By Eq. (3.14), this is equivalent to the condition $\sqrt{d_a}\|\nabla_a f_x\|_{\mathrm{op}} \leq \varepsilon \operatorname{Tr}\rho$ for all $a \in [k]$. By Proposition 3.5, and assuming the bound $\operatorname{Tr}\rho \geq \frac{7}{8}$ from above, the latter occurs with failure probability $ke^{-\Omega(\frac{nD}{kd_{\max}})}$ provided $n \geq Ck\frac{d_{\max}^2}{D}$ for a universal constant $C > 0$.

Finally, we describe an event that ensures that $\|\Phi^{(ab)}\|_{\mathrm{op}} \leq \lambda \frac{\operatorname{Tr}\rho}{\sqrt{d_a d_b}}$ for $\lambda = \frac{1}{20k}$ for any fixed $a \neq b$, which is the other condition needed for quantum expansion. Recall that each $\Phi^{(ab)}$ is distributed as $\frac{1}{nD}\Phi_A$, where A is a tuple of $\frac{nD}{d_a d_b}$ many $d_a \times d_b$ matrices with independent standard Gaussian entries. Thus, taking $t^2 = O(\frac{\lambda\sqrt{nD}}{d_a + d_b})$ and again assuming that $\operatorname{Tr}\rho \geq \frac{7}{8}$, we have $\|\Phi^{(ab)}\|_{\mathrm{op}} \leq \lambda \frac{\operatorname{Tr}\rho}{\sqrt{d_a d_b}}$ with failure probability at most $(\frac{\sqrt{nD}}{kd_{\max}})^{-\Omega(d_{\min})}$.

By the union bound, we conclude that all $\Phi^{(ab)}$ for $a \neq b$ are (ε, λ)-quantum expanders and that $\operatorname{Tr}\rho \in (\frac{7}{8}, \frac{9}{8})$, up to a failure probability of at most

$$e^{-\Omega(nD)} + ke^{-\Omega\left(\frac{nD}{kd_{\max}}\right)} + k^2\left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})}.$$

The final term dominates and so we obtain the desired bound on the failure probability. □

We now show our second strong convexity result, namely that if our function is strongly convex at the identity then it is also strongly convex in an operator norm ball about the identity. The proof is given in Appendix B.

LEMMA 3.13 (Robust strong convexity). *There is a universal constant $0 < \varepsilon_0 < 1$ such that if $\|\nabla_a f_x(I_D)\|_{\mathrm{op}} \leq \varepsilon_0/\sqrt{d_a}$ for all $a \in [k]$ and $|\nabla_0 f_x(I_D)| \leq \varepsilon_0$, then*

$$\|\nabla^2 f_x(\Theta) - \nabla^2 f_x\|_{op} = O(\delta)$$

for any $\Theta \in \mathbb{P}$ such that $\delta := \|\log\Theta\|_{\mathrm{op}}(I_D) \leq \varepsilon_0$. In particular, for any $\lambda > 0$, if f_x is λ-strongly convex at I_D then f_x is $(\lambda - O(\delta))$-strongly convex at Θ.

Finally we obtain our strong convexity result near the identity.

THEOREM 3.14 (Strong convexity near identity). *There are universal constants $C, c > 0$ such that the following holds. Let $x = (x_1, \ldots, x_n)$ be independent standard Gaussian random variables in $\mathbb{R}^D$, where $n \geq Ck\frac{d_{\max}^2}{D}$. Then, with probability at least $1 - k^2(\frac{\sqrt{nD}}{kd_{\max}})^{-\Omega(d_{\min})}$, the function f_x is $\frac{1}{2}$-strongly convex at any point $\Theta \in \mathbb{P}$ such that $\|\log\Theta\|_{\mathrm{op}} \leq c$.*

PROOF. We can choose $C > 0$ such that both Proposition 3.5 and Theorem 3.12 apply (the former with $\varepsilon \leq \varepsilon_0/9$, where ε_0 is the universal constant from Lemma 3.13). Then the assumptions of Lemma 3.13 are satisfied for $\lambda = \frac{3}{4}$ with failure probability at most

$$2(k+1)e^{-\varepsilon^2\frac{nD}{8d_{\max}}} + k^2\left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})},$$

where the latter term dominates, and there exists a constant $0 < c \leq \varepsilon_0$ such that f is $\frac{1}{2}$-strongly convex at any point Θ such that $\|\log\Theta\|_{\mathrm{op}} \leq c$. □

REMARK 3.15. *While Theorem 3.14 uses the operator norm to quantify closeness to the identity, we can easily translate it into a statement in terms of the geodesic distance on $\mathbb{P}$. Namely, under the same hypotheses it holds that f_x is $\frac{1}{2}$-strongly convex on the geodesic ball $B_r(I_D)$ of radius $r = \frac{c}{\sqrt{(k+1)d_{\max}}}$.*

3.5. *Proof of Theorem 2.1.* We are now ready to prove the main result of this section according to the plan outlined in Section 3.2. We restate the theorem for convenience.

THEOREM 2.1 (Tensor normal Frobenius error, restated). *There is a universal constant $C > 0$ such that the following holds. Suppose*

$$\max\{1, \varepsilon^2\} \le \frac{nD}{Ck^2 d_{\max}^2 \max\{k, d_{\max}\}}. \tag{2.2}$$

Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \cdots \otimes \widehat{\Theta}_k$ for n independent samples of the tensor normal model with precision matrix $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k$ satisfies

$$D_F(\widehat{\Theta}_a \| \Theta_a) = O\left(k^{1/2} \frac{\sqrt{d_a}\, d_{\max}}{\sqrt{nD}} \varepsilon\right) \quad \text{for all } a \in [k],$$
$$D_F(\widehat{\Theta} \| \Theta) = O\left(k^{3/2} \frac{d_{\max}}{\sqrt{n}} \varepsilon\right),$$

where the Kronecker factors of $\widehat{\Theta}$ and Θ are chosen such that $\det \widehat{\Theta}_1 = \cdots = \det \widehat{\Theta}_k$ and $\det \Theta_1 = \cdots = \det \Theta_k$, with probability at least

$$1 - ke^{-\Omega\left(\varepsilon^2 d_{\max}\right)} - k^2 \left(\frac{\sqrt{nD}}{k d_{\max}}\right)^{-\Omega(d_{\min})},$$

PROOF. By step 1 in Section 3.2, it is enough to prove the theorem assuming $\Theta = I_D$. Assuming this, we now show that the minimizer of f_x exists and is close to $\Theta = I_D$ with high probability. Let $C, c > 0$ be the constants from Theorem 3.14. For

$$\delta = \frac{\sqrt{82k}\, d_{\max}}{\sqrt{nD}} \varepsilon, \tag{3.15}$$

consider the following two events:

1. $\|\nabla f_x\|_F \le \delta$.
2. f_x is $\frac{1}{2}$-strongly convex on the geodesic ball $B_r(I_D)$ of radius $r = \frac{c}{\sqrt{(k+1)d_{\max}}}$.

By our assumption (2.2), if we choose the constant $C > 0$ large enough, both Proposition 3.5 (with the parameter ε in the proposition set to $\delta/\sqrt{82k}$) and Theorem 3.14 apply, with the former showing the first and the latter showing the second event. The desired success probability then follows from the union bound.

By Lemma 3.1 and Corollary 3.2, these two events, together with our assumption (2.2) (again for $C > 0$ large enough), imply the MLE $\widehat{\Theta}$ exists and is unique. Moreover, $\widehat{\Theta}$ is of geodesic distance at most 4δ from I_D and satisfies

$$D_F(\widehat{\Theta}_a \| I_a) \le 8\sqrt{d_a}\delta = O\left(\sqrt{k}\, \frac{\sqrt{d_a}\, d_{\max}}{\sqrt{nD}} \varepsilon\right)$$

and $D_F(\widehat{\Theta} \| I_D) \le 8ke^{8\delta k}\sqrt{D}\delta = O\left(k^{3/2} \frac{d_{\max}}{\sqrt{n}} \varepsilon\right)$. □

4. Improvements for the matrix normal model. We now prove Theorem 2.3, which improves over Theorem 2.1 in the case of the matrix normal model ($k = 2$). Our results for the matrix normal model are stronger in that

1. the MLE is shown to be close to the truth *in spectral norm* rather than the looser Frobenius norm,

2. the errors are *tight for the individual factors*, and
3. the failure probability is *inverse exponential* in the number of samples rather than inverse polynomial.

The proof plan is similar to that in Section 3.2, the main difference being that we now work directly with quantum expansion instead of translating into strong convexity. Our main tool is a bound by Kwok, Lau and Ramachandran (2019) which uses the notion of a *spectral gap*.

DEFINITION (Spectral gap). Let $\Phi\colon \mathrm{Mat}(d_b) \to \mathrm{Mat}(d_a)$ be a completely positive map. Say Φ has *spectral gap* γ if

$$\sigma_2(\Phi) \le (1-\gamma)\frac{\operatorname{Tr}\Phi(I_{d_b})}{\sqrt{d_a d_b}} \tag{4.1}$$

where σ_2 denotes the second largest singular value of Φ.

Recall that by the variational formula for singular values, if we let $X_1 \in \mathrm{Mat}(d_b)$ be the first (right) singular vector of Φ, we can rewrite the above condition as

$$\sigma_2(\Phi) = \max_{\langle H, X_1\rangle = 0} \frac{\|\Phi(H)\|_F}{\|H\|_F} \le (1-\gamma)\frac{\operatorname{Tr}\Phi(I_{d_b})}{\sqrt{d_a d_b}}.$$

On the other hand, the definition of an (ε,λ)-quantum expander is given in Eq. (3.12) as

$$\|\Phi\|_0 := \max_{\langle H, I_{d_b}\rangle = 0} \frac{\|\Phi(H)\|_F}{\|H\|_F} \le \lambda\frac{\operatorname{Tr}\Phi(I_{d_b})}{\sqrt{d_a d_b}}.$$

Due to the ε-doubly balanced condition in Eq. (3.11), it turns out that these two notions are closely related, as the following lemma shows.

LEMMA 4.1 (Lemma A.3 in Franks and Moitra (2020)). *There exists a universal constant $c > 0$ with the following property. If Φ is an (ε,λ)-quantum expander and $\varepsilon \le c(1-\lambda)$, then Φ has spectral gap $1-\lambda-O(\varepsilon)$.*

We now state the bound of Kwok, Lau and Ramachandran (2019) in our language. Because $k = 2$, the gradient and Hessian are completely described by the completely positive map $\Phi^{(12)}$ (compare Lemmas 3.3 and 3.8 with Eq. (3.8)).

THEOREM 4.2 (Theorem 1.8, Proof of Theorem 3.22 in Kwok, Lau and Ramachandran (2019)). *There is a universal constant $C > 0$ such that the following holds. If the completely positive map $\Phi^{(12)}$ defined in Eq. (3.8) is ε-doubly balanced and has spectral gap γ, where $\gamma^2 \ge C\varepsilon\log d_{\min}$, then, restricted to $\mathrm{SPD}(d_1)\times\mathrm{SPD}(d_2)$, where $\mathrm{SPD}(d)$ denotes the $d\times d$ positive definite matrices of unit determinant, the function f_x has a unique minimizer (P_1,P_2) such that*

$$\max\{\|P_1 - I_{d_1}\|_{\mathrm{op}}, \|P_2 - I_{d_2}\|_{\mathrm{op}}\} = O\Big(\frac{\varepsilon\log d_{\min}}{\gamma}\Big).$$

Moreover, $f_x(P_1,P_2) \ge (1-\frac{4\varepsilon^2}{\lambda})\frac{\|x\|^2}{nD}$.

COROLLARY 4.3. *There is a universal constant $C > 0$ such that the following holds. Let $0 \le \varepsilon,\gamma < 1$ and suppose the completely positive map $\Phi^{(12)}$ defined in Eq. (3.8) is ε-doubly balanced and has spectral gap γ, where $\gamma^2 \ge C\varepsilon\log d_{\min}$. Further suppose that $\|x\|^2/nD = 1+\delta$ where $|\delta| \le .5$. Then the MLE's $\widehat{\Theta}_1, \widehat{\Theta}_2$ satisfy*

$$\max\{\|\widehat{\Theta}_1 - I_{d_1}\|_{\mathrm{op}}, \|\widehat{\Theta}_2 - I_{d_2}\|_{\mathrm{op}}\} = O\Big(\delta + \frac{\varepsilon\log d_{\min}}{\gamma}\Big).$$

PROOF. To compute the MLE, we reparametrize by $\Theta_1 = P_1$ and $\Theta_2 = \lambda P_2$ where $P_1, P_2 \in \mathrm{SPD}(d_1) \times \mathrm{SPD}(d_2)$ and $\lambda \in \mathbb{R}_{>0}$. Plugging this reparametrization into f_x shows λ, P_1, P_2 solve

$$\arg\min_{\lambda, P_1, P_2} \lambda f_x(P_1, P_2) - \log \lambda.$$

In particular, the MLE $\widehat{\Theta}_1, \widehat{\Theta}_2$ exists uniquely if f_x restricted to $\mathrm{SPD}(d_1) \times \mathrm{SPD}(d_2)$ has unique minimizers P_1, P_2. Such unique minimizers exist by Theorem 4.2. Given P_1, P_2, solving the simple one-dimensional optimization problem for λ yields $\lambda = 1/f_x(P_1, P_2)$. We have $\widehat{\Theta} = P_1$, and

$$\|\widehat{\Theta}_2 - I_{d_2}\|_{op} = \|\lambda P_2 - I_{d_2}\|_{op} \leq \lambda \|P_2 - I_{d_2}\|_{op} + |1 - \lambda|.$$

By Theorem 4.2, $nD/\|x\|^2 \leq \lambda \leq (1 + 4\varepsilon^2/\gamma)^{-1}(nD/\|x\|^2)$. By our assumptions on γ, ε, we have $\varepsilon^2/\gamma \leq \varepsilon/\gamma \leq \varepsilon/\gamma^2$. Thus $\lambda = (1 + O(\varepsilon^2/\gamma))(1 + O(\delta)) = (1 + O(\varepsilon^2/\gamma + \delta)) = O(1)$ if C is large enough, so $\lambda \|P_2 - I_{d_2}\|_{op} = O(\|P_2 - I_{d_2}\|_{op}) = O(\varepsilon^2 \log d_{\min}/\gamma)$ by Theorem 4.2. $|1 - \lambda| = O(\varepsilon^2/\gamma + \delta) = O(\varepsilon^2 \log d_{\min}/\gamma)$. Combining these bounds completes the proof. □

Lemma 4.1 and Theorem 4.2, along with what we have shown so far, already imply a preliminary version of Theorem 2.3. In proving Theorem 3.12, we've already shown that Φ is an $(\varepsilon\sqrt{\frac{d_{\max}}{nd_{\min}}}, 1 - \lambda)$-quantum expander with failure probability

$$O(e^{-\Omega(d_{\max}\varepsilon^2)}) + \left(\frac{\sqrt{nD}}{d_{\min}}\right)^{-\Omega(d_{\min})}.$$

By Theorem 4.2, we immediately have that with the above failure probability the MLEs satisfy

$$D_{\mathrm{op}}(\Theta_a' \| \Theta_a) = O\left(\varepsilon\sqrt{\frac{d_{\max}}{nd_{\min}}} \log d_{\min}\right),$$

which matches Theorem 2.3 for the larger Kronecker factor.

One of the main results of this section is the following theorem, which shows that the expansion constant λ of Φ can be made constant with *exponentially small* failure probability. Recall that for the matrix model, the samples x_i can be viewed as $d_1 \times d_2$-matrices, which we denote by X_i.

THEOREM 4.4. *There are universal constants $C > 0$ and $0 < \lambda < 1$ such that the following holds. Let $X = (X_1, \ldots, X_n)$ be random $d_1 \times d_2$ matrices with independent standard Gaussian entries, where $n \geq C\frac{d_{\max}}{d_{\min}} \max\{\log \frac{d_{\max}}{d_{\min}}, \frac{1}{\varepsilon^2}, \varepsilon^2\}$. Then, Φ_X is an $(\varepsilon\sqrt{\frac{d_{\max}}{nd_{\min}}}, \lambda)$-quantum expander with probability at least $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$.*

We will prove Theorem 4.4 in Appendix C using techniques similar to Franks and Moitra (2020) using Cheeger's inequality. This also improves our result on strong convexity (Theorem 3.14), which will be useful in the analysis of the flip-flop algorithm. Indeed, for $k = 2$, using Theorem 4.4 (in place of Theorem 3.11) with Lemma 3.10 in the proof of Theorem 3.12 improves the failure probability in Theorem 3.12 to $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$. As in the proof of Theorem 3.14, combining this failure probability bound with Lemma 3.13 yields the next corollary.

COROLLARY 4.5. *There are universal constants $C, c > 0$ and $\lambda \in (0, 1)$ such that the following holds. Let $x = (x_1, \ldots, x_n)$ be independent standard Gaussian random variables in $\mathbb{R}^{d_1 d_2}$, where $n \geq C\frac{d_{\max}}{d_{\min}} \max\{\log \frac{d_{\max}}{d_{\min}}, \frac{1}{\varepsilon^2}\}$. Then, with probability at least $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$, the function f_x is $(1 - \lambda)$-strongly convex at any point $\Theta \in \mathbb{P}$ such that $\|\log \Theta\|_{\mathrm{op}} \leq c$.*

4.1. *Proof of Theorem 2.3.* We now use Theorem 4.4 as well as some more refined concentration inequalities to prove Theorem 2.3. The additional concentration is required to obtain the tighter bounds on the smaller Kronecker factor. Throughout this section, we assume without loss of generality that $d_1 \leq d_2$.

The idea of the proof is to apply one step of the flip-flop algorithm to "renormalize" the samples such that the second partial trace is proportional to I_{d_2}. This has the effect of making the second component of the gradient ∇f_x equal to zero. We will show that the first component still enjoys the same concentration exploited in Proposition 3.5 even after the step of flip-flop – thus the total gradient has become smaller, but only the second component of the estimate has changed. Thus, intuitively, the total change in the first component will be small. Using Lemma 3.13 to control the change induced in the minimal eigenvalue of the Hessian by the first step of the flip-flop and applying Lemma 3.1 results in a Frobenius error proportional to the new gradient after flip-flop, which gives us the tighter bound. To obtain a relative spectral error bound, we employ a similar strategy but with Theorem 4.2 instead of Lemma 3.1.

We now discuss the concentration bound. Let $X_1, \ldots, X_n$ be random $d_1 \times d_2$ matrices with independent standard Gaussian entries. Consider new random variables Y_i obtained by right-multiplying X_i by the square root of the result of one step of the flip-flop algorithm for the second, larger Kronecker factor. That is:

$$Y_i = X_i \left(\frac{1}{nd_1} \sum_{i=1}^{n} X_i^T X_i \right)^{-1/2}.$$

The completely positive map $\Phi^{(12)}$ corresponding to the "renormalized" samples $Y_1, \ldots, Y_n$ is $\frac{1}{nD}\Phi_Y$. By construction, it satisfies

$$\frac{1}{nD}\Phi_Y(I_{d_2}) = \frac{1}{d_2} \sum_{i=1}^{n} X_i \left(\sum_{i=1}^{n} X_i^T X_i \right)^{-1} X_i^T \quad \text{and} \quad \frac{1}{nD}\Phi_Y^*(I_{d_1}) = \frac{I_{d_2}}{d_2}. \tag{4.2}$$

Note also that $\operatorname{Tr} \Phi_Y(I_{d_2}) = \operatorname{Tr} \Phi_Y^*(I_{d_1}) = \|Y\|^2 = nD$. Thus Φ_Y is δ-doubly balanced if and only if $\|\frac{1}{nD}\Phi_Y(I_{d_2}) - \frac{1}{d_1} I_{d_1}\|_{op} \leq \frac{\delta}{d_1}$.

LEMMA 4.6 (Concentration after flip-flop). *There is a universal constant $C > 0$ such that the following holds. Let $X_1, \ldots, X_n$ be random $d_1 \times d_2$ matrices with independent standard Gaussian entries, where $d_1 \leq d_2$. If $n \geq \frac{d_2}{d_1}$ and $\varepsilon \geq C$, then for Φ_Y as in Eq. (4.2) we have, with probability at least $1 - \exp(-\Omega(\varepsilon^2 d_1))$,*

$$\left\| \frac{1}{nD}\Phi_Y(I_{d_2}) - \frac{I_{d_1}}{d_1} \right\|_{\text{op}} \leq \varepsilon \sqrt{\frac{1}{nD}}.$$

By the remarks preceding the lemma, this condition implies Φ_Y is $\varepsilon\sqrt{\frac{d_1}{nd_2}}$-doubly balanced.

The proof of this lemma uses a result of Hayden, Leung and Winter (2006) on the overlap of two random projections, combined with a standard net argument. The details can be found in Appendix D. With this bound in hand, we may now prove Theorem 2.3.

THEOREM 2.3 (Matrix normal spectral error, restated). *There is a universal constant $C > 0$ with the following property. Suppose $d_1 \leq d_2$ and $n \geq C\frac{d_2}{d_1}\max\{\log \frac{d_2}{d_1}, \frac{\log^2 d_1}{\varepsilon^2}, \varepsilon^2\}$. Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \widehat{\Theta}_2$ for n independent samples from the matrix normal model with precision matrix $\Theta = \Theta_1 \otimes \Theta_2$ satisfies*

$$D_{\text{op}}(\widehat{\Theta}_1 \| \Theta_1) = O\left(\varepsilon \sqrt{\frac{d_1}{nd_2}} \log d_1 \right) \quad \textit{and} \quad D_{\text{op}}(\widehat{\Theta}_2 \| \Theta_2) = O\left(\varepsilon \sqrt{\frac{d_2}{nd_1}} \right),$$

with probability at least $1 - O(e^{-\Omega(d_1\varepsilon^2)})$.

PROOF. As in the proof of Corollary 4.3, recall that we define $\widehat{\Theta}_1(x), \widehat{\Theta}_2(x)$ to be $\arg\min_{\det \widehat{\Theta}_1 = 1} f_x(\widehat{\Theta}_1 \otimes \widehat{\Theta}_2)$. By the equivariance discussed in Section 3.2, it is enough to prove Theorem 2.3 under the assumption $\Theta_a = I_{d_a}$ for $a \in \{1, 2\}$.
Let x be the tuple of our random samples and Φ_x be the off-diagonal operator from the Hessian, defined by $\Phi_x := \dfrac{1}{\sqrt{d_1 d_2}} \cdot \nabla^2_{12} f_x$ (that is, $\Phi_x = \Phi^{(12)}$ in our notation from Lemma 3.8). Additionally, let y be the result of normalizing the samples on the second tensor factor; that is:

$$y := (\sqrt{nd_1} I_{d_1} \otimes \Phi^*_x(I_2)^{-1/2})x =: (I_{d_1} \otimes B)x.$$

Define Φ_y similarly as we defined Φ_x.

Consider the following events:

1. The operator Φ_x is a $(\varepsilon\sqrt{d_2/nd_1}, 1-\lambda)$-quantum expander.
2. The operator Φ_y is $\varepsilon\sqrt{d_1/nd_2}$-balanced.
3. $|1 - \frac{\|X\|}{\sqrt{nD}}| \le \frac{\varepsilon}{\sqrt{d_2}}$.

These events occur with failure probability $1 - O(e^{-\Omega(d_1\varepsilon^2)})$. Indeed, by the assumption

$$n \ge C(d_2/d_1)\max\{\log(d_2/d_1), \frac{\log^2 d_1}{\varepsilon^2}, \varepsilon^2\} \ge C(d_2/d_1)\max\{\log(d_2/d_1), \varepsilon^{-2}, \varepsilon^2\},$$

Item 1 occurs with probability $1 - O(e^{-\Omega(d_1\varepsilon^2)})$ by Theorem 4.4. By Lemma 4.6, Item 2 occurs with probability $1 - O(e^{-\Omega(d_1\varepsilon^2)})$. Finally, Item 3 occurs with probability $1 - e^{-\Omega(nd_1\varepsilon^2)}$ by Proposition 3.7. By the union bound all the events occur with probability $1 - O(e^{-\Omega(d_1\varepsilon^2)})$.

Let x satisfy all three properties. We now use a lemma relating the quantum expansion of Φ_y and Φ_x; this is analogous to Lemma 3.13 guaranteeing the strong convexity of f in a neighborhood of the origin. By Lemma 4.4 in Franks and Moitra (2020), if $\kappa(B^2) - 1 \le \delta < c$, then Φ_y is an $(\varepsilon\sqrt{d_2/nd_1} + O(\delta), 1 - \lambda + O(\delta))$-quantum expander. We have

$$\kappa(B^2) = \kappa(\|x\|^{-2}\Phi^*_x(I_{d_1})) = O(\|\|x\|^{-2}\Phi^*_x(I_{d_1}) - I_{d_2}\|_{op}) = O(\varepsilon\sqrt{d_2/nd_1})$$

by the balancedness of Φ_x. Thus Φ_y is an $(\varepsilon\sqrt{d_1/nd_2}, 1 - \lambda/2)$-quantum expander provided $\varepsilon\sqrt{d_2/nd_1}$ is small enough compared to λ.

By our choice of n, we have $\varepsilon\sqrt{d_1/nd_2} \le c\lambda^2 \log d_1$ provided $\varepsilon \le c\lambda^2$. Thus Theorem 4.2 applies to Φ_y and so the MLE $\widehat{\Theta}_1(y), \widehat{\Theta}_2(y)$ satisfy

$$\|\widehat{\Theta}_1(y) - I_{d_1}\|_{op}, \|\widehat{\Theta}_2(y) - I_{d_2}\|_{op} = O\left(\varepsilon\sqrt{\frac{d_1}{nd_2}}\log d_1\right).$$

By the equivariance of the MLE we have $\widehat{\Theta}_1(x) = \widehat{\Theta}_1(y)$ and $\widehat{\Theta}_2(x) = B\widehat{\Theta}_2(y)B$. This immediately yields the bound $D_{op}(\widehat{\Theta}_1(x)\|I_{d_1}) \le \varepsilon\sqrt{\frac{d_1}{nd_2}}\log d_1$. To bound $D_{op}(\widehat{\Theta}_2(x)\|I_{d_2})$, use invariance of D_{op} and the approximate triangle inequality (Lemma E.1) to write

$$D_{op}(\widehat{\Theta}_2(x)\|I_{d_2}) = D_{op}(\widehat{\Theta}_2(y)\|B^{-2}) \le D_{op}(\widehat{\Theta}_2(y)\|I_{d_2}) + D_{op}(B^{-2}\|I_{d_2}).$$

We have already shown that the first term is $O(\varepsilon\sqrt{\frac{d_1}{nd_2}})$. Writing $B^{-2} = \frac{\|x\|^2}{nd_1d_2}\frac{d_2\Phi^*(I_{d_1})}{\|x\|^2}$, the second term is $O(\varepsilon\sqrt{\frac{d_2}{nd_1}})$ by Item 3 and Item 1, completing the proof. By setting the constant C in the statement of the theorem large enough compared to $1/\lambda, 1/c$, all the constraints will be satisfied. □

5. Convergence of flip-flop algorithms. In this section we prove that the flip-flop algorithms for the matrix and tensor normal models converge quickly to the MLE with high probability. Throughout this section, we will use $\log$ to denote logarithm on base 2. We begin by describing the flip-flop algorithm.

Input: Samples $x = (x_1, \dots, x_n)$ where $x_i \in \mathbb{R}^D$ is sampled from a (unknown) centered normal distribution $\mathcal{N}(0, \Sigma)$, where each entry of x_i is encoded in binary, with bit size $\leq b$.
Approximation parameter $0 < \delta < 1$, given in binary representation.
Output: $\overline{\Theta} \in \mathbb{P}$ such that $\|\nabla_a f_x(\overline{\Theta})\|_F \leq \delta$ for each $a \in [k]$
Algorithm:

1. Set $\overline{\Theta}_a = I_a$ for each $a \in [k]$.
2. For $t = 1, \dots, T = 24k \cdot \log(1/\delta)$, repeat the following:
   - Compute each component of the gradient $\nabla f_x(\overline{\Theta}_1, \dots, \overline{\Theta}_k)$, denoting $\nabla_a := \nabla_a f_x(\overline{\Theta}_1, \dots, \overline{\Theta}_k)$, and find the index $a \in [k]$ for which $\|\nabla_a\|_F$ is largest. To compute the gradient, see Remark 3.4.
   - If $\|\nabla_a\|_F < \delta$, output $\overline{\Theta}$ and return.
   - Otherwise, set $\overline{\Theta}_a \leftarrow \frac{1}{d_a} \cdot \overline{\Theta}_a^{1/2} (\rho^{(a)})^{-1} \cdot \overline{\Theta}_a^{1/2}$, where $\rho = \frac{1}{nD} \cdot \overline{\Theta}^{1/2} \left(\sum_{i=1}^n x_i x_i^\dagger \right) \overline{\Theta}^{1/2}$ and $\rho^{(a)}$ is the partial trace (see Section 3.3).

Algorithm 1: Generic flip-flop algorithm

Before we analyze the convergence of the flip-flop algorithms for the tensor and matrix normal models, we discuss the straightforward generalization of convergence of general descent methods whenever the objective function is strongly geodesically convex. The next lemma shows that any descent method which manages to decrease the value of the function with respect to the gradient by a constant (more precisely the parameter α), if starting from a sublevel set where the function is strongly convex, will converge quickly to the optimum. The proof of the lemma is the same as the one from (Franks and Moitra, 2020, Lemma 4.8), but we write it here for completeness.

LEMMA 5.1. *Let $f : \mathbb{P} \to \mathbb{R}$ be λ-strongly geodesically convex in a sublevel set containing x_0, and let $\alpha, \beta, \nu > 0$ be constants. Suppose that $\|\nabla f(x_0)\|_F^2 \leq \nu \leq \beta$. If $\{x_t\}$ is a descent sequence which satisfies*

$$f(x_{t+1}) \leq f(x_t) - \alpha \cdot \min\{\beta, \|\nabla f(x_t)\|_F^2\},$$

then in T iterations we must have an element x_r with $r \leq T$ such that

$$\|\nabla f(x_r)\|_F^2 \leq \nu \cdot 2^{-T\alpha\lambda}.$$

PROOF. Let f^* be the minimum value of the function f and let S be the sublevel set of f containing x_0 over which f is λ-strongly geodesically convex. Since $\{x_t\}$ is a descent sequence, we know that each $x_t \in S$. Since f is λ-strongly geodesically convex in S, we have

$$f^* \geq f(x) - \frac{1}{2\lambda} \cdot \|\nabla f(x)\|_F^2$$

for any $x \in S$, and in particular for any x_t in our descent sequence.

We will show that for any x_t such that $\|\nabla f(x_t)\|_F^2 \leq \varepsilon \leq \beta$, in $\ell \leq 1/\alpha\lambda$ steps we must have an element $x_{t+\ell}$ such that $\|\nabla f(x_{t+\ell})\|_F^2 \leq \varepsilon/2$. This is enough to conclude the proof of the lemma, as with this claim we see that we halve the squared norm of the gradient at every sequence of $1/\alpha\lambda$ steps and will remain less than β.

To see this, assume that $\|\nabla f(x_{t+\ell})\|_F^2 \geq \varepsilon/2$ for $0 \leq \ell \leq m$. Then, from our descent property we have

$$f(x_{t+a+1}) \leq f(x_{t+a}) - \alpha \cdot \min\{\beta, \|\nabla f(x_{t+a})\|_F^2\} \leq f(x_{t+a}) - \alpha \cdot \varepsilon/2$$

and therefore $f(x_{t+m}) \leq f(x_t) - m \cdot \alpha \cdot \varepsilon/2$.

On the other hand, our assumption that $\|\nabla f(x_t)\|_F^2 \leq \varepsilon$, together with strong geodesic convexity of f and minimality of f^* imply

$$f(x_t) - \frac{\varepsilon}{2\lambda} \leq f(x_t) - \frac{1}{2\lambda} \cdot \|\nabla f(x_t)\|_F^2 \leq f^* \leq f(x_{t+m})$$

and therefore we have

$$f(x_t) - \frac{\varepsilon}{2\lambda} \leq f(x_{t+m}) \leq f(x_t) - m \cdot \alpha \cdot \varepsilon/2$$

which implies $m \leq \frac{1}{\alpha\lambda}$. This concludes our proof. □

The following proposition shows that after the first step of the flip-flop algorithm, the first term $\frac{1}{nD}\sum_{i=1}^n x_i^T \Theta x_i$ of the log-likelihood remains unchanged by the flip-flop algorithm. Thus, the proposition below proves that $\nabla_0 f_x(\Theta) = 0$ after the first iteration.

PROPOSITION 5.2 (Trace Invariance). *Let Υ be a scaling produced by the flip-flop algorithm (Algorithm 1) in any step other than the first, and let*

$$\rho_\Upsilon := \frac{1}{nD} \cdot \Upsilon^{1/2} \cdot \sum_{i=1}^n x_i x_i^\dagger \cdot \Upsilon^{1/2}.$$

Then $\operatorname{Tr} \rho_\Upsilon = 1$.

PROOF. Suppose Θ is the scaling coming before Υ. Then there is $a \in [k]$ such that

$$\Upsilon_a = \frac{1}{d_a} \cdot \Theta_a^{1/2} (\rho_\Theta^{(a)})^{-1} \Theta_a^{1/2}$$

and $\Upsilon_b = \Theta_b$ for all $b \neq a$. Let $E_{(a)} = I_1 \otimes \cdots \otimes I_{a-1} \otimes (\rho_\Theta^{(a)})^{-1} \otimes I_{a+1} \otimes \cdots \otimes I_k$. Thus, we have:

$$\begin{aligned}
\operatorname{Tr} \rho_\Upsilon &= \frac{1}{nD} \cdot \operatorname{Tr}\left[\Upsilon^{1/2} \sum_{i=1}^n x_i x_i^\dagger \cdot \Upsilon^{1/2}\right] = \frac{1}{nD} \cdot \operatorname{Tr}\left[\Upsilon \sum_{i=1}^n x_i x_i^\dagger\right] \\
&= \frac{1}{nD} \cdot \operatorname{Tr}\left[\frac{1}{d_a} \cdot \Theta^{1/2} E_{(a)} \Theta^{1/2} \cdot \sum_{i=1}^n x_i x_i^\dagger\right] = \frac{1}{d_a} \cdot \operatorname{Tr}\left[E_{(a)} \rho_\Theta\right] \\
&= \frac{1}{d_a} \cdot \operatorname{Tr}\left[(\rho_\Theta^{(a)})^{-1} \rho_\Theta^{(a)}\right] = 1
\end{aligned}$$

□

LEMMA 5.3 (Descent Lemma). *Let $k \geq 2$. If Θ, Υ are successive scalings from the flip-flop algorithm, such that $\nabla_0 f(\Theta) = \nabla_0 f(\Upsilon) = 0$, then we have:*

$$f_x(\Upsilon) \leq f_x(\Theta) - \frac{1}{6} \cdot \min\left\{\frac{1}{d_{\max}}, \frac{\|\nabla f_x(\Theta)\|_F^2}{k-1}\right\}$$

PROOF. Let

$$\rho_\Theta := \frac{1}{nD} \cdot \Theta^{1/2} \cdot \left(\sum_{i=1}^n x_i x_i^\dagger \right) \cdot \Theta^{1/2}.$$

Additionally, let $a \in [k]$ be such that $\nabla_a := \nabla_a f_x(\Theta)$ has largest norm. As Υ is the successive scaling, we have that $\Upsilon_b = \Theta_b$ when $b \neq a$ and

$$\Upsilon_a = \frac{1}{d_a} \cdot \Theta_a^{1/2} \cdot (\rho_\Theta^{(a)})^{-1} \cdot \Theta_a^{1/2}.$$

In particular, the above means that we can write Υ in the following way:

$$\Upsilon = \Theta^{1/2} \cdot E_{(a)} \cdot \Theta^{1/2}$$

where $E_{(a)} = I_1 \otimes \cdots \otimes I_{a-1} \otimes \left(\frac{1}{d_a} \cdot (\rho_\Theta^{(a)})^{-1} \right) \otimes I_{a+1} \otimes \cdots \otimes I_k$. Moreover, $\nabla_0 f(\Theta) = 0$ implies $1 = \operatorname{Tr} \rho_\Theta$ which implies $\operatorname{Tr} \rho_\Upsilon = 1$ for ρ_Υ as in Proposition 5.2. Hence, we have:

$$\begin{aligned}
f_x(\Upsilon) &= \frac{1}{nD} \sum_{i=1}^n \langle x_i, \Upsilon x_i \rangle - \frac{1}{D} \log\det(\Upsilon) \\
&= \operatorname{Tr} \rho_\Upsilon - \frac{1}{D} \log\det(\Upsilon) \\
&= \operatorname{Tr} \rho_\Theta - \frac{1}{D} \log\det(\Theta^{1/2} \cdot E_{(a)} \cdot \Theta^{1/2}) \\
&= \operatorname{Tr} \rho_\Theta - \frac{1}{D} \left(\log\det(E_{(a)}) + \log\det(\Theta) \right) \\
&= f_x(\Theta) + \frac{1}{D} \cdot \log\det\left(E_{(a)}^{-1} \right) \\
&= f_x(\Theta) + \frac{1}{d_a} \cdot \log\det\left(d_a \cdot \rho_\Theta^{(a)} \right)
\end{aligned}$$

Lemma 5.1 from Garg et al. (2019) states that for any d-dimensional PSD matrix Z of trace d, the following inequality holds:

$$\log\det(Z) \leq \max\left\{ -\frac{\|Z - I_d\|_F^2}{6}, -\frac{1}{6} \right\}.$$

Since $\operatorname{Tr} \rho_\Theta^{(a)} = \operatorname{Tr} \rho_\Theta = 1$, applying the above with $Z = d_a \cdot \rho_\Theta^{(a)}$ implies

$$\begin{aligned}
\frac{1}{d_a} \cdot \log\det\left(d_a \cdot \rho_\Theta^{(a)} \right) &\leq \max\left\{ -\frac{\|d_a \cdot \rho_\Theta^{(a)} - I_a\|_F^2}{6 d_a}, -\frac{1}{6 d_a} \right\} \\
&= \max\left\{ -\frac{\|\nabla_a\|_F^2}{6}, -\frac{1}{6 d_a} \right\} \\
&\leq \max\left\{ -\frac{\|\nabla f_x(\Theta)\|_F^2}{6(k-1)}, -\frac{1}{6 d_{\max}} \right\}
\end{aligned}$$

The equality is from Lemma 3.3 and Remark 3.4. In the last inequality we used that $\nabla_0 f(\Theta) = 0$ and that a is the index (out of $k-1$ indices where the gradient is nonzero) where the gradient has largest norm. □

We now have all the lemmas we need to prove that, given appropriate initial conditions on the input samples, the flip-flop algorithm will converge quickly to the MLE.

LEMMA 5.4 (Fast Convergence from Initial Conditions). *If the samples $x_1, \dots, x_n \in \mathbb{R}^D$ and the numbers $\lambda, r, \nu > 0$ satisfy the following conditions:*

1. *f_x is λ-strongly geodesically convex on the ball $B_r(I_D)$, and*
2. $\|\nabla f_x(I_D)\|_F \leq \nu^{1/2} < \dfrac{r\lambda}{2} \leq \min\left\{1, \sqrt{\frac{k}{d_{\max}}}\right\}$.

Then, in $T = \dfrac{12k}{\lambda} \cdot \log\left(\dfrac{4\nu^{1/2} d_{\max}^{1/2}}{\varepsilon\lambda}\right)$ iterations, Algorithm 1 outputs an estimator $\overline{\Theta}$ such that the equal-determinant factors $\widehat{\Theta}_a, \overline{\Theta}_a$ satisfy

$$D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon, \quad \forall\, a \in [k].$$

PROOF. The initial conditions above imply that Lemma 3.1 applies, and therefore we have that the sublevel set $\{\Upsilon \mid f(\Upsilon) \leq f(I_D)\}$ is contained in the ball $B_r(I_D)$. In particular, the above condition on the sublevel set implies that Lemma 5.3 applies, and thus we have that each step of the flip-flop algorithm will decrease the value of the objective function in accordance with the requirements of Lemma 5.1, with parameters $\alpha = 1/6k$, $\beta = \frac{k}{d_{\max}}$ and $\nu > 0$.

Thus, after $T = \dfrac{12k}{\lambda} \cdot \log(\nu^{1/2}/\delta)$ steps, Lemma 5.1 guarantees us that we will encounter a point $\overline{\Theta}$ such that

$$\|\nabla f_x(\overline{\Theta})\|_F \leq \delta.$$

Setting $\delta = \dfrac{\lambda\varepsilon}{4d_{\max}^{1/2}}$, when we find such a point with gradient $\leq \delta$, Corollary 3.2 implies that for each $a \in [k]$, the component-wise distance from $\overline{\Theta}$ to $\widehat{\Theta}$ is bounded by

$$D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon.$$

This in particular implies that $D_{op}(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon$. □

5.1. *Tensor flip-flop: Proof of Theorem 2.5.* We are now ready to state the tensor flip-flop algorithm and prove its fast convergence to the MLE.

Input: Samples $x = (x_1, \dots, x_n)$ where $x_i \in \mathbb{R}^D$ is sampled from a (unknown) centered normal distribution $\mathcal{N}(0, \Sigma)$, where each entry of x_i is encoded in binary, with bit size $\leq b$.
Approximation parameter $0 < \varepsilon < 1$, given in binary representation.

Output: $\overline{\Theta} \in \mathbb{P}$ such that $D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) < \varepsilon$, for each $a \in [k]$, where $\widehat{\Theta}$ is the MLE for the precision matrix of Σ.

Algorithm:

1. Set $\overline{\Theta}_a = I_a$ for each $a \in [k]$, and $\delta = \dfrac{\varepsilon}{8 d_{\max}^{1/2}}$.
2. For $t = 1, \dots, T = 24kd_{\max} \cdot \log(1/\delta)$, repeat the following:
   - Compute each component of the gradient $\nabla f_x(\overline{\Theta}_1, \dots, \overline{\Theta}_k)$, denoting $\nabla_a := \nabla_a f_x(\overline{\Theta}_1, \dots, \overline{\Theta}_k)$, and find the index $a \in [k]$ for which $\|\nabla_a\|_F$ is largest.
   - If $\|\nabla_a\|_F < \delta$, output $\overline{\Theta}$ and return.
   - Otherwise, set $\overline{\Theta}_a \leftarrow \dfrac{1}{d_a} \cdot \overline{\Theta}_a^{1/2} (\rho^{(a)})^{-1} \cdot \overline{\Theta}_a^{1/2}$, where $\rho = \dfrac{1}{nD} \cdot \overline{\Theta}^{1/2} \left(\sum_{i=1}^n x_i x_i^\dagger \right) \overline{\Theta}^{1/2}$.

Algorithm 2: Tensor flip-flop algorithm

LEMMA 5.5 (Initial Conditions for Tensor Normal Model). *There exist absolute constants $\Gamma > 0$, $4 \geq \gamma > 0$ such that the following holds. When the number of samples $n \geq \Gamma \cdot k^2 \cdot d_{\max}^3 / D$, with probability at least $1 - k^2 \cdot \left(\dfrac{\sqrt{nD}}{kd_{\max}} \right)^{-\Omega(d_{\min})} - 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}$ we have that the following conditions hold:*

1. $\|\nabla f_x(I)\|_F < \dfrac{\gamma}{4\sqrt{(k+1)d_{\max}}}$
2. *f_x is $\frac{1}{2}$-strongly geodesically convex at $B_r(I_D)$, where $r = \dfrac{\gamma}{\sqrt{(k+1)d_{\max}}}$*

PROOF. The lemma follows from the observation that Proposition 3.5 implies condition 1, and Theorem 3.14 implies condition 2. So all we need to do is to check the parameters.

By Theorem 3.14, if we set $\gamma = c$ and if the number of samples $n \geq Ck\dfrac{d_{\max}^2}{D}$, where $c, C > 0$ are the constants from Theorem 3.14, then the second condition fails to hold with probability at most

$$k^2 \cdot \left(\frac{\sqrt{nD}}{kd_{\max}} \right)^{-\Omega(d_{\min})}.$$

By Proposition 3.5 with parameter $\varepsilon = \frac{\gamma}{100kd_{\max}^{1/2}}$, if the number of samples satisfies $n \geq \dfrac{10^4 k^2 d_{\max}^3}{\gamma^2 \cdot D}$ then the first condition fails to hold probability at most

$$2k \cdot \exp\left(-\frac{nD\gamma^2}{128(k+1)d_{\max}^2} \right) = 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}.$$

Letting $\Gamma = \max\{10^4/\gamma^2, C\}$, having $n \geq \Gamma k^2 d_{\max}^3 / D$ samples gives a sample upper bound that holds for both situations above. Thus, by the union bound, one of the conditions 1

or 2 fails with probability at most

$$k^2 \cdot \left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})} + 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}.$$

This concludes the proof. □

THEOREM 2.5 (Tensor flip-flop convergence, restated). *If $\widehat{\Theta}$ denotes the MLE estimator for Θ, then provided $n = \Omega(k^2 \cdot d_{\max}^3/D)$, the flip-flop algorithm computes $\overline{\Theta}$ with*

$$D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon$$

in $O(k\log(1/\varepsilon))$ iterations with probability at least

$$1 - k^2 \cdot \left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})} - 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}.$$

PROOF. When the number of samples is $n = \Omega(k^2 \cdot d_{\max}^3/D)$, with probability

$$1 - k^2 \cdot \left(\frac{\sqrt{nD}}{kd_{\max}}\right)^{-\Omega(d_{\min})} - 2k \cdot e^{-\Omega(nD/kd_{\max}^2)}$$

we have that the hypothesis of Lemma 5.5 applies, which implies that there exists a constant $4 \geq \gamma > 0$ such that our objective function f_x is $\frac{1}{2}$-strongly geodesically convex at a ball $B_r(I)$ for $r = \dfrac{\gamma}{\sqrt{(k+1)d_{\max}}}$ and $\|\nabla f_x(I)\|_F \leq \nu^{1/2} := \dfrac{\gamma}{4\sqrt{(k+1)d_{\max}}} < \sqrt{\frac{k}{d_{\max}}}$.

Thus, by Lemma 5.4, we have that in $T = 24k \cdot \log(8\nu d_{\max}^{1/2}/\varepsilon) = O(k\log(1/\varepsilon))$ iterations the flip-flop algorithm converges to an estimator such that $D_F(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) \leq \varepsilon$ for all $a \in [k]$. This concludes the proof. □

5.2. *Matrix flip-flop: Proof of Theorem 2.6.* We are now ready to state the matrix flip-flop algorithm and prove its fast convergence to the MLE. The proof strategy of this section is a bit different from the tensor normal model case, as now the number of samples is not large enough to guarantee that the initial conditions from Lemma 5.4 will apply with high probability.

However, we can proceed as in Franks and Moitra (2020) and use the results from Kwok, Lau and Ramachandran (2019) to show that the MLE is in a constant size operator norm ball around the identity. Hence, by Theorem 3.14, the log-likelihood is strongly geodesically convex in a small geodesic ball around the MLE. This implies (Franks and Moitra, 2020, Lemma 4.7) that any point with sufficiently small gradient of the log-likelihood function is contained in a sublevel set on which the log-likelihood is strongly geodesically convex. Such a point can be found by applying the flip-flop algorithm for polynomially many iterations, at which point Lemma 5.1 applies to yield an ε-minimizer in $O(\log(1/\varepsilon))$ further iterations.

Input: Samples $x = (x_1, \dots, x_n)$ where $x_i \in \mathbb{R}^{d_1 \times d_2}$ is sampled from a (unknown) centered normal distribution $\mathcal{N}(0, \Sigma)$, where each entry of x_i is encoded in binary, with bit size $\leq b$.
Approximation parameter $\varepsilon > 0$, given in binary representation.
Output: $\overline{\Theta} \in \mathbb{P}$ s.th. $D_{\mathrm{op}}(\widehat{\Theta}_a \,||\, \overline{\Theta}_a) < \varepsilon$, for $a \in \{1, 2\}$, where $\widehat{\Theta}$ is the MLE for the precision matrix Θ.
Algorithm:

1. Let γ, λ be the constants from Lemma 5.6. Set $\overline{\Theta}_a = I_a$ for each $a \in \{1, 2\}$, and $\delta = \dfrac{\varepsilon\lambda}{4\sqrt{d_{\max}}}$.
2. For $t = 1, \dots, T = 60 d_{\max} \cdot \left(\dfrac{16}{\lambda\gamma}\right)^2 + \dfrac{24 d_{\max}}{\lambda} \cdot \log(1/\delta)$, repeat the following:
   - Compute each component of the gradient $\nabla f_x(\overline{\Theta}_1, \overline{\Theta}_2)$, denoting $\nabla_a := \nabla_a f_x(\overline{\Theta}_1, \overline{\Theta}_2)$, and find the index $a \in \{1, 2\}$ for which $\|\nabla_a\|_{op}$ is non-zero.
   - If $\|\nabla_a\|_{op} < \delta$, output $\overline{\Theta}$ and return.
   - Otherwise, set $\overline{\Theta}_a \leftarrow \dfrac{1}{d_a} \cdot \overline{\Theta}_a^{1/2} \cdot (\rho^{(a)})^{-1} \cdot \overline{\Theta}_a^{1/2}$, where $\rho = \dfrac{1}{n d_1 d_2} \overline{\Theta}^{1/2} \cdot \left(\sum_{i=1}^n x_i x_i^\dagger\right) \cdot \overline{\Theta}^{1/2}$.

Algorithm 3: Matrix flip-flop algorithm

LEMMA 5.6 (Initial Conditions Matrix Normal Model). *There exist absolute constants $\Gamma > 0$ and $\gamma, \lambda \in (0, 1]$ such that the following holds. When the number n of samples satisfies*

$$n \geq \Gamma \cdot \frac{d_{\max}}{d_{\min}} \cdot \max\left\{\log \frac{d_{\max}}{d_{\min}}, \frac{\log^2 d_{\min}}{\varepsilon^2}\right\},$$

then with probability at least $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$ the following conditions hold:

1. $|\nabla_0 f_x(I)| \leq \frac{\gamma}{2}$
2. *The MLE's* $\widehat{\Theta}_1, \widehat{\Theta}_2$ *satisfy* $\|\widehat{\Theta}_i - I_{d_i}\|_{op} \leq \gamma/2$, *for* $i \in \{1, 2\}$
3. f_x *is* λ*-strongly geodesically convex at any* $\Theta \in \mathbb{P}$ *such that* $\|\log \Theta\|_{op} \leq \gamma$.

PROOF. By Proposition 3.7 and the fact that $|\nabla_0 f_x(I)| = |1 - \|x\|^2/nD|$, Condition 1 holds with probability $1 - 2e^{-\gamma^2 nD/8}$.

By Proposition 3.5, if $n \geq \dfrac{4 d_{\max}}{d_{\min}\gamma^2}$, condition 1 fails to hold with probability at most $6e^{-\gamma^2 n d_{\min}/32}$. If Γ is larger than the universal constant C of Theorem 2.3, this theorem implies that condition 2 fails to hold with probability at most $O(e^{-\Omega(d_1\varepsilon^2)})$. Finally, if Γ is

larger than the universal constant C of Corollary 4.5 and γ is smaller than the universal constant c from Corollary 4.5, this corollary implies that there is a universal constant $\lambda \in (0,1)$ such that condition 3 fails to hold with probability at most $e^{-\Omega(d_{\min}\varepsilon^2)}$. Thus, choosing Γ larger than the foregoing universal constants, as well as $\Gamma > 4/\gamma^2$, by a union bound we satisfy the hypotheses for all three results with the desired probability. □

To prove that flip-flop converges once the initial conditions are satisfied, we need the following general lemma (Franks and Moitra, 2020, Lemma 4.7) on strongly geodesically convex functions, which tells us that once the gradient is small then the point must be inside a sublevel set of our function which is contained in a ball where our function is strongly convex. Though their lemma is stated for the manifold of positive definite matrices of determinant one, the proof uses no specific properties of this manifold beyond the fact that it is a Hadamard manifold. Thus their lemma holds for the manifold $\mathbb{P}$ as well.

LEMMA 5.7 (Franks and Moitra (2020)). *Suppose that $f : \mathbb{P} \to \mathbb{R}$ is geodesically λ-strongly convex on the geodesic ball of radius κ about θ, and that $\nabla f(\theta) = 0$. If $\|\nabla f(\Upsilon)\|_F < \lambda\kappa/8$, then Υ is contained in a sublevel set of f on which f is geodesically λ-strongly convex.*

Now we must show that the flip-flop algorithm reaches a point with small enough gradient relatively quickly, which is given by the following lemma, which follows the analysis given by Garg et al. (2019):

LEMMA 5.8. *Given $\delta > 0$ and samples $x = (x_1, \dots, x_n)$, let $f^* := \min_{\overline{\Theta} \in \mathbb{P}} f_x(\overline{\Theta})$ be the minimizer of the log-likelihood function. The flip-flop algorithm, in at most*

$$T = 12(f_x(I_D) - f^* + |\nabla_0 f_x(I)|) \cdot \max\left\{\frac{d_{\max}}{2}, \frac{1}{\delta^2}\right\} \text{ iterations,}$$

outputs $\overline{\Theta}$ such that $\|\nabla f_x(\overline{\Theta})\|_F < \delta$.

PROOF. Let $\overline{\Theta}^{(i)}$ be the iterates of the flip-flop algorithm, where $\overline{\Theta}^{(0)} = I_D$. After we perform the first normalization step of the flip-flop algorithm, we obtain a scaling $\overline{\Theta}^{(1)}$ such that $\nabla_0 f_x(\overline{\Theta}^{(1)}) = 0$. By Proposition 5.2, after each subsequent step, that is $\overline{\Theta}^{(i)}$ for $i \geq 2$, we maintain that $\nabla_0 f_x(\overline{\Theta}^{(i)}) = 0$. Thus, Lemma 5.3 applies and we obtain that

$$f_x(\overline{\Theta}^{(1)}) - f_x(\overline{\Theta}^{(T)}) \geq \frac{1}{12}\sum_{i=1}^{T-1} \min\left\{\frac{2}{d_{\max}}, \|\nabla f_x(\overline{\Theta}^{(i)})\|_F^2\right\}.$$

Let $\alpha^2 := \frac{1}{nD}\sum_{i=1}^n x_i^\dagger x_i$. Lemma 5.3 and $\nabla_0 f_x(\alpha I) = 0$ imply $f_x(\alpha I_D) - f_x(\overline{\Theta}^{(1)}) \geq 0$. Moreover, since $f_x(I_D) - f_x(\overline{\Theta}^{(1)}) = f_x(I_D) - f_x(\alpha I_D) + f_x(\alpha I_D) - f_x(\overline{\Theta}^{(1)})$, we have

$$f_x(I_D) - f_x(\overline{\Theta}^{(1)}) \geq f_x(I_D) - f_x(\alpha I_D) \geq -|\nabla_0 f_x(I)|.$$

The above inequalities imply that

$$f_x(I_D) - f^* \geq f_x(I_D) - f_x(\overline{\Theta}) \geq -|\nabla_0 f_x(I)| + \frac{1}{12}\sum_{i=1}^{T-1} \min\left\{\frac{2}{d_{\max}}, \|\nabla f_x(\overline{\Theta}^{(i)})\|_F^2\right\}.$$

Thus, for some $1 \leq t \leq T = 12(f_x(I_D) - f^* + |\nabla_0 f_x(I)|) \cdot \max\left\{\frac{d_{\max}}{2}, \frac{1}{\delta^2}\right\}$ steps, we must reach a scaling $\overline{\Theta}^{(t)}$ such that $\|\nabla f_x(\overline{\Theta}^{(t)})\|_F^2 < \delta^2$. □

We are now ready to prove our main theorem of this subsection: it says that under the hypotheses of Theorem 2.3, the flip-flop converges exponentially fast to the minimizer.

THEOREM 2.6 (Matrix flip-flop convergence, restated). *Let* $(\widehat{\Theta}_1, \widehat{\Theta}_2)$ *denote the MLE for* (Θ_1, Θ_2). *There exists a universal constant* $\Gamma > 0$ *such that when given*

$$n \geq \Gamma \cdot \frac{d_{\max}}{d_{\min}} \cdot \max\left\{\log\left(\frac{d_{\max}}{d_{\min}}\right), \frac{\log^2 d_{\min}}{\varepsilon^2}\right\}$$

samples in the matrix normal model, the flip-flop algorithm computes $(\overline{\Theta}_1, \overline{\Theta}_2)$ *with*

$$D_{op}(\overline{\Theta}_a, \widehat{\Theta}_a) \leq \varepsilon$$

for $a \in \{1,2\}$ *in* $O\left(d_{\max} + \log(1/\varepsilon)\right)$ *iterations with probability at least* $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$.

PROOF. If we take Γ to be the universal constant according to Lemma 5.6, with probability at least $1 - e^{-\Omega(d_{\min}\varepsilon^2)}$ we have that the conditions of Lemma 5.6 are satisfied. Thus, there exist constants $\lambda, \gamma \in (0,1]$ such that:

1. $|\nabla_0 f_x(I)| \leq \frac{\gamma}{2}$
2. The MLE's $\widehat{\Theta}_1, \widehat{\Theta}_2$ satisfy $\left\|\widehat{\Theta}_i - I_{d_i}\right\|_{op} \leq \gamma/2$, for $i \in \{1,2\}$
3. f_x is λ-strongly geodesically convex at any $\Theta \in \mathbb{P}$ such that $\left\|\log \Theta\right\|_{op} \leq \gamma$.

In particular, we have that our function f_x is lower bounded by

$$f^* := f_x(\widehat{\Theta}_1, \widehat{\Theta}_2) = 1 - \frac{1}{D}\log\det(\widehat{\Theta}_1 \otimes \widehat{\Theta}_2) \geq 1 - \frac{1}{D}\log\det((1+\gamma/2)I_D) \geq -\gamma/2,$$

and we also know that $f_x(I_D) = 1 + \nabla_0 f_x(I_D) \leq 1 + \gamma/2$. Thus, by Lemma 5.8 with parameter $\lambda\gamma/32\sqrt{d_{\max}}$, we know that in at most $30 \cdot \max\left\{d_{\max} \cdot (32/\lambda\gamma)^2, d_{\max}/2\right\} = O(d_{\max})$ steps, we will find a point $\overline{\Theta}$ with $\left\|\nabla f_x(\overline{\Theta})\right\|_F < \lambda\gamma/32\sqrt{d_{\max}}$.

For the $\overline{\Theta}$ we just obtained, we will now show that Lemma 5.7 applies. Note that conditions 2 and 3 imply that f_x is λ-strongly geodesically convex in an operator norm ball of radius $\gamma/2$ around the MLE $\widehat{\Theta}$. Thus, Remark 3.15 implies that f_x is λ-strongly geodesically convex in the ball $B_\kappa(\widehat{\Theta})$ with $\kappa = \dfrac{\gamma}{2\sqrt{3d_{\max}}}$. Since $\left\|\nabla f_x(\overline{\Theta})\right\|_F < \dfrac{\lambda\gamma}{32\sqrt{d_{\max}}} < \kappa\lambda/8$, the conditions of Lemma 5.7 apply.

Therefore, f_x is λ-strongly convex in a ball that contains the sublevel set $\{\Upsilon \mid f_x(\Upsilon) \leq f_x(\overline{\Theta})\}$. By Lemma 5.3, Lemma 5.1 applies with $\alpha = \frac{1}{6}, \beta = 1/d_{\max}$, and $0 \leq \nu \leq \beta$. Thus in $T = \dfrac{6}{\lambda}\log_2(\nu/\delta^2)$ we obtain a $\overline{\Theta}'$ such that $\left\|\nabla_x(\overline{\Theta}')\right\|_F \leq \delta$. Setting $\delta = \dfrac{\varepsilon\lambda}{4\sqrt{d_{\max}}}$, Corollary 3.2 implies that $D_{op}(\widehat{\Theta}_a \mid\mid \overline{\Theta}'_a) \leq \varepsilon$ for $a \in \{1,2\}$. In particular, Algorithm 3 correctly returns a scaling ε-close to the MLE. The number of iterations follows from plugging in the chosen values for δ, ν. □

6. Lower bounds. Here we discuss known lower bounds for estimating unstructured precision matrices (i.e., the case $k = 1$ of the tensor normal model). Afterwards we prove a new lower bound on the matrix normal model.

6.1. *Lower bounds for unstructured precision matrices.* Here we briefly recall and, for completeness, prove well-known lower bounds on the accuracy of any estimator for the precision matrix in the Frobenius and operator error from independent samples of a Gaussian. The lower bounds follow from Fano's method and the relationship between the Frobenius

error and the relative entropy (which is proportional to Stein's loss). Informally, both bounds imply no estimator for a $d \times d$ precision matrix from n samples can have accuracy smaller than $d/\sqrt{n}$ (resp. $\sqrt{d/n}$) in Frobenius error or relative Frobenius error (resp. operator norm error or relative operator norm error) with probability more than $1/2$.

PROPOSITION 6.1 (Frobenius and operator error). *There is $c > 0$ such that the following holds. Let $X = X_\Theta \in \mathbb{R}^{dn}$ denote n independent samples from a Gaussian with precision matrix $\Theta \in \mathrm{PD}(d)$. Consider any estimator $\widehat{\Theta} = \widehat{\Theta}(X)$ for the precision Θ, and let $B \subset \mathrm{PD}(d)$ denote the ball about I_d of radius $1/2$ in the operator norm.*

1. *Let $\delta^2 = c \min\left\{1, d^2/n\right\}$. Then*

$$\sup_{\Theta \in B} \Pr\left[\|\widehat{\Theta} - \Theta\|_F \geq \delta\right] \geq \frac{1}{2}. \tag{6.1}$$

2. *Let $\delta^2 = c \min\{1, d/n\}$. Then*

$$\sup_{\Theta \in B} \Pr\left[\|\widehat{\Theta} - \Theta\|_{op} \geq \delta\right] \geq \frac{1}{2}. \tag{6.2}$$

As a consequence, we have

$$\sup_{\Theta \in B} \mathbb{E}[\|\widehat{\Theta} - \Theta\|_F^2] = \Omega\left(\min\left\{\frac{d^2}{n}, 1\right\}\right)$$

$$\text{and } \sup_{\Theta \in B} \mathbb{E}[\|\widehat{\Theta} - \Theta\|_{op}^2] = \Omega\left(\min\left\{\frac{d}{n}, 1\right\}\right).$$

The proof uses Fano's inequality with mutual information bounded by relative entropy, as in Yang and Barron (1999).

LEMMA 6.2 (Fano's method). *Let $\{P_i : i \in [m]\}$ be a finite set of probability distributions over a set $\mathcal{X}$, and let $T : \mathcal{X} \to [m]$ be an estimator for i from a sample of P_i. Then*

$$\max_{i \in [m]} \Pr_{x \sim P_i}[T(x) \neq i] \geq 1 - \frac{-\log 2 + \max_{i,j \in [m]} D_{KL}(P_i || P_j)}{\log m}.$$

PROOF OF PROPOSITION 6.1. We first prove Eq. (6.1), the lower bound on estimation in the Frobenius norm. We begin by the standard reduction from estimation to testing; let V_0 be a 1-separated set in the Frobenius ball B_F of radius 1 in the $d \times d$ symmetric matrices, i.e. the set $B_F = \{A : A \text{ Symmetric}, \|A\|_F \leq 1\}$. We may take V_0 to have cardinality $m \geq 2^{d(d+1)/2}$ because B_F is a Euclidean ball of radius 1 in the linear subspace of $d \times d$ symmetric matrices, which has dimension $d(d+1)/2$, and hence the greedy $1/2$-packing of B_F in the Frobenius norm has cardinality at least $2^{d(d+1)/2}$. Let $0 \leq \delta \leq 1/2$, and let $V = I_d + \delta V_0 = \{I_d + \delta v : v \in V_0\}$. Write $V = \{\Theta_1, \ldots, \Theta_m\}$. Note that V is contained within the operator norm ball B. Let P_i be the distribution of $X(\Theta_i)$ on $\mathbb{R}^{dn}$, i.e $\mathcal{N}(0, \Theta_i^{-1})^{\otimes n}$. Define the estimator T by

$$T(x) = \arg\min_{i \in [m]} \|\Theta_i - \widehat{\Theta}(x)\|_F,$$

so that $\Pr[T(X) = i] \geq \Pr[\|\widehat{\Theta} - \Theta_i\|_F \leq \delta/2]$ because V is δ-separated. In order to apply the local Fano's method, we use the well-known fact that $D_{KL}(P_i||P_j) = nD_{KL}(\mathcal{N}(0, \Theta_i^{-1}) \,||\, \mathcal{N}(0, \Theta_j^{-1})) = O(nD_F(\Theta_j \,||\, \Theta_i)^2)$ when $\Theta_i^{-1}\Theta_j$ has eigenvalues uniformly bounded away from zero by Proposition E.2. This condition on the eigenvalues holds because $I_d/2 \preceq \Theta_j, \Theta_j \preceq 3I_d/2$ for $i, j \in [m]$ by our assumption that $\delta \leq 1/2$. Moreover, for

$i \in [m]$, we have $\|\Theta_i^{-1}\|_{op} \leq 2$ and so $D_F(\Theta_j||\Theta_i) = O(\|\Theta_i - \Theta_j\|_F) = O(\delta)$. We now have $D_{KL}(P_i||P_j) \leq Cn\delta^2$ for some absolute constant C. By Fano's lemma,

$$\min_{i\in[m]} \Pr_{x\sim P_i}[T(X) = i] \leq \frac{-\log 2 + Cn\delta^2}{d(d+1)(\log 2)/2}.$$

If $\delta^2 = c\min\{\frac{d^2}{n}, 1\}$, the right-hand side is bounded by $\frac{1}{2}$ and the assumption $\delta \leq 1/2$ is satisfied provided c is a small enough absolute constant. On the other hand, we showed $\Pr_{x\sim P_i}[T(X) = i] \geq \Pr[\|\widehat{\Theta} - \Theta_i\|_F \leq \delta/2]$. Thus,

$$\min_{i\in[m]} \Pr[\|\widehat{\Theta} - \Theta_i\|_F \leq \delta/2] \leq 1/2.$$

Because $V \subset B$, this proves Eq. (6.1). To obtain Eq. (6.2), the lower bound in operator norm, instead start with a packing V_0 of the unit operator norm ball of cardinality $m \geq 2^{d(d+1)/2}$. We modify the proof by bounding $D_{KL}(P_i||P_j) = O(n\|\Theta_i - \Theta_j\|_F^2) = O(nd\|\Theta_i - \Theta_j\|_{op}^2) \leq Cnd\delta^2$. Proceeding as before, we find that for $\delta = c\min\{\frac{d}{n}, 1\}$,

$$\min_{i\in[m]} \Pr[\|\widehat{\Theta} - \Theta_i\|_{op} \leq \delta/2] \leq 1/2.$$

Again, we have $\Theta_1, \ldots, \Theta_m \in B$, so Eq. (6.2) follows. □

We remark that the above proof shows the necessity of a scale-invariant dissimilarity measure to obtain error bounds that are independent of the ground truth precision matrix Θ. Indeed, replacing the packing V by κV for some $\kappa \to \infty$ shows that $\sup_{\Theta\in\kappa B} \Pr\left[\|\widehat{\Theta} - \Theta\|_F \geq \kappa\delta\right] \geq \frac{1}{2}$. That is, no fixed bound can be obtained with probability $1/2$.

We now use the result just obtained to prove bounds on the relative Frobenius and operator error. Because $I_d/2 \preceq \Theta \preceq 3I_d/2$ for $\Theta \in B$, the bounds $\|\Theta - \widehat{\Theta}\|_F \leq \|\Theta\|_{op} D_F(\widehat{\Theta}||\Theta)$ and $\|\Theta - \widehat{\Theta}\|_{op} \leq \|\Theta\|_{op} D_{op}(\widehat{\Theta}||\Theta)$ together with Proposition 6.1 imply the following corollary.

COROLLARY 6.3 (Relative Frobenius and operator error). *There is $c > 0$ such that the following holds for $X, \widehat{\Theta}, B$ as in Proposition 6.1.*

1. *Let $\delta^2 = c\min\{1, d^2/n\}$. Then*

$$\sup_{\Theta\in B} \Pr\left[D_F(\widehat{\Theta}||\Theta) \geq \delta\right] \geq \frac{1}{2}. \tag{6.3}$$

2. *Let $\delta^2 = c\min\{1, d/n\}$. Then*

$$\sup_{\Theta\in B} \Pr\left[D_{op}(\widehat{\Theta}||\Theta) \geq \delta\right] \geq \frac{1}{2}. \tag{6.4}$$

As a consequence, we have

$$\sup_{\Theta\in B} \mathbb{E}[D_F(\widehat{\Theta}||\Theta)^2] = \Omega\left(\min\left\{\frac{d^2}{n}, 1\right\}\right)$$

$$\text{and } \sup_{\Theta\in B} \mathbb{E}[D_{op}(\widehat{\Theta}||\Theta)^2] = \Omega\left(\min\left\{\frac{d}{n}, 1\right\}\right).$$

6.2. *Lower bounds for the matrix normal model.* If Θ_2 is known, then we can compute $(I \otimes \Theta_2^{1/2})X$, which is distributed as nd_2 independent samples from a Gaussian with precision matrix Θ_1. In this case, one can estimate Θ_1 in operator norm with an RMSE rate of $O(\sqrt{d_1/nd_2})$. One could hope that this rate holds for Θ_1 even when Θ_2 is not known. Here we show that, to the contrary, the rate for Θ_1 cannot be better than $O(\sqrt{d_1/n\min(nd_1, d_2)})$. Thus, for $d_2 > nd_1$, it is impossible to estimate Θ_1 as well as one could if Θ_2 were known. Note that, in this regime, there is no hope of recovering Θ_2 even if Θ_1 is known.

THEOREM 6.4 (Lower bound for matrix normal models). *There is $c > 0$ such that the following holds. Let $\widehat{\Theta}_1$ be any estimator for Θ_1 from a tuple X of n samples of the matrix normal model with precision matrices Θ_1, Θ_2. Let $B \subset \mathrm{PD}(d_1)$ denote the ball about I_{d_1} of radius $1/2$ in the operator norm.*

1. *Let $\delta^2 = c\min\left\{1, \frac{d_1^2}{n\min\{nd_1, d_2\}}\right\}$. Then*

$$\sup_{\Theta_1 \in B, \Theta_2 \in \mathrm{PD}(d_2)} \Pr\left[D_F(\widehat{\Theta}_1 || \Theta_1) \geq \delta\right] \geq \frac{1}{2}. \tag{6.5}$$

2. *Let $\delta^2 = c\min\left\{1, \frac{d_1}{n\min\{nd_1, d_2\}}\right\}$. Then*

$$\sup_{\Theta_1 \in B, \Theta_2 \in \mathrm{PD}(d_2)} \Pr\left[D_{op}(\widehat{\Theta}_1 || \Theta_1) \geq \delta\right] \geq \frac{1}{2}. \tag{6.6}$$

As a consequence, we have

$$\sup_{\Theta_1 \in B, \Theta_2 \in \mathrm{PD}(d_2)} \mathbb{E}[D_F(\widehat{\Theta}_1 || \Theta_1)^2] = \Omega\left(\min\left\{\frac{d_1^2}{n\min\{nd_1, d_2\}}, 1\right\}\right)$$

$$\textit{and} \sup_{\Theta_1 \in B, \Theta_2 \in \mathrm{PD}(d_2)} \mathbb{E}[D_{op}(\widehat{\Theta}_1 || \Theta_1)^2] = \Omega\left(\min\left\{\frac{d_1}{n\min\{nd_1, d_2\}}, 1\right\}\right).$$

Intuitively, the above theorem holds because we can choose Σ_2 to zero out all but nd_1 columns of each X_i, which allows access to at most $n \cdot nd_1$ samples from a Gaussian with precision Θ_1. However, this does not quite work because Σ_2 would not be invertible and hence the precision matrix Θ_2 would not exist. We must instead choose Σ_2 to be approximately equal to a random projection of rank nd_1. The resulting construction allows us to deduce the same lower bounds for estimating Θ_1 as the Gaussian case with at most $n\min\{d_2, nd_1\}$ independent samples.

One might ask why the rank of the random projection cannot be taken to be even less than nd_1, yielding an even stronger bound. If the rank is less than nd_1, then the support of Σ_2 can be estimated. This would allow one to approximately diagonalize Σ_2 so that the n samples can be treated as nd_2 independent samples in $\mathbb{R}^{d_1}$, yielding the rate $\sqrt{d_1/nd_2}$ using, e.g., Tyler's M estimator. We now state the lower bound.

LEMMA 6.5. *Suppose $d_2 > nd_1$. Let X denote a tuple of n samples from the matrix normal model with precision matrices Θ_1, Θ_2. Let Y be a tuple of $n\min\{nd_1, d_2\}$ Gaussians on $\mathbb{R}^{d_1}$ with precision matrix Θ_1. Let $\widehat{\Theta}_1$ be any estimator for Θ_1. For every $\delta > 0$, there is a distribution on Θ_2 and an estimator $\tilde{\Theta}$ such that the distribution of $\widehat{\Theta}_1(X)$ and the distribution of $\tilde{\Theta}(Y)$ differ by at most δ in total variation distance.*

PROOF. If $d_2 \leq nd_1$, then setting $\Theta_2 = I(d_2)$ shows that $\widehat{\Theta}_1$ has access to precisely nd_2 samples from a Gaussian $\mathbb{R}^{d_1}$ with precision matrix Θ_1. Thus we may take $\tilde{\Theta} = \widehat{\Theta}_1$ in that case, completing the proof. The harder case is $d_2 > nd_1$.

Let B be any $d_2 \times d_2$ matrix such that the last $d_2 - nd_1$ columns are zero. Given access to the tuple X of n samples $\sqrt{\Sigma_1}X_iB^T$, where X_i are i.i.d Gaussian $d_1 \times d_2$ matrices, clearly $\widehat{\Theta}_2$ has access to at most n^2d_1 samples of the Gaussian on $\mathbb{R}^{d_1}$ with precision matrix Θ_1 because X_iB^T depends only on the first d_1 columns of each X_i.

However, we must supply *invertible* B in order for $\Theta_2 = (BB^T)^{-1}$ to exist. Let $\delta \geq 0$. Let B_δ be the random matrix obtained by choosing the first nd_1 columns of B_δ uniformly at random among the collections of nd_1 orthonormal vectors in $\mathbb{R}^{d_2}$. Let the remaining entries be i.i.d uniform in $[-\delta, \delta]$ (the precise distribution of the remaining entries does not matter as long as they are independent, continuous, and small). Let $Y_\delta := (\sqrt{\Sigma_1}X_1B_\delta^T, \ldots, \sqrt{\Sigma_1}X_nB_\delta^T)$ denote the resulting random variable with B_δ and X chosen independently. If $\delta = 0$, then, by the argument above, with access to the random variable $Y_\delta := (\sqrt{\Sigma_1}X_1B_\delta^T, \ldots, \sqrt{\Sigma_1}X_nB_\delta^T)$ the estimator $\widehat{\Theta}_1$ has access to at most n^2d_1 samples of a Gaussian on $\mathbb{R}^{d_1}$ with precision matrix Θ_1. We claim that as $\delta \to 0$, the distribution of Y_δ tends to that of Y_0 in total variation distance. Thus the distribution of $\widehat{\Theta}_1(Y_\delta)$ converges to that of $\widehat{\Theta}_1(Y_0)$ in total variation. Since Y_0 only depends on n^2d_1 samples to the Gaussian on $\mathbb{R}^{d_1}$ with precision matrix Θ_1, which we call Y, defining $\tilde{\Theta}(Y) = \widehat{\Theta}_1(Y_0)$ proves the theorem. Actually, as B has a probability zero chance of being singular, the final family of densities Y'_δ we will use is Y_δ conditioned on B being invertible. As B is invertible with probability 1 for $\delta > 0$, the total variation distance between Y'_δ, Y_δ is zero for all $\delta > 0$ and hence Y'_δ converges to Y_0 in total variation distance provided Y_δ does.

It remains to prove that Y_δ converges to Y_0 in total variation distance. First note that $Y_\delta = Y_0 + \delta Z$ where $Z_i = \sqrt{\Theta_1}X_iC^T$, where C is a random matrix where the first nd_1 columns are zero and the last $d_2 - nd_1$ columns have entries i.i.d uniform on $[-1, 1]$. Because of the zero patterns of B_0 and C and the fact that the entries of X are i.i.d., the random variables Y_0 and Z are independent. If we can show that Y_0 has a density with respect to the Lebesgue measure on $\mathbb{R}^{nd_1d_2}$, then $Y_0 + \delta Z$ converges to Y_0 in total variation distance as $\delta \to 0$. This follows because $Y_0 + \delta Z$ has a density obtained by convolving the density of Y_0, an L_1 function, by the law of δZ. The density of $Y_0 + \delta Z$ then converges to that of Y_0 in L_1 by the continuity of the translation operator in L_1. We thank Oliver Diaz for communicating a proof of this fact.

By invertibility of Σ_1, it is enough to show that Y_0 has a density when $\Sigma_1 = I(d_1)$. Consider $Y_0 = (X_iB_0^T, \ldots, X_nB_0^T)$. We may think of Y_0 as the $d_2 \times nd_1$ random matrix obtained by horizontally concatenating the matrices $B_0X_i^T$. Almost every matrix of these dimensions has rank nd_1, but if we had set even more of the columns of B_0 to zero then Y_0 would have rank *less* than nd_1 with probability 1 and hence would not have a density. This is why we cannot push this argument any further.

Now consider the nd_1 random vectors in $\mathbb{R}^{d_2}$ that are the columns of the matrices $B_0X_i^T$, for $i \in \{1, \ldots, n\}$. Because B_0 is supported only in its first nd_1 columns, the joint distribution of these random vectors may be obtained by sampling nd_1 independent standard Gaussian vectors v_j on $\mathbb{R}^{nd_1}$ and then multiplying them by the $d_2 \times nd_1$ matrix B' that is the restriction of B_0 to its first nd_1 columns. We have chosen B' such that it is an isometry into a uniformly random subspace of $\mathbb{R}^{d_2}$ of dimension nd_1. Thus $Bv_j/\|v_j\|$ are nd_1 many independent, random unit vectors in $\mathbb{R}^{d_2}$. As the $\|v_j\|$ are also independent, Bv_j are thus independent. Each marginal Bv_i has a density; one may sample it by choosing a uniformly random vector and then choosing the length $\|v_i\|$, hence the density is a product density in spherical coordinates. The joint density of the Bv_j is then the product density of the marginal densities. □

The above lemma combined with Corollary 6.3 immediately implies Theorem 6.4. We remark that the below proof uses no properties about d_F; a lower bound on any error metric for estimating a Gaussian with $n\min\{nd_1, d_2\}$ samples will transfer to the matrix normal model. In particular, Theorem 6.4 holds true when d_F is replaced by the Frobenius error and d_{op} replaced by the operator norm error.

PROOF OF THEOREM 6.4. To show Item 1, let $\delta^2 \le c\min\left\{1, \frac{d_1^2}{n\min\{nd_1,d_2\}}\right\}$. Let Θ_2 be the distributed as in Lemma 6.5 so that, as guaranteed by Lemma 6.5 there is an estimator $\tilde{\Theta}$ with access to a tuple Y of $n\min\{nd_1, d_2\}$ samples of a Gaussian on $\mathbb{R}^{d_1}$ with precision matrix Θ_1 satisfying $d_{TV}(\widehat{\Theta}_1(X), \tilde{\Theta}(Y)) \le \delta_0$. Corollary 6.3 implies

$$\sup_{\Theta\in B} \Pr_Y\left[d_F(\tilde{\Theta}||\Theta_1) \ge \delta\right] \ge \frac{1}{2}.$$

On the other hand, the total variation distance bound implies

$$\sup_{\Theta\in B,\Theta_2\in \mathrm{PD}(d_2)} \Pr_X\left[d_F(\widehat{\Theta}_1||\Theta_1) \ge \delta\right] \ge \sup_{\Theta\in B} \Pr_{\Theta_2,X}\left[d_F(\widehat{\Theta}_1||\Theta_1) \ge \delta\right] \ge \frac{1}{2} - \delta_0.$$

Allowing $\delta_0 \to 0$ implies the theorem. The proof of Item 2 is similar. □

7. Numerics and regularization. In the undersampled regime, most effort so far has focused on the sparse case. Existing estimators, such as the Gemini estimator Zhou (2014) and KGlasso estimator Tsiligkaridis, Hero and Zhou (2013) enforce sparsity by adding a regularizer proportional to the ℓ_1 norm of the precision matrices to encourage sparsity. We refer to these as Glasso-type estimators. We propose a new, shrinkage-based estimator that is simple to compute and experimentally outperforms Gemini and KGlasso in a natural generative model.

To describe our estimator, we remind the reader that the maximum likelihood estimator is the optimum of a function depending on the sample covariance matrix. Namely, we choose $\Theta \in \mathbb{P}$ optimizing $\operatorname{Tr}\Theta S - \frac{1}{D}\log\det\Theta$ where $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ is the sample covariance matrix[2]. In our estimator, S is replaced by a shrinkage estimator for the covariance matrix. Namely, we replace S by $\tilde{S} := (1-\alpha)S + \alpha\frac{\operatorname{Tr} S}{D} I$ for some $\alpha \in [0,1]$. This is known as a shrinkage estimator with ridge regularization . The estimator, which we call the shrinkage-based flip-flop estimator (or ShrinkFlop for short), seems closely related to the shrinkage estimator considered in Goes et al. (2020) and the Frobenius penalty considered in Tang and Allen (2018).

We consider a generative model in which the covariance matrices Θ_i are distributed as a low-rank Wishart matrix plus a small multiple of the identity matrix to ensure invertibility; we refer to them as "spiked" covariance matrices. We also show that, even when Θ_1 is sparse, our shrinkage-based estimator can outputerform Gemini and KGlasso when Σ_2 is spiked. Moreover, we observe that our regularized estimator is significantly faster to compute than the GLASSO-type estimators. All three estimators require parameter tuning, so when possible we compare throughout a plausible range of parameters for each of them. We leave determination of α by cross-validation for future work.

To define our estimator more precisely, we consider the following objective function. Let $\rho := \frac{1}{nD}\sum_{i=1}^n x_i x_i^T$. Define

$$f_x^\alpha(\Theta) := \operatorname{Tr}\Theta\left((1-\alpha)\rho + \alpha\frac{\operatorname{Tr}\rho}{D} I\right) - \frac{1}{D}\log\det\Theta,$$

[2]or, more accurately, the matrix of second moments

where $\alpha \in [0,1]$ is a parameter which must be tuned. Set $\widehat{\Theta}_x^\alpha = \arg\min_{\Theta \in \mathbb{P}} \ell_x^\alpha(\Theta)$ and $\tilde{\rho} := (1-\alpha)\rho + \alpha \frac{\mathrm{Tr}\,\rho}{D} I$. Observe the following:

1. If $\alpha > 0$, then $\tilde{\rho}$ is invertible and therefore $\widehat{\Theta}_x^\alpha$ exists uniquely for every x. This is because, whenever $\tilde{\rho}$ is invertible, the function f_x^α has a unique minimizer even over the larger domain $\mathrm{PD}(D)$,[3] and is geodesically convex. Thus it has a unique minimizer over the geodesically convex domain $\mathbb{P}$.
2. $\widehat{\Theta}_x^\alpha$ can be approximately computed using the flip-flop algorithm with ρ (as in Algorithm 1) replaced by $\tilde{\rho}$. Due to the special form of $\tilde{\rho}$, one need not look at the entirety of $\tilde{\rho}$ to perform each step of the flip-flop algorithm; see Algorithm 4. The only modification is that at each flip-flop step the estimate of the covariance is replaced by a convex combination with a scaled identity matrix.

Input: Samples $x = (x_1, \ldots, x_n)$ where $x_i \in \mathbb{R}^D$ is sampled from a (unknown) centered normal distribution $\mathcal{N}(0, \Sigma)$ and a real number $\varepsilon > 0$.
Output: Regularized estimators $\overline{\Theta} \in \mathbb{P}$.
Algorithm:

1. Set $\overline{\Theta}_a = I_a$ for each $a \in [k]$, and define $\overline{\Theta} := \otimes_a \overline{\Theta}_a$.
2. Repeat the following:
   - Compute the index a for which $\|\nabla_b f_x^\alpha(\overline{\Theta})\|_F$ is largest (see Eq. (7.1)). If $\|\nabla_a f_x^\alpha(\overline{\Theta})\|_F < \varepsilon$, end loop and **return** $\overline{\Theta}$.
   - Otherwise, set

$$\overline{\Theta}_a^{-1} \leftarrow d_a \left((1-\alpha)\overline{\Theta}_a^{-1/2} \rho^{(a)} \cdot \overline{\Theta}_a^{-1/2} + \frac{\alpha \|x\|^2}{D} \Big(\prod_{b \neq a} \mathrm{Tr}\,\Theta_b \Big) I_{d_a} \right),$$

   where $\rho = \frac{1}{nD} \cdot \overline{\Theta}^{1/2} \left(\sum_{i=1}^n x_i x_i^\dagger \right) \overline{\Theta}^{1/2}$.

Algorithm 4: Shrinkage-based flip-flop algorithm

The gradient of f_x^α is similar to that of f_x:

$$\nabla_a f_x^\alpha(\Theta) = (1-\alpha)\rho^{(a)} + \frac{\|x\|^2 \alpha}{D} \Big(\prod_{b \neq a} \mathrm{Tr}\,\overline{\Theta}_b \Big) \overline{\Theta}_a - \frac{1}{d_a} I_{d_a}, \tag{7.1}$$

for ρ as in Algorithm 4.

7.1. *Spiked, dense covariances.* Here we compare the performance of our shrinkage based estimator with Zhou's single step estimator (Gemini) for the matrix normal model assuming Σ_1, Σ_2 are dense, spiked covariance matrices. Fig. 1 was generated by setting $d_1 = 25, d_2 = 50$ and $n = 1$, and for each choice of regularization parameter independently generating 5 different pairs $\Sigma_1 \sim I_{d_1} + 10 v_1 v_1^T$ and $\Sigma_2 \sim I_{d_2} + 10 v_2 v_2^T$ where the v_i are standard d_i-dimensional Gaussian vectors and then normalizing each to have trace d_i, respectively. As in Zhou (2014), we always normalize the trace of $\widehat{\Theta}_1$ to match that of Θ_1 to focus on the core difficulties rather than estimating the overall scale of the data (which is easy to do). For each of the 5 pairs, we computed the relative squared error (the squared Frobenius error from Θ_1 to $\widehat{\Theta}_1$ divided by the squared Frobenius norm of Θ_1) for samples drawn from this distribution 5 times. Finally we averaged all 25 errors. Note that with $d_1 = 25, d_2 = 50$ and $n = 1$, the MLE never exists

[3] This is a consequence of the AM-GM inequality

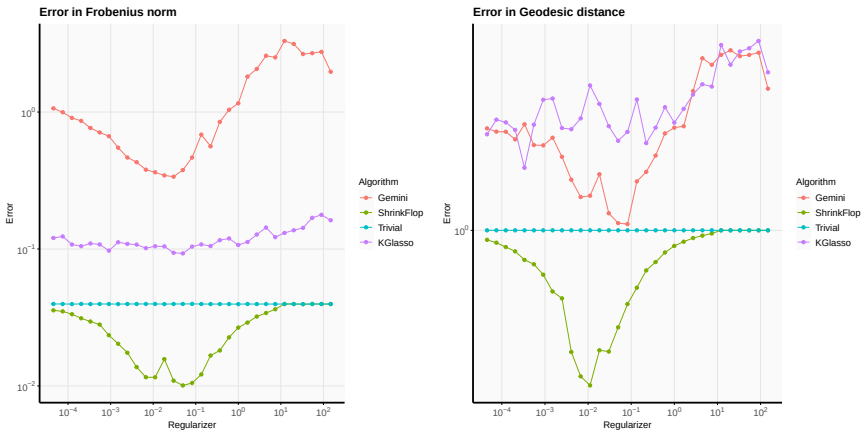

Figure 1: Average Frobenius error with $d_1 = 25, d_2 = 50, n = 1$ for spiked, dense covariance matrices. "Gemini" refers to the Gemini estimator of Zhou (2014), "KGlasso" refers to the Kronecker Glasso algorithm Tsiligkaridis, Hero and Zhou (2013), "ShrinkFlop" refers to our shrinkage-based flip-flop estimator, and "Trivial" refers to the estimator that always outputs the identity matrix. The choice of regularizer α is given by $\alpha = \arctan x$, where x is the value on the x-axis in the figures above.

and so we cannot compare with flip-flop without shrinkage. To compute the error in geodesic distance, we compute the squared geodesic distance between the estimator and the truth and divide by the geodesic distance between the truth and the identity.

Fig. 1 demonstrates that in the spiked case, for all choices of regularization parameter, the Gemini and KGlasso estimator were outperformed by the "trivial" estimator which always outputs the identity matrix. For a broad range of regularization parameters, our regularized estimator outperforms both the trivial estimator and Gemini. The poor performance of Gemini and KGlasso in this case are to be expected because the true precision matrices are dense.

7.2. *Sparse and partially sparse precision matrices.* We now compare the performance of our estimator with other leading estimators in the case when one or more of the precision matrices Θ_1, Θ_2 is sparse. We find that when both Θ_1 and Θ_2 are sparse, the Gemini estimator outperforms the regularized Sinkhorn algorithm in Frobenius error; see Fig. 2. However, when Θ_2 is spiked, we find that the shrinkage-based flip-flop estimator outperforms the Gemini estimator; see Fig. 3. In practice, Θ_2 is often considered a nuisance parameter and Θ_1 is the object that should be interpretable (e.g. sparse). Thus ill-conditioned and dense nuisance parameters Θ_2 can break GLASSO-type estimators even when Θ_1 is sparse.

The figures were generated in the same manner as Fig. 1, apart from the generative model. The sparse matrices Θ_i were generated by adding $\frac{1}{2}I(d_1)$ to the Laplacian of a random multigraph with $.4d_i$ edges and normalizing to have trace d_i, and the spiked covariance matrix Σ_2 for Fig. 3 was drawn according to $\Sigma_2 \sim I_{d_2} + 100 v_2 v_2^T$ where v_2 has i.i.d. Gaussian coordinates, and then normalized to have trace d_2.

7.3. *Performance as a function of the number of samples.* In this subsection we examine how the number of samples affects the error. We found best performance when the shrinkage parameter scales inverse exponentially with the number of samples, in this case $2^{-1.1n}$. The regularization parameter for Gemini was chosen to scale as $\sqrt{(\log d_1)/d_2 n}$ as suggested in Zhou (2014).

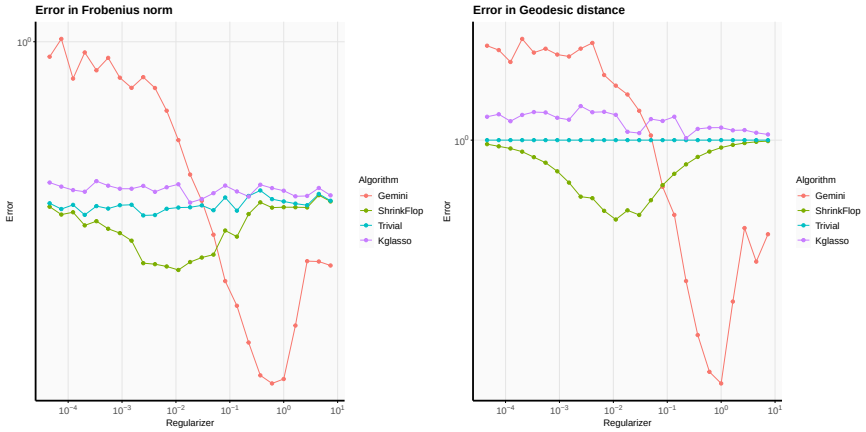

Figure 2: Average error with $d_1 = 25, d_2 = 50, n = 1$ for both precision matrices sparse. The choice of regularizer α for ShrinkFlop is given by $\alpha = \arctan x$, where x is the value on the x-axis in the figures above.

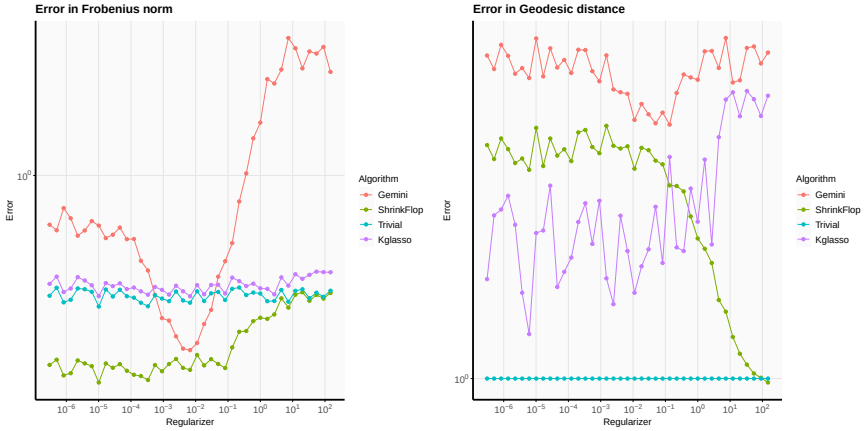

Figure 3: Average error with $d_1 = 25, d_2 = 50, n = 1$ for Θ_1 sparse and Θ_2 spiked. Interestingly, for Frobenius distance the shrinkage-based estimator performs best at the bottom of the regularization path and for the geodesic distance it performs best for larger regularization parameters. The choice of regularizer α for ShrinkFlop is given by $\alpha = \arctan x$, where x is the value on the x-axis in the figures above.

In Fig. 4 we see that Gemini outperforms flip-flop with a single sample, and shrinkage-based flip-flop outperforms both. When the number of samples are increased, the error for shrinkage-based flip-flop approaches that of flip-flop from below.

7.4. *Computational aspects.* Here we experimentally demonstrate that our shrinkage-based estimator is much faster to compute than the Glasso-type estimators. We considered $d_1 = 100, d_2 = 200$ with $n = 1$ samples. We considered regularized flip-flop and Gemini on spiked data, and chose the best regularization parameters for both using the computation from Fig. 1. We did not consider KGLasso because it consistently took far longer than Gemini. Out of 2 draws of data from 2 instances, the average time of completion was 0.078 seconds for regularized flip-flop and 30 seconds for Gemini. Both computations were done in R using a Macbook pro with an Apple M1 chip with 16 gigabytes of RAM.

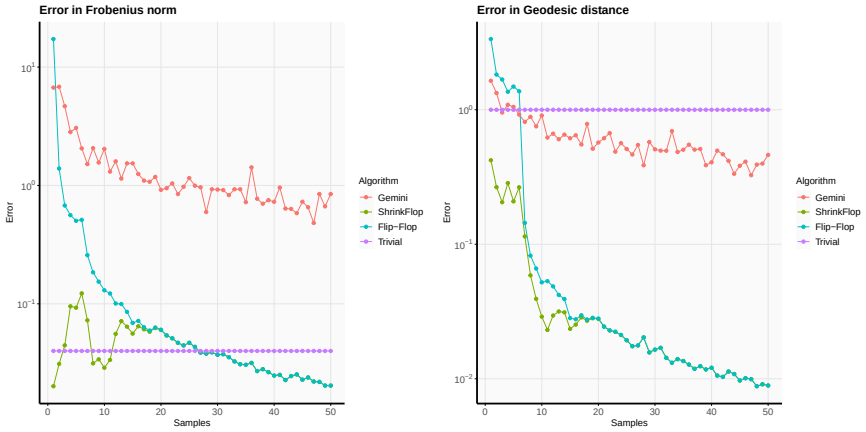

Figure 4: Average error with $d_1 = 25, d_2 = 25$, and number of samples n ranging from 1 to 50 for both precision matrices spiked. It was necessary to choose $d_1 = d_2$ so that the flip-flop estimator converged for $n = 1$. We did not include KGlasso here because its running time was prohibitive.

8. Conclusion and open problems. Though there has been a large volume of work on estimating the covariance in the matrix and tensor normal models under further assumptions like sparsity and well-conditionedness, some fundamental questions concerning estimation without further assumptions were still open prior to our work. Contrary to the state of the art for unstructured covariance estimation (i.e. $k = 1$), all previous existing results depended on the condition number of the true covariance matrices (in the case of the tensor normal model under Frobenius norm) or had suboptimal sample complexity (the matrix normal model under operator norm). Using strong convexity in the geometry induced by the Fisher information metric, we are largely able to remedy these issues and obtain nearly optimal estimates in the strongest possible metrics, namely the *relative* operator and Frobenius norms. As a consequence, we can also control other equivariant statistical distances such as relative entropy, Fisher-Rao distance, and total variation distance.

In particular, we showed that the maximum likelihood estimator (MLE) for the covariance matrix in the matrix normal models has optimal sample complexity up to logarithmic factors in the dimensions. We showed that the MLE for tensor normal models with a constant number of tensor factors has optimal sample complexity in the regime where it is information-theoretically possible to recover the covariance matrix to within a constant Frobenius error. Whenever the number of samples is large enough for either of the aforementioned statistical results to hold, we show that the flip-flop algorithm converges to the MLE exponentially quickly. Hence, the output of the flip-flop algorithm with $O\left(d_{\max} + \log n\right)$ iterations is an efficiently computable estimator with statistical guarantees comparable to those we show for the MLE.

We also observed empirically that under a certain natural generative model of ill-conditioned, dense covariance matrices, the flip-flop algorithm combined with a very simple shrinkage technique can outperform existing estimators designed for the sparse case (Gemini and KGLASSO). We view our empirical results as evidence that, in some cases, flip-flop combined with shrinkage provides the fastest, simplest and most statistically accurate known method to estimate the covariances. More work is needed in the future to rigorously understand the statistical guarantees for flip-flop with shrinkage.

Our main theoretical open question is whether the assumption $n = \Omega(d_{\max}^3/D)$ for Theorem 2.1 can be weakened to $n = \Omega(d_{\max}^2/D)$ for $k \geq 3$. Equivalently, do the guarantees of Theorem 2.1 hold even when one cannot hope to estimate the Kronecker factors to constant

Frobenius error, but only to constant *operator norm* error? In the case $k = 1$ (i.e. unstructured covariance estimation) the weaker assumption is well-known to suffice, and for $k = 2$ the same follows (up to logarithmic factors) by our Theorem 2.3. Filling in this final gap will place the tensor normal model on the same sound theoretical footing as unstructured covariance estimation.

APPENDIX A: PISIER'S PROOF OF QUANTUM EXPANSION

In this appendix we discuss the proof of Theorem 3.11. Pisier's original theorem is proved in a slightly different language:

THEOREM A.1 (Pisier). *Let $A_1, \dots, A_N, A$ be independent $n \times m$ random matrices with independent standard Gaussian entries. For any $t \geq 2$, with probability at least $1 - t^{-\Omega(m+n)}$,*

$$\left\| \left(\sum_{i=1}^{N} A_i \otimes A_i \right) \circ \Pi \right\|_{\mathrm{op}} \leq O\left(t^2 \sqrt{N} \left(\mathbb{E} \|A\|_{\mathrm{op}} \right)^2 \right),$$

where Π denotes the orthogonal projection onto the traceless subspace of $\mathbb{R}^m \otimes \mathbb{R}^m$, that is, onto the orthogonal complement of $\mathrm{vec}(I_m)$.

We first explain how Theorem A.1, along with the following estimate of the operator norm of a standard Gaussian $n \times m$ random matrix A (see Theorem 5.32 in Vershynin (2010)),

$$\mathbb{E} \|A\|_{\mathrm{op}} \leq \sqrt{n} + \sqrt{m}, \tag{A.1}$$

implies Theorem 3.11.

PROOF OF THEOREM 3.11. Choose $n = d_a$ and $m = d_b$. Observe that

$$\|\Phi_A\|_0 = \max_{\substack{H \text{ traceless symmetric} \\ \|H\|_F = 1}} \|\Phi(H)\|_F \leq \max_{\substack{H \in \mathrm{Mat}(m) \\ \|H\|_F = 1}} \|\Phi(\Pi(H))\|_F = \|\Phi \circ \Pi\|_{\mathrm{op}}.$$

Here we identified $\mathrm{Mat}(m) \cong \mathbb{R}^m \otimes \mathbb{R}^m$, so Π identifies with the orthogonal projection onto the traceless matrices, and we used that $\|\Pi(H)\|_F \leq \|H\|_F$, since Π is an orthogonal projection. Using Eq. (3.10) and Theorem A.1, it follows that

$$\|\Phi_A\|_0 \leq O\left(t^2 \sqrt{N} \left(\mathbb{E} \|A\|_{\mathrm{op}} \right)^2 \right),$$

with the desired probability. Using Eq. (A.1), we can bound the right-hand side operator norm,

$$\left(\mathbb{E} \|A\|_{\mathrm{op}} \right)^2 = O\left(\left(\sqrt{n} + \sqrt{m} \right)^2 \right) = O\left(n + m \right),$$

which concludes the proof. □

In the remainder we discuss the proof of Theorem A.1. Our setting required the result on rectangular matrices with strong error bounds, so we follow the proof in Pisier (2012) with these minor modifications and claim no originality. The proof proceeds by a symmetrization trick, followed by the trace method. We will first state a necessary concentration result and then give the proof of Theorem A.1.

THEOREM A.2 (Theorem 1.5 in Pisier (1986)). *Let A be a centered Gaussian random variable that takes values in a separable Banach space with norm $\|\cdot\|$. Then $\|A\|$ concentrates with parameter $\sigma^2 := \sup\{\mathbb{E}\langle \xi, A \rangle^2 \mid \|\xi\|_* \leq 1\}$, where $\|\xi\|_*$ denotes the dual norm. That is:*

$$\forall t > 0 : \quad \Pr\left(\left| \|A\| - \mathbb{E}\|A\| \right| \geq t \right) \leq 2 \exp\left(-\frac{\Omega(t^2)}{\sigma^2} \right).$$

Another equivalent definition of sub-Gaussianity is (see Lemma 5.5 of Vershynin (2010))

$$\forall p \geq 2: \quad (\mathbb{E}\|A\|^p)^{\frac{1}{p}} \leq \mathbb{E}\|A\| + O\left(\sqrt{\frac{p}{\sigma^2}}\right). \tag{A.2}$$

We calculate the sub-Gaussianity parameter for our setting below.

COROLLARY A.3. *Let A be an $n \times m$ matrix with independent standard Gaussian entries. Then $\|A\|_{\mathrm{op}}$ is sub-Gaussian with parameter $\sigma^2 = 1$.*

PROOF. Note that the dual norm is the trace norm $\|\cdot\|_1$, hence the concentration parameter can be estimated as

$$\sigma^2 = \sup\left\{\mathbb{E}\langle \xi, A\rangle^2 \mid \|\xi\|_1 \leq 1\right\} = \sup\left\{\|\xi\|_F^2 \mid \|\xi\|_1 \leq 1\right\} = 1,$$

where we first used that $\langle \xi, A\rangle$ is distributed the same as $\|\xi\|_F A_{11}$ by orthogonal invariance, and then that the trace norm dominates the Frobenius norm, with equality attained for example by $\xi = E_{11}$. □

We can now prove a lower bound that complements Eq. (A.1).

LEMMA A.4. *Let $n \geq m$ and A be an $n \times m$ matrix with independent standard Gaussian entries. Then there is some universal constant $c > 0$ such that*

$$\mathbb{E}\|A\|_{\mathrm{op}} \geq c(\sqrt{n} + \sqrt{m}).$$

PROOF. Note that

$$\mathbb{E}\|A\|_{\mathrm{op}} \geq \mathbb{E}\|Ae_1\|_2 \geq \frac{\sqrt{n}}{2} \geq \frac{\sqrt{m} + \sqrt{n}}{4},$$

where we used the lower bound on norm of standard Gaussian $g \in \mathbb{R}^n$ by standard integration $\mathbb{E}\sqrt{X}$ for $X = \|g\|_2^2$ distributed as a chi-square variable with n degrees of freedom (see Section 3.1 of Chandrasekaran et al. (2012)), and in the final step we used $n \geq m$. □

We will also use the Hölder inequality for the *Schatten p-norm* $\|A\|_p = (\mathrm{Tr}\left[(A^T A)^{\frac{p}{2}}\right])^{\frac{1}{p}}$, where $p \geq 1$:

$$\left|\mathrm{Tr}\prod_{i=1}^{p} A_i\right| \leq \prod_{i=1}^{p}\|A_i\|_p, \tag{A.3}$$

PROOF OF THEOREM A.1. The operator we want to control has entries which are dependent in complicated ways. We first begin with a standard symmetrization trick to linearize (compare the proof of Lemma 4.1 in Pisier (2014)). A single entry of $A_i \otimes A_i$ is either a product gg' of two independent standard Gaussians, or the square g^2 of a single standard Gaussian. In expectation, we have $\mathbb{E}gg' = 0, \mathbb{E}g^2 = 1$, and so the expected matrix is

$$\mathbb{E}\left(\sum_{i=1}^{N} A_i \otimes A_i\right) = N\,\mathrm{vec}(I_n)\,\mathrm{vec}(I_m)^T.$$

Accordingly, after projection we have

$$\mathbb{E}\left(\sum_{i=1}^{N} A_i \otimes A_i\right) \circ \Pi = 0.$$

Therefore we may add an independent copy: Let $B_1, \dots, B_N$ be independent $n \times m$ random matrices with standard Gaussian entries, that are also independent from $A_1, \dots, A_N$. Then,

$$\left(\sum_{i=1}^N A_i \otimes A_i\right) \circ \Pi = \mathbb{E}_B \left(\sum_{i=1}^N A_i \otimes A_i - \sum_{i=1}^N B_i \otimes B_i\right) \circ \Pi$$

and hence, for any $p \geq 1$,

$$\mathbb{E}_A \left\| \left(\sum_{i=1}^N A_i \otimes A_i\right) \circ \Pi \right\|_{\mathrm{op}}^p \leq \mathbb{E}_{A,B} \left\| \left(\sum_{i=1}^N A_i \otimes A_i - \sum_{i=1}^N B_i \otimes B_i\right) \circ \Pi \right\|_{\mathrm{op}}^p$$

by Jensen's inequality, as $\|\cdot\|_{\mathrm{op}}^p$ is convex as the composition of the norm $\|\cdot\|_{\mathrm{op}}$ with the convex and nondecreasing function $x \to x^p$. Now note (A_i, B_i) has the same distribution as $(\frac{A_i+B_i}{\sqrt{2}}, \frac{A_i-B_i}{\sqrt{2}})$, so the right-hand side is equal to

$$\begin{aligned} &\mathbb{E} \left\| \frac{1}{2} \left(\sum_{i=1}^N (A_i + B_i) \otimes (A_i + B_i) - \sum_{i=1}^N (A_i - B_i) \otimes (A_i - B_i)\right) \circ \Pi \right\|_{\mathrm{op}}^p \\ &= \mathbb{E} \left\| \left(\sum_{i=1}^N A_i \otimes B_i + \sum_{i=1}^N B_i \otimes A_i\right) \circ \Pi \right\|_{\mathrm{op}}^p \leq 2^p \, \mathbb{E} \left\| \sum_{i=1}^N A_i \otimes B_i \right\|_{\mathrm{op}}^p \end{aligned}$$

Thus, we have proved that

$$\mathbb{E} \left\| \left(\sum_{i=1}^N A_i \otimes A_i\right) \circ \Pi \right\|_{\mathrm{op}}^p \leq 2^p \, \mathbb{E} \left\| \sum_{i=1}^N A_i \otimes B_i \right\|_{\mathrm{op}}^p . \tag{A.4}$$

Note that we have lost the projection and removed the dependencies. Next we use the trace method to bound the right-hand side of Eq. (A.4). That is, we approximate the operator norm by the Schatten p-norm for a large enough p and control these Schatten norms using concentration of moments of Gaussians (compare the proof of Theorem 16.6 in Pisier (2012)). For any $q \geq 1$,

$$\begin{aligned} \mathbb{E} \left\| \sum_{i=1}^N A_i \otimes B_i \right\|_{2q}^{2q} &= \mathbb{E} \operatorname{Tr} \left(\sum_{i,j \in [N]} A_i^T A_j \otimes B_i^T B_j \right)^q \\ &= \sum_{i,j \in [N]^q} \mathbb{E} \operatorname{Tr} \left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q} \otimes B_{i_1}^T B_{j_1} \cdots B_{i_q}^T B_{j_q} \right) \\ &= \sum_{i,j \in [N]^q} \mathbb{E} \operatorname{Tr} \left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q} \right) \mathbb{E} \operatorname{Tr} \left(B_{i_1}^T B_{j_1} \cdots B_{i_q}^T B_{j_q} \right) \end{aligned}$$

where we used the independence of $\{A_i\}$ and $\{B_i\}$ in the last step. Now, the expectation of a monomial of independent standard Gaussian random variables is always nonnegative. Thus the same is true for $\mathbb{E} \operatorname{Tr}(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q})$, so we can upper bound the sum term by term as

$$\begin{aligned} &\sum_{i,j \in [N]^q} \mathbb{E} \operatorname{Tr} \left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q} \right) \mathbb{E} \operatorname{Tr} \left(B_{i_1}^T B_{j_1} \cdots B_{i_q}^T B_{j_q} \right) \\ &\leq \sum_{i,j \in [N]^q} \mathbb{E} \operatorname{Tr} \left(A_{i_1}^T A_{j_1} \cdots A_{i_q}^T A_{j_q} \right) \mathbb{E} \left(\|B_{i_1}\|_{2q} \|B_{j_1}\|_{2q} \cdots \|B_{i_q}\|_{2q} \|B_{j_q}\|_{2q} \right) \end{aligned}$$

$$\leq \sum_{i,j\in[N]^q} \mathbb{E}\operatorname{Tr}\left(A_{i_1}^T A_{j_1}\cdots A_{i_q}^T A_{j_q}\right)\mathbb{E}\left(\|B_1\|_{2q}^{2q}\right)$$

$$= \left(\mathbb{E}\|\sum_{i=1}^N A_i\|_{2q}^{2q}\right)\left(\mathbb{E}\|A\|_{2q}^{2q}\right) = N^q\left(\mathbb{E}\|A\|_{2q}^{2q}\right)^2$$

In the first step we used Hölder's inequality (A.3) for the Schatten norm. The second step holds since $\mathbb{E}\|B_i\|_{2q}^k \leq (\mathbb{E}\|B_i\|_{2q}^{2q})^{\frac{k}{2q}}$ by Jensen's inequality, so we can collect like terms together. Next, we used that the B_i have the same distribution as A. In the last step, we used that $\sum_i A_i$ has the same distribution as $\sqrt{N}A$. Accordingly, we have proved

$$\text{(A.5)}\quad \mathbb{E}\left\|\sum_{i=1}^N A_i\otimes B_i\right\|_{\text{op}}^{2q} \leq \mathbb{E}\left\|\sum_{i=1}^N A_i\otimes B_i\right\|_{2q}^{2q} \leq N^q\left(\mathbb{E}\|A\|_{2q}^{2q}\right)^2 \leq N^q m^2\left(\mathbb{E}\|A\|_{\text{op}}^{2q}\right)^2,$$

In the third inequality we used that $A\in\operatorname{Mat}(n,m)$ has rank $\leq m$, and therefore $\|A\|_{2q}^{2q}\leq m\|A\|_{\text{op}}^{2q}$. To bound the right-hand side, we wish to apply Theorem A.2 for the Banach space $\operatorname{Mat}(n,m)$ with the operator norm $\|\cdot\|_{\text{op}}$. Thus, using Eqs. (A.4) and (A.5) and Theorem A.2, we obtain for $q\geq 1$,

$$\mathbb{E}\left\|\left(\sum_{i=1}^N A_i\otimes A_i\right)\circ\Pi\right\|_{\text{op}}^{2q} \leq 2^{2q}N^q m^2\left(\mathbb{E}\|A\|_{\text{op}}^{2q}\right)^2 \leq (4N)^q m^2\left(\mathbb{E}\|A\|_{\text{op}}+C\sqrt{q}\right)^{4q},$$

where $C>0$ is the universal constant from the big-O notation in Eq. (A.2).

Finally, we can use Markov's inequality to see that, for some large universal constant $C'\geq 1$ that we choose later, the event

$$\left\|\left(\sum_{i=1}^N A_i\otimes A_i\right)\circ\Pi\right\|_{\text{op}} \leq (C't)^2\sqrt{4N}\left(\mathbb{E}\|A\|_{\text{op}}\right)^2$$

holds up to failure probability at most

$$\frac{\mathbb{E}\left\|\left(\sum_{i=1}^N A_i\otimes A_i\right)\circ\Pi\right\|_{\text{op}}^{2q}}{\left((C't)^2\sqrt{4N}\left(\mathbb{E}\|A\|_{\text{op}}\right)^2\right)^{2q}} \leq \frac{m^2\left(\mathbb{E}\|A\|_{\text{op}}+C\sqrt{q}\right)^{4q}}{(C't)^{4q}\left(\mathbb{E}\|A\|_{\text{op}}\right)^{4q}}.$$

We choose $q=\max\{1, C^{-2}(\mathbb{E}\|A\|_{\text{op}})^2\}$, which by Lemma A.4 is $\Omega(m+n)$.

We now have two cases. If $\mathbb{E}\|A\|_{\text{op}}\leq C$, and therefore $q=1$, then we can bound the failure probability as

$$\frac{m^2\left(\mathbb{E}\|A\|_{\text{op}}+C\sqrt{q}\right)^{4q}}{(C't)^{4q}\left(\mathbb{E}\|A\|_{\text{op}}\right)^{4q}} \leq m^2\left(\frac{2C}{C't\mathbb{E}\|A\|_{\text{op}}}\right)^{4q} \leq m^2 t^{-4q} = t^{-\Omega(m+n)}$$

The inequality follows by choosing C' large enough, as $\mathbb{E}\|A\|_{\text{op}}$ is bounded below by a universal constant (according to the proof of Lemma A.4). In the final step we used that $q=\Omega(m+n)$, as well as the fact that $t\geq 2$, so the prefactor m^2 can be absorbed at the cost of slightly changing the constant in the exponent.

If instead $C\sqrt{q}\leq\mathbb{E}\|A\|_{\text{op}}$, then we bound the failure probability

$$\frac{m^2\left(\mathbb{E}\|A\|_{\text{op}}+C\sqrt{q}\right)^{4q}}{(C't)^{4q}\left(\mathbb{E}\|A\|_{\text{op}}\right)^{4q}} \leq m^2\left(\frac{2\mathbb{E}\|A\|_{\text{op}}}{C't\mathbb{E}\|A\|_{\text{op}}}\right)^{4q} \leq m^2 t^{-4q} = t^{-\Omega(m+n)}$$

Here the inequality follows by choosing $C' \geq 2$. In the final step we used that $q = \Omega(m+n)$, as well as the fact that $t \geq 2$, so the prefactor m^2 can be absorbed at the cost of slightly changing the constant in the exponent. □

APPENDIX B: PROOF OF THE ROBUSTNESS LEMMA

In this section we give a proof of Lemma 3.13, which shows that strong convexity at a particular point implies strong convexity nearby. First note that by Remark 3.9, we have $\nabla^2 f_x(\Theta) = \nabla^2 f_{x'}$ where $x' = \Theta^{1/2}x$. Thus we need only bound the difference between f_x and $f_{x'}$ for $\|\log \Theta\|_{op}$ small, $\Theta \in \mathbb{P}$. For a matrix δ in $\mathrm{Mat}(d_a)$, we adopt the abuse of notation

$$e^{\delta_{(a)}} = I_{d_1} \otimes \cdots \otimes I_{d_{a-1}} \otimes e^{\delta} \otimes I_{d_{a+1}} \otimes \cdots \otimes I_{d_k}.$$

We will write $\Theta^{1/2}$ as e^{δ}, where $\delta = \sum_{a=1}^k \delta_{(a)}$. We now have $\Theta^{1/2} = e^{\delta} = \otimes_{a=1}^k e^{\delta_a}$, and $\frac{1}{2}\|\log \Theta\|_{op} = \|\delta\|_{op} = \sum_{a=1}^k \|\delta_a\|_{op}$. To bound the difference between $\nabla^2 f_{x'}$ and $\nabla^2 f_x$, we will show each component of the Hessian $\nabla f_{x'}$ (as presented in Lemma 3.8) only changes (from ∇f_x) by a small amount under the perturbation $x' := e^{\delta}x \leftarrow x$. In particular we will give bounds on each block under each component-wise perturbation $e^{\delta_{(a)}}x \leftarrow x$, and write the overall perturbation as a sequence of such component-wise perturbations. For convenience, we adopt the short-hand

$$\rho_x := \frac{1}{nD} x x^T.$$

We begin with an easy fact relating the exponential map and the operator norm.

FACT B.1. *For all symmetric $d \times d$ matrices A such that $\|A\|_{op} \leq 1$, we have*

$$\|e^A - I\|_{op} \leq 2\|A\|_{op}.$$

The 00 component of the Hessian is a scalar $\nabla^2_{00} f = \mathrm{Tr}[\rho]$, and for $a \geq 1$ we think of each $0a$ component as a vector:

$$\sum_a \langle z_0, (\nabla^2_{0a} f) Z_a \rangle = z_0 \langle \rho, \sum_a \sqrt{d_a} Z_{(a)} \rangle$$

The diagonal components involve only one-body marginals of ρ:

$$\langle Z_a, (\nabla^2_{aa} f) Z_a \rangle = \langle d_a \rho^{(a)}, Z_a^2 \rangle$$

And the off-diagonal components involve two-body marginals:

$$\langle Z_a, (\nabla^2_{ba} f) Z_b \rangle = \langle \sqrt{d_a d_b}\rho^{(ab)}, Z_a \otimes Z_b \rangle.$$

Therefore in Lemma B.2 and Lemma B.3, we will prove perturbation bounds on one-body marginals, and in Lemma B.6 we will prove perturbation bounds on two-body marginals. This will allow us to bound the change in the $0a$ components, diagonal components, and the off-diagonal components, respectively. Then, following the structure of the proof of Theorem 3.12, we will collect all the term-wise bounds to prove an overall bound at the end of the section.

LEMMA B.2. *For $x \in \mathbb{R}^{D \times n}$ and a symmetric matrix $\delta \in \mathrm{Mat}(d_a)$ such that $\|\delta\|_{op} \leq 1$, if we denote $x' := e^{\delta_{(a)}}x$ then*

$$\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} \leq 8\|\delta\|_{op}\|\rho^{(a)}_x\|_{op}.$$

PROOF. By definition, $\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} = \sup_{\|Z\|_1 \le 1} \langle Z_{(a)}, \rho_{x'} - \rho_x \rangle$. Let $\delta' := e^{\delta} - I_a$. Note that $\|\delta'\|_{op} \le 2\|\delta\|_{op}$ by Fact B.1 and our assumption $\|\delta\|_{op} \le 1$. Now

$$\begin{aligned}\langle Z_{(a)}, \rho_{x'} - \rho_x \rangle &= \langle Z_{(a)}, (I+\delta')_a \rho_x (I+\delta')_a - \rho_x \rangle \\ &= \langle Z, \delta' \rho^{(a)}_x \rangle + \langle Z, \rho^{(a)}_x \delta' \rangle + \langle Z, \delta' \rho^{(a)}_x \delta' \rangle \\ &\le (2\|\delta'\|_{op} + \|\delta'\|^2_{op}) \|\rho^{(a)}\|_{op} \|Z\|_1 \\ &\le 8\|\delta\|_{op} \|\rho^{(a)}\|_{op}.\end{aligned}$$

□

LEMMA B.3. *For $x \in \mathbb{R}^{D\times n}$ and symmetric matrix $\delta \in \mathrm{Mat}(d_b)$ such that $\|\delta\|_{op} \le 1$, if we denote $x' := e^{\delta_{(b)}} x$ then for $b \neq a$:*

$$\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} \le 2\|\delta\|_{op} \|\rho^a_x\|_{op}.$$

PROOF. By definition, $\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} = \sup_{\|Z\|_1 \le 1, Z \succeq 0} |\langle Z_{(a)}, \rho_{x'} - \rho_x \rangle|$.
Let $\delta' := e^{\delta} - I_b$. Note that $\|\delta'\|_{op} \le 2\|\delta\|_{op}$ by Fact B.1 and our assumption $\|\delta\|_{op} \le 1$. Now

$$\begin{aligned}|\langle Z_{(a)}, \rho_{x'} - \rho_x \rangle| &= |\langle Z_{(a)}, e^{\delta_{(b)}} \rho_x e^{\delta_{(b)}} - \rho_x \rangle| \\ &= |\langle Z_{(a)} \delta'_{(b)}, \rho_x \rangle| = |\langle Z \otimes \delta', \rho^{(ab)}_x \rangle| \\ &\le \langle \|\delta'\|_{op} Z \otimes I_b, \rho^{(ab)}_x \rangle \\ &= \|\delta'\|_{op} \langle Z, \rho^{(a)}_x \rangle \le 2\|\delta\|_{op} \|Z\|_1 \|\rho^{(a)}_x\|_{op}.\end{aligned}$$

□

This is already enough to prove a bound on $0a$ and aa terms:

COROLLARY B.4. *Let $x \in \mathbb{R}^{D\times n}$ be such that $\|d_a \rho^{(a)}_x\|_{op} \le 1 + \frac{1}{20}$, and for $b \in [k]$ let $\delta_b \in \mathrm{Mat}(d_b)$ be a symmetric matrix such that $\sum_b \|\delta_b\|_{op} \le \frac{1}{8}$. Denoting $\delta_{(b)} := (\delta_b)_{(b)}$, $\delta = \sum_b \delta_{(b)}$ and $x' = e^{\delta} x$, for $a \ge 1$ we have*

$$\|\nabla^2_{aa} f(e^{2\delta}) - \nabla^2_{aa} f(I)\|_{op} \le 25\|\delta\|_{op}.$$

PROOF. Recall from Lemma 3.8 that $\langle Y, (\nabla^2_{aa} f_x) Y \rangle = \langle d_a \rho^{(a)}_x, Y^2 \rangle$; thus it is enough to show that $\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} \le 25\|\delta\|_{op}/d_a$. We treat the perturbation e^{δ} as the composition of k perturbations:

$$x_{(0)} := x \to x_{(1)} := e^{\delta_{(1)}} x_{(0)} \to \ldots \to x_{(k)} := e^{\delta_{(k)}} x_{(k-1)} = x'.$$

We can use Lemma B.2 to handle $e^{\delta_{(a)}}$ and Lemma B.3 for the rest. Let Z be a symmetric matrix.

$$\begin{aligned}|\langle \rho^{(a)}_{x'} - \rho^{(a)}_x, Z \rangle| &\le \sum_{j=1}^{k} |\langle \rho^{(a)}_{x_{(j)}} - \rho^{(a)}_{x_{(j-1)}}, Z \rangle| \\ &\le \sum_{j=1}^{k} 8\|\delta_j\|_{op} \|\rho^{(a)}_{x_{(j-1)}}\|_{op} \|Z\|_1.\end{aligned}$$

Where the last inequality is due to Lemmas B.2 and B.3. To bound each term in the right-hand side, note that by Lemmas B.2 and B.3 we have

$$\|\rho^{(a)}_{x_{(j)}}\|_{op} \leq \|\rho^{(a)}_{x_{(j)}} - \rho^{(a)}_{x_{(j-1)}}\|_{op} + \|\rho^{(a)}_{x_{(j-1)}}\|_{op} \leq (1 + 8\|\delta_j\|_{op})\|\rho^{(a)}_{x_{(j-1)}}\|_{op}$$

and hence by induction the j^{th} term in the sum is at most

$$8\|\delta_j\|_{op}\left(\prod_{l=1}^{k}(1 + 8\|\delta_l\|_{op})\right)\|\rho^{(a)}_x\|_{op}\|Z\|_1.$$

By our assumption that $\sum_l \|\delta_l\|_{op} \leq 1/8$, this is at most $8\|\delta_j\|_{op}e^{8\sum\|\delta_l\|_{op}}\|\rho^{(a)}_x\|_{op}\|Z\|_1 \leq 8e\|\delta_j\|_{op}\|\rho^{(a)}_x\|_{op}\|Z\|_1$. Adding up the terms and using that $\|\delta\|_{op} = \sum\|\delta_{(c)}\|_{op}$, the overall sum is then at most $8e\|\delta\|_{op}\|\rho^{(a)}_x\|_{op}\|Z\|_1$. Using our assumption on $\|d_a\rho^{(a)}_x\|_{op}$ completes the proof. □

COROLLARY B.5. *Let $x \in \mathbb{R}^{D\times n}$ be such that $\|d_a\rho^{(a)}_x\|_{op} \leq 1 + \frac{1}{20}$, and for $b \in [k]$ let δ_b be symmetric matrices such that $\|\sum_b \delta_{(b)}\|_{op} = \sum_b \|\delta_b\|_{op} \leq \frac{1}{8}$, where once again we denote $\delta_{(b)} := (\delta_b)_{(b)}$ and $\delta := \sum_b \delta_{(b)}$. Denoting $x' := e^{\delta}x$, for $a \geq 1$ we have*

$$|\nabla^2_{00}f_{x'} - \nabla^2_{00}f_x| \leq 5\|\delta\|_{op}$$

$$\textit{and } \|\nabla^2_{0a}f_{x'} - \nabla^2_{0a}f_x\|_{op} \leq 25\|\delta\|_{op}.$$

PROOF. Recall from Lemma 3.8 that the 00 component of the Hessian is just the scalar $\operatorname{Tr}\rho$. The assumption that $\|d_a\rho^{(a)}_x\|_{op} \leq 1 + \frac{1}{20}$ implies $\operatorname{Tr}[\rho_x] = \operatorname{Tr}\rho^{(a)}_x \leq 1 + 1/20$. Now we can use the approximation for e^{δ} in Fact B.1:

$$|\operatorname{Tr}[\rho_{x'} - \rho_x]| = |\langle\rho_x, e^{2\delta} - I\rangle| \leq \operatorname{Tr}[\rho_x]\|e^{2\delta} - I\|_{op} \leq 5\|\delta\|_{op}$$

In the last step we used our bound on $\operatorname{Tr}[\rho_x]$. The $0a$ component is a vector, so it is enough to bound the inner product with any traceless matrix Z of unit Frobenius norm:

$$|\langle\rho^{(a)}_{x'} - \rho^{(a)}_x, \sqrt{d_a}Z\rangle| \leq \|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op}\sqrt{d_a}\|Z\|_1.$$

In the proof of Corollary B.4 we showed under the same assumptions we have $\|\rho^{(a)}_{x'} - \rho^{(a)}_x\|_{op} \leq 25\|\delta\|_{op}/d_a$, from which it follows that the above is at most $25\|\delta\|_{op}\|Z\|_F$. □

The off-diagonal components require the following two lemmata on bipartite marginals:

LEMMA B.6. *For $x \in \mathbb{R}^{D\times n}$ and a symmetric matrix $\delta \in \operatorname{Mat}(d_c)$ such that $\|\delta\|_{op} \leq \frac{1}{8}$; if we denote $x' := e^{\delta_{(c)}}x$, then for $c \in \{a, b\}$ we have*

$$\sup_{Y\in S^0_{d_a}, Z\in S^0_{d_b}} \frac{|\langle\rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle|}{\|Y\|_F\|Z\|_F} \leq 3\|\delta\|_{op} \sup_{Y\in S_{d_a}, Z\in S_{d_b}} \frac{\langle\rho^{(ab)}_x, Y \otimes Z\rangle}{\|Y\|_F\|Z\|_F}.$$

Note that S^0_d are traceless symmetric matrices, whereas S_d are symmetric matrices.

PROOF. By taking adjoints, we can assume w.l.o.g. that $c = b$. Let $R : \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_b)$ be defined as $R(Z) := e^{\delta}Ze^{\delta}$. Then

$$|\langle\rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle| = |\langle\rho^{(ab)}_x, Y \otimes (R(Z) - Z)\rangle|$$

The subspace $S^0_{d_b}$ is not invariant under R, but we show $R \approx I$. Let $\delta' := e^{\delta} - I$; by Fact B.1, $\|\delta'\|_{op} \le \frac{1}{4}$. Now

$$\|R(Z) - Z\|_F \le 2\|\delta' Z\|_F + \|\delta' Z \delta'\|_F \le (2\|\delta'\|_{op} + \|\delta'\|_{op}^2)\|Z\|_F \le 3\|\delta\|_{op}\|Z\|_F.$$

We combine these inequalities and apply a change of variables $R(Z) - Z \leftarrow Z'$ to finish the proof.

$$\begin{aligned}\sup_{Y \in S^0_{d_a}, Z \in S^0_{d_b}} \frac{|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle|}{\|Y\|_F \|Z\|_F} &= \sup_{Y \in S^0_{d_a}, Z \in S^0_{d_b}} \frac{|\langle \rho^{(ab)}_x, Y \otimes (R(Z) - Z)\rangle|}{\|Y\|_F \|Z\|_F} \\ &\le \sup_{Y \in S^0_{d_a}, Z' \in S_{d_b}} \frac{|\langle \rho^{(ab)}_x, Y \otimes Z'\rangle| \cdot 3\|\delta\|_{op}}{\|Y\|_F \|Z'\|_F}.\end{aligned}$$

□

LEMMA B.7. *For $x \in \mathbb{R}^{D \times n}$ and a symmetric matrix $\delta \in \mathrm{Mat}(d_c)$ such that $\|\delta\|_{op} \le \frac{1}{8}$; if we denote $x' := e^{\delta_{(c)}} x$, then for $c \notin \{a, b\}$ we have*

$$\sup_{Y \in S^0_{d_a}, Z \in S^0_{d_b}} \frac{|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle|}{\|Y\|_F \|Z\|_F} \le 4\|\delta\|_{op} \sup_{Y \in S_{d_a}, Z \in S_{d_b}} \frac{\langle \rho^{(ab)}_x, Y \otimes Z\rangle}{\|Y\|_F \|Z\|_F}.$$

PROOF. Let $\delta' := e^{2\delta} - I_c$ so that $|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle| = |\langle \rho^{(abc)}_x, Y \otimes Z \otimes \delta'\rangle|$. We first assume $Y, Z \succeq 0$, and without loss of generality we assume that $\|Y\|_F = \|Z\|_F = 1$. Because $\rho^{(abc)}_x, Y, Z \succeq 0$, and $\delta' \preceq \|\delta'\|_{op} \cdot I_c$, we have

$$\begin{aligned}|\langle \rho^{(abc)}_x, Y \otimes Z \otimes \delta'\rangle| &\le \langle \rho^{(abc)}_x, Y \otimes Z \otimes \|\delta'\|_{op} \cdot I_c\rangle \\ &\le \|\delta'\|_{op} \langle \rho^{(ab)}_x, Y \otimes Z\rangle \le 2\|\delta\|_{op} \langle \rho^{(ab)}_x, Y \otimes Z\rangle,\end{aligned}$$

where the last inequality is by Fact B.1. To finish the proof we decompose $Y = Y_+ - Y_-, Z = Z_+ - Z_-$, where Y_+, Y_-, Z_+, Z_- are all positive semidefinite, and bound

$$\begin{aligned}|\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle| &\le \sum_{s,t \in \{+,-\}} |\langle \rho^{(ab)}_{x'} - \rho^{(ab)}_x, Y \otimes Z\rangle| \\ &\le \sum_{s,t \in \{+,-\}} 2\|\delta\|_{op} \langle \rho^{(ab)}_x, Y_s \otimes Z_t\rangle \\ &\le 2 \left(\sup_{Y \in S_{d_a}, Z \in S_{d_b}} \frac{\langle \rho^{(ab)}_x, Y \otimes Z\rangle}{\|Y\|_F \|Z\|_F} \right) \|\delta\|_{op} \sum_{s,t \in \{+,-\}} \|Y_s\|_F \|Z_t\|_F\end{aligned}$$

The Cauchy Schwarz inequality allows us to bound the summation:

$$\sum_{s,t \in \{+,-\}} \|Y_s\|_F \|Z_t\|_F \le (2\|Y_+\|_F^2 + 2\|Y_-\|_F^2)^{1/2} (2\|Z_+\|_F^2 + 2\|Z_-\|_F^2)^{1/2} = 2\|Y\|_F \|Z\|_F.$$

Plugging this bound in to the supremum on the left-hand side in the statement of the lemma completes the proof. □

The following definition will be helpful for translating the above lemmas into statements about the Hessian.

DEFINITION. For a linear map $M : \mathrm{Mat}(d) \to \mathrm{Mat}(d')$, we let $\|M\|_0$ denote the $F \to F$ norm of its restriction to the traceless subspaces $S_d^0 \to S_{d'}^0$, i.e.

$$\|M\|_0 = \sup_{Z \in \mathcal{S}_d^0} \frac{\|M(Z) - \frac{\operatorname{Tr} M(Z)}{d'} I_{d'}\|_F}{\|Z\|_F}.$$

The following lemma will be helpful.

LEMMA B.8 (Kwok, Lau and Ramachandran (2019)). *For $x \in \mathbb{R}^{D \times n}$,*

$$\|\nabla_{ab}^2 f_x\|_{F \to F}^2 \leq \|d_a \rho_x^{(a)}\|_{op} \|d_b \rho_x^{(b)}\|_{op}.$$

Analogously to the proof of Corollary B.4, we can now combine Lemma B.6 and Lemma B.7 to bound the effect of a perturbation with more than one nontrivial tensor factor.

COROLLARY B.9. *Let $x \in \mathbb{R}^{D \times n}$ be such that $\|d_a \rho_x^{(a)}\|_{op}, \|d_b \rho_x^{(b)}\|_{op} \leq 1 + \frac{1}{20}$, and for $c \in [k]$ let δ_c be a symmetric matrix such that $\|\sum_c \delta_{(c)}\|_{op} = \sum_c \|\delta_c\|_{op} \leq \frac{1}{8}$. Denoting $x' := e^{\delta} x$, we have*

$$\|\nabla_{ab}^2 f_{x'} - \nabla_{ab}^2 f_x\|_0 \leq 21 \|\delta\|_{op}$$

PROOF. First, using Lemma 3.8, we write the left-hand and right-hand sides of the inequalities in Lemma B.6 and Lemma B.7 in terms of the Hessian:

$$\sup_{Y \in S_{d_a}^0, Z \in S_{d_b}^0} \frac{\langle \rho_{x'}^{(ab)} - \rho_x^{(ab)}, Y \otimes Z \rangle}{\|Y\|_F \|Z\|_F} = \frac{\|\nabla_{ab}^2 f_{x'} - \nabla_{ab}^2 f_x\|_0}{\sqrt{d_a d_b}},$$

$$\text{and} \quad \sup_{Y \in S_{d_a}, Z \in S_{d_b}} \frac{\langle \rho_x^{(ab)}, Y \otimes Z \rangle}{\|Y\|_F \|Z\|_F} = \frac{\|\nabla_{ab}^2 f_x\|_{F \to F}}{\sqrt{d_a d_b}}.$$

Using the same iterative strategy as in the proof of Corollary B.4 for the left-hand sides of the above identities, we have

$$|\langle Y, (\nabla_{ab}^2 f_{x'} - \nabla_{ab}^2 f_x) Z \rangle| \leq 20 \|\delta\|_{op} \|\nabla_{ab}^2 f_x\|_{F \to F} \|Y\|_F \|Z\|_F,$$

using Lemma B.6 for a and b and Lemma B.7 for the rest. Finally, we may rewrite Lemma B.8 using Lemma 3.8 to find $\|\nabla_{ab}^2 f_x\|_{F \to F}^2 \leq \|d_a \rho_x^{(a)}\|_{op} \|d_a \rho_x^{(a)}\|_{op}$. Using our assumption that $\|d_a \rho_a\|_{op}, \|d_b \rho_b\|_{op} \leq 1 + \frac{1}{20}$ completes the proof. □

Now we can combine the above term-by-term bounds to bound the change in the Hessian.

PROOF OF LEMMA 3.13. The above corollaries (B.5,B.4,B.9) require $\|d_a \rho^{(a)}\|_{op} \leq 1 + \frac{1}{20}$, which are implied by our assumption on the gradient:

$$\|d_a \rho^{(a)}\|_{op} \leq 1 + |\operatorname{Tr} \rho - 1| + \|d_a \rho^{(a)} - (\operatorname{Tr} \rho) I_{d_a}\|_{op}$$

$$= 1 + |\nabla_0 f| + \|\sqrt{d_a} \nabla_a f\|_{op} \leq 1 + 2\varepsilon_0,$$

so choosing $\varepsilon_0 \leq \frac{1}{40}$ suffices. Recall the expression of the Hessian as a quadratic form evaluated on $Z = (z_0, Z_1, \ldots, Z_k)$:

$$\langle Z, (\nabla^2 f) Z \rangle =$$

$$z_0 (\nabla_{00}^2 f) z_0 + 2 \sum_a \langle z_0, (\nabla_{0a}^2 f) Z_a \rangle + \sum_a \langle Z_a, (\nabla_{aa}^2 f) Z_a \rangle + \sum_{a \neq b} \langle Z_a, (\nabla_{ab}^2 f) Z_b \rangle.$$

Let $x' := e^{\delta}x$. Then by Corollary B.5 we have a bound on the $0a$ terms:

$$|z_0^2(\nabla_{00}^2 f_{x'} - \nabla_{00}^2 f_x) + 2\sum_a \langle z_0, (\nabla_{0a}^2 f_{x'} - \nabla_{0a}^2 f_x) Z_a\rangle|$$

$$\leq 5\|\delta\|_{op} z_0^2 + (2|z_0|)25\|\delta\|_{op}\sum_a \|Z_a\|_F \leq \|\delta\|_{op}(17kz_0^2 + 25\sum_a \|Z_a\|_F^2)$$

In the last step we used Young's inequality ($2pq \leq p^2 + q^2$) for each term with $p = z_0$, $q = \|Z_a\|_F$.

By Corollary B.4 we have a bound on the diagonal terms, and by Corollary B.9 we have a bound on the off-diagonal terms:

$$|\sum_{ab}\langle Z_a, (\nabla_{ab}^2 f_{x'} - \nabla_{ab}^2 f_x) Z_b\rangle| \leq \|\delta\|_{op}\left(25\sum_a \|Z_a\|_F^2 + 21\sum_{a\neq b}\|Z_a\|_F\|Z_b\|_F\right)$$

$$\leq (25 + 21(k-1))\|\delta\|_{op}\left(\sum_a \|Z_a\|_F^2\right)$$

So combining all three terms we see:

$$|\langle Z, (\nabla^2 f_{x'} - \nabla^2 f_x) Z\rangle| \leq \|\delta\|_{op}\left(17kz_0^2 + (25 + 25 + 21(k-1))\sum_a \|Z_a\|_F^2\right)$$

$$\leq 50k\|\delta\|_{op}\left(z_0^2 + \sum_a \|Z_a\|_F^2\right) = 50k\|\delta\|_{op}\|Z\|^2.$$

Note that this also gives an upper bound for $\|\nabla^2 f_{x'}\|_{op}$. □

APPENDIX C: THE CHEEGER CONSTANT OF A RANDOM OPERATOR

This section concerns completely positive maps; see Eq. (3.9) for a definition. In particular, we will consider $\Phi = \Phi_X$ where $X_1, \ldots, X_n$ are i.i.d. from $\mathcal{N}(0, I_{d_1} \otimes I_{d_2})$. To prove Theorem 4.4, we first define the Cheeger constant of completely positive map. This is similar to a concept defined in Hastings (2007).

DEFINITION. Let $\Phi : \mathrm{Mat}(d_1) \to \mathrm{Mat}(d_2)$ be a completely positive map. The Cheeger constant $\mathrm{ch}(\Phi)$ is given by

$$\mathrm{ch}(\Phi) := \min_{\Pi_1,\Pi_2:\mathrm{vol}(\Pi_1,\Pi_2)\leq \mathrm{Tr}\,\Phi(I)} \phi(\Pi_1, \Pi_2)$$

where $\Pi_1 : \mathbb{C}^{d_1} \to \mathbb{C}^{d_1}$ and $\Pi_1 : \mathbb{C}^{d_2} \to \mathbb{C}^{d_2}$ are orthogonal projections that are not both zero, and the *conductance* $\phi(\Pi_1, \Pi_2)$ of the "cut" Π_1, Π_2 is defined to be

$$\phi(\Pi_1, \Pi_2) := \frac{\mathrm{cut}(\Pi_1, \Pi_2)}{\mathrm{vol}(\Pi_1, \Pi_2)}$$

where

$$\mathrm{vol}(\Pi_1, \Pi_2) := \mathrm{Tr}\,\Phi(\Pi_1) + \mathrm{Tr}\,\Phi^*(\Pi_2)$$

and

$$\mathrm{cut}(\Pi_1, \Pi_2) := \mathrm{Tr}\,\Pi_2\Phi(I_{d_1} - \Pi_1) + \mathrm{Tr}(I_{d_2} - \Pi_2)\Phi(\Pi_1).$$

For intuition, consider a weighed bipartite graph G on $[d_1] \cup [d_2]$. The projections Π_1 and Π_2 are analogous to subsets of $A \subset [d_1]$ and $B \subset [d_2]$, respectively. The quantity $\mathrm{vol}(\Pi_1, \Pi_2)$ is analogous to the total mass of the edges adjacent to A plus that of the edges adjacent to B, which is the volume of $A \cup B$ considered as a cut of G. The quantity $\mathrm{cut}(\Pi_1, \Pi_2)$ corresponds to the total mass of the edges between $A \cup B$ and its complement, or the size of the cut $A \cup B$. In fact, if the Cheeger constant were defined with Π_1 and Π_2 restricted to be coordinate projections, it would be exactly the Cheeger constant of the bipartite graph on $[d_1]$ and $[d_2]$ with edge i, j weighted by $\operatorname{Tr} e_i e_i^T \Phi(e_j e_j^T)$, and the volume and the cut would be the same as the volume and the cut on that bipartite graph. Note that if $\Phi = \Phi_X$ for X as above, then each edge-weight $\operatorname{Tr} e_i e_i^T \Phi(e_j e_j^T)$ of the bipartite graph is a χ^2 random variable with n degrees of freedom.

It was recently shown that a large Cheeger constant implies quantum expansion:

LEMMA C.1 (Franks and Moitra (2020), Remark 5.5). *There exist absolute constants c, C such if $\varepsilon < c\,\mathrm{ch}(\Phi)^2$ and Φ is ε-balanced, then Φ is an*

$$\left(\varepsilon, \max\left\{1/2, 1 - \mathrm{ch}(\Phi)^2 + C\frac{\varepsilon}{\mathrm{ch}(\Phi)^2}\right\}\right) - \textit{quantum expander}.$$

We proceed to bound the Cheeger constant of a random operator. The Cheeger constant of an operator is scale-invariant, so for convenience we let $\Phi = \Phi_X$ where $X_1, \ldots, X_n$ are i.i.d. from $\mathcal{N}(0, I_{d_1} \otimes I_{d_2})$. Our main observation is the following.

LEMMA C.2. *Let Φ be distributed as above. Let $\Pi_1 : \mathbb{C}^{d_1} \to \mathbb{C}^{d_1}, \Pi_2 : \mathbb{C}^{d_2} \to \mathbb{C}^{d_2}$ be orthogonal projections, of rank r_1, r_2, respectively. Then $\mathrm{cut}(\Pi_1, \Pi_2), \mathrm{vol}(\Pi_1, \Pi_2), \mathrm{vol}(I_{d_1}, I_{d_2})$ is jointly distributed as*

$$R_1, R_1 + 2R_2, 2R_1 + 2R_2 + 2R_3$$

where R_1, R_2, R_3 are independent χ^2 random variables with $F_1 := nr_1(d_2 - r_2) + nr_2(d_1 - r_1), F_2 := nr_1r_2, F_3 := n(d_1 - r_1)(d_2 - r_2)$ degrees of freedom, respectively.

PROOF. As the distribution of Φ_X is invariant under the action $(U_1, U_2) \cdot \Phi_X(Y) = U_1 \Phi_X(U_2^T Y U_2) U_1^T$ of unitary matrices U_1, U_2, the distribution of $\mathrm{cut}(\Pi_1, \Pi_2), \mathrm{vol}(\Pi_1, \Pi_2)$ depends only on the rank of Π_1, Π_2. Thus we may compute in the case that Π_1, Π_2 are coordinate projections, in which case one may verify the fact straightforwardly; see the intuition after the definition of the Cheeger constant. □

We next show a sufficient condition for the Cheeger constant being bounded away from 1 that is amenable to the previous lemma.

LEMMA C.3. *Let r_1, r_2 not both zero be the ranks of projections $\Pi_1 : \mathbb{C}^{d_1} \to \mathbb{C}^{d_1}, \Pi_2 : \mathbb{C}^{d_2} \to \mathbb{C}^{d_2}$, and let $F_1 := nr_1(d_2 - r_2) + nr_2(d_1 - r_1)$ and $F_2 := nr_1r_2$. If*

- *for all Π_1, Π_2 such that $F_2 \geq (4/9)nd_1d_2$ we have*

$$\mathrm{vol}(\Pi_1, \Pi_2) \geq (101/200 - \delta)\,\mathrm{vol}(I_{d_1}, I_{d_2}) = (1.01 - 2\delta)\operatorname{Tr}\Phi(I_{d_1}), \tag{C.1}$$

and

- *for all Π_1, Π_2 such that $F_2 < (4/9)nd_1d_2$, we have*

$$\mathrm{vol}(\Pi_1, \Pi_2) \leq (4/3 + \delta)(F_1 + 2F_2) \text{ and } \mathrm{cut}(\Pi_1, \Pi_2) \geq (2/3 - \delta)F_1, \tag{C.2}$$

then $\mathrm{ch}(\Phi) \geq 1/6 - O(\delta)$ for $\delta < .005$.

PROOF. The first assumption implies we only need to reason about the case $F_2 < (4/9)nd_1d_2$, i.e. $r_1r_2 < (4/9)d_1d_2$. This is because the infimum in the definition of the Cheeger constant does not include Π_1, Π_2 such that $\mathrm{vol}(\Pi_1, \Pi_2) > 1/2 \operatorname{Tr} \Phi(I_{d_1})$.

It remains to show that $F_1/(F_1 + 2F_2) \geq 1/3$ provided $r_1r_2 < (4/9)d_1d_2$. Indeed, if either $r_1 = 0$ or $r_2 = 0$, then $F_2 = 0$ and $F_1 > 0$ and the claim holds, else

$$\begin{aligned} F_1/(F_1 + 2F_2) &= \frac{r_1d_2 + r_2d_1 - 2r_1r_2}{r_1d_2 + r_2d_1} \\ &= 1 - 2\sqrt{\frac{r_1r_2}{d_1d_2}} \frac{1}{\sqrt{r_1d_2/r_2d_1} + \sqrt{r_2d_1/r_1d_2}} \\ &\geq 1 - \sqrt{4/9} = 1/3. \end{aligned}$$

In the last inequality we used that $a + a^{-1} \geq 2$ for all $a \in \mathbb{R}_+$ and that $r_1r_2 < (4/9)d_1d_2$. □

Next we use Lemma C.2 to show that for fixed Π_1, Π_2, with high probability the events in Lemma C.3 hold.

LEMMA C.4. *Let r_1, r_2 not both zero be the ranks of projections $\Pi_1 : \mathbb{C}^{d_1} \to \mathbb{C}^{d_1}, \Pi_2 : \mathbb{C}^{d_2} \to \mathbb{C}^{d_2}$, and let $F_1 := nr_1(d_2 - r_2) + nr_2(d_1 - r_1)$ and $F_2 = nr_1r_2$. Then*

- *if $F_2 \geq (4/9)nd_1d_2$, then Eq. (C.1) holds with $\delta = 0$ with probability at least $1 - e^{-\Omega(nd_1d_2)}$.*
- *if $F_2 < (4/9)nd_1d_2$, then Eq. (C.2) holds with $\delta = 0$ with probability at least $1 - e^{-\Omega(F_1)}$.*
- *Finally, $\mathrm{vol}(\Pi_1, \Pi_2) \geq \frac{1}{2} \operatorname{Tr} \Phi(I_{d_1})(d_1/d_2)$ with probability at least $1 - e^{-\Omega(nr_1d_2 + nr_2d_1)}$.*

PROOF. Recall from Lemma C.2 that, $\mathrm{cut}(\Pi_1, \Pi_2), \mathrm{vol}(\Pi_1, \Pi_2), \mathrm{vol}(I_{d_1}, I_{d_2})$ are jointly distributed as $R_1, R_1 + 2R_2, 2R_1 + 2R_2 + 2R_3$ for R_1, R_2, R_3 each an independent χ^2 random variable with F_1, F_2, F_3 degrees of freedom, respectively. When $\delta = 0$, Eq. (C.1) holds if $R_2 > \frac{1}{99}R_1 + \frac{101}{99}R_3$ and Eq. (C.2) holds if $R_1 + R_2 + R_3 \leq (4/3)(F_1 + F_2 + F_3)$. Thus it is enough to show

- if $F_2 = nr_1r_2 \geq (4/9)nd_1d_2$, then with probability $1 - e^{-\Omega(nd_1d_2)}$ we have $R_2 > \frac{1}{99}R_1 + \frac{101}{99}R_3$, and
- if $F_2 = nr_1r_2 \leq (2/3)nd_1d_2 =$, then with probability $1 - e^{-\Omega(F_1)}$ we have $R_1 \geq (2/3)F_1$ and $R_1 + 2R_2 \leq (4/3)(F_1 + 2F_2)$,
- and with probability $1 - e^{-\Omega(F_1 + 2F_2)}$, $R_1 + 2R_2 \geq (2/3)(F_1 + 2F_2) = (2/3)n(r_1d_2 + r_2d_1)$ and $R_1 + R_2 + R_3 \leq (4/3)(F_1 + F_2 + F_3) = (4/3)nd_1d_2$.

All three follow from standard results for concentration of χ^2 random variables; see e.g. Wainwright (2019). We only prove the first item; the second and third items are straightforward. To prove the first item, note that $F_1 + 2F_2 \geq (4/3)(F_1 + F_2 + F_3)$, because

$$\begin{aligned} (F_1 + 2F_2)/(F_1 + F_2 + F_3) &= \frac{r_1}{d_1} + \frac{r_2}{d_2} \\ &= \sqrt{\frac{r_1r_2}{d_1d_2}} \left(\sqrt{\frac{r_1d_2}{r_2d_1}} + \sqrt{\frac{r_2d_1}{r_1d_2}} \right) \geq (2/3) \cdot 2 \geq 4/3. \end{aligned}$$

In particular, $F_2 \geq (2/3)(F_2 + F_3)$ and $F_2 \geq (F_1 + F_2)/6$. We first reason about the ratio between R_2 and R_3 using the first inequality. With probability $1 - e^{-cF_2} \geq 1 - e^{-cnd_1d_2}$, $R_2 \geq (8/9)F_2$ and $R_2 + R_3 \leq (10/9)(F_2 + F_3)$, so $R_2 \geq (8/9)(2/3)(9/10)(R_2 + R_3)$, or $R_2 \geq 8/7R_3$. We next apply the same reasoning with the inequality $F_2 \geq (F_1 + F_2)/6$ for F_1

and F_2. With probability $1 - e^{-cF_2}$, we have $R_2 \geq (8/9)F_2$ and $R_1 + R_2 \leq (10/9)(F_1 + F_2)$. Thus $R_1 \geq (8/9)(1/6)(9/10)(R_1 + R_2)$, or $R_2 \geq (2/13)R_1$. Calculation shows this implies $R_2 > \frac{1}{99}R_1 + \frac{101}{99}R_3$. □

Finally, we show using a net argument that the Cheeger constant is large for *all* projections.

LEMMA C.5 (Franks and Moitra (2020)). *There is a δ-net N of the rank r orthogonal projections $\Pi : \mathbb{C}^d \to \mathbb{C}^d$ with $|N| = \exp(O(dr|\ln\delta|))$.*

As a corollary, the set of pairs of projections Π_1, Π_2 of rank r_1, r_2 has a δ-net of size on the order of $(r_1 d_1 + r_2 d_2)|\ln\delta|$.

LEMMA C.6 (A net suffices). *Suppose $\|\Pi_1' - \Pi_2\|_F, \|\Pi_2' - \Pi_2\|_F \leq \delta$. Then*

$$|\operatorname{cut}(\Pi_1, \Pi_2) - \operatorname{cut}(\Pi_1', \Pi_2')| \leq 4\delta \operatorname{Tr}\Phi(I_{d_1})$$
$$\textit{and } |\operatorname{vol}(\Pi_1, \Pi_2) - \operatorname{vol}(\Pi_1', \Pi_2')| \leq 4\delta \operatorname{Tr}\Phi(I_{d_1}).$$

PROOF. We begin with the first inequality.

$$\begin{aligned}|\operatorname{cut}(\Pi_1', \Pi_2') - \operatorname{cut}(\Pi_1, \Pi_2)| \leq{}& |\operatorname{Tr}\Pi_2'\Phi(I_{d_1} - \Pi_1') - \operatorname{Tr}\Pi_2\Phi(I_{d_1} - \Pi_1)| \\ &+ |\operatorname{Tr}(I_{d_2} - \Pi_2')\Phi(\Pi_2') - \operatorname{Tr}(I_{d_2} - \Pi_2)\Phi(\Pi_2)|.\end{aligned}$$

Consider the first term:

$$\begin{aligned}&|\operatorname{Tr}\Pi_2'\Phi(I_{d_1} - \Pi_1') - \operatorname{Tr}\Pi_2\Phi(I_{d_1} - \Pi_1)| \\ &= |\operatorname{Tr}(\Pi_2' - \Pi_2)\Phi(I_{d_1} - \Pi_1') + \operatorname{Tr}\Pi_2\Phi(\Pi_1 - \Pi_1')| \\ &\leq \delta\|\Phi(I_{d_1} - \Pi_1')\|_F + \delta\|\operatorname{Tr}\Phi^*(\Pi_2)\|_F \\ &\leq 2\delta \operatorname{Tr}\Phi(I_{d_1}).\end{aligned}$$

The same inequality for the second term follows by symmetry. The proof of the second inequality is similar. □

LEMMA C.7 (Applying union bound). *Let $d_1 < d_2$. Suppose $n \geq C\frac{d_2}{d_1}\log(d_2/d_1)$. Then $\operatorname{ch}(\Phi) = \Omega(1)$ with failure probability $O(e^{-\Omega(nd_1)})$.*

PROOF. Let $\delta' \leq cd_1/d_2$. Let $\mathcal{N}(r_1, r_2)$ be a δ'-net for the pairs of projections of rank r_1, r_2, respectively, with $|\mathcal{N}(r_1, r_2)| = e^{O((d_1 r_1 + d_2 r_2)\log(1/\delta'))}$, and $N = \bigcup_{r_1, r_2}\mathcal{N}(r_1, r_2)$. We claim that it is enough to show that with probability $\exp(-cnd_1)$, for all r_1, r_2 not both zero we have

1. Eq. (C.1) holds with $\delta = 0$ for every $\Pi_1, \Pi_2 \in \mathcal{N}(r_1, r_2)$ when $r_1 r_2 \geq (4/9)d_1 d_2$,
2. and Eq. (C.2) holds with $\delta = 0$ for all $\Pi_1, \Pi_2 \in \mathcal{N}(r_1, r_2)$ when $r_1 r_2 < (4/9)d_1 d_2$,
3. and $\operatorname{vol}(\Pi_1, \Pi_2) \geq \operatorname{Tr}\Phi(I)(d_1/d_2)$.

Let us check that the hypotheses of Lemma C.3 with $\delta \leq c$ (for some small enough constant c) are implied by these three items; this will imply that conditioned on the three items we have $\operatorname{ch}(\Phi) \geq \Omega(1)$. Because every pair (Π_1', Π_2') of projections of ranks r_1, r_2 is most δ' far from some element (Π_1, Π_2) of $\mathcal{N}(r_1, r_2)$, by Lemma C.6 (and the inequality $\operatorname{vol}(\Pi_1, \Pi_2) \geq \operatorname{Tr}\Phi(I)(d_1/d_2)$) we have

$$(1 - 4\delta' \cdot d_2/d_1)\operatorname{vol}(\Pi_1, \Pi_2) \leq \operatorname{vol}(\Pi_1', \Pi_2') \leq (1 + 4\delta' \cdot d_2/d_1)\operatorname{vol}(\Pi_1, \Pi_2).$$

By assumption, $4\delta' \cdot d_2/d_1 \le c$. This shows Eq. (C.1) holds with $\delta \le c$ when $r_1 r_2 \ge (4/9) d_1 d_2$. It remains to show that Eq. (C.2) holds when $r_1 r_2 < (4/9) d_1 d_2$. Firstly, when $r_1 r_2 < (4/9) d_1 d_2$ we have

$$\operatorname{vol}(\Pi_1', \Pi_2') \le (1+c) \operatorname{vol}(\Pi_1, \Pi_2) \le (1+c)(4/3)(F_1 + 2F_2). \tag{C.3}$$

Next, observe that

$$\operatorname{cut}(\Pi_1', \Pi_2') \ge \operatorname{cut}(\Pi_1, \Pi_2) - c \operatorname{vol}(\Pi_1, \Pi_2).$$

In the proof of Lemma C.3 it is shown that if $r_1 r_2 < (4/9) d_1 d_2$ then $F_1 \ge \frac{1}{3}(F_1 + 2F_2)$, in which case

$$\begin{aligned} \operatorname{cut}(\Pi_1', \Pi_2') &\ge \operatorname{cut}(\Pi_1, \Pi_2) - c \operatorname{vol}(\Pi_1, \Pi_2) \ge \\ &\ge (2/3) F_1 - c(4/3)(F_1 + 2F_2) \ge (2/3 - c) F_1. \end{aligned} \tag{C.4}$$

Taken together, Eqs. (C.3) and (C.4) show that Eq. (C.2) holds when $r_1 r_2 < (4/9) d_1 d_2$. We next show that the three conditions hold with the desired probability. We show that for fixed r_1, r_2, each item holds with probability at least $1 - e^{n(r_1 d_2 + r_2 d_1)}$. This implies the conditions hold for all r_1, r_2 with the desired probability because the sum of $e^{-\Omega(n(r_1 d_2 + r_2 d_1))}$ over all $0 \le r_1 \le d_1, 0 \le r_2 \le d_2$ apart from $r_1 = r_2 = 0$ is $O(e^{-\Omega(n d_1)})$. By our choice of n we have $(d_1 r_1 + d_2 r_2) \log(1/\delta') \le cn(r_1 d_2 + r_2 d_1)$ for r_1, r_2 not both zero.

We first bound the failure probability for the first item. By Lemma C.4, if $r_1 r_2 \ge (4/9) d_1 d_2$ then Eq. (C.1) holds for every $\Pi \in \mathcal{N}(r_1, r_2)$ with probability

$$\begin{aligned} 1 - |\mathcal{N}(r_1, r_2)| e^{-\Omega(n d_1 d_2)} &= 1 - |\mathcal{N}(r_1, r_2)| e^{-\Omega(n(r_2 d_1 + r_1 d_2))} \\ &= 1 - e^{-\Omega(n(r_2 d_1 + r_1 d_2))}. \end{aligned}$$

Next we bound the probability for the second item. By Lemma C.4, Eq. (C.2) holds for fixed $\Pi \in \mathcal{N}(r_1, r_2)$ with probability $1 - e^{-\Omega(F_1)}$, but as the proof of Lemma C.3 shows, we have $F_1 \ge \frac{1}{3}(F_1 + 2F_2)$ when $r_1 r_2 < (4/9) d_1 d_2$. Thus $F_1 = \Omega(n(r_1 d_2 + r_2 d_1))$. Now, by the union bound and the lower bound on n, Eq. (C.2) holds for every element of $\mathcal{N}(r_1, r_2)$ with probability $1 - |\mathcal{N}(r_1, r_2)| e^{-\Omega(n(r_1 d_2 + r_2 d_1)} = 1 - e^{-\Omega(n(r_1 d_2 + r_2 d_1)}$.

The third item holds with probability $1 - e^{-\Omega(n(r_1 d_2 + r_2 d_1))}$ by Lemma C.4, so by a similar application of the union bound and our choice of n it holds for all elements of $\mathcal{N}(r_1, r_2)$ with probability $1 - e^{-\Omega(n(r_1 d_2 + r_2 d_1))}$. □

PROOF OF THEOREM 4.4. Let $\Phi := \Phi_X$. To prove Theorem 4.4, we apply Lemma C.1 using Proposition 3.5 to bound the balancedness of Φ and Lemma C.7 to bound $\operatorname{ch}(\Phi)$. Indeed, $\|\nabla_a f\|_{op} \le \varepsilon_0$ for $a \in \{1, 2\}$ if and only if Φ is ε_0-balanced, so by Proposition 3.5 the operator Φ is ε_0-balanced with probability $1 - e^{-\Omega(n d_1 \varepsilon_0^2)} - e^{-\Omega(n d_2 \varepsilon_0^2)} \ge 1 - 2e^{-\Omega(n d_1 \varepsilon_0^2)}$ provided $n \ge C \varepsilon_0^{-2} d_2/d_1$. Setting $\varepsilon_0 = \varepsilon \sqrt{d_2/n d_1}$ proves the balancedness claim. For the expansion, Lemma C.7 shows $\operatorname{ch}(\Phi) = \Omega(1)$ with failure probability $O(e^{-\Omega(n d_1)}) = O(e^{-\Omega(d_2 \varepsilon^2)})$. By making C large enough we can ensure that $\varepsilon_0 \le c$ for any constant $c > 0$. Hence Lemma C.1 applies with balanced-ness ε_0, which implies that Φ is an $(\varepsilon \sqrt{d_2/n d_1}, 1 - \lambda)$-quantum expander for some absolute constant λ. □

APPENDIX D: PROOF OF CONCENTRATION FOR MATRIX NORMAL MODEL

PROOF OF LEMMA 4.6. For convenience, we consider the differently normalized random variable $Z = Y/\sqrt{n d_1}$. Note that Z satisfies $Z_i = X_i \Phi_X^*(I_{d_1})^{-1/2} = X_i (\sum_{i=1}^n X_i^\dagger X_i)^{-1/2}$. Thus we need to bound the random matrix

$$\sum_{i=1}^n Z_i Z_i^\dagger - \frac{d_2}{d_1} I_{d_1} = \sum_{i=1}^n X_i \left(\sum_j X_i^\dagger X_i \right)^{-1} X_i^\dagger - \frac{d_2}{d_1} I_{d_1}. \tag{D.1}$$

Since we are interested in the spectral norm of Eq. (D.1), we will consider the random variable $\langle \xi, \sum_{i=1}^n Z_i Z_i^\dagger \xi \rangle$ for a fixed unit vector $\xi \in \mathbb{R}^{d_1}$. We will show that this variable ξ is highly concentrated, and apply a union bound over a net of the unit vectors. To show the concentration, we first cast $\langle \xi, \sum_{i=1}^n Z_i Z_i^\dagger \xi \rangle$ as the inner product between a random orthogonal projection and a fixed one.

Considering Z_i as a $d_1 \times d_2$ matrix, we can consider Z as an $nd_1 \times d_2$ matrix by vertically concatenating the Z_i. By definition of the flip-flop step, $Z^\dagger Z = \sum Z_i^\dagger Z_i = I_{d_2}$, so the columns of Z are an orthonormal basis of $\mathbb{R}^{nd_1}$. Here $\dagger$ denotes transpose for $nd_1 \times d_2$ matrices. In fact, the columns of Z are a *uniformly random* orthonormal basis of a d_2 dimensional subspace $\mathbb{R}^{nd_1}$; that is, they are a uniformly random element of the Steifel manifold. Thus, $ZZ^\dagger$ is a uniformly random rank d_2 orthonormal projection on $\mathbb{R}^{nd_1}$. We can now write

$$\langle \xi, \sum_{i=1}^n Z_i Z_i^\dagger \xi \rangle = \langle ZZ^\dagger, \xi\xi^\dagger \otimes I_n \rangle.$$

The matrix $\xi\xi^* \otimes I_n$ is a rank n projection on $\mathbb{R}^{nd_1}$. We use the following result on the inner product of random projections.

THEOREM D.1 (Lemma III.5 in Hayden, Leung and Winter (2006)). *Let P be a uniformly random orthogonal projection of rank a on $\mathbb{R}^m$ and let Q be a fixed orthogonal projection of rank b on $\mathbb{R}^m$. Then*

$$\Pr\left[\langle P, Q \rangle \notin (1 \pm \varepsilon)\frac{ab}{m}\right] = e^{-\Omega(ab\varepsilon^2)}.$$

We may apply the above theorem with $Q = \xi\xi^* \otimes I_n$, $m = nd_1$, $a = d_2$, and $b = n$ to obtain

$$\Pr\left[\langle \xi, \sum_{i=1}^n Z_i Z_i^\dagger \xi \rangle \notin (1 \pm \varepsilon_0)\frac{d_2}{d_1}\right] = e^{-\Omega(nd_2\varepsilon_0^2)}. \tag{D.2}$$

Next we apply a standard net argument for the unit vectors over $\mathbb{R}^{nd_1}$. We use the following lemma.

LEMMA D.2 (Lemma 5.4 Vershynin (2010)). *Let A be a symmetric $d \times d$ matrix, and let $\mathcal{N}_\delta$ be an δ-net of $\mathbb{S}^{d-1}$ for some $\delta \in [0, 1)$. Then*

$$\|A\|_{op} \leq (1 - 2\delta)^{-1} \sup_{\xi \in \mathcal{N}_\delta} |\langle \xi, A\xi \rangle|.$$

We apply the above lemma with $A = \sum_{i=1}^n Z_i Z_i^\dagger - \frac{d_2}{d_1} I_{d_1}$ and $d = d_1$. Fix a net $\mathcal{N} = \mathcal{N}_\delta$ for $\delta = 1/4$; by standard packing bounds (e.g. Lemma 4.2 in Vershynin (2010)) we may take $|\mathcal{N}| \leq 9^{d_1}$. By Eq. (D.2) and the union bound, with failure probability $9^{d_1} e^{-\Omega(nd_2\varepsilon_0^2)}$ we have that $|\langle \xi, A\xi \rangle| \leq \frac{d_2}{d_1}\varepsilon_0$ for all $\xi \in \mathcal{N}$, and by Lemma D.2 this event implies $\|A\|_{op} \leq 2\frac{d_2}{d_1}\varepsilon_0$.

It remains to translate our bound on A to our desired bound on Φ_Y. Because $Z = Y/\sqrt{nd_1}$, $\Phi_Y = nd_1\Phi_Z$. Thus $A = \frac{1}{nd_1}\Phi_Y(I_{d_2}) - \frac{d_2}{d_1} I_{d_1}$, so $\|\frac{1}{nD}\Phi_Y(I_{d_2}) - \frac{1}{d_1} I_{d_1}\|_{op} = \frac{1}{d_2}\|A\|_{op}$. Setting $\varepsilon_0 = \varepsilon\sqrt{\frac{d_1}{4nd_2}}$ shows $\|\frac{1}{nD}\Phi_Y(I_{d_2}) - \frac{1}{d_1} I_{d_1}\|_{op} \leq \frac{\varepsilon}{\sqrt{nD}}$ with failure probability at most $9^{d_1} e^{-\Omega(d_1\varepsilon^2)}$, which is at most $e^{-\Omega(d_1)\varepsilon^2}$ provided ε is a large enough constant. □

APPENDIX E: RELATIVE ERROR METRICS

In this section we discuss the properties of our relative error metrics d_F, d_{op}. First note that they can be related to the usual norms by following inequalities:

$$\|B^{-1}\|_{op}^{-1} D_F(A\|B) \le \|A-B\|_F \le \|B\|_{\text{op}} D_F(A\|B)$$

$$\|B^{-1}\|_{op}^{-1} D_{op}(A\|B) \le \|A-B\|_{\text{op}} \le \|B\|_{\text{op}} D_{\text{op}}(A\|B).$$

Next we state the approximate triangle inequality, also called a *local* triangle inequality in Yang and Barron (1999), and approximate symmetry for our relative error metrics.

LEMMA E.1. *Let $A, B, C \in \text{PD}(d)$. Let $D \in D_{op}, D_F$. Provided $D(A||B), D(B||C)$ are at most an absolute constant c, we have*

$$D(A||C) = O(D(A||B) + D(B||C)), \tag{E.1}$$

$$D(B||A) = O(D(A||B)), \text{ and} \tag{E.2}$$

$$D(A^{-1}||B^{-1}) = O(D(A||B)). \tag{E.3}$$

For D_{op}, the result is (Franks and Moitra, 2020, Lemma C.1). For D_F, the result holds because because $D_F(A||B) = \Theta(d(A,B))$ if either is at most some absolute constant (shown below). Because $d(A,B)$ is a metric, it automatically satisfies Eqs. (E.1) and (E.2). Furthermore, $d(A,B) = d(A^{-1}, B^{-1})$ by direct calculation. We next consider the relationship between D_F and other dissimilarity measures in the statistics literature.

PROPOSITION E.2 (Relationships between dissimilarity measures.). *There is a constant $c > 0$ such that the following holds. If any of $D_F(\Theta_1||\Theta_2)$, $d_{TV}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1}))$, $d_{KL}(\mathcal{N}(0,\Theta_1^{-1})||\mathcal{N}(0,\Theta_2^{-1})$ or the Fisher-Rao distance $d(\Theta_1,\Theta_2)$ is at most c, then*

$$D_F(\Theta_1||\Theta_2) \asymp d_{TV}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1})^2 \asymp d_{KL}(\mathcal{N}(0,\Theta_1^{-1})||\mathcal{N}(0,\Theta_2^{-1}) \asymp d(\Theta_1,\Theta_2).$$

PROOF. For the relationship between $D_F(\Theta_1||\Theta_2)^2$ and the Fisher-Rao distance $d(\Theta_1,\Theta_2)^2$, observe that the former is $\sum_{i=1}^d (\lambda_i - 1)^2$ and the latter is $\sum_{i=1}^d (\log \lambda_i)^2$ for the eigenvalues λ_i of $\Theta_2^{-1}\Theta_1$. The relationship follows because in any fixed interval not containing 0, $\lambda - 1 \asymp \log \lambda$.

To relate D_F to the total variation distance, we use the following bound from Devroye, Mehrabian and Reddad (2018):

$$.01 \le \frac{d_{TV}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1})}{D_F(\Theta_1||\Theta_2)} \le 1.5.$$

This implies that if either the numerator or denominator is a small enough constant, then they are on the same order.

Next we reason for the relative entropy. If $D_{KL}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1}) \le c$ or $D_{op}(\Theta_2||\Theta_1) \le 1/2$, then we have

$$D_{KL}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1}) \asymp D_F(\Theta_2||\Theta_1)^2.$$

This bound can be seen explicitly from the formula

$$\begin{aligned} D_{KL}(\mathcal{N}(0,\Theta_1^{-1}), \mathcal{N}(0,\Theta_2^{-1}) &= \frac{1}{2}\operatorname{Tr}\Theta_1^{-1}\Theta_2 - \frac{1}{2}\log\det(\Theta_1^{-1}\Theta_2) - \frac{d}{2} \\ &= \frac{1}{2}\sum_{i=1}^d (\lambda_i - 1 - \log\lambda_i) \end{aligned}$$

where $\lambda_i \in 1 \pm D_{op}(\Theta_2||\Theta_1)$ are the eigenvalues of $\Theta_1^{-1}\Theta_2$ and the fact that $\lambda - 1 - \log\lambda \asymp (\lambda-1)^2$ on $[1/2, 3/2]$. Choose c small enough that $\frac{1}{2}(\lambda - 1 - \log\lambda) \leq c$ implies $\lambda \in [1/2, 3/2]$. Finally, there is some absolute constant c such that $D_F(\Theta_1||\Theta_2) \leq c$ or $d(\Theta_1, \Theta_2) \leq c$ then $D_F(\Theta_1||\Theta_2) \asymp d(\Theta_1, \Theta_2)$. □

Acknowledgements. This work was supported in part by NWO Veni grant no. 680-47-459 and NWO grant OCENW.KLEIN.267.

REFERENCES

ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics* **4** 764.

AMÉNDOLA, C., KOHN, K., REICHENBACH, P. and SEIGAL, A. (2020). Invariant theory and scaling algorithms for maximum likelihood estimation. *arXiv preprint arXiv:2003.13662*.

BACÁK, M. (2014). *Convex analysis and optimization in Hadamard spaces* **22**. Walter de Gruyter GmbH & Co KG.

BHATIA, R. (2009). *Positive definite matrices* **24**. Princeton university press.

BÜRGISSER, P., GARG, A., OLIVEIRA, R., WALTER, M. and WIGDERSON, A. (2018). Alternating Minimization, Scaling Algorithms, and the Null-Cone Problem from Invariant Theory. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)* (A. R. KARLIN, ed.). *Leibniz International Proceedings in Informatics (LIPIcs)* **94** 24:1–24:20. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

BÜRGISSER, P., FRANKS, C., GARG, A., OLIVEIRA, R., WALTER, M. and WIGDERSON, A. (2019). Towards a theory of non-commutative optimization: geodesic 1st and 2nd order methods for moment maps and polytopes. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)* 845–861. IEEE.

CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics* **12** 805–849.

DERKSEN, H. and MAKAM, V. (2020). Maximum likelihood estimation for matrix normal models via quiver representations. *arXiv preprint arXiv:2007.10206*.

DERKSEN, H., MAKAM, V. and WALTER, M. (2020). Maximum likelihood estimation for tensor normal models via castling transforms. *arXiv preprint arXiv:2011.03849*.

DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2018). The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*.

DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **64** 105–123.

FRANKS, C. and MOITRA, A. (2020). Rigorous Guarantees for Tyler's M-estimator via quantum expansion. *arXiv preprint arXiv:2002.00071*.

GARG, A., GURVITS, L., OLIVEIRA, R. and WIGDERSON, A. (2019). Operator scaling: theory and applications. *Foundations of Computational Mathematics* 1–68.

GOES, J., LERMAN, G., NADLER, B. et al. (2020). Robust sparse covariance estimation by thresholding Tyler's M-estimator. *The Annals of Statistics* **48** 86–110.

GURVITS, L. (2004). Classical complexity and quantum entanglement. *Journal of Computer and System Sciences* **69** 448–484.

HASTINGS, M. (2007). Random unitaries give quantum expanders. *Physical Review A* **76** 032315.

HAYDEN, P., LEUNG, D. W. and WINTER, A. (2006). Aspects of generic entanglement. *Communications in mathematical physics* **265** 95–117.

JAMES, W. and STEIN, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics* 443–460. Springer.

KWOK, T. C., LAU, L. C. and RAMACHANDRAN, A. (2019). Spectral Analysis of Matrix Scaling and Operator Scaling. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)* 1184–1204. IEEE.

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics* **40** 1024–1060.

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics* **239** 37–49.

MARDIA, K. V. and GOODALL, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics* **6** 347–385.

PISIER, G. (1986). Probabilistic methods in the geometry of Banach spaces. *Letta G., Pratelli M. (eds) Probability and Analysis* **1206**.

PISIER, G. (2012). Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society* **49** 237–323.

PISIER, G. (2014). Quantum expanders and geometry of operator spaces. *Journal of the European Mathematical Society* **16** 1183–1219.

SKOVGAARD, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scandinavian journal of statistics* 211–223.

SUN, W., WANG, Z., LIU, H. and CHENG, G. (2015). Non-convex statistical optimization for sparse tensor graphical model. *Advances in Neural Information Processing Systems* **28** 1081–1089.

TANG, T. M. and ALLEN, G. I. (2018). Integrated Principal Components Analysis. *arXiv preprint arXiv:1810.00832*.

TSILIGKARIDIS, T., HERO, A. O. I. and ZHOU, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Transactions on Signal Processing* **61** 1743–1755.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing* **56** 478–491.

WIESEL, A. (2012). Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing* **60** 6182–6189.

YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599.

ZHOU, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics* **42** 532–562.