The Annals of Statistics
2026, Vol. 54, No. 1, 93–119
https://doi.org/10.1214/25-AOS2539

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

BY COLE FRANKS[1,a], RAFAEL OLIVEIRA[2,b], AKSHAY RAMACHANDRAN[3,c] AND MICHAEL WALTER[4,d]

[1]*Department of Mathematics, Massachusetts Institute of Technology,* [a]*franks@mit.edu*
[2]*Cheriton School of Computer Science, University of Waterloo,* [b]*rafael@uwaterloo.ca*
[3]*Computer Science Department, University of British Columbia,* [c]*aramach@cs.ubc.ca*
[4]*Ludwig-Maximilians-Universität München,* [d]*michael.walter@lmu.de*

The matrix normal model, that is, the family of Gaussian matrix-variate distributions whose covariance matrices are 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 normal models.

For the above models, we show that the maximum likelihood estimator (MLE) achieves *nearly optimal nonasymptotic sample complexity* and *nearly tight error rates* in the Fisher–Rao and Thompson metrics. In contrast to prior work, our results do not rely on the factors being well conditioned or sparse, nor do we need to assume an accurate enough initial guess. For the matrix normal model, all our bounds are minimax optimal up to logarithmic factors, and for the tensor normal model our bounds for the largest factor and for overall covariance matrix are minimax optimal up to constant factors provided there are enough samples for any estimator to obtain constant Frobenius error. In the same regimes as our sample complexity bounds, we show that the flip-flop algorithm, a practical and widely used iterative procedure to compute the MLE, converges linearly with high probability.

Our main technical insight is that, given enough samples, the negative log-likelihood function is *strongly geodesically convex* in the geometry on positive-definite matrices induced by the Fisher information metric. This strong convexity is determined by the expansion of certain random quantum channels.

1. Introduction. Covariance matrix estimation is an important task in statistics, machine learning and the empirical sciences. We consider covariance estimation for centered matrix-variate and tensor-variate Gaussian data, that is, when individual data points are matrices or tensors. Matrix and tensor-variate data arise naturally in numerous applications, such as gene microarrays, clinical trials, spatiotemporal data, signal processing and brain imaging (see [4, 16, 17, 22] and references therein). A significant challenge in this setting is that the dimensionality of these problems is 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* [9, 16, 22]. Here, the matrix follows a multivariate Gaussian distribution and the covariance between any two entries in the matrix is a product of an interrow factor and an intercolumn factor. In spatiotemporal statistics, this is referred to as a separable covariance structure [16]. Formally, if a matrix normal random variable X takes values in

Received November 2021; revised January 2025.
MSC2020 subject classifications. Primary 62F12; secondary 62F30.
Key words and phrases. Covariance estimation, matrix normal model, tensor normal model, maximum likelihood estimation, geodesic convexity, operator scaling, quantum expansion.

the space of $d_1 \times d_2$ 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 dimensions $d_1 \times d_1$ 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 $\Sigma_1, \ldots, \Sigma_k$. Hence, a centered tensor normal distribution is denoted by $\mathcal{N}(0, \Sigma_1 \otimes \cdots \otimes \Sigma_k)$. In this work, we study the estimation of the covariance factors $\Sigma_1, \ldots, \Sigma_k$ or (equivalently) the precision factors $\Theta_1 := \Sigma_1^{-1}, \ldots, \Theta_k := \Sigma_k^{-1}$ from n samples of $\mathcal{N}(0, \Sigma_1 \otimes \cdots \otimes \Sigma_k)$. We emphasize that the goal is to estimate *each of the factors*, rather than estimating the overall product $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k$ or $\Sigma := \Sigma_1 \otimes \cdots \otimes \Sigma_k$ by an arbitrary precision or covariance matrix (that may not be of tensor product form).

This problem falls into the field of estimation theory: for a family $\mathcal{P} := \{p_\Theta\}_{\Theta \in \mathbb{P}}$ of distributions with parameter space $\mathbb{P}$, given samples from an unknown distribution $X_1, \ldots, X_n \sim p_\Theta$, compute an estimate $\hat{\Theta} \approx \Theta$ of the true parameter value. The quality of our estimate depends on some *error measure*, chosen based on the downstream application of the estimation problem. Our parameter space $\mathbb{P}$ is the set of Kronecker products of k precision matrices, each of dimension d_i, which will be taken from the space of positive definite matrices (denoted $\mathrm{PD}(d_i)$).

The error measures in our work will be given by the *Fisher–Rao* and *Thompson* metrics. These are the relevant error metrics for statistical applications, as they are intimately tied to error measures for the corresponding distributions, such as total variation and relative entropy. Further theoretical justification is given by Chentsov's theorem ([8], Theorem 3), which states that for smooth parameter manifolds, the Fisher information metric[1] is the unique Riemannian metric that preserves all relevant information with respect to parameter estimation. We refer the reader to [10], Section A, in the Supplementary Material for further details on these metrics, as well as their connection to other natural metrics used for the matrix and tensor normal models.

DEFINITION 1.1 (Fisher–Rao and Thompson distances). The Fisher–Rao distance for centered Gaussians parameterized by their precision matrices is given by

$$d_{\mathrm{FR}}(\widehat{\Theta}, \Theta) = \frac{1}{\sqrt{2}} \left\| \log\left(\Theta^{-1/2} \widehat{\Theta} \Theta^{-1/2}\right) \right\|_F. \tag{1.1}$$

The Thompson distance is given by

$$d_{\mathrm{op}}(\widehat{\Theta}, \Theta) := \left\| \log\left(\Theta^{-1/2} \widehat{\Theta} \Theta^{-1/2}\right) \right\|_{\mathrm{op}}. \tag{1.2}$$

We have the following simple relation between the two metrics that follows directly from the same relation between the operator and Frobenius norms.

FACT 1.2. *For A, B positive definite matrices of dimension d, that is, $A, B \in \mathrm{PD}(d)$, the Fisher–Rao and Thompson metrics are related by*

$$d_{\mathrm{op}}(A, B) \le \sqrt{2} \cdot d_{\mathrm{FR}}(A, B) \le \sqrt{d} \cdot d_{\mathrm{op}}(A, B).$$

Now that we are equipped with our error measures, we can formally ask the foundational questions on the parameter estimation problem for the tensor normal model.[2] We begin with the sample complexity questions.

[1] The Fisher–Rao distance is the distance function arising from the Fisher information metric.

[2] Since the matrix normal model is a special case of the tensor normal model (when $k = 2$), we will refer to our model as the tensor normal model whenever we treat the general case.

PROBLEM 1.3 (Sample complexity upper bound). Let $\varepsilon > 0$ be an error parameter and $\delta \in (0, 1)$ be a failure parameter. Given sample access to an unknown tensor normal distribution $\mathcal{N}(0, \Theta_1^{-1} \otimes \cdots \otimes \Theta_k^{-1})$, how many samples $n(\varepsilon, \delta)$ are ***sufficient*** for the existence of estimator $\widehat{\Theta}_a$ satisfying, with probability $1 - \delta$,

$$d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a) \leq \varepsilon \quad \text{for all } a \in [k]?$$

In practical settings often the number of samples n is fixed, so many results in the literature give bounds on the error ε and failure probability δ for fixed value of n. The first consideration for such a result is its ***sample threshold***: this is the number of samples n_0 that is required in order for the proposed estimator to give any nontrivial guarantees, that is, better than an arbitrary guess in $\mathbb{P}$. The second consideration is the ***error rate*** achieved by the proposed estimator, that is, how fast the error decreases as the number of samples grows.

Problem 1.3 is only concerned with *upper bounds* on the number of samples needed to obtain good enough estimates for the true precision factors. It is natural to ask what is the *optimal* upper bound on the number of samples, that is, the minimum number of samples required to estimate the precision factors. This leads us to the following problem.

PROBLEM 1.4 (Sample complexity lower bound). Let $\varepsilon > 0$ be an error parameter and $\delta \in (0, 1)$ be a failure parameter. How many samples from a distribution $\mathcal{N}(0, \Theta_1^{-1} \otimes \cdots \otimes \Theta_k^{-1})$ are ***necessary*** for existence of estimator $\widehat{\Theta}_a$ such that, with probability $1 - \delta$

$$d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a) \leq \varepsilon \quad \text{for all } a \in [k]?$$

REMARK 1.5. The above notion of sample complexity lower bound can be used to derive a minimax lower bound as follows: if $n \geq n(\varepsilon, \delta)$ samples are required to achieve $d_{\mathrm{FR}}(\hat{\Theta}_a, \Theta_a) \leq \varepsilon$ error with probability at least $1 - \delta$, then given $n < n(\varepsilon, \delta)$ samples,

$$\inf_{\hat{\Theta}} \sup_{\Theta \in \mathbb{P}} \mathbb{E}\Big[\max_{i \in [k]} d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a)\Big] \geq \delta \cdot \varepsilon,$$

where inf is over all possible estimators $\hat{\Theta}$, the sup is over the parameter space $\mathbb{P}$, and the expectation is over the distribution corresponding to parameter Θ.

A complete solution to the sample complexity problem requires one to prove tight *upper* and *lower* bounds on the number of samples to estimate the factors of the covariance matrix for a given error and probability guarantee.

The above questions are concerned with the mathematical existence of an estimator with a prescribed number of samples, which accurately estimates the true precision factors. However, a more relevant question for practical purposes is whether the estimator proposed for Problem 1.3 can be computed efficiently. More succinctly, one can ask whether there is a gap between statistical estimation versus computational estimation. This is captured by the following computational variant of Problem 1.3.

PROBLEM 1.6. Let $\varepsilon > 0$ be an error parameter and $\delta \in (0, 1)$ be a failure parameter. Given sample access to an unknown tensor normal distribution $\mathcal{N}(0, \Theta_1^{-1} \otimes \cdots \otimes \Theta_k^{-1})$, how many samples from the above distribution are *sufficient* for there to exist estimators $\widehat{\Theta}_a$ that are ***efficiently computable*** and satisfy, with probability $1 - \delta$,

$$d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a) \leq \varepsilon \quad \text{for all } a \in [k]?$$

Moreover, give an algorithm to compute this estimator, which runs in polynomial time and achieves the above error bounds and success probability.

This work fully addresses the three problems above for the matrix and tensor normal models.

Our solution to Problem 1.3 comes from the analysis of the most natural candidate: the *maximum likelihood estimator* (MLE). Informally, we give the following sample complexity bounds for this estimator.

THEOREM (Sample complexity, tensor normal model). *Let $\mathcal{N}(0, \Theta_1^{-1} \otimes \cdots \otimes \Theta_k^{-1})$ be a tensor normal distribution with $k \geq 2$, where each Θ_i is a positive definite matrix of dimension d_i, and let $D := \prod_{i=1}^k d_i$. Given a number of samples n respecting the sample threshold $n \gtrsim \frac{k^2 d_{\max}^3}{D}$, the MLE achieves minimax optimal error rate in Fisher–Rao distance*

$$d_{\mathrm{FR}}(\hat{\Theta}, \Theta) \lesssim \sqrt{\frac{k d_{\max}^2}{n}}, \qquad d_{\mathrm{FR}}(\hat{\Theta}_a, \Theta_a) \lesssim \sqrt{\frac{k d_a d_{\max}^2}{nD}}$$

with high probability. Further, for the matrix normal model (i.e., $k = 2$), the sample threshold is improved to $n \gtrsim \frac{d_{\max}^2 \log^2 d_{\min}}{D}$, and the error can be bounded in the Thompson metric as

$$d_{\mathrm{op}}(\hat{\Theta}_a, \Theta_a) \lesssim \sqrt{\frac{d_a^2 \log^2 d_{\min}}{nD}}.$$

Our estimation guarantees are *distribution independent*, in particular the above bounds hold regardless of condition number or sparsity or other properties of the true precision matrix. This means that they apply to the most general model where the precision factors are allowed to be arbitrary positive definite matrices with no restrictions.

By Fact 1.2, the d_{op} bound for the matrix normal model recovers the d_{FR} error rate for the tensor normal bound up to logarithmic factors; furthermore, it implies strong estimation guarantees in the operator norm, which are useful in spectral applications (see [3]).

The above guarantees are *tight* compared to classical lower bounds (see Proposition 4.1), matching the sample complexity lower bounds even for the simpler $k = 1$ setting. The d_{FR} error rate for the full precision matrix as well as the largest tensor factor are tight up to the factor $\sqrt{k}$. And the sample threshold matches the lower bound for estimating the largest tensor factor up to a single $d_{\max}$ factor. In the $k = 2$ matrix normal model, the error rate is tight in the more refined d_{op} metric, matching the classical lower bound for estimating a single tensor up to log factors. The sample threshold matches the classical lower bound up to log factors.

We solve Problem 1.4 by extending the lower bound for the unstructured Gaussian estimation problem to the matrix and tensor normal model. While the above results are near-optimal for estimation of the largest tensor factor (via the classical lower bound), one could hope for better results for the smaller tensor factors,[3] as they intuitively receive more information from each tensor data. Our next contribution is a stronger sample complexity lower bound, which shows this is not the case.

THEOREM 1.7 (Lower bound for matrix normal models). *Let $\widehat{\Theta}_1$ be any estimator for Θ_1 given n samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$. For $d_1 \leq d_2$, there exist $\Theta_1 \in \mathrm{PD}(d_1)$ and $\Theta_2 \in \mathrm{PD}(d_2)$ such that*

$$d_{\mathrm{FR}}(\widehat{\Theta}_1, \Theta_1) \gtrsim \sqrt{\frac{d_1^2}{n \cdot \min\{n d_1, d_2\}}}, \qquad d_{\mathrm{op}}(\widehat{\Theta}_1, \Theta_1) \gtrsim \sqrt{\frac{d_1}{n \cdot \min\{n d_1, d_2\}}}$$

with constant probability.

[3] In certain applications, such as brain fMRI, one is interested only in the smaller factor, whereas the larger factor is treated as a nuisance parameter.

When $nd_1 \ll d_2$, our lower bound is significantly stronger than the classical lower bound for estimating Θ_1 *assuming* Θ_2 *is known*, namely $\sqrt{d_1^2/nd_2}$ for d_{FR} and $\sqrt{d_1/nd_2}$ for d_{op}. Our result generalizes naturally to the tensor normal model, as we discuss further in Section 4. This implies that the matrix and tensor estimation problems are strictly harder than separate instances of the classical Gaussian estimation problem. We are also able to show that a simple modification of the MLE obtains a matching upper bound for the matrix normal model.

Lastly, our solution to Problem 1.6 comes from analyzing the *flip-flop algorithm* to compute the MLE. This is the *first rigorous convergence analysis* of the flip-flop algorithm, which was proposed in the independent works [4, 9, 16] and is widely used in practice.

THEOREM (Computational estimation, informal). *With high probability, the MLE can be computed efficiently. Namely, the flip-flop algorithm enjoys exponential convergence rate* $\log(1/\delta)$ *to achieve a* δ *approximation to the MLE.*

For a full comparison and relation between our results above and previous works, we refer the reader to [10], Section B, in the Supplementary Material.

Technical contributions and overview. We now discuss the main conceptual ideas and principles behind our results. In the matrix and tensor normal models (i.e., $k \geq 2$ case), the MLE is a solution to an explicit optimization problem over the space of tensor products of positive definite matrices, which we denote by $\mathbb{P}$. When we endow the parameter space $\mathbb{P}$ with a natural Riemannian metric induced by the Fisher information, the negative log-likelihood becomes a *geodesically convex* function of the parameter space (first observed in [23]). In this work, we use *geodesic convexity* of the negative log-likelihood function to show that the MLE for the tensor normal model indeed recovers all the benefits of the unstructured Gaussian setting ($k = 1$). Our strategy, as we outline in Section 2.2, proceeds as follows: provided one is given enough samples, we prove that the negative log-likelihood function is *strongly geodesically convex*, and the gradient at the true precision matrix is small. With these two facts, we are able to conclude our bounds via a generalization of the usual argument that with a strongly convex function, any point with a small enough gradient (in our case the true precision matrix) is close to the optimizer (the MLE).

The global geodesic perspective is also key when analyzing algorithms to compute the MLE. Inspired by recent research in computer science [5–7, 12], we view the flip-flop algorithm as a natural geodesic extension of the block-coordinate geodesic gradient descent method, which is a standard convex optimization method. Once we establish strong geodesic convexity of the negative log-likelihood function, we can show that the iterates of the flip-flop algorithm converge exponentially quickly to the MLE once the gradient of our current guess is sufficiently small. Our proof generalizes to any descent method with reasonable guarantees.

This geodesic geometry perspective induces a natural error metric under which our analysis becomes linearly invariant, and this allows us to prove sample complexity and error bounds that are independent of condition number. Furthermore, by using global geodesic convexity of the negative log-likelihood function, we are able to decouple our analysis of the estimator from our algorithm to compute the MLE and, therefore, we are able to remove the initial guess assumption from our error bounds. The bounds we achieve are tight in general, as we show in Section 4, and our bounds even improve upon the previous results in the sparse setting as soon as the condition number or initialization error becomes moderately large (square root of the maximum dimension of the Kronecker factors). For detailed comparison of our bounds with prior work, we point the reader to [10], Section B, in the Supplementary Material.

We believe that the strength of the derived bounds, along with the principled analysis of a very simple and practical algorithm, make strong arguments in favor of the geodesic

perspective for understanding the tensor normal model. We now present the formal definitions of our problems and state our main results.

1.1. *Formal definitions and our results.* We write $\mathrm{Mat}(d)$ for the space of real $d \times d$ matrices and $\mathrm{PD}(d)$ for the convex cone of $d \times d$ real symmetric positive definite matrices; $\mathrm{GL}(d)$ denotes the group of real invertible $d \times d$ matrices. We write $\succeq$ for the Löwner order. For matrices A and B, $\|A\|_{\mathrm{op}}$ denotes the operator norm, $\|A\|_F = (\operatorname{Tr} A^T A)^{\frac{1}{2}}$ the Frobenius norm and $\langle A, B\rangle = \operatorname{Tr} A^T B$ the Hilbert–Schmidt inner product. We say A is a traceless matrix if $\operatorname{Tr} A = 0$. We denote by $\kappa(A) = \|A\|_{\mathrm{op}}\|A^{-1}\|_{\mathrm{op}}$ the condition number of A. For functions $f, g\colon S \to \mathbb{R}$ on any set S, we say $f = O(g)$ if there is a constant $C > 0$ such that $f(x) \le Cg(x)$ for all $x \in S$, and similarly $f = \Omega(g)$ if there is a constant $c > 0$ such that $f(x) \ge cg(x)$ for all $x \in S$. If $f = O(g)$ and $g = O(f)$, we write $f = \Theta(g)$. In case C, c depend on another parameter λ, we write O_λ and Ω_λ, respectively. We abbreviate $[k] = \{1, \dots, k\}$ for $k \in \mathbb{N}$. All other notation is introduced in the remainder of the text as needed.

We can now formally define the tensor normal model, of which the matrix normal model is a particular case.

DEFINITION 1.8. For dimensions $d_1, \dots, d_k \in \mathbb{N}$, the *tensor normal model* is the family of centered multivariate Gaussian distributions with covariance matrix given by a Kronecker product $\Sigma = \Sigma_1 \otimes \cdots \otimes \Sigma_k$ of positive definite matrices, with $\Sigma_a \in \mathrm{PD}(d_a)$, $a \in [k]$, that is, the distributions $\mathcal{N}(0, \Sigma_1 \otimes \cdots \otimes \Sigma_k)$. For $k = 2$, this is known as the *matrix normal model*.

Note that each Σ_a is a $d_a \times d_a$ matrix and Σ is a $D \times D$-matrix, where $D = d_1 \cdots d_k$. Our goal is to estimate k Kronecker factors $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_k$ such that $\widehat{\Sigma}_a \approx \Sigma_a$ for each $a \in [k]$ given access to n i.i.d. random samples $x_1, \dots, x_n \in \mathbb{R}^D$ drawn from the model. A weaker requirement is to only approximate the full covariance, that is, $\widehat{\Sigma}_1 \otimes \cdots \otimes \widehat{\Sigma}_k \approx \Sigma$.

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 matrix with d_1 rows and D/d_1 columns, where a column is indexed by a tuple $(i_2 \in [d_2], \dots, i_k \in [d_k])$ and given by the vector in $\mathbb{R}^{d_1}$ with i_1^{st} entry equal to $(x_j)_{i_1,\dots,i_k}$. 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. In an analogous way, we may flatten it to a $d_2 \times d_1 d_3 \cdots d_k$ matrix, 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 $d_2 \cdots d_k$ many. The MLE attempts to decorrelate the columns to obtain rates such as those one would obtain if the columns were independent.

The MLE is easier to describe in terms of the precision matrices, which we now define.

DEFINITION 1.9 (Precision matrices). For a $D \times D$-covariance matrix Σ arising in the tensor normal model, we refer to $\Theta = \Sigma^{-1}$ as the *precision matrix*. We also define the *Kronecker factor precision matrices* $\Theta_1, \dots, \Theta_k$ as the unique positive-definite matrices such that $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k$ and $(\det \Theta_a)^{1/d_a}$ is the same for each $a \in [k]$. In other words, we choose $\Theta_a = \lambda \Theta'_a$ where $\det \Theta'_a = 1$ and $\lambda > 0$ is a constant (not depending on $a \in [k]$). We make this choice because the Kronecker factors of Θ are determined only up to a scalar.

Let $\mathbb{P}$ denote the parameter space of all precision matrices Θ for the tensor normal model with fixed dimensions $d_1, \dots, d_k$, that is,

$$\mathbb{P} = \{\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k : \Theta_a \in \mathrm{PD}(d_a)\}.$$

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

$$f_x(\Theta) = \frac{1}{nD} \sum_{i=1}^{n} x_i^T \Theta x_i - \frac{1}{D} \log\det \Theta. \tag{1.3}$$

The *maximum likelihood estimator (MLE)* for Θ is then defined as

$$\widehat{\Theta} := \operatorname*{arg\,min}_{\Theta \in \mathbb{P}} f_x(\Theta) \tag{1.4}$$

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 $(\widehat{\Theta}_1, \dots, \widehat{\Theta}_k)$ is *an* MLE for $(\Theta_1, \dots, \Theta_k)$ if $\bigotimes_{a=1}^k \widehat{\Theta}_a = \widehat{\Theta}$. Note that $\mathbb{P}$ is *not* a convex domain under the Euclidean geometry on the $D \times D$ matrices.

To state our results, and throughout this paper, we write $d_{\min} = \min_{1 \le a \le k} d_a$, $d_{\max} = \max_{1 \le a \le k} d_a$, and $D = \prod_{i=1}^k d_a$. Recall that we identify factors $\Theta_1, \dots, \Theta_k$ from Θ using the convention $\det \Theta_1^{1/d_1} = \cdots = \det \Theta_k^{1/d_k}$, and likewise for the MLE $\widehat{\Theta}$.

1.2. *Results on sample complexity and error bounds.* We begin with our result on the sample complexity for the tensor normal model.

THEOREM 1.10 (Tensor normal model sample complexity upper bounds). *There are universal constants $C, c_1, c_2 > 0$ such that the following holds. Suppose that $t \ge 1$ and*

$$n \ge Ck^2 \frac{d_{\max}^3}{D} t^2. \tag{1.5}$$

Then, with probability at least $1 - ke^{-c_1 t^2 d_{\max}} - k^2 (\frac{\sqrt{nD}}{k d_{\max}})^{-c_2 d_{\min}}$, the MLE $\widehat{\Theta}$ for n independent samples of the tensor normal model with precision matrix Θ is unique and satisfies

$$d_{\mathrm{FR}}(\widehat{\Theta}, \Theta) = O\left(\frac{\sqrt{k}\, d_{\max}}{\sqrt{n}} t\right) \quad \text{and} \quad d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a) = O\left(\frac{\sqrt{k d_a}\, d_{\max}}{\sqrt{nD}} t\right) \quad \text{for all } a \in [k].$$

Our error guarantees are tight for both the full precision matrix and the largest factor, as they match the lower bound for the simpler Gaussian estimation problem described in Proposition 4.1 up to the factor $\sqrt{k}$. Also note that the parameter t allows a trade-off between error guarantees and probabilistic guarantees. In particular, choosing $t^2 \approx \log n$ guarantees vanishing failure probability as $n \to \infty$.

For the matrix normal model $k = 2$, we obtain a stronger result:[4] first, we improve the *sample threshold* by a polynomial factor; second, we are able to bound the *error rate* for the individual factors in the tighter Thompson metric; and finally we improve the dependence on the *failure probability* from polynomial to exponential. Recall that we identify Θ_1, Θ_2 from Θ using the convention $\det \Theta_1^{1/d_1} = \det \Theta_2^{1/d_2}$.

[4]The key technical tool we use for our matrix normal model result is a sophisticated analysis of operator scaling from [14]. In order to lift this to the tensor normal model, we would need a similar analysis of the tensor scaling problem. This is significantly more difficult, as is discussed in more detail in, for example, [5].

THEOREM 1.11 (Matrix normal model sample complexity upper bounds). *There are universal constants $c, C > 0$ with the following property. Suppose $t \geq 1$ and*

$$n \geq C \frac{d_{\max}}{d_{\min}} \max\{\log d_{\max}, t^2 \log^2 d_{\min}\}.$$

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

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

with probability at least $1 - e^{-cd_{\min}t^2}$.

We again note that the parameter t allows for a trade-off between error and probabilistic guarantees, so in particular we can achieve vanishing failure probability as $n \to \infty$ by choosing, for example, $t^2 \approx \log n$. Further, we emphasize that the above error guarantees are tight for *both tensor factors*, matching the classical Gaussian lower bound in Proposition 4.1 for each individual tensor factor up to $\log d_{\min}$ factors.

Recalling Fact 1.2, we see that this stronger d_{op} guarantee recovers the optimal d_{FR} error rate for the tensor normal model up to $\log d_{\min}$ factors. Further, the sample threshold is also tight up to $\log d_{\min}$ factors, matching the known lower bound for Gaussian estimation. Finally, the guarantee in the Thompson metric gives much stronger accuracy for spectral applications such as PCA (see, e.g., [3]).

In applications such as brain fMRI, one is interested only in Θ_1, and Θ_2 is treated as a nuisance parameter. If the nuisance parameter Θ_2 were known, we could 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 4, we show a new lower bound for the matrix normal model that implies this better rate cannot hold. 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 distributed according to the matrix normal model with covariance matrix $\Sigma_1 \otimes \Sigma_2'$ for some $\Sigma_2' \in \text{PD}(d_2')$, the MLE for Y obtains a matching upper bound.

COROLLARY 1.12 (Estimating only Θ_1). *There is a universal constant $C > 0$ with the following property. Let $\Theta_1 \in \text{PD}(d_1)$, $\Theta_2 \in \text{PD}(d_2)$, X be distributed according to $\mathcal{N}(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$, and suppose that $1 < d_1 \leq d_2$ and $t \geq 1$. Let $Y = (Y_1, \ldots, Y_n)$ be the random variable obtained by removing all but*

$$d_2' = \min\left\{d_2, \frac{nd_1}{C \max\{\log n, t^2 \log^2 d_1\}}\right\}$$

columns of X_i for each $i \in [n]$. Then the MLE $\widehat{\Theta} = \widehat{\Theta}_1 \otimes \widehat{\Theta}_2$ for Y satisfies

$$d_{\text{op}}(\widehat{\Theta}_1, \Theta_1) = O\left(t\sqrt{\frac{d_1}{nd_2'}} \log d_1\right),$$

with probability $1 - e^{-\Omega(d_1 t^2)}$. This rate is tight up to factors of $\log d_1$ and $t^2 \log^2 d_1$.

TABLE 1
Worst-case sample requirements and error rates of estimators

Work	Setting	Sample threshold	Error rate (above sample threshold)
[21], Theorem 3	general, $k=2$	$\max\{1, \frac{\kappa^2}{d}\}\kappa^2 \min\{\kappa, d\} d \log d$	$\frac{\|\hat{\Theta}^{(3)}-\Theta\|_F}{\|\Theta_a\|_{\text{op}}} \lesssim \kappa^2\sqrt{\frac{d^2 \log d}{n}}$
[21], Theorem 4	$s \lesssim d$, $k=2$	$\max\{1, \frac{\kappa^2}{d}\}\kappa^2 \min\{\kappa, d\} \log d$	$\frac{\|\hat{\Theta}-\Theta\|_F}{\|\Theta\|_{\text{op}}} \lesssim \kappa^2\sqrt{\frac{d \log d}{n}}$
[25], Theorem 3.1	$k=2$, $s \le d^2$	$\max\left\{1, \frac{\kappa^2}{d}\right\}\kappa^2 \min\{\kappa, d\} \frac{(s+1)\log d}{d}$	$\frac{\|\hat{\Theta}-\Theta\|_{\text{op}}}{\|\Theta\|_{\text{op}}} \lesssim \kappa^2\sqrt{\frac{(s+1)\log d}{n}}$
[15]	$k \ge 2$ $s \le d^2$	$k^2(\min\{\kappa, d\})^{k-1} \max\{1, \frac{\kappa^2}{d}\}\kappa^2 \frac{(s+d)\log d}{d^{k-1}}$	$\frac{\|\hat{\Theta}_a-\Theta_a\|_F}{\|\Theta_a\|_F} \lesssim \kappa\sqrt{\frac{d(s+d)\log d}{nd^k}}$
Theorem 1.11	general, $k=2$	$\log^2 d$	$d_{\text{op}}(\hat{\Theta}_a, \Theta_a) \lesssim \sqrt{\frac{\log^2 d}{n}}$
Theorem 1.10	general, $k \ge 3$	$\frac{k^2 d^3}{d^k}$	$d_{\text{FR}}(\hat{\Theta}_a, \Theta_a) \lesssim \sqrt{\frac{kd^3}{nd^k}}$

Table 1 provides a high-level comparison of the above results and previous works. For clarity, we consider the simplified setting where all dimensions of the Kronecker factors are equal to d, all precision matrices are sparse with row sparsity r (which implies total sparsity $s \le rd$), and all condition numbers of precision factors are upper bounded by κ. A detailed comparison with all relevant parameters can be found in [10], Tables B.3 and B.4.

As can be seen from the table, our sample threshold and error rates are independent of condition number factors, and our error measures d_{FR} and d_{op} are tighter than those used in previous works, as can be seen in [10], Remark A.4, and [10], Proposition A.8. While prior works are able to give improved guarantees for sparse inputs, we note that they also have polynomial dependence on condition number. This becomes significant even for moderate values of condition number (e.g., $\kappa = d^2$), and so our estimator gives improved guarantees in the most general setting.

The table above has simplified and crude upper bounds. These can be improved by considering more precise quantities of each covariance. We give more detailed bounds in the Supplementary Material and omit the exact results, which the reader can find in the cited works directly.

1.3. *Results on the flip-flop algorithm for MLE estimation.* The MLEs for the matrix and tensor normal models can be computed by a natural iterative procedure that is known as the *flip-flop algorithm*. In Algorithm 1 below, we describe it for the matrix normal model ($k=2$), where the samples x_i can be viewed as $d_1 \times d_2$ matrices X_i. The general flip-flop algorithm is described in Algorithm 2 in Section 5.

We can motivate the flip-flop algorithm by noting that if in the first step we already have $\overline{\Theta}_2 = \Theta_2$ (the true precision factor), 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 of the sample covariance would be a good estimator for Θ_1. As we do not know Θ_2, the flip-flop algorithm instead uses $\overline{\Theta}_2$ as our current best guess, with the hope that each iteration will improve the next guess.

For the general tensor normal model (Algorithm 2), in each step the flip-flop algorithm chooses one of the dimensions $a \in [k]$ and uses the ath flattening of the samples x_i (which are just X_i and X_i^T in the matrix case) to update $\overline{\Theta}_a$.

Input: Samples $X = (X_1, \dots, X_n)$, where $X_i \in \mathbb{R}^{d_1 \times d_2}$, initial guess $\widetilde{\Theta} \in \mathbb{P}$. Parameters $T \in \mathbb{N}$ and $\delta > 0$.
Output: An estimate $\overline{\Theta} = \overline{\Theta}_1 \otimes \overline{\Theta}_2 \in \mathbb{P}$ of the MLE.
Algorithm:

1. Set $\overline{\Theta}_1 = \widetilde{\Theta}_1$ and $\overline{\Theta}_2 = \widetilde{\Theta}_2$.

2. For $t = 1, \dots, T$, repeat the following:

   - If t is odd, set $a = 1$ and $\Upsilon = \frac{1}{nd_2} \sum_{i=1}^n X_i \overline{\Theta}_2 X_i^T$. If t is even, set $a = 2$ and $\Upsilon = \frac{1}{nd_1} \sum_{i=1}^n X_i^T \overline{\Theta}_1 X_i$.
   - If $t > 1$ and $\|\nabla_a f_x(\overline{\Theta})\|_F \le \delta$, return $\overline{\Theta}$
   - Update $\overline{\Theta}_a \leftarrow \Upsilon^{-1}$

ALGORITHM 1. *Flip-flop algorithm for the matrix normal model.*

The advantage of flip-flop over other estimators are twofold: it directly converges to the MLE, as opposed to regularization approaches that trade-off accuracy for speed; and it has small iteration complexity. Each iteration of flip-flop is extremely fast to compute (one matrix inversion), whereas (most) other works have expensive complexity per iteration (solving a convex program). See details in [10], Section B.5.

Our next results show that the flip-flop algorithm can efficiently compute the MLE when the hypotheses of Theorem 1.10 or Theorem 1.11 hold. We state our result for the tensor normal model and then give an improved version for the matrix normal model.

THEOREM 1.13 (Tensor normal flip-flop). *There are universal constants $C, c, c_1, c_2 > 0$ such that the following holds. Suppose $x = (x_1, \dots, x_n)$ are $n \ge Ck^2 d_{\max}^3 / D$ independent samples from $\mathcal{N}(0, \Theta^{-1})$, where $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k$. Then, with probability at least*

$$1 - k\, e^{-c_1 \frac{nD}{k^2 d_{\max}^2}} - k^2 \left(\frac{\sqrt{nD}}{k d_{\max}} \right)^{-c_2 d_{\min}},$$

the MLE $\widehat{\Theta}$ exists, and for any $0 < \delta < \frac{c}{\sqrt{(k+1)d_{\max}}}$, the number of iterations T needed for Algorithm 2 to output $\overline{\Theta}$ with $d_{\mathrm{FR}}(\overline{\Theta}_a, \widehat{\Theta}_a) \le \sqrt{2d_a} \cdot \delta$ for all $a \in [k]$, is:

1. *when the initial guess is $\widetilde{\Theta}$ with $\nabla_0 f_x(\widetilde{\Theta}) = 0$,*

$$T = O\left(k^2 d_{\max} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) + k \log \frac{1}{\delta} \right)$$

2. *when the initial guess is $\widetilde{\Theta}$ with $\nabla_0 f_x(\widetilde{\Theta}) = 0$ and $d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) = O(\frac{1}{k d_{\max}})$,*

$$T = O\left(k \log\left(\frac{\sqrt{k d_{\max}} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta})}{\delta} \right) \right) = O\left(k \log \frac{1}{\delta} \right)$$

3. *without any initial guess (and starting from $\frac{1}{f_x(I_D)} \cdot I_D$),*

$$T = O\left(k^2 d_{\max} (1 + \log \kappa(\Theta)) + k \log \frac{1}{\delta} \right)$$

THEOREM 1.14 (Matrix normal flip-flop). *There are universal constants $C, c, c_1 > 0$ such that the following holds. Let $1 < d_1, d_2$. Suppose $x_1, \dots, x_n \in \mathbb{R}^{d_1 d_2}$ are*

$$n \ge C \frac{d_{\max}}{d_{\min}} \max\{\log d_{\max}, \log^2 d_{\min}\}$$

independent samples from $\mathcal{N}(0, (\Theta_1 \otimes \Theta_2)^{-1})$. With probability at least $1 - \exp(-\frac{c_1 \cdot n d_{\min}^2}{d_{\max} \log^2 d_{\min}})$, the MLE $\widehat{\Theta}$ exists, and for every $0 < \delta < \frac{c}{\sqrt{d_{\max}}}$, the number of iterations

TABLE 2
Performance of estimators without any assumptions, from initial guess $\widetilde{\Theta}$

Work	Setting	Main subroutine
[21], Theorem 3	$k = 2$, general	matrix inversion
[21], Theorem 4	$k = 2$, $s_a \lesssim d_a$	convex program
[25], Theorem 3.1	$k = 2$, general s_a	convex program
[25], Theorem 3.3	$k = 2$, $r_{s,a} \lesssim \sqrt{d_a}$	linear program
[24]	$k \geq 4$, general $r_{s,a}$	truncated gradient descent
[15]	$k \geq 2$, general s_a	convex program
Theorems 1.13 and 1.14	$k \geq 2$	matrix inversion

T needed for Algorithms 1 and 2 to output $\overline{\Theta}$ with $d_{\mathrm{FR}}(\overline{\Theta}_a, \widehat{\Theta}_a) = O(\sqrt{d_a}\delta)$ for $a \in \{1, 2\}$, is:

1. *when the initial guess is $\widetilde{\Theta}$ with $\nabla_0 f_x(\widetilde{\Theta}) = 0$,*
$$T = O\left(d_{\max} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) + \log \frac{1}{\delta}\right)$$
2. *when the initial guess is $\widetilde{\Theta}$ with $\nabla_0 f_x(\widetilde{\Theta}) = 0$ and $d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) = O(\frac{1}{d_{\max}})$,*
$$T = O\left(\log\left(\frac{\sqrt{d_{\max}} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta})}{\delta}\right)\right) = O\left(\log \frac{1}{\delta}\right)$$
3. *without any initial guess (and starting from $\frac{1}{f_x(I_D)} \cdot I_D$),*
$$T = O\left(d_{\max}(1 + \log \kappa(\Theta_1 \otimes \Theta_2)) + \log \frac{1}{\delta}\right)$$

Plugging in the error rates for the MLE from Theorems 1.10 and 1.11 into Theorems 1.13 and 1.14 (with $t = 1$) shows that the output of the flip-flop algorithm with $O(k^2 d_{\max}(1 + \log \kappa(\Theta)) + k \log(n))$ iterations is an efficiently computable estimator with the same statistical guarantees as we have shown for the MLE.

Table 2 summarizes the iteration complexity of previous works and of the above theorems, in the most general setting where one is not given any assumptions about the initial guess. We give a detailed comparison of performance in [10], Section B. Note that, while the number of iterations of the flip-flop algorithm is larger than in previous works, each iteration is much faster in our case (matrix inversion) than in previous works (which need to solve a convex program). This justifies the better performance of flip-flop in practical settings.

A key contribution of this work is that our estimator, the MLE, is well-defined independent of any additional information. In particular, we have decoupled our sample complexity analysis from the algorithmic analysis of our estimator. Thus, our initial guess assumption only affects the runtime of the algorithm, and not the sample complexity.

In the above, we see that the iteration complexity of the flip-flop algorithm depends on the condition number of the precision matrix, when we do not have any assumption on the initial guess (case 3 in Theorems 1.13 and 1.14). However, if we assume that we have an initial guess which is "close to the true precision matrix" (case 2) we show that Algorithms 1 and 2 achieve much faster convergence to the MLE. Note that we state in the above theorems that the initial guess is close enough to the MLE, but in the sample regimes of the above theorems, Theorems 1.10 and 1.11 tell us that the MLE is very close to the true precision matrix. This allows us to do a full comparison between the performance of flip-flop and other proposed estimators in several previously considered settings. For details, see [10], Section B.5.

2. Geodesic convexity, sample complexity and error bounds. We now explain how we use geodesic convexity, following a strategy similar to [11], to prove Theorem 1.10. The detailed proofs of all results in this section can be found in [10], Section D.

2.1. *Geodesic convexity.* The negative log-likelihood for the tensor normal model, that is, equation (1.4), is an optimization problem over the parameter space $\mathbb{P}$, which is a subset of the space $\mathrm{PD}(D)$ of positive-definite real symmetric $D \times D$ matrices. As we have discussed in the previous section, we will consider the Riemannian metric on $\mathrm{PD}(D)$ that arises from the Fisher information metric on centered Gaussians parametrized by their covariance matrices [20].[5] When we endow $\mathrm{PD}(D)$ with this metric, we see that the 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. Moreover, if A is an invertible matrix, the transformation $\Theta \mapsto A\Theta A^T$ is an isometry with respect to this metric, that is, it preserves the geodesic distance. This invariance is natural and desirable, as 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 distance, and Kullback–Leibler divergence are unchanged.

Another very useful property is that our domain $\mathbb{P}$ is a *totally geodesic submanifold* of $\mathrm{PD}(D)$: for any two points $A, B \in \mathbb{P}$, the entire geodesic between A and B remains in our domain $\mathbb{P}$. Thus, the negative log-likelihood is truly an optimization problem over the Riemannian manifold $\mathbb{P}$ under the Fisher information metric.

As $\mathbb{P}$ is a totally geodesic submanifold of $\mathrm{PD}(D)$, the invariance properties described above for $\mathrm{PD}(D)$ are directly inherited by $\mathbb{P}$. The manifold $\mathbb{P}$ carries a natural action by the group

$$\mathbb{G} = \{A = A_1 \otimes \cdots \otimes A_k : A_a \in \mathrm{GL}(d_a)\}.$$

Namely, if $\Theta \in \mathbb{P}$ and $A \in \mathbb{G}$ then $A\Theta A^T \in \mathbb{P}$. Thus, as discussed above, the map $\Theta \mapsto A\Theta A^T$ is an isometry of the Riemannian manifold $\mathbb{P}$, thereby preserving statistical relationship between the corresponding Gaussians.

As observed by [23], the negative log-likelihood function (equation (1.4)) is convex when restricted to geodesics of the Fisher information metric. In other words, the negative log-likelihood is *geodesically convex* on our manifold $\mathbb{P}$. To see this fact, we will now formally describe the structure of the manifold $\mathbb{P}$ and define geodesic convexity.

In the manifold $\mathbb{P}$, the tangent space at any point $\Theta \in \mathbb{P}$ is given by

$$\mathfrak{p} := \left\{ \sum_{i=1}^{k} I_{d_1} \otimes \cdots \otimes I_{d_{i-1}} \otimes \log(\Gamma_i) \otimes I_{d_{i+1}} \otimes \cdots I_{d_k} \;\middle|\; \Theta^{1/2} \Gamma \Theta^{1/2} \in \mathbb{P} \right\}$$

which can be identified with the real vector space

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

equipped with the following inner product and norm:

$$\langle H, K \rangle := H_0 K_0 + \sum_{a=1}^{k} \mathrm{Tr}\, H_a^T K_a, \qquad \|H\|_F := \langle H, H \rangle^{1/2}.$$

The direction $(1; 0, \ldots, 0)$ changes Θ by an overall scalar, and tangent directions supported only in the ath component for $a \in [k]$ only change Θ_a (subject to its determinant staying fixed). In order to make this inner product agree with the natural Frobenius inner product on the tangent space $\mathfrak{p}$, we parametrize the exponential map as in the following definition.

[5] This is the same as the metric arising from the Hessian of the log-determinant [2], Chapter 6.

DEFINITION 2.1 (Exponential map and geodesics). The *exponential map* $\exp_\Theta : \mathbb{H} \to \mathbb{P}$ at $\Theta = \Theta_1 \otimes \cdots \otimes \Theta_k \in \mathbb{P}$ is defined by

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

By definition, 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 $\mathbb{P}$.

The geodesics on $\mathbb{P}$ defined above are simply the geodesics of the Fisher information metric on $\mathrm{PD}(D)$, reparametrized in terms of the identification of the tangent space $\mathbb{H}$ given above.

We take the convention that the geodesics have unit speed if $\|H\|_F^2 = 1$. The geodesic distance $d(\Theta, \Theta')$ between two points Θ and $\Theta' = \exp_\Theta(H)$ is therefore equal to $\|H\|_F$ that can also be computed as $D^{-1/2} \|\log \Theta^{-1/2} \Theta' \Theta^{-1/2}\|_F$, which we will take to be our notion of geodesic distance. To summarize, we have the following.

DEFINITION 2.2 (Geodesic distance and balls). The *geodesic distance* $d(\Theta, \Theta')$ between two points Θ and Θ' of $\mathbb{P}$ is given by

$$d(\Theta, \Theta') := \frac{1}{\sqrt{D}} \left\|\log \Theta^{-1/2} \Theta' \Theta^{-1/2}\right\|_F = \sqrt{\frac{2}{D}} \cdot d_{\mathrm{FR}}(\Theta, \Theta'), \tag{2.1}$$

where log denotes the matrix logarithm and d_{FR} is the Fisher–Rao distance defined in equation (1.1).

The closed *(geodesic) ball* of radius $r > 0$ about Θ is defined as

$$B_r(\Theta) = \left\{\exp_\Theta(H) : H \in \mathbb{H}, \|H\|_F \le r\right\},$$

The manifold $\mathrm{PD}(D)$, and hence $\mathbb{P}$, is a *Hadamard manifold*, that is, a complete, simply connected Riemannian manifold of nonpositive sectional curvature [1]. 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 $t \in [0, 1]$.

The definition of geodesics yields the following notion of geodesic convexity of functions.

DEFINITION 2.3 (Geodesic convexity). Given a geodesically convex domain $\Gamma \subseteq \mathbb{P}$, a function f is *(strictly) geodesically convex* on Γ if, and only if, the function $t \mapsto f(\gamma(t))$ is (strictly) convex on $[0, 1]$ for any geodesic $\gamma(t)$ with $\gamma(0), \gamma(1) \in \Gamma$.

The function f is λ-*strongly* geodesically convex if $t \mapsto f(\gamma(t))$ is λ-strongly convex along every unit-speed geodesic γ with endpoints in Γ.

For a twice differentiable function $f : \mathbb{P} \to \mathbb{R}$, we say that it is λ-strong geodesically convex *at* Θ if $\partial_{t=0}^2 f(\exp_\Theta(tH)) \ge \lambda \|H\|_F^2$ for all $H \in \mathbb{H}$, and we say it is λ-strong geodesically convex on Γ if it is λ-strong geodesically convex for every $\Theta \in \Gamma$.

EXAMPLE 2.4. It is instructive to consider the case $k = 1$, or $\mathbb{P} = \mathrm{PD}(D)$. The geodesics through Θ are the curves $t \mapsto \sqrt{\Theta} e^{\sqrt{D} \cdot Ht} \sqrt{\Theta}$ where $H \in \mathbb{H}$. As an example of a geodesically convex function, consider the likelihood for the precision matrix of a Gaussian with data $x_1, \ldots, x_n$. 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 (1.3) as

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

We claim that $f_x(\Theta)$ is always geodesically convex, and in fact *strictly* geodesically convex whenever ρ is invertible. Indeed,

$$\partial_{t=0}^2 f_x\big(\sqrt{\Theta}e^{\sqrt{D}\cdot tH}\sqrt{\Theta}\big) = D \cdot \operatorname{Tr}\sqrt{\Theta}\rho\sqrt{\Theta}H^2 \geq 0$$

with strict inequality whenever ρ is invertible (and H nonzero).

The computation in the example easily generalizes to the tensor normal model, which allows us to prove geodesic convexity of the negative log-likelihood function in our setting.

We now formally define the *Riemannian* gradient and Hessian.

DEFINITION 2.5 (Gradient and Hessian). Let $f : \mathbb{P} \to \mathbb{R}$ be a differentiable function and $\Theta \in \mathbb{P}$. The *(Riemannian) gradient* $\nabla f(\Theta)$ is the unique element in $\mathbb{H}$ such that

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

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

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

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,

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

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, while $\nabla_{a0} f$ 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, 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 $\nabla^2 f(\Theta) \succeq \lambda I_{\mathbb{H}}$, that is, the Hessian is positive definite with eigenvalues larger than or equal to λ.

2.2. *Proof outline*. With the above definitions, we are able to state a proof plan for Theorem 1.10. Proofs of all claims not proved in this subsection can be found in the Supplementary Material [10]. The proof is a Riemannian version of the standard approach using strong convexity, and it goes by the following steps:

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 $x_1, \dots, x_n$ are distributed as n independent samples from $\mathcal{N}(0, I_D)$. By equivariance of the likelihood function, the MLE $\widehat{\Theta}(x')$ for the former is exactly $\Theta^{1/2}\widehat{\Theta}(x)\Theta^{1/2}$. By invariance of the Fisher–Rao metric, $d_{\mathrm{FR}}(\widehat{\Theta}(x'), \Theta) = d_{\mathrm{FR}}(\widehat{\Theta}(x), I_D)$; the same is true for d_{op}. This shows that to prove Theorem 1.10 it is enough to consider the case that $\Theta = I_D$, that is, the standard Gaussian.

2. *Bound the gradient*: Show that the gradient $\nabla f_x(I_D)$ is small with high probability.
3. *Strong convexity*: 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 with small gradient implies the optimizer cannot be far. Since in step 2 we show that the gradient *at the true covariance* is small, our approach will prove that the optimizer (i.e., the MLE) is not far from the true covariance.

We begin by formally stating the fact given in step 1, as we will use it in later sections.

FACT 2.6. *Let $x := (x_1, \dots, x_n)$ be a tuple of n independent samples of $\mathcal{N}(0, I_D)$, and $x'_i := \Theta^{-1/2} x_i$ be the corresponding samples of $\mathcal{N}(0, \Theta^{-1})$, with $x' := (x'_1, \dots, x'_n)$. If $\widehat{\Theta}(x)$, $\widehat{\Theta}(x')$ are the MLE's for the samples x, x', respectively, then $\widehat{\Theta}(x') = \Theta^{1/2}\widehat{\Theta}(x)\Theta^{1/2}$.*
Thus, $d_{\mathrm{FR}}(\widehat{\Theta}(x'), \Theta) = d_{\mathrm{FR}}(\widehat{\Theta}(x), I_D)$ and $d_{\mathrm{op}}(\widehat{\Theta}(x'), \Theta) = d_{\mathrm{op}}(\widehat{\Theta}(x), I_D)$.

The following lemma shows that strong convexity in a ball about a point where the gradient is sufficiently small implies the optimizer cannot be far. This lemma thus ensures that if we prove steps 2 and 3, then Theorem 1.10 follows.

LEMMA 2.7. *Let $f : \mathbb{P} \to \mathbb{R}$ be geodesically convex and twice differentiable. Let $\Theta \in \mathbb{P}$ be such that $\|\nabla f(\Theta)\|_F \le \delta$, and f is λ-strongly geodesically convex in a ball $B_r(\Theta)$ of radius $r > \frac{2\delta}{\lambda}$. Then the sublevel set $\{\Upsilon \in \mathbb{P} : f(\Upsilon) \le f(\Theta)\}$ is contained in the ball $B_{2\delta/\lambda}(\Theta)$, f has a unique minimizer $\widehat{\Theta}$, where $\widehat{\Theta} \in B_{\delta/\lambda}(\Theta)$, and $f(\widehat{\Theta}) \ge f(\Theta) - \frac{\delta^2}{2\lambda}$.*

Hence, we now need to carry out steps 2 and 3 in the plan above.

2.3. *Bounding the gradient.* Proceeding as step 2 of the plan from Section 2.2, we now compute the gradient of the objective function and bound it using 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. As in Example 2.4, let $\rho := \frac{1}{nD}\sum_i x_i x_i^T$ denote the "second sample moments," and rewrite the objective function (1.3) as

$$f_x(\Theta) = \operatorname{Tr}\rho\,\Theta - \frac{1}{D}\log\det\Theta. \tag{2.2}$$

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

$$\operatorname{Tr}\rho^{(J)} H = \operatorname{Tr}\rho\, H_{(J)} \tag{2.3}$$

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 (in the order determined by J) and as the identity on the rest. This property uniquely determines $\rho^{(J)}$. We write $\rho^{(a)}$ and $\rho^{(ab)}$ if $J = \{a\}$ and $J = \{a, b\}$, respectively.

If ρ is positive (semi)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 $\rho^{(J)}$ can be computed analogously to the discussion in Section 1.1: "flatten" the data x by regarding it as a $d_J \times N_J$ matrix $x^{(J)}$, where $N_J = \frac{nD}{d_J}$; then $\rho^{(J)} = \frac{1}{nD} x^{(J)}(x^{(J)})^T$. The gradient can be readily computed in terms of partial traces.

LEMMA 2.9 (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*

$$\nabla_a f_x = \sqrt{d_a}\left(\rho^{(a)} - \frac{\operatorname{Tr}\rho}{d_a} I_{d_a}\right) \quad \text{for } a \in [k], \tag{2.4}$$

$$\nabla_0 f_x = \operatorname{Tr}\rho - 1. \tag{2.5}$$

REMARK 2.10 (Gradient at other points from equivariance). 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. (2.4) and (2.5) 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 bounds on the norm of the gradient, as outlined in step 2 of Section 2.2.

PROPOSITION 2.11 (Gradient bound). *Let* $x = (x_1, \ldots, x_n)$ *consist of independent standard Gaussian random variables in* $\mathbb{R}^D$. *Suppose that* $0 < \varepsilon < 1$ *and* $n \geq \frac{d_{\max}^2}{D\varepsilon^2}$. *Then the following occurs with probability at least* $1 - 2(k+1)e^{-\varepsilon^2 nD/(8d_{\max})}$:

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

$$|\nabla_0 f_x| \leq \varepsilon.$$

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

2.4. *Strong convexity from expansion.* In this section, we establish our strong convexity result, step 3 of the plan from Section 2.2, in Proposition 2.18. The proposition 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 using quantum expansion techniques (Proposition 2.17), and then (in the Supplementary Material) we bound how the Hessian changes away from the identity; see [10], Lemma D.3. We then combine these results to prove Proposition 2.18.

Similar to our gradient calculations, we compute the components of the Hessian in terms of partial traces, but now we also need to consider two coordinates at a time.

LEMMA 2.12 (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*

$$\langle H, (\nabla_{aa}^2 f_x) H\rangle = d_a \operatorname{Tr}\rho^{(a)} H^2,$$

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

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*

$$\nabla_{0a}^2 f_x \mathrel{\widehat{=}} \sqrt{d_a}\left(\rho^{(a)} - \frac{\operatorname{Tr}\rho}{d_a} I_{d_a}\right) \mathrel{\widehat{=}} \nabla_{a0}^2 f_x \quad (\forall a \in [k]),$$

$$\nabla_{00}^2 f_x = \operatorname{Tr}\rho.$$

Here, we use the conventions from Definition 2.5. In particular, we identify $\nabla_{a0}^2 f_x$, which is a linear operator from $\mathbb{R}$ to the traceless symmetric matrices, with a traceless symmetric matrix, and similarly for its adjoint $\nabla_{0a}^2 f_x$. The notation $\mathrel{\widehat{=}}$ reminds us of these identifications.

REMARK 2.13 (Hessian at other points from equivariance). Analogously to Remark 2.10, 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 2.12 with ρ replaced by $\Theta^{1/2}\rho\,\Theta^{1/2}$.

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

$$\Phi^{(ab)}\colon \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_a) \quad \text{given by} \quad \langle H, \Phi^{(ab)}(K)\rangle = \operatorname{Tr}\rho^{(ab)}(H \otimes K) \tag{2.6}$$

to the traceless symmetric matrices. Equation (2.6) is a special case of a *completely positive map*, which is a linear map of the form

$$\Phi_A\colon \operatorname{Mat}(d_b) \to \operatorname{Mat}(d_a), \qquad \Phi_A(Z) = \sum_{i=1}^{N} A_i Z A_i^T \tag{2.7}$$

for $d_a \times d_b$ matrices $A_1, \ldots, A_N$. Completely positive maps are quantum analogues of nonnegative matrices. To see that $\Phi^{(ab)}$ is completely positive, 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 (2.7) can also be written as

$$\operatorname{vec}\big(\Phi_A(Z)\big) = \sum_{i=1}^{N} (A_i \otimes A_i)\operatorname{vec}(Z). \tag{2.8}$$

Let $\Phi^*\colon \operatorname{Mat}(d_a) \to \operatorname{Mat}(d_b)$ be 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*. Quantum expansion is a quantum analogue of expansion of a nonnegative matrix viewed as a bipartite graph.

DEFINITION 2.14 (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

$$\left\| \frac{\Phi(I_{d_b})}{\operatorname{Tr}\Phi(I_{d_b})} - \frac{I_{d_a}}{d_a} \right\|_{\mathrm{op}} \leq \frac{\varepsilon}{d_a} \quad \text{and} \quad \left\| \frac{\Phi^*(I_{d_a})}{\operatorname{Tr}\Phi^*(I_{d_a})} - \frac{I_{d_b}}{d_b} \right\|_{\mathrm{op}} \leq \frac{\varepsilon}{d_b}. \tag{2.9}$$

The map Φ is an *(ε, η)-quantum expander* if Φ is ε-doubly balanced and

$$\|\Phi\|_0 := \max_{\substack{H \in \operatorname{Mat}(d_a) \\ \text{traceless symmetric}}} \max_{\substack{K \in \operatorname{Mat}(d_b) \\ \text{traceless symmetric}}} \frac{\langle H, \Phi(K)\rangle}{\|H\|_F \|K\|_F} \leq \eta \frac{\operatorname{Tr}\Phi(I_{d_b})}{\sqrt{d_a d_b}}. \tag{2.10}$$

A $(0, \eta)$-quantum expander is called a *η-quantum expander*.

Quantum expanders originate in quantum information theory and quantum computation [13]. There one typically takes $d_a = d_b$ and $\varepsilon = 0$, so that equation (2.10) simplifies to $\|\Phi\|_0 \leq \eta$. Here, we follow the definitions of [11, 14], who recognized the connection between quantum expansion and spectral gaps of the Hessian for operator scaling.[6] The following lemma allows us to translate quantum expansion properties into strong convexity.

[6]Definition 2.14 is invariant under rescaling $\Phi \mapsto c\Phi$ for $c > 0$. We note that some of the above can be slightly simplified if one opts for a nonscale invariant definition.

LEMMA 2.15 (Strong convexity from expansion). *If the completely positive maps $\Phi^{(ab)}$ defined in equation (2.6) are (ε, η)-quantum expanders for every $a \neq b \in [k]$, then*

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

Assuming $k \geq 3$, the right-hand side is at most $k(\eta + \varepsilon)$.

We are concerned with $\Phi^{(ab)}$ that arise from random Gaussians. Just like random graphs give rise to good expanders, random completely positive maps (choosing Kraus operators at random from well-behaved distributions) yield good quantum expanders. When the Kraus operators are standard Gaussians, we have the following result by [18, 19].[7]

THEOREM 2.16 (Pisier). *Let $A_1, \ldots, A_N$ 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 equation (2.7), satisfies*

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

PROOF. Observe that

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

Here, we identify $\mathrm{Mat}(d_b) \cong \mathbb{R}^{d_b} \otimes \mathbb{R}^{d_b}$, 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 equation (2.8), the result now follows from [10], Theorem C.1, with $n = d_a$ and $m = d_b$. □

When the samples $x = (x_1, \ldots, x_n)$ are independent standard Gaussians in $\mathbb{R}^D$, the random completely positive maps $\Phi^{(ab)}$ have the same distribution as $\frac{1}{nD}\Phi_A$, where the Kraus operators $A_1, \ldots, A_N$ 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 2.16 once doubly 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 2.9, and similarly for the adjoint. Therefore, the completely positive maps $\Phi^{(ab)}$ are ε-doubly balanced if and only if, for all $a \in [k]$,

$$\sqrt{d_a} \|\nabla_a f_x\|_{\mathrm{op}} \leq \varepsilon \operatorname{Tr} \rho = (1 + \nabla_0 f_x)\varepsilon, \tag{2.11}$$

hence double balancedness can be controlled using the gradient bounds in Proposition 2.11.

Using Theorem 2.16, we can prove the following strong convexity result at the identity.

PROPOSITION 2.17 (Strong convexity at identity). *There is a universal constant $C > 0$ 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 $1 - k^2 (\frac{\sqrt{nD}}{k d_{\max}})^{-\Omega(d_{\min})}$,*

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

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

[7] Pisier's technical result is slightly different. We state and prove our variant of Pisier's theorem in [10], Theorem C.1, in the Supplementary Material.

We also prove a robustness result for the Hessian ([10], Lemma D.3), which implies that when our function is strongly convex at the identity then it is also strongly convex in an *operator norm* (Thompson metric—the d_{op} defined in Definition 1.1) ball about the identity. Accordingly, we obtain the following proposition.

PROPOSITION 2.18 (Strong convexity near identity). *There are constants $C, c > 0$ 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 $1 - k^2(\frac{\sqrt{nD}}{kd_{\max}})^{-\Omega(d_{\min})}$, the function f_x is $\frac{1}{2}$-strongly convex at any point $\Theta \in \mathbb{P}$ such that $d_{\mathrm{op}}(\Theta, I_D) \leq c$.*

While Proposition 2.18 uses d_{op} to quantify closeness, we can easily translate it into a statement in terms of the geodesic distance. Namely, under the same hypotheses f_x is $\frac{1}{2}$-strongly convex on the geodesic ball $B_r(I_D)$ of radius $r = c/\sqrt{(k+1)d_{\max}}$, where $c > 0$ is the universal constant from Proposition 2.18. This follows from the following lemma.

LEMMA 2.19. *For any $\Theta \in \mathbb{P}$, we have $d_{\mathrm{op}}(\Theta, I_D) \leq \sqrt{(k+1)d_{\max}} \cdot d(\Theta, I_D)$.*

2.5. *Tensor normal model: Sample complexity and error bounds.* We have all ingredients to prove Theorem 1.10 according to the plan in Section 2.2. Since the objective is strongly convex and its gradient is small with high probability, Lemma 2.7 implies the next result, which bounds the geodesic distance between the MLE and the true precision matrix.

PROOF OF THEOREM 1.10. By Fact 2.6, we may prove the theorem assuming $\Theta = I_D$. Assuming this, we now show that the minimizer of f_x is unique and is close to $\Theta = I_D$ with high probability. Recall from equation (1.5) that $n \geq Ck^2\frac{d_{\max}^3}{D}t^2$.

Let $c > 0$ be the constant from Proposition 2.18. Consider the two events:

1. $\|\nabla f_x\|_F \leq \delta := \sqrt{82k}\frac{d_{\max}}{\sqrt{nD}}t$.
2. f_x is λ-strongly convex over $B_r(I_D)$, where $\lambda = \frac{1}{2}$ and $r := \frac{c}{\sqrt{(k+1)d_{\max}}}$.

By our choice of parameters, where C is a large enough constant, we have

$$\frac{\delta}{\sqrt{82k}} < 1, \qquad n \geq \frac{d_{\max}^2}{D(\frac{\delta}{\sqrt{82k}})^2}, \qquad n \geq Ck\frac{d_{\max}^2}{D} \quad \text{and} \quad r > \frac{2\delta}{\lambda}.$$

Thus, Proposition 2.11, with $\varepsilon = \frac{\delta}{\sqrt{82k}}$, applies and it shows that the first event holds up to a failure probability of at most

$$2(k+1)e^{-(\frac{\delta}{\sqrt{82k}})^2\frac{nD}{8d_{\max}}} = ke^{-\Omega(t^2 d_{\max})}.$$

Moreover, Proposition 2.18 and Lemma 2.19 also apply, showing that the second event holds up to a failure probability of at most

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

By the above and the union bound, both events hold simultaneously with the claimed success probability. Thus, as the above two events hold, Lemma 2.7 applies (with our choice of δ and λ) and shows that the MLE $\widehat{\Theta}$ exists, is unique, and satisfies $d(\widehat{\Theta}, \Theta) \leq \frac{\delta}{\lambda} = 2\delta$.

Since $d(\widehat{\Theta}_a, \Theta_a) \leq d(\widehat{\Theta}, \Theta)$, by the relationship between geodesic distance and Fisher–Rao distance, we get

$$d_{\mathrm{FR}}(\widehat{\Theta}, \Theta) = \sqrt{\frac{D}{2}} \cdot d(\widehat{\Theta}, \Theta) \quad \text{and} \quad d_{\mathrm{FR}}(\widehat{\Theta}_a, \Theta_a) \leq \sqrt{\frac{d_a}{2}} \cdot d(\widehat{\Theta}_a, \Theta_a) \leq \sqrt{\frac{d_a}{2}} \cdot d(\widehat{\Theta}, \Theta),$$

which imply the desired distance bounds. □

3. Matrix normal model: Improved sample complexity and error bounds. We can prove a stronger result for the matrix normal model ($k = 2$). Theorem 1.11 improves over Theorem 1.10 in the following aspects:

1. it works over a better (i.e., *smaller*) sample threshold,
2. we obtain tight error bounds for the individual factors in *spectral distance* d_{op},
3. the failure probability is *inverse exponential* in the number of samples.

Recall that when $k = 2$, the samples can be viewed as $d_1 \times d_2$-matrices, denoted by X_i. From the samples, we construct the completely positive map $\Phi_X : \mathrm{Mat}(d_2) \to \mathrm{Mat}(d_1)$ defined as $\Phi_X(Z) := \sum_{i=1}^n X_i Z X_i^T$. The above improvements come from working directly with quantum expansion, via the spectral gap of the completely positive map Φ_X, instead of translating it into strong convexity.

One of our main technical results is the following theorem, which shows that the expansion parameter of the map can be made *constant* with *exponentially small* failure probability.

THEOREM 3.1 (Improved expansion). *There are universal constants $C > 0$ and $\eta \in (0, 1)$ such that the following holds. For $d_1 \le d_2$, $d_2 > 1$, let $X = (X_1, \ldots, X_n)$ be random $d_1 \times d_2$ matrices with independent standard Gaussian entries, where $n \ge C \frac{d_2}{d_1} \max\{\log d_2, t^2\}$ and $t \ge 1$. Then Φ_X is a $(t\sqrt{\frac{d_2}{nd_1}}, \eta)$-quantum expander with probability at least $1 - e^{-\Omega(d_2 t^2)}$.*

We prove Theorem 3.1 in [10], Section C.2, by the use of Cheeger's inequality. Our techniques are similar to the ones used in [11].[8]

To obtain our error bounds, we combine the above result on the quantum expansion with the work of [14], which gives us bounds in operator norm on how far the MLE is from our true precision matrices as a function of the expansion.

The above takes care of aspect 1 (estimating in operator norm with a reduced sample threshold) and aspect 3 (inverse exponential failure probability), as well as tight error bounds on the larger Kronecker factor of the precision matrix. Now, we need to work a bit more to get tight bounds on the smaller factor of the precision matrix. To get a better control on the smaller factor, the idea is to apply one step of the flip-flop algorithm to "renormalize" the samples such that the second (larger-dimensional) 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. In [10], Proposition E.6, we show that, even after the first step of flip-flop, the first component still enjoys the same concentration exploited in Proposition 2.11—thus, the total gradient has become smaller, but only the second component of the MLE estimate has changed. Thus, intuitively, the total change in the first component will be small. Combining [10], Proposition E.6, with [10], Lemma E.9, which shows robustness of quantum expansion, we are able to control the quantum expansion of the new completely positive map. Hence, we are again in position to employ [10], Corollary E.4, to get tight error bounds for the smaller Kronecker factor.

The detailed proof of Theorem 1.11 and the necessary claims are given in [10], Section E.

4. Lower bounds. In this section, we prove new lower bounds for estimating precision matrices in the matrix and tensor normal models. Proofs of all claims not proved in this section can be found in the Supplementary Material [10], Section F. We begin by stating a well-known lower bound for estimating unstructured precision matrices (the case $k = 1$).

[8] Theorem 3.1 also improves our result on strong convexity (Propositions 2.17 and 2.18) for $k = 2$. Indeed, for $k = 2$, using Theorem 3.1 in place of Theorem 2.16 improves the failure probability to $1 - e^{-\Omega(d_2 t^2)}$. However, we cannot use this to improve our results for $k \ge 3$ because Theorem 3.1 is not capable of proving subconstant quantum expansion. We need quantum expansion less than $1/(k-1)$ to obtain a nontrivial result from Lemma 2.15.

PROPOSITION 4.1 (Lower bound for unstructured Gaussians). *There is $c > 0$ such that the following holds. Let $\widehat{\Theta}$ be any estimator for $\Theta \in \mathrm{PD}(d)$ from a tuple X of n samples from $\mathcal{N}(0, \Theta^{-1})$. Let $B \subset \mathrm{PD}(d)$ be the operator norm ball about I_d of radius $1/2$. Then*:

1. *Let $\delta^2 = c \min\{1, \frac{d^2}{n}\}$. Then $\sup_{\Theta \in B} \Pr[d_{\mathrm{FR}}(\widehat{\Theta}, \Theta) \geq \delta] \geq \frac{1}{2}$.*
2. *Let $\delta^2 = c \min\{1, \frac{d}{n}\}$. Then $\sup_{\Theta \in B} \Pr[d_{\mathrm{op}}(\widehat{\Theta}, \Theta) \geq \delta] \geq \frac{1}{2}$.*

As a consequence (*see Remark* 1.5), *we have*

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

Having the lower bound above in mind, we now discuss what is needed to prove a lower bound for the matrix normal model. In this section, we assume, without loss of generality, that $d_2 \geq d_1 \geq 1$ and we are given samples $X_1, \ldots, X_n \in \mathbb{R}^{d_1 \times d_2}$ distributed according to $\mathrm{vec}(X) \sim N(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$. If Θ_1 was known, we could compute $Y := \Theta_1^{1/2} X$, which "decorrelates" the rows of X and, therefore, we could treat the rows of Y as nd_1 independent samples from $N(0, \Theta_2^{-1})$. If $nd_1 \leq cd_2$ for some small enough $c > 0$ (i.e., $n < cd_2/d_1$), the above $k = 1$ lower bound would imply that we cannot estimate Θ_2 to constant accuracy in the operator norm even if we had complete knowledge of Θ_1.

Since $d_2 \geq d_1$, we could hope for better results for estimating Θ_1, since we intuitively have more samples for this mode. Namely, assume we knew Θ_2 and pre-process $Y := X\Theta_2^{1/2}$ to "decorrelate" the columns of X, which means we could treat the columns of Y as nd_2 independent samples from $N(0, \Theta_1^{-1})$. In this case, we could estimate Θ_1 in operator norm with 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 $n \ll d_2/d_1$, it is impossible to estimate Θ_1 as well as one could if Θ_2 were known.

THEOREM 4.2 (Lower bound for matrix normal models). *There is $c > 0$ such that the following holds. Let $d_1 \leq d_2$, $\Theta_1 \in \mathrm{PD}(D_1)$, $\Theta_2 \in \mathrm{PD}(d_2)$ and $\widehat{\Theta}_1$ be any estimator for Θ_1 from a tuple X of n samples of $\mathcal{N}(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$. Let $B \subset \mathrm{PD}(d_1)$ denote the ball about I_{d_1} of radius $1/2$ in the operator norm. Then*:

1. *Let $\delta^2 = c \min\{1, \frac{d_1^2}{n \min\{nd_1, d_2\}}\}$. Then $\sup_{\substack{\Theta_1 \in B \\ \Theta_2 \in \mathrm{PD}(d_2)}} \Pr[d_{\mathrm{FR}}(\widehat{\Theta}_1, \Theta_1) \geq \delta] \geq \frac{1}{2}$.*
2. *Let $\delta^2 = c \min\{1, \frac{d_1}{n \min\{nd_1, d_2\}}\}$. Then $\sup_{\substack{\Theta_1 \in B \\ \Theta_2 \in \mathrm{PD}(d_2)}} \Pr[d_{\mathrm{op}}(\widehat{\Theta}_1, \Theta_1) \geq \delta] \geq \frac{1}{2}$.*

As a consequence, we have

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

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

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. This 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 $\sqrt{d_1/nd_2}$ for Θ_1 in the operator norm using, for example, Tyler's M-estimator [11]. We now state the main tool in establishing the lower bound.

LEMMA 4.3. *Let X denote a tuple of n samples from $\mathcal{N}(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$ and let $\widehat{\Theta}_1(X)$ be any estimator for Θ_1. Let Y be a tuple of $n \min\{nd_1, d_2\}$ samples from $\mathcal{N}(0, \Theta_1^{-1})$. For every $\delta > 0$, there is a distribution on Θ_2 and an estimator $\tilde{\Theta}(Y)$ 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.*

We will use this lemma to show Theorem 4.2 in the contrapositive: if there was a good estimator for the matrix normal model, then we could use this to produce a good estimator for Gaussian estimation. Namely, given Gaussian samples $Y \sim N(0, \Theta_1^{-1})$, we could simulate samples $X \sim N(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$ from the matrix normal model by considering $X_i := (Y_{i,1} \cdots Y_{i,d_2})\sqrt{\Theta_2}$, that is, grouping d_2 columns into a matrix and applying $\sqrt{\Theta_2}$ on the right. Then by the above lemma, if $\hat{\Theta}_1(X)$ is a good estimator of Θ_1, then $\tilde{\Theta}(Y)$ is also a good estimator for Θ_1. We now give the formal proof of Theorem 4.2.

PROOF OF THEOREM 4.2. To show claim 1, let $\delta^2 \le c \min\{1, \frac{d_1^2}{n \min\{nd_1, d_2\}}\}$. Let Θ_2 be distributed as in Lemma 4.3 so that, as guaranteed by Lemma 4.3 for $n \min\{nd_1, d_2\}$ samples $Y \sim N(0, \Theta_1^{-1})$ there is an estimator $\tilde{\Theta}(Y)$ satisfying $D_{\mathrm{TV}}(\widehat{\Theta}_1(X), \tilde{\Theta}(Y)) \le \delta_0$. Here, X is distributed according to the matrix normal model $X \sim N(0, \Theta_1^{-1} \otimes \Theta_2^{-1})$. Proposition 4.1 implies

$$\sup_{\Theta_1 \in B} \Pr_Y\big[d_{\mathrm{FR}}\big(\tilde{\Theta}(Y), \Theta_1\big) \ge \delta\big] \ge \frac{1}{2}.$$

Clearly, we have

$$\sup_{\substack{\Theta_1 \in B,\\ \Theta_2 \in \mathrm{PD}(d_2)}} \Pr_X\big[d_{\mathrm{FR}}\big(\widehat{\Theta}_1(X), \Theta_1\big) \ge \delta\big] \ge \sup_{\Theta_1 \in B} \Pr_{\Theta_2, X}\big[d_{\mathrm{FR}}\big(\widehat{\Theta}_1(X), \Theta_1\big) \ge \delta\big].$$

On the other hand, since the distributions of $\hat{\Theta}_1(X)$ and $\tilde{\Theta}(Y)$ differ by at most δ_0 in total variation distance, this implies

$$\begin{aligned}\sup_{\Theta_1 \in B} \Pr_{\Theta_2, X}\big[d_{\mathrm{FR}}\big(\widehat{\Theta}_1(X), \Theta_1\big) \ge \delta\big] &\ge \sup_{\Theta_1 \in B} \Pr_Y\big[d_{\mathrm{FR}}\big(\tilde{\Theta}(Y), \Theta_1\big) \ge \delta\big] - \delta_0 \\ &\ge \frac{1}{2} - \delta_0.\end{aligned}$$

Allowing $\delta_0 \to 0$ implies claim 1. To prove claim 2, replace d_{FR} by d_{op} in the above. □

We remark that the proof of Theorem 4.2 uses no properties about d_{FR} or d_{op}. Therefore, the above proof implies that any lower bound for estimating a Gaussian with $n \min\{nd_1, d_2\}$ samples transfers similar to the matrix normal model. The above strategy can clearly be lifted to the tensor normal model by considering more components.

THEOREM 4.4 (Lower bound for tensor normal models). *There is $c > 0$ such that the following holds. Let $\Theta_1 \in \mathrm{PD}(d_1)$, $\Theta_a \in \mathrm{PD}(d_a)$ for $a \in [k]$ and $\widehat{\Theta}_1$ be any estimator for Θ_1 from a tuple X of n samples of $\mathcal{N}(0, \bigotimes_{a\in[k]} \Theta_a^{-1})$. Let $B \subset \mathrm{PD}(d_1)$ denote the ball about I_{d_1} of radius $1/2$ in the operator norm. Then*:

1. *Let* $\delta^2 = c \min\{1, \frac{d_1^2}{n \min\{nd_1, D/d_1\}}\}$. *Then* $\sup_{\substack{\Theta_1\in B \\ \Theta_a\in\mathrm{PD}(d_a), 1\neq a\in[k]}} \Pr[d_{\mathrm{FR}}(\widehat{\Theta}_1, \Theta_1) \geq \delta] \geq \frac{1}{2}$.
2. *Let* $\delta^2 = c \min\{1, \frac{d_1}{n \min\{nd_1, D/d_1\}}\}$. *Then* $\sup_{\substack{\Theta_1\in B \\ \Theta_a\in\mathrm{PD}(d_a), 1\neq a\in[k]}} \Pr[d_{\mathrm{op}}(\widehat{\Theta}_1, \Theta_1) \geq \delta] \geq \frac{1}{2}$.

As a consequence, we have

$$\sup_{\Theta_1\in B, \Theta_a\in\mathrm{PD}(d_a), 1\neq a\in[k]} \mathbb{E}\big[d_{\mathrm{FR}}(\widehat{\Theta}_1, \Theta_1)^2\big] = \Omega\left(\min\left\{\frac{d_1^2}{n \min\{nd_1, D/d_1\}}, 1\right\}\right) \quad \textit{and}$$

$$\sup_{\Theta_1\in B, \Theta_a\in\mathrm{PD}(d_a), 1\neq a\in[k]} \mathbb{E}\big[d_{\mathrm{op}}(\widehat{\Theta}_1, \Theta_1)^2\big] = \Omega\left(\min\left\{\frac{d_1}{n \min\{nd_1, D/d_1\}}, 1\right\}\right).$$

5. Iteration complexity of the flip-flop algorithm. We now prove Theorems 1.13 and 1.14, which state fast convergence of the flip-flop algorithm to the MLE with high probability. Detailed proofs of our main technical result, Theorem 5.2, along with the claims needed to prove it, can be found in [10], Section G, in the Supplementary Material.

We state the flip-flop algorithm for the general tensor normal model in Algorithm 2. It generalizes Algorithm 1 presented earlier in Section 1.3 for the matrix normal.

REMARK 5.1 (Matrix flip-flop from tensor flip-flop). To see how Algorithm 1 arises from Algorithm 2, note that if we update $\overline{\Theta}_a$ in the tth iteration, then the corresponding gradient component vanishes in the subsequent iteration. Since for the matrix normal model there are only two gradient components to consider, this means that the algorithm will necessarily alternate between updating $\overline{\Theta}_1$ and $\overline{\Theta}_2$. In other words, for the matrix normal model the algorithm truly "flip-flops" between the two coordinates. Moreover, [10], Lemma G.1, shows that $\operatorname{Tr}\rho_t = 1$ from the second iteration of Algorithm 2 onwards. Therefore, Algorithm 1 agrees with Algorithm 2 except that in the first iteration we skip the stopping condition and always update $\overline{\Theta}_1$. This will not impact the analysis, as one can see in [10], Lemma G.5.

Input: Samples $x = (x_1, \ldots, x_n)$, where each $x_i \in \mathbb{R}^D = \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k}$. Parameters $T \in \mathbb{N}$ and $\delta > 0$. Initial guess $\widetilde{\Theta} \in \mathbb{P}$ satisfying $\operatorname{Tr}[\rho\widetilde{\Theta}] = 1$.
Output: An estimate $\overline{\Theta} = \overline{\Theta}_1 \otimes \cdots \otimes \overline{\Theta}_k \in \mathbb{P}$ of the MLE.
Algorithm:

1. Set $\overline{\Theta}_a = \widetilde{\Theta}_a$ for each $a \in [k]$.

2. For $t = 1, \ldots, T$, repeat the following:

   - Compute $\rho_t = \frac{1}{nD} \cdot \overline{\Theta}^{1/2}(\sum_{i=1}^n x_i x_i^T)\overline{\Theta}^{1/2}$, where $\overline{\Theta} = \overline{\Theta}_1 \otimes \cdots \otimes \overline{\Theta}_k$.
   - Compute each component of the gradient using the formula $\nabla_a f_x(\overline{\Theta}) = \sqrt{d_a}(\rho_t^{(a)} - \operatorname{Tr}(\rho_t)\frac{I_{d_a}}{d_a})$, where $\rho_t^{(a)}$ denotes the partial trace (Definition 2.8), and find the index $a \in [k]$ for which $\|\nabla_a f_x(\overline{\Theta})\|_F$ is largest.
   - If $\|\nabla_a f_x(\overline{\Theta})\|_F \leq \delta$, return $\overline{\Theta}$
   - Update $\overline{\Theta}_a \leftarrow \frac{1}{d_a}\overline{\Theta}_a^{1/2}(\rho_t^{(a)})^{-1}\overline{\Theta}_a^{1/2}$.

ALGORITHM 2. *Flip-flop algorithm for the tensor normal model* ($k \geq 2$).

The key insight is that given appropriate initial conditions on the samples (which we later show to hold under the same sample requirements as for our results on the MLE), the flip-flop algorithm will converge quickly to the MLE. Namely, we show that the MLE is in a constant size operator norm ball around the true precision matrix, where the negative log-likelihood function f_x is strongly geodesically convex. This implies that f_x is strongly geodesically convex in a small geodesic ball around the MLE. Hence, any point with sufficiently small gradient of f_x is contained in a sublevel set on which f_x is strongly geodesically convex ([10], Lemma G.2). Such a point is found in polynomially many iterations of the flip-flop algorithm ([10], Lemma G.5). Then strong convexity implies that a δ-minimizer is found in $O(\log(1/\delta))$ further iterations ([10], Lemma G.3). Thus, we obtain the main technical result of this section.

THEOREM 5.2 (Convergence from initial conditions). *Let $\Theta \in \mathbb{P}$ be our true precision matrix, $x_1, \dots, x_n \in \mathbb{R}^D$ our samples, $\lambda > 0$ and $0 < \zeta \le \min\{1, 16\sqrt{(k+1)(k-1)}/\lambda\}$ such that*:

1. *f_x is λ-strongly geodesically convex at any $\Theta' \in \mathbb{P}$ such that $d_{\mathrm{op}}(\Theta', \Theta) \le \zeta$.*
2. *$|\nabla_0 f_x(\Theta)| \le 1/2$.*
3. *The MLE $\widehat{\Theta}$ exists and satisfies $d_{\mathrm{op}}(\widehat{\Theta}, \Theta) \le \zeta/2$.*

Then, for every $0 < \delta < \lambda\zeta/16\sqrt{(k+1)d_{\max}}$, the number of iterations T needed for Algorithm 2 to output $\overline{\Theta}$ with $d_{\mathrm{FR}}(\overline{\Theta}_a, \widehat{\Theta}_a) \le \sqrt{\frac{d_a}{2}} \cdot \frac{\delta}{\lambda}$ for all $a \in [k]$ is:

1. *when the initial guess is $\widetilde{\Theta}$ with $\nabla_0 f_x(\widetilde{\Theta}) = 0$,*

$$T = O\left(\frac{k^2 d_{\max}}{\zeta^2\lambda^2} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) + \frac{k}{\lambda}\log\left(\frac{\lambda\zeta}{\delta \cdot \sqrt{k d_{\max}}}\right)\right)$$

2. *if the initial guess $\widetilde{\Theta}$ satisfies $\nabla_0 f_x(\widetilde{\Theta}) = 0$ and $d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta}) \le \frac{\lambda\zeta}{100 d_{\max}\sqrt{k(k+1)}}$, then*

$$T = O\left(\frac{k}{\lambda}\log\left(\frac{\sqrt{k d_{\max}} \cdot d_{\mathrm{op}}(\widetilde{\Theta}, \widehat{\Theta})}{\delta}\right)\right) = O\left(k \log\frac{1}{\delta}\right)$$

3. *with initial guess $\frac{1}{f_x(I_D)} \cdot I_D$,*

$$T = O\left(\frac{k^2 d_{\max}}{\zeta^2\lambda^2} \cdot \log\kappa(\Theta) + \frac{k}{\lambda}\log\left(\frac{\lambda\zeta}{\delta \cdot \sqrt{k d_{\max}}}\right)\right)$$

With Theorem 5.2 at hand, fast convergence of the flip-flop algorithm for both the matrix and tensor normal models follow simply by proving that the initial conditions above will be satisfied with high probability, given a high enough number of samples. More precisely, we show that the sample complexity results of Section 2 already imply the conditions of Theorem 5.2, thereby proving Theorems 1.13 and 1.14.

PROOF OF THEOREM 1.13. For $\lambda = \frac{1}{2}$, $0 < \zeta < 1$ a sufficiently small universal constant, and $r = \frac{\zeta}{\sqrt{(k+1)d_{\max}}}$, consider the following events (i.e., the conditions of Theorem 5.2):

1. f_x is λ-strongly geodesically convex at any $\Theta' \in \mathbb{P}$ such that $d_{\mathrm{op}}(\Theta', \Theta) \le \zeta$. In particular, f_x is λ-strongly geodesically convex on the geodesic ball $B_r(\Theta)$.
2. $\|\nabla f_x(\Theta)\|_F < \frac{r\lambda}{2}$. In particular, $|\nabla_0 f_x(\Theta)| < \frac{1}{2}$.
3. The MLE $\widehat{\Theta}$ exists and satisfies $d(\widehat{\Theta}, \Theta) \le r/2$. In particular, $d_{\mathrm{op}}(\widehat{\Theta}, \Theta) \le \zeta/2$.

We first bound the success probability of these events similar to the proof of Theorem 1.10. For this, we may assume without loss of generality that $\Theta = I_D$ by Remarks 2.10 and 2.13. Then the first event holds with probability at least $1 - k^2(\frac{\sqrt{nD}}{kd_{\max}})^{-\Omega(d_{\min})}$ by Proposition 2.18 and Lemma 2.19, provided we choose C large enough and ζ small enough universal constants. For the second event, we apply Proposition 2.11 with

$$\varepsilon = \frac{1}{10\sqrt{k}}\frac{r\lambda}{2} = \frac{\zeta}{40\sqrt{k(k+1)d_{\max}}},$$

which satisfies $\varepsilon < 1$ and $n \geq \frac{d_{\max}^2}{D\varepsilon^2}$ provided we choose ζ sufficiently small and C sufficiently large universal constants. With these choices, the second event holds with probability at least

$$1 - 2(k+1)e^{-\varepsilon^2 \frac{nD}{8d_{\max}}} = 1 - k\, e^{-\Omega(\frac{nD}{k^2 d_{\max}^2})}.$$

Thus, the two events hold simultaneously with the desired success probability by the union bound. Moreover, by Lemma 2.7, the events 1 and 2 together also imply event 3. The above shows that the conditions of Theorem 5.2 are satisfied. Thus, the iteration complexity of Algorithm 2 follows from Theorem 5.2. □

PROOF OF THEOREM 1.14. Consider the events below for constants $\lambda, \zeta \in (0, 1)$:

1. f_x is λ-strongly geodesically convex at any $\Theta' \in \mathbb{P}$ such that $d_{\mathrm{op}}(\Theta', \Theta) \leq \zeta$.
2. $|\nabla_0 f_x(\Theta)| \leq 1/2$.
3. The MLE $\widehat{\Theta}$ exists and satisfies $d_{\mathrm{op}}(\widehat{\Theta}, \Theta) \leq \zeta/2$.

To bound the success probability of these events, we may assume without loss of generality that $\Theta = I_D$ by Remarks 2.10 and 2.13. We will also assume that $d_1 \leq d_2$.

If $\lambda \in (0, 1)$ is a suitable universal constant, C is a large enough universal constant, and ζ is a small enough universal constant, by [10], Corollary E.5, with $t^2 = nd_1/d_2$, the first event holds with probability at least $1 - e^{-\Omega(nd_1)} \geq 1 - e^{-\Omega(nd_1^2/d_2 \log^2 d_1)}$ in view of our assumption on n.

The second event holds with probability at least $1 - e^{-\Omega(nD)}$ by [10], Proposition D.2. Finally, by Theorem 1.11 with $t^2 = nd_1/d_2 \log^2 d_1$ (which can be made larger than 1 by our assumption on n assuming C is large enough), the third event holds with probability at least $1 - e^{-\Omega(nd_1^2/d_2 \log^2 d_1)}$.

Event 3 follows from Theorem 1.11 via the fact that $d_{\mathrm{op}}(\widehat{\Theta}_1 \otimes \widehat{\Theta}_2, \Theta_1 \otimes \Theta_2) \leq d_{\mathrm{op}}(\widehat{\Theta}_1, \Theta_1) + d_{\mathrm{op}}(\widehat{\Theta}_2, \Theta_2)$. Thus, with probability at least $1 - e^{-\Omega(nd_1^2/d_2 \log^2 d_1)}$ all three events hold simultaneously, by the union bound, meaning the conditions of Theorem 5.2 are satisfied. Thus, the iteration complexity of Algorithms 1 and 2 follows from Theorem 5.2. □

6. Conclusion and open problems. In this work, we almost optimally address the fundamental question of parameter estimation for the matrix and tensor normal models, as well as the question of efficient computation of this estimator. Contrary to the state of the art for unstructured covariance estimation (i.e., $k = 1$), all previous existing results (in their sample complexity bounds as well as the error rates and guarantees of their estimators) depended on the *condition number* of the true covariance matrices and on a *sufficiently accurate starting guess* and, therefore, had suboptimal guarantees in the general case. By proving strong convexity in the geometry induced by the Fisher information metric, we remedy these issues and obtain nearly optimal estimates (*without* dependence on condition number) in the strongest possible metrics, namely the Fisher–Rao and Thompson distances. As a consequence, we

also control other equivariant statistical distances such as relative entropy and total variation distance.

In particular, we showed that the maximum likelihood estimator (MLE) for the covariance matrix in the matrix normal model 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 $O(d_{\max}(1 + \log \kappa(\Theta)) + \log n)$ iterations (see the discussion after Theorem 1.14) is an efficiently computable estimator with statistical guarantees comparable to those we show for the MLE.

Our main open question is whether the sample threshold requirement $n = \Omega(k^2 d_{\max}^3 / D)$ for Theorem 1.10 can be weakened to $n = \Omega(k^2 d_{\max}^2 / D)$ for $k \geq 3$. Equivalently, do the guarantees of Theorem 1.10 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 1.11. Filling in this gap will place the tensor normal model on the same sound theoretical footing as unstructured covariance estimation.

Acknowledgments. CF acknowledges Ankur Moitra for interesting discussions and Shuheng Zhou for sharing the code for her Gemini estimator. All authors would like to thank the anonymous reviewers for their reviews and suggestions.

MW is also affiliated with the Faculty of Computer Science of Ruhr-Universität Bochum and the Korteweg-de Vries Institute for Mathematics and QuSoft at the University of Amsterdam.

Funding. AR and MW acknowledge support by the Dutch Research Council (NWO grant OCENW.KLEIN.267). MW furthermore acknowledges supported by the European Union (ERC Grant Agreement No. 101040907) and the German Federal Ministry of Research, Technology and Space (QuBRA, 13N16135; QuSol, 13N17173).

SUPPLEMENTARY MATERIAL

Supplement to "Near optimal sample complexity for matrix and tensor normal models via geodesic convexity" (DOI: 10.1214/25-AOS2539SUPP; .pdf). The full proofs of the claims mentioned above and a full discussion and comparison between this work and previous works can be found in [10].

REFERENCES

[1] Bačák, M. (2014). *Convex Analysis and Optimization in Hadamard Spaces*. *De Gruyter Series in Nonlinear Analysis and Applications* **22**. de Gruyter, Berlin. MR3241330 https://doi.org/10.1515/9783110361629

[2] Bhatia, R. (2009). *Positive Definite Matrices*. *Princeton Series in Applied Mathematics*. Princeton Univ. Press, Princeton, NJ. MR2284176

[3] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[4] Brown, P. J., Kenward, M. G. and Bassett, E. E. (2001). Bayesian discrimination with longitudinal data. *Biostatistics* **2** 417–432.

[5] Bürgisser, P., Franks, C., Garg, A., Oliveira, R., Walter, M. and Wigderson, A. (2018). Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes. In 59*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2018 883–897. IEEE Computer Soc., Los Alamitos, CA. MR3899650 https://doi.org/10.1109/FOCS.2018.00088

[6] 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* 60*th Annual Symposium on Foundations of Computer Science* 845–861. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228202 https://doi.org/10.1109/FOCS.2019.00055

[7] 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 9*th Innovations in Theoretical Computer Science*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **94** Art. No. 24, 20 pp. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3761760

[8] ČENCOV, N. N. (1978). Algebraic foundation of mathematical statistics. *Math*. *Oper.forsch*. *Stat*., *Ser. Stat*. **9** 267–276. MR0512264 https://doi.org/10.1080/02331887808801428

[9] DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J*. *Stat*. *Comput*. *Simul*. **64** 105–123.

[10] FRANKS, C., OLIVEIRA, R., RAMACHANDRAN, A. and WALTER, M. (2026). Supplement to "Near optimal sample complexity for matrix and tensor normal models via geodesic convexity." https://doi.org/10.1214/25-AOS2539SUPP

[11] FRANKS, W. C. and MOITRA, A. (2020). Rigorous guarantees for Tyler's M-estimator via quantum expansion. In *Conference on Learning Theory* 1601–1632. PMLR.

[12] GARG, A., GURVITS, L., OLIVEIRA, R. and WIGDERSON, A. (2020). Operator scaling: Theory and applications. *Found*. *Comput*. *Math*. **20** 223–290. MR4081171 https://doi.org/10.1007/s10208-019-09417-z

[13] HASTINGS, M. B. (2007). Random unitaries give quantum expanders. *Phys*. *Rev. A* (3) **76** 032315, 11 pp. MR2486279 https://doi.org/10.1103/PhysRevA.76.032315

[14] KWOK, T. C., LAU, L. C. and RAMACHANDRAN, A. (2019). Spectral analysis of matrix scaling and operator scaling. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 1184–1204. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228221 https://doi.org/10.1109/FOCS.2019.00074

[15] LYU, X., WEI SUN, W., WANG, Z., LIU, H., YANG, J. and CHENGK, G. (2020). Tensor graphical model: Non-convex optimization and statistical inference. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **8** 2024–2037.

[16] MARDIA, K. V. and GOODALL, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*. *North-Holland Ser. Statist*. *Probab*. **6** 347–386. North-Holland, Amsterdam. MR1268443

[17] MITCHELL, T. M., HUTCHINSON, R., NICULESCU, R. S., PEREIRA, F., WANG, X., JUST, M. and NEWMAN, S. (2004). Learning to decode cognitive states from brain images. *Mach*. *Learn*. **57** 145–175.

[18] PISIER, G. (2012). Grothendieck's theorem, past and present. *Bull*. *Amer. Math*. *Soc*. (*N.S*.) **49** 237–323. MR2888168 https://doi.org/10.1090/S0273-0979-2011-01348-9

[19] PISIER, G. (2014). Quantum expanders and geometry of operator spaces. *J*. *Eur. Math*. *Soc*. (*JEMS*) **16** 1183–1219. MR3226740 https://doi.org/10.4171/JEMS/458

[20] SKOVGAARD, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scand*. *J*. *Stat*. **11** 211–223. MR0793171

[21] TSILIGKARIDIS, T., HERO, A. O. III and ZHOU, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Trans*. *Signal Process*. **61** 1743–1755. MR3038387 https://doi.org/10.1109/TSP.2013.2240157

[22] WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans*. *Signal Process*. **56** 478–491. MR2445531 https://doi.org/10.1109/TSP.2007.907834

[23] WIESEL, A. (2012). Geodesic convexity and covariance estimation. *IEEE Trans*. *Signal Process*. **60** 6182–6189. MR3006411 https://doi.org/10.1109/TSP.2012.2218241

[24] XU, P., ZHANG, T. and GU, Q. (2017). Efficient algorithm for sparse tensor-variate Gaussian graphical models via gradient descent. In *Artificial Intelligence and Statistics* 923–932. PMLR.

[25] ZHOU, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann*. *Statist*. **42** 532–562. MR3210978 https://doi.org/10.1214/13-AOS1187