The Product of Gaussian Matrices is Close to Gaussian

Yi Li

Division of Mathematical Sciences, Nanyang Technological University

David P. Woodruff

Department of Computer Science, Carnegie Mellon University

Abstract

We study the distribution of the *matrix product* $G_1G_2\cdots G_r$ of r independent Gaussian matrices of various sizes, where G_i is $d_{i-1}\times d_i$, and we denote $p=d_0$, $q=d_r$, and require $d_1=d_{r-1}$. Here the entries in each G_i are standard normal random variables with mean 0 and variance 1. Such products arise in the study of wireless communication, dynamical systems, and quantum transport, among other places. We show that, provided each d_i, $i=1,\dots,r$, satisfies $d_i\geq Cp\cdot q$, where $C\geq C_0$ for a constant $C_0>0$ depending on r, then the matrix product $G_1G_2\cdots G_r$ has variation distance at most δ to a $p\times q$ matrix G of i.i.d. standard normal random variables with mean 0 and variance $\prod_{i=1}^{r-1}d_i$. Here $\delta\to 0$ as $C\to\infty$. Moreover, we show a converse for constant r that if $d_i<C'\max\{p,q\}^{1/2}\min\{p,q\}^{3/2}$ for some i, then this total variation distance is at least δ', for an absolute constant $\delta'>0$ depending on C' and r. This converse is best possible when $p=\Theta(q)$.

2012 ACM Subject Classification Mathematics of computing → Probability and statistics

Keywords and phrases random matrix theory, total variation distance, matrix product

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2021.35

Funding *Yi Li*: Supported in part by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013

David P. Woodruff: Supported in part by Office of Naval Research (ONR) grant N00014-18-1-256 and a Simons Investigator Award

Acknowledgements D. Woodruff would like to thank Sébastien Bubeck, Sitan Chen, and Jerry Li for many helpful discussions.

1 Introduction

Random matrices play a central role in many areas of theoretical, applied, and computational mathematics. One particular application is dimensionality reduction, whereby one often chooses a rectangular random matrix $G\in\mathbb{R}^{m\times n}$, $m\ll n$, and computes $G\cdot x$ for a fixed vector $x\in\mathbb{R}^n$. Indeed, this is the setting in compressed sensing and sparse recovery [12], randomized numerical linear algebra [18, 20, 36], and sketching algorithms for data streams [25]. Often G is chosen to be a Gaussian matrix, and in particular, an $m\times n$ matrix with entries that are i.i.d. normal random variables with mean 0 and variance 1, denoted by $N(0,1)$. Indeed, in compressed sensing, such matrices can be shown to satisfy the Restricted Isometry Property (RIP) [10], while in randomized numerical linear algebra, in certain applications such as support vector machines [29] and non-negative matrix factorization [19], their performance is shown to often outperform that of other sketching matrices.

Our focus in this paper will be on understanding the *product* of two or more Gaussian matrices. Such products arise naturally in different applications. For example, in the over-constrained ridge regression problem $\min_x\|Ax-b\|_2^2+\lambda\|x\|_2^2$, the design matrix $A\in\mathbb{R}^{n\times d}$, $n\gg d$, is itself often assumed to be Gaussian (see, e.g., [26]). In this case, the "sketch-and-solve" algorithmic framework for regression [32] would compute $G\cdot A$ and $G\cdot b$ for an $m\times n$

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021).
Editors: Mary Wootters and Laura Sanità; Article No. 35; pp. 35:1–35:22
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Gaussian matrix G with $m \approx sd_\lambda$, where sd_λ is the so-called statistical dimension [2], and solve for the x which minimizes $\|G \cdot Ax - G \cdot b\|_2^2 + \lambda\|x\|_2^2$. While computing $G \cdot A$ is slower than computing the corresponding matrix product for other kinds of sketching matrices G, it often has application-specific [29, 19] as well as statistical benefits [31]. Notice that $G \cdot A$ is the product of two independent Gaussian matrices, and in particular, G has a small number of rows while A has a small number of columns – this is precisely the rectangular case we will study below. Other applications in randomized numerical linear algebra where the product of two Gaussian matrices arises is when one computes the product of a Gaussian sketching matrix and Gaussian noise in a spiked identity covariance model [37].

The product of two or more Gaussian matrices also arises in diverse fields such as multiple-input multiple-output (MIMO) wireless communication channels [24]. Indeed, similar to the above regression problem in which one wants to reconstruct an underlying vector x, in such settings one observes the vector $y = G_1 \cdots G_r \cdot x + \eta$, where x is the transmitted signal and η is background noise. This setting corresponds to the situation in which there are r scattering environments separated by major obstacles, and the dimensions of the G_i correspond to the number of "keyholes" [24]. To determine the mutual information of this channel, one needs to understand the singular values of the matrix $G_1 \cdots G_r$. If one can show the distribution of this product is close to that of a Gaussian distribution in total variation distance, then one can use the wide range of results known for the spectrum of a single Gaussian matrix (see, e.g., [35]). Other applications of products of Gaussian matrices include disordered spin chains [11, 3, 15], stability of large complex dynamical systems [22, 21], symplectic maps and Hamiltonian mechanics [11, 4, 28], quantum transport in disordered wires [23, 13], and quantum chromodynamics [27]; we refer the reader to [14, 1] for an overview.

The main question we ask in this work is:

What is the distribution of the product $G_1 G_2 \cdots G_r$ of r independent Gaussian matrices of various sizes, where G_i is $d_{i-1} \times d_i$?

Our main interest in the question above will be when G_1 has a small number $p = d_0$ of rows, and G_r has a small number $q = d_r$ of columns. Despite the large body of work on random matrix theory (see, e.g., [34] for a survey), we are not aware of any work which attempts to bound the total variation distance of the entire distribution of $G_1 G_2 \cdots G_r$ to a Gaussian distribution itself.

1.1 Our Results

Formally, we consider the problem of distinguishing the product of normalized Gaussian matrices

$$A_r = \left(\frac{1}{\sqrt{d_1}} G_1\right)\left(\frac{1}{\sqrt{d_2}} G_2\right) \cdots \left(\frac{1}{\sqrt{d_{r-1}}} G_{r-1}\right)\left(\frac{1}{\sqrt{d_1}} G_r\right)$$

from a single normalized Gaussian matrix

$$A_1 = \frac{1}{\sqrt{d_1}} G_1.$$

We show that, when r is a constant, with constant probability we cannot distinguish the distributions of these two random matrices when $d_i \gg p, q$ for all i; and, conversely, with constant probability, we can distinguish these two distributions when the d_i are not large enough.

▶ **Theorem 1** (Main theorem). *Suppose that $d_i \geq \max\{p,q\}$ for all i and $d_{r-1} = d_1$.*

(a) *It holds that*

$$d_{TV}(A_r, A_1) \leq C_1 \sum_{i=1}^{r-1} \sqrt{\frac{pq}{d_i}},$$

where $d_{TV}(A_r, A_1)$ denotes the total variation distance between A_r and A_1, and $C_1 > 0$ is an absolute constant.

(b) *If $p, q, d_1, \ldots, d_r$ further satisfy that*

$$\sum_{j=1}^{r-1} \frac{1}{d_j} \geq \frac{C_2^r}{\max\{p,q\}^{\frac{1}{2}} \min\{p,q\}^{\frac{3}{2}}},$$

where $C_2 > 0$ is an absolute constant, then $d_{TV}(A_r, A_1) \geq 2/3$.

Part (a) states that $d_{TV}(A_r, A_1) < 2/3$ when $d_i \geq C_1' pq$ for all i for a constant C_1' depending on r. The converse in (b) implies that $d_{TV}(A_r, A_1) \geq 2/3$ when $d_i \leq C_2' \max\{p,q\}^{1/2} \min\{p,q\}^{3/2}$ for some i for a constant C_2' depending on r. When $p = \Theta(q)$ and r is a constant, we obtain a dichotomy (up to a constant factor) for the conditions on p, q and d_i.

1.2 Our Techniques

Upper Bound. We start by explaining our main insight as to why the distribution of a product $G_1 \cdot G_2$ of a $p \times d$ matrix G_1 of i.i.d. $N(0,1)$ random variables and a $d \times q$ matrix G_2 of i.i.d. $N(0,1)$ random variables has low variation distance to the distribution of a $p \times q$ matrix A of i.i.d. $N(0,d)$ random variables. One could try to directly understand the probability density function as was done in the case of Wishart matrices in [7, 30], which corresponds to the setting when $G_1 = G_2$. However, there are certain algebraic simplifications in the case of the Wishart distribution that seem much less tractable when manipulating the density function of the product of independent Gaussians [9]. Another approach would be to try to use entropic methods as in [8, 6]. Such arguments try to reveal entries of the product $G_1 \cdot G_2$ one-by-one, arguing that for most conditionings of previous entries, the new entry still looks like an independent Gaussian. However, the entries are clearly not independent – if $(G_1 \cdot G_2)_{i,j}$ has large absolute value, then $(G_1 \cdot G_2)_{i,j'}$ is more likely to be large in absolute value, as it could indicate that the i-th row of G_1 has large norm. One could try to first condition on the norms of all rows of G_1 and columns of G_2, but additional issues arise when one looks at submatrices: if $(G_1 \cdot G_2)_{i,j}, (G_1 \cdot G_2)_{i,j'}$, and $(G_1 \cdot G_2)_{i',j}$ are all large, then it could mean the i-th row of G_1 and the i'-th row of G_1 are correlated with each other, since they both are correlated with the j-th column of G_2. Consequently, since $(G_1 \cdot G_2)_{i,j'}$ is large, it could make it more likely that $(G_1 \cdot G_2)_{i',j'}$ has large absolute value. This makes the entropic method difficult to apply in this context.

Our upper bound instead leverages beautiful work of Jiang [16] and Jiang and Ma [17] which bounds the total variation distance between the distribution of an $r \times \ell$ submatrix of a random $d \times d$ orthogonal matrix (orthonormal rows and columns) and an $r \times \ell$ matrix with i.i.d. $N(0, 1/d)$ entries. Their work shows that if $r \cdot \ell / d \to 0$ as $d \to \infty$, then the total variation distance between these two matrix ensembles goes to 0. It is not immediately clear how to apply such results in our context. First of all, which submatrix should we be looking at? Note though, that if V^T is a $p \times d$ uniformly random (Haar measure) matrix with orthonormal rows, and E is a $d \times q$ uniformly random matrix with orthonormal columns, then by rotational invariance, $V^T E$ is identically distributed to a $p \times q$ submatrix of a $d \times d$ random orthonormal matrix. Thus, setting $r = p$ and $\ell = q$ in the above results, they imply that $V^T E$ is close in variation distance to a $p \times q$ matrix H with i.i.d. $N(0, 1/d)$ entries. Given G_1 and G_2, one could then write them in their *singular value decomposition*, obtaining $G_1 = U\Sigma V^T$ and $G_2 = E\mathrm{T}F^T$. Then V^T and E are independent and well-known to be uniformly random $p \times d$ and $d \times q$ orthonormal matrices, respectively. Thus $G_1 \cdot G_2$ is close in total variation distance to $U\Sigma H \mathrm{T} F^T$. However, this does not immediately help either, as it is not clear what the distribution of this matrix is. Instead, the "right" way to utilize the results above is to (1) observe that $G_1 \cdot G_2 = U\Sigma V^T G_2$ is identically distributed as $U\Sigma X$, where X is a matrix of i.i.d. normal random variables, given the rotational invariance of the Gaussian distribution. Then (2) X is itself close to a product $W^T Z$ where W^T is a random $p \times d$ matrix with orthonormal rows, and Z is a random $d \times q$ matrix with orthonormal columns, by the above results. Thus, $G_1 \cdot G_2$ is close to $U\Sigma W^T Z$. Then (3) $U\Sigma W^T$ has the same distribution as G_1, so $U\Sigma W^T Z$ is close to $G_1' Z$, where G_1' and G_1 are identically distributed, and G_1' is independent of Z. Finally, (4) $G_1' Z$ is identically distributed as a matrix A_1 of standard normal random variables because G_1' is Gaussian and Z has orthonormal columns, by rotational invariance of the Gaussian distribution.

We hope that this provides a general method for arguments involving Gaussian matrices - in step (2) we had the quantity $U\Sigma X$, where X was a Gaussian matrix, and then viewed X as a product of a short-fat random orthonormal matrix W^T and a tall-thin random orthonormal matrix Z. Our proof for the product of more than 2 matrices recursively uses similar ideas, and bounds the growth in variation distance as a function of the number r of matrices involved in the product.

Lower Bound. For our lower bound for constant r, we show that the fourth power of the Schatten 4-norm of a matrix, namely, $\|X\|_{S_4}^4 = \mathrm{tr}((X^T X)^2)$, can be used to distinguish a product A_r of r Gaussian matrices and a single Gaussian matrix A_1. We use Chebyshev's inequality, for which we need to find the expectation and variance of $\mathrm{tr}((X^T X)^2)$ for $X = A_r$ and $X = A_1$.

Let us consider the expectation first. An idea is to calculate the expectation recursively, that is, for a fixed matrix M and a Gaussian random matrix G we express $\mathbb{E}\,\mathrm{tr}(((MG)^T(MG))^2)$ in terms of $\mathbb{E}\,\mathrm{tr}((M^T M)^2)$. The real situation turns out to be slightly more complicated. Instead of expressing $\mathbb{E}\,\mathrm{tr}(((MG)^T(MG))^2)$ in terms of $\mathbb{E}\,\mathrm{tr}((M^T M)^2)$ directly, we decompose $\mathbb{E}\,\mathrm{tr}(((MG)^T(MG))^2)$ into the sum of expectations of a few functions in terms of M, say,

$$\mathbb{E}\,\mathrm{tr}(((MG)^T(MG))^2) = \mathbb{E}\, f_1(M) + \mathbb{E}\, f_2(M) + \cdots + \mathbb{E}\, f_s(M)$$

and build up the recurrence relations for $\mathbb{E}\, f_1(MG), \ldots, \mathbb{E}\, f_s(MG)$ in terms of $\mathbb{E}\, f_1(M)$, $\mathbb{E}\, f_2(M)$, ..., $\mathbb{E}\, f_s(M)$. It turns out that the recurrence relations are all linear, i.e.,

$$\mathbb{E}\, f_i(MG) = \sum_{j=1}^{s} a_{ij}\, \mathbb{E}\, f_j(M), \quad i = 1, \ldots, s, \tag{1}$$

whence we can solve for $\mathbb{E}\, f_i(A_r)$ and obtaining the desired expectation $\mathbb{E}\,\mathrm{tr}((A_r^T A_r)^2)$.

Now we turn to variance. One could try to apply the same idea of finding recurrence relations for $\mathrm{Var}(Q) = \mathbb{E}(Q^2) - (\mathbb{E}\, Q)^2$ (where $Q = \mathrm{tr}(((MG)^T(MG))^2)$), but it quickly becomes intractable for the $\mathbb{E}(Q^2)$ term as it involves products of eight entries of M, which all need to be handled carefully as to avoid any loose bounds; note, the subtraction of $(\mathbb{E}\, Q)^2$

is critically needed to obtain a small upper bound on $\mathrm{Var}(Q)$ and thus loose bounds on $\mathbb{E}(Q^2)$ would not suffice. For a tractable calculation, we keep the product of entries of M to 4th order throughout, without involving any terms of 8th order. To do so, we invoke the law of total variance,

$$\operatorname*{Var}_{M,G}(\mathrm{tr}((MG)^T(MG))^2) = \operatorname*{\mathbb{E}}_{M}\left(\operatorname*{Var}_{G}(\mathrm{tr}((G^TM^TMG)^2))\middle| M\right)+\operatorname*{Var}_{M}\left(\operatorname*{\mathbb{E}}_{G}\mathrm{tr}((G^TM^TMG)^2)\middle| M\right). \quad (2)$$

For the first term on the right-hand side, we use Poincaré's inequality to upper bound it. Poincaré's inequality for the Gaussian measure states that

$$\operatorname*{Var}_{g\sim N(0,I_m)}(f(g)) \leq C \operatorname*{\mathbb{E}}_{g\sim N(0,I_m)} \|\nabla f(g)\|_2^2$$

for a differentiable function f on $\mathbb{R}^m$. Here we can simply let $f(X) = \mathrm{tr}((MX)^T(MX))^2)$ and calculate $\mathbb{E}\|\nabla f(G)\|_2^2$. This is tractable since $\mathbb{E}\|\nabla f(G)\|_2^2$ involves the products of at most 4 entries of M, and we can use the recursive idea for the expectation above to express

$$\mathbb{E}\|\nabla f(G)\|_2^2 = \sum_i a_{ij}\,\mathbb{E}\, g_i(M)$$

for a few functions g_i's and establish a recurrence relation for each g_i.

The second term on the right-hand side of (2) can be dealt with by plugging in (1), and turns out to depend on a new quantity $\mathrm{Var}(\mathrm{tr}^2(M^TM))$. We again apply the recursive idea and the law of total variance to

$$\operatorname*{Var}_{M,G}(\mathrm{tr}^2(G^TM^TMG)) = \operatorname*{\mathbb{E}}_{M}\left(\operatorname*{Var}_{G}(\mathrm{tr}^2((G^TM^TMG))\middle| M\right) + \operatorname*{Var}_{M}\left(\operatorname*{\mathbb{E}}_{G}\mathrm{tr}^2(G^TM^TMG)\middle| M\right).$$

Again, the first term on the right-hand side can be handled by Poincaré's inequality and the second-term turns out to depend on $\mathrm{Var}(\mathrm{tr}((M^TM)^2))$, which is crucial. We have now obtained a double recurrence involving inequalities on $\mathrm{Var}(\mathrm{tr}((M^TM)^2))$ and $\mathrm{Var}(\mathrm{tr}^2((M^TM)^2))$, from which we can solve for an upper bound on $\mathrm{Var}(\mathrm{tr}(A_r^TA_r)^2)$. This upper bound, however, grows exponentially in r, which is impossible to improve due to our use of Poincaré's inequality.

2 Preliminaries

Notation. For a random variable X and a probability distribution $\mathcal{D}$, we use $X \sim \mathcal{D}$ to denote that X is subject to $\mathcal{D}$. For two random variables X and Y defined on the same sample space, we write $X \overset{d}{=} Y$ if X and Y are identically distributed.

We use $\mathcal{G}_{m,n}$ to denote the distribution of $m\times n$ Gaussian random matrices of i.i.d. entries $N(0,1)$ and $\mathcal{O}_{m,n}$ to denote the uniform distribution (Haar) of an $m \times n$ random matrix with orthonormal rows. For a distribution $\mathcal{D}$ on a linear space and a scaling factor $\alpha \in \mathbb{R}$, we use $\alpha\mathcal{D}$ to denote the distribution of αX, where $X \sim \mathcal{D}$.

For two probability measures μ and ν on the Borel algebra $\mathcal{F}$ of $\mathbb{R}^m$, the total variation distance between μ and ν is defined as

$$d_{TV}(\mu,\nu) = \sup_{A\in\mathcal{F}} |\mu(A)-\nu(A)| = \frac{1}{2}\int_{\mathbb{R}^m}\left|\frac{d\mu}{d\nu}-1\right| d\nu.$$

If ν is absolutely continuous with respect to μ, one can define the Kullback-Leibler Divergence between μ and ν as

$$D_{\mathrm{KL}}(\mu\|\nu) = \int_{\mathbb{R}^m}\frac{d\mu}{d\nu}\log_2\frac{d\mu}{d\nu}d\nu.$$

If ν is not absolutely continuous with respect to μ, we define $D_{\mathrm{KL}}(\mu\|\nu) = \infty$.

When μ and ν correspond to two random variables X and Y, respectively, we also write $d_{TV}(\mu,\nu)$ and $D_{\mathrm{KL}}(\mu\|\nu)$ as $d_{TV}(X,Y)$ and $D_{\mathrm{KL}}(X\|Y)$, respectively.

The following is the well-known relation between the Kullback-Leibler divergence and the total variation distance between two probability measures.

▶ **Lemma 2** (Pinsker's Inequality [5, Theorem 4.19])**.** $d_{TV}(\mu,\nu) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mu\|\nu)}$.

The following result, concerning the distance between the submatrix of a properly scaled Gaussian random matrix and a submatrix of a random orthogonal matrix, is due to Jiang and Ma [17].

▶ **Lemma 3** ([17])**.** *Let $G \sim \mathcal{G}_{d,d}$ and $Z \sim \mathcal{O}_{d,d}$. Suppose that $p,q \le d$ and $\hat{G}$ is the top-left $p \times q$ block of G and $\hat{Z}$ the top-left $p \times q$ block of Z. Then*

$$d_{\mathrm{KL}}\left(\frac{1}{\sqrt{d}}\hat{G}\middle\|\hat{Z}\right) \le C\frac{pq}{d}, \tag{3}$$

where $C > 0$ is an absolute constant.

The original paper [17] does not state explicitly the bound in (3) and only states that the Kullback-Leibler divergence tends to 0 as $d \to \infty$. A careful examination of the proof of [17, Theorem 1(i)], by keeping track of the order of the various $o(1)$ terms, reveals the quantitative bound (3).

Useful Inequalities. We list two useful inequalities below.

▶ **Lemma 4** (Poincaré's inequality for Gaussian measure [5, Theorem 3.20])**.** *Let $X \sim N(0, I_n)$ be the standard n-dimensional Gaussian distribution and $f : \mathbb{R}^n \to \mathbb{R}$ be any continuously differentiable function. Then*

$$\mathrm{Var}(f(X)) \le \mathbb{E}\left(\|\nabla f(X)\|_2^2\right).$$

▶ **Lemma 5** (Trace inequality, [33])**.** *Let A and B be symmetric, positive semidefinite matrices and k be a positive integer. Then*

$$\mathrm{tr}((AB)^k) \le \min\left\{\|A\|_{op}^k \mathrm{tr}(B^k), \|B\|_{op}^k \mathrm{tr}(A^k)\right\}.$$

3 Upper Bound

Let $r \ge 2$ be an integer. Suppose that $G_1,\dots,G_r$ are independent Gaussian random matrices, where $G_i \sim \mathcal{G}_{d_{i-1},d_i}$ and $d_0 = p$, $d_r = q$ and $d_{r-1} = d_1$. Consider the product of normalized Gaussian matrices

$$A_r = \left(\frac{1}{\sqrt{d_1}}G_1\right)\left(\frac{1}{\sqrt{d_2}}G_2\right)\cdots\left(\frac{1}{\sqrt{d_{r-1}}}G_{r-1}\right)\left(\frac{1}{\sqrt{d_1}}G_r\right)$$

and a single normalized Gaussian random matrix

$$A_1 = \frac{1}{\sqrt{d_1}}G_1'$$

where $G_1' \sim \mathcal{G}_{p,q}$. In this section, we shall show that when $p,q \ll d_i$ for all i, we cannot distinguish A_r from A_1 with constant probability.

For notational convenience, let $W_i = \frac{1}{\sqrt{d_i}}G_i$ for $i \le r$ and $W_r = \frac{1}{\sqrt{d_1}}G_r$. Assume that $pq \le \beta d_i$ for some constant β for all i. Our question is to find the total variation distance between the matrix product $W_1W_2\cdots W_r$ and the product W_1W_r of two matrices.

▶ **Lemma 6.** *Let p, q, d, d' be positive integers satisfying that $pq \le \beta d$ and $pq \le \beta d'$ for some constant $\beta < 1$. Suppose that $A \in \mathbb{R}^{p\times d}$, $G \sim \frac{1}{\sqrt{d}}\mathcal{G}_{d,d'}$, and $L \sim \mathcal{O}_{d',d}$. Further suppose that G and L are independent. Let $Z \sim \mathcal{O}_{q,d}$ be independent of A, G and L. Then*

$$d_{TV}(AGL, AZ^T) \le C\sqrt{\frac{pq}{d}},$$

where $C > 0$ is an absolute constant.

Proof. Let $A = U\Sigma V^T$ be its singular value decomposition, where V has dimension $d \times p$. Then

$$AGL = U\Sigma(V^T GL) \overset{d}{=} U\Sigma X,$$

where X is a $p \times q$ random matrix of i.i.d. $N(0, 1/d)$ entries. Suppose that $\tilde{Z}$ consists of the top p rows of Z^T. Then

$$AZ^T = U\Sigma(V^T Z^T) \overset{d}{=} U\Sigma\tilde{Z}.$$

Note that X and Z are independent of U and Σ. It follows from Lemma 3 that

$$d_{\mathrm{KL}}(AGL\|AZ^T) = d_{\mathrm{KL}}(U\Sigma X\|U\Sigma\tilde{Z}) = d_{\mathrm{KL}}(X\|\tilde{Z}) \le C\frac{pq}{d},$$

where $C > 0$ is an absolute constant. The result follows from Pinsker's inequality (Lemma 2). ◀

The next theorem follows from the lemma above.

▶ **Theorem 7.** *It holds that*

$$d_{TV}(W_1 \cdots W_r, W_1 W_r) \le C\sum_{i=1}^{r}\sqrt{\frac{pq}{d_i}},$$

where $C > 0$ is an absolute constant.

Proof. Let $W_r = U\Sigma V^T$ and $X_i \sim \mathcal{O}_{q,d_i}$, independent from each other and from the W_i's. Applying the preceding lemma with $A = W_1 \cdots W_{r-2}$, $G = W_{r-1}$ and $L = U$, we have

$$d_{TV}(W_1 \cdots W_{r-2}W_{r-1}W_r, W_1 \cdots W_{r-2}X_{r-1}^T\Sigma V^T) \le C\sqrt{\frac{pq}{d_{r-1}}},$$

Next, applying the preceding lemma with $A = W_1 \cdots W_{r-3}$, $G = W_{r-1}$ and $L = X_r$, we have

$$d_{TV}(W_1 \cdots W_{r-2}X_r\Sigma V^T, W_1 \cdots W_{r-3}X_{r-2}^T\Sigma V^T) \le C\sqrt{\frac{pq}{d_{r-2}}},$$

Iterating this procedure, we have in the end that

$$d_{TV}(W_1 W_2 X_3\Sigma V^T, W_1 X_2\Sigma V^T) \le C\sqrt{\frac{pq}{d_2}}.$$

Since U, Σ and V are independent and $X_2 \overset{d}{=} U$, it holds that $X_2\Sigma V^T \overset{d}{=} W_r$. Therefore,

$$d_{TV}(W_1 \cdots W_r, W_1 W_r) \le C\sum_{i=2}^{r-1}\sqrt{\frac{pq}{d_i}}.$$ ◀

Repeating the same argument for $W_1 W_r$, we obtain the following corollary immediately.

▶ **Corollary 8.** *It holds that*

$$d_{TV}(A_r, A_1) \le C\sum_{i=1}^{r-1}\sqrt{\frac{pq}{d_i}},$$

where $C > 0$ is an absolute constant.

4 Lower Bound

Suppose that r is a constant. We shall show that one can distinguish the product of r Gaussian random matrices

$$A_r = \left(\frac{1}{\sqrt{d_1}}G_1\right)\left(\frac{1}{\sqrt{d_2}}G_2\right)\cdots\left(\frac{1}{\sqrt{d_{r-1}}}G_{r-1}\right)\left(\frac{1}{\sqrt{d_1}}G_r\right),$$

from one Gaussian random matrix

$$A_1 = \frac{1}{\sqrt{d_1}}G_1'$$

when the intermediate dimensions $d_1, \ldots, d_{r-1}$ are not large enough. Considering $h(X) = \operatorname{tr}((X^T X)^2)$, it suffices to show that one can distinguish $h(A_r)$ and $h(A_1)$ with a constant probability for constant r. By Chebyshev's inequality, it suffices to show that

$$\max\left\{\sqrt{\operatorname{Var}(h(A_1))}, \sqrt{\operatorname{Var}(h(A_r))}\right\} \leq c(\mathbb{E}\, h(A_r) - \mathbb{E}\, h(A_1))$$

for a small constant c. We calculate that:

▶ **Lemma 9.** *Suppose that r is a constant, $d_i \geq \max\{p, q\}$ for all $i = 1, \ldots, r$. When $p, q, d_1, \ldots, d_r \to \infty$,*

$$\mathbb{E}\, h(A_r) = \frac{pq(p+q+1)}{d_r^2} + (1 + o(1))\frac{pq(p-1)(q-1)}{d_r^2}\sum_{j=1}^{r-1}\frac{1}{d_j}.$$

▶ **Lemma 10.** *Suppose that r is a constant, $d_i \geq \max\{p, q\}$ for all $i = 1, \ldots, r$. There exists an absolute constant C such that, when $p, q, d_1, \ldots, d_r$ are sufficently large,*

$$\operatorname{Var}(h(A_r)) \leq \frac{C^r(p^3 q + pq^3)}{d_1^4}.$$

We conclude with the following theorem, which can be seen as a tight converse to Corollary 8 up to a constant factor on the conditions for $p, q, d_1, \ldots, d_r$.

▶ **Theorem 11.** *Suppose that r is a constant and $d_i \geq \max\{p, q\}$ for all $i = 1, \ldots, r$. Further suppose that $d_1 = d_r$. When $p, q, d_1, \ldots, d_r$ are sufficiently large and satisfy that*

$$\sum_{j=1}^{r-1}\frac{1}{d_j} \geq \frac{C^r}{\max\{p,q\}^{\frac{1}{2}}\min\{p,q\}^{\frac{3}{2}}},$$

where $C > 0$ is some absolute constant, with probability at least $2/3$, one can distinguish A_r from A_1.

4.1 Calculation of the Mean

Suppose that A is a $p \times q$ random matrix, and is rotationally invariant under left- and right-multiplication by orthogonal matrices. We define

$$S_1(p,q) = \mathbb{E}\, A_{11}^4 \quad \text{(diagonal)}$$
$$S_2(p,q) = \mathbb{E}\, A_{21}^4 \quad \text{(off-diagonal)}$$
$$S_3(p,q) = \mathbb{E}\, A_{i1}^2 A_{j1}^2 \quad (i \neq j) \quad \text{(same column)}$$

$$S_4(p,q) = \mathbb{E}\, A_{1i}^2 A_{1j}^2 \quad (i \neq j) \quad \text{(same row)}$$
$$S_5(p,q) = \mathbb{E}\, A_{1i}^2 A_{2j}^2 \quad (i \neq j)$$
$$S_6(p,q) = \mathbb{E}\, A_{ik} A_{il} A_{jk} A_{jl} \quad (i \neq j, k \neq l) \quad \text{(rectangle)}$$

Since A is left- and right-invariant under rotations, these quantities are well-defined. Then

$$\begin{aligned}\mathbb{E}\operatorname{tr}((A^TA)^2) = \mathbb{E}\sum_{1\leq i,j\leq q}(A^TA)_{ij}^2 &= \sum_{i=1}^q \mathbb{E}(A^TA)_{ii}^2 + \sum_{1\leq i,j\leq q, i\neq j}\mathbb{E}(A^TA)_{ij}^2 \\ &= q\,\mathbb{E}(A^TA)_{11}^2 + q(q-1)\,\mathbb{E}(A^TA)_{12}^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(A^TA)_{11}^2 = \mathbb{E}\left(\sum_{i=1}^p A_{i1}^2\right)^2 &= \sum_{i=1}^p \mathbb{E}\, A_{i1}^4 + \sum_{1\leq i,j\leq p, i\neq j}\mathbb{E}\, A_{i1}^2 A_{j1}^2 \\ &= \mathbb{E}\, A_{11}^4 + (p-1)\,\mathbb{E}\, A_{21}^4 + p(p-1)\,\mathbb{E}\, A_{11}^2 A_{21}^2 \\ &=: S_1(p,q) + (p-1)S_2(p,q) + p(p-1)S_3(p,q)\end{aligned}$$

$$\begin{aligned}\mathbb{E}(A^TA)_{12}^2 = \mathbb{E}\left(\sum_{i=1}^p A_{i1}A_{i2}\right)^2 &= \sum_{i=1}^p \mathbb{E}\, A_{i1}^2 A_{i2}^2 + \sum_{1\leq i,j\leq p, i\neq j}\mathbb{E}\, A_{i1}A_{i2}A_{j1}A_{j2} \\ &= pS_4(p,q) + p(p-1)S_6(p,q).\end{aligned}$$

When $S_1(p,q) = S_2(p,q)$, we have

$$\begin{aligned}\mathbb{E}\operatorname{tr}((A^TA)^2) &= q(pS_1(p,q) + p(p-1)S_3(p,q)) + q(q-1)(pS_4(p,q) + p(p-1)S_6(p,q)) \\ &= pqS_1(p,q) + pq(p-1)S_3(p,q) + pq(q-1)S_4(p,q) + p(p-1)q(q-1)S_6(p,q).\end{aligned}$$

When $A = G$, we have

$$S_1(p,q) = S_2(p,q) = 3, \quad S_3(p,q) = S_4(p,q) = S_5(p,q) = 1, \quad S_6(p,q) = 0$$

and so

$$\mathbb{E}\operatorname{tr}((A^TA)^2) = 3pq + pq(p-1) + pq(q-1) = pq(p+q+1).$$

Next, consider $A = BG$, where B is a $p \times d$ random matrix and G a $d \times q$ random matrix of i.i.d. $N(0,1)$ entries. The following proposition is easy to verify, and its proof is postponed to Appendix A.

▶ **Proposition 12.** *It holds that* $\mathbb{E}\, A_{21}^4 = \mathbb{E}\, A_{11}^4$.

Suppose that the associated functions of B are named $T_1, T_2, T_3, T_4, T_6, T_5$. Then we can calculate that (detailed calculations can be found in Appendix B)

$$S_1(p,q) = 3dT_1(p,d) + 3d(d-1)T_4(p,d)$$
$$S_3(p,q) = 3dT_3(p,d) + d(d-1)T_5(p,d) + 2d(d-1)T_6(p,d)$$
$$S_4(p,q) = dT_1(p,d) + d(d-1)T_4(p,d)$$
$$S_5(p,q) = dT_3(p,d) + d(d-1)T_5(p,d)$$
$$S_6(p,q) = dT_3(p,d) + d(d-1)T_6(p,d)$$

It is clear that S_1, S_3, S_4, S_5, S_6 depend only on d (not on p and q) if T_1, T_3, T_4, T_5, T_6 do so. Furthermore, if $T_1 = 3T_4$ then we have $S_1 = 3S_4$ and thus $S_4 = d(d+2)T_4$. If $T_3 = 2T_6 + T_5$ then $S_3 = d(d+2)T_3$ and $S_3 = 2S_6 + S_5$. Hence, if $T_3 = T_4$ then $S_3 = S_4$. We can verify that all these conditions are satisfied with one Gaussian matrix and we can iterate it to obtain these quantities for the product of r Gaussian matrices with intermediate dimensions $d_1, d_2, \ldots, d_{r-1}$. We have that

$$S_3 = S_4 = \prod_{i=1}^{r-1} d_i(d_i+2), \quad S_1 = 3S_4, \quad S_6 = \sum_{j=1}^{r-1} \left(\prod_{i=1}^{j-1} d_i(d_i+2)\right) d_j \left(\prod_{i=j+1}^{r-1} d_i(d_i-1)\right).$$

Therefore, normalizing the i-th matrix by $1/\sqrt{d_i}$, that is,

$$A = \left(\frac{1}{\sqrt{d_1}}G_1\right)\left(\frac{1}{\sqrt{d_2}}G_2\right)\cdots\left(\frac{1}{\sqrt{d_{r-1}}}G_{r-1}\right)\left(\frac{1}{\sqrt{d_1}}G_r\right),$$

we have for constant r that

$$\begin{aligned} \mathbb{E}\,\mathrm{tr}((A^TA)^2) &= \frac{1}{d_1^2 d_2^2 \cdots d_{r-1}^2 d_1^2}\left(pq(p+q+1)S_3 + pq(p-1)(q-1)S_6\right) \\ &\approx \frac{pq(p+q+1)}{d_r^2} + \frac{pq(p-1)(q-1)}{d_r^2}\sum_{j=1}^{r-1}\frac{1}{d_j}. \end{aligned} \tag{4}$$

4.2 Calculation of the Variance

Let $M \in \mathbb{R}^{p\times p}$ be a random symmetric matrix, and let $G \in \mathbb{R}^{p\times q}$ be a random matrix of i.i.d. $N(0,1)$ entries. We want to find the variance of $\mathrm{tr}((G^TMG)^2)$. The detailed calculations of some steps can be found in Appendix C.

Our starting point is the law of total variance, which states that

$$\mathrm{Var}(\mathrm{tr}((G^TMG)^2)) = \mathop{\mathbb{E}}_M\left(\mathop{\mathrm{Var}}_G(\mathrm{tr}((G^TMG)^2))\middle|\, M\right) + \mathop{\mathrm{Var}}_M\left(\mathop{\mathbb{E}}_G \mathrm{tr}((G^TMG)^2)\middle|\, M\right) \tag{5}$$

Step 1a. We shall handle each term separately. Consider the first term, which we shall bound using the Poincaré inequality for Gaussian measures. Define $f(X) = \mathrm{tr}((X^TMX)^2)$, where $X \in \mathbb{R}^{p\times q}$. We shall calculate ∇f.

$$f(X) = \left\|X^TMX\right\|_F^2 = \sum_{1\le i,j\le q} (X^TMX)_{ij}^2 = \sum_{1\le i,j\le q}\left(\sum_{1\le k,l\le p} M_{kl}X_{ki}X_{lj}\right)^2.$$

Then

$$\frac{\partial f}{\partial X_{rs}} = \sum_{1\le i,j\le q} 2\Bigg(\sum_{1\le u,v\le p} M_{uv}X_{ui}X_{vj}\Bigg)\Bigg(\sum_{1\le k,l\le p}\frac{\partial}{\partial X_{rs}}(M_{kl}X_{ki}X_{lj})\Bigg).$$

Note that

$$\frac{\partial}{\partial X_{rs}}(M_{kl}X_{ki}X_{lj}) = \begin{cases} M_{kl}X_{lj}, & (k,i) = (r,s) \text{ and } (l,j) \ne (r,s) \\ M_{kl}X_{ki}, & (k,i) \ne (r,s) \text{ and } (l,j) = (r,s) \\ 2M_{rr}X_{rs}, & (k,i) = (r,s) \text{ and } (l,j) = (r,s) \\ 0, & \text{otherwise.} \end{cases}$$

we have that

$$\frac{\partial f}{\partial X_{rs}} = 4\left(\sum_{1\le u,v\le p} M_{uv}X_{us}X_{vs}\right) M_{rr}X_{rs} + 2\sum_{(l,j)\ne(r,s)}\left(\sum_{1\le u,v\le p} M_{uv}X_{us}X_{vj}\right) M_{rl}X_{lj}$$

$$+ 2 \sum_{(k,i)\neq(r,s)} \left(\sum_{1\le u,v\le p} M_{uv} X_{ui} X_{vs} \right) M_{kr} X_{ki}$$

$$= 4 \left[\left(\sum_{1\le u,v\le p} M_{uv} X_{us} X_{vs} \right) M_{rr} X_{rs} + \sum_{(l,j)\neq(r,s)} \left(\sum_{u,v} M_{uv} X_{us} X_{vj} \right) M_{rl} X_{lj} \right]$$

$$= 4 \sum_{l,j} \left(\sum_{u,v} M_{uv} X_{us} X_{vj} \right) M_{rl} X_{lj}.$$

Next we calculate $\mathbb{E}(\partial f/\partial X_{rs})^2$ when X is i.i.d. Gaussian.

$$\left(\frac{1}{4} \frac{\partial f}{\partial X_{rs}} \right)^2 = \sum_{\substack{l,j \\ l',j'}} \sum_{\substack{u,v \\ u',v'}} M_{uv} M_{u'v'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{us} X_{u's} X_{vj} X_{lj} X_{v'j'} X_{l'j'}$$

We discuss different cases of j, j', s.

When $j \neq j' \neq s$, it must hold that $u = u'$, $v = l$ and $v' = l'$ for a possible nonzero contribution, and the total contribution in this case is at most $q(q-1)B^{(1)}_{r,s}$, where

$$B^{(1)}_{r,s} = \sum_{1\le l,l'\le p} \sum_{u} M_{ul} M_{ul'} M_{rl} M_{rl'} = \sum_{u} \langle M_{u,\cdot}, M_{r,\cdot} \rangle^2.$$

When $j = j' \neq s$, it must hold that $u = u'$ for a possible nonzero contribution, and the total contribution in this case is at most $(q-1)B^{(2)}_{r,s}$, where

$$B^{(2)}_{r,s} = \sum_{l,l'} \sum_{u,v,v'} M_{uv} M_{uv'} M_{rl} M_{rl'} \, \mathbb{E}\, X^2_{us} X_{vj} X_{lj} X_{v'j} X_{l'j}$$

$$= \|M\|_F^2 \, \|M_{r,\cdot}\|_2^2 + 2 \sum_{u} \langle M_{u,\cdot}, M_{r,\cdot} \rangle^2.$$

When $j = s \neq j'$, it must hold that $v' = l'$ for possible nonzero contribution, and the total contribution in this case is at most $(q-1)B^{(3)}_{r,s}$, where

$$B^{(3)}_{r,s} = \sum_{j'\neq s} \left[\sum_{l,l'} \sum_{u,v} M_{uv} M_{u'l'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{us} X_{u's} X_{vs} X_{ls} X^2_{l'j'} \right]$$

$$= \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot} \rangle + \operatorname{tr}(M) M_{ll'}) M_{rl} M_{rl'}.$$

When $j = j' = s$, the nonzero contribution is

$$B^{(4)}_{r,s} = \sum_{l,l'} \sum_{\substack{u,v \\ u',v'}} M_{uv} M_{u'v'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{us} X_{u's} X_{vs} X_{ls} X_{v's} X_{l's}.$$

Since u, u', v, v', l, l' needs to be paired, the only case which is not covered by $B^{(1)}_{rs}, B^{(3)}_{rs}$ and $B^{(3)}_{rs}$ is when $u = v$, $u' = v'$ and $l = l'$, in which case the contribution is at most

$$\sum_{l} \sum_{u,u'} M_{uu} M_{u'u'} M^2_{rl} \, \mathbb{E}\, X^2_{us} X^2_{u's} X^2_{ls} \lesssim \operatorname{tr}^2(M) \, \|M_{r,\cdot}\|_2^2 \, .$$

Hence

$$B^{(4)}_{r,s} \lesssim B^{(1)}_{r,s} + B^{(2)}_{r,s} + B^{(3)}_{r,s} + \operatorname{tr}^2(M) \, \|M_{r,\cdot}\|_2^2 \, .$$

It follows that

$$\sum_{r,s} B^{(1)}_{r,s} = q \sum_{u,r} \langle M_{u,\cdot}, M_{r,\cdot}\rangle^2 = q\,\mathrm{tr}(M^4)$$

$$\sum_{r,s} B^{(2)}_{r,s} = q \sum_{r} \|M\|_F^2 \|M_{r,\cdot}\|_2^2 + 2q \sum_{u,r} \langle M_{u,\cdot}, M_{r,\cdot}\rangle^2 = q\,\|M\|_F^4 + 2q\,\mathrm{tr}(M^4)$$

$$\begin{aligned}\sum_{r,s} B^{(3)}_{r,s} &= \sum_{r,s} \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot}\rangle + \mathrm{tr}(M) M_{l'l'}) M_{rl} M_{rl'} \\ &\le 2q\,\mathrm{tr}(M^4) + q\,\mathrm{tr}(M)\,\|M\|_F \sqrt{\mathrm{tr}(M^4)}\end{aligned}$$

Note that $\mathrm{tr}(M^4) \le \mathrm{tr}^2(M^2) = \|M\|_F^4$. Hence

$$\begin{aligned}\frac{1}{16}\,\mathbb{E}\,\|\nabla f\|_2^2 &\le \sum_{r,s} ((q-1)(q-2)B^{(1)}_{rs} + (q-1)B^{(2)}_{rs} + (q-1)B^{(3)}_{rs} + B^{(4)}_{rs}) \\ &\lesssim \sum_{r,s} (q^2 B^{(1)}_{rs} + qB^{(2)}_{rs} + qB^{(3)}_{rs} + \mathrm{tr}^2(M)\,\|M_{r,\cdot}\|_2^2) \\ &\lesssim q^3\,\mathrm{tr}(M^4) + q^2\,\|M\|_F^4 + q^2\,\mathrm{tr}(M)\,\|M\|_F \sqrt{\mathrm{tr}(M^4)} + q\,\mathrm{tr}^2(M)\,\|M\|_F^2 .\end{aligned}$$

By the Gaussian Poincaré inequality,

$$\begin{aligned}&\mathop{\mathrm{Var}}_G(\mathrm{tr}((G^TMG)^2)|M) \\ &\lesssim \mathbb{E}\,\|\nabla f\|_2^2 \\ &\lesssim q^3\,\mathrm{tr}(M^4) + q^2\,\|M\|_F^4 + q^2\,\mathrm{tr}(M)\,\|M\|_F \sqrt{\mathrm{tr}(M^4)} + q\,\mathrm{tr}^2(M)\,\|M\|_F^2 .\end{aligned} \tag{6}$$

For the terms on the right-hand side, we calculate that (using the trace inequality (Lemma 5))

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}((G^TMG)^4) = \mathbb{E}\,\mathrm{tr}((MGG^T)^4) \le \mathbb{E}\,\left\|GG^T\right\|_{op}^4 \mathrm{tr}(M^4) &= \mathbb{E}\,\|G\|_{op}^8\,\mathrm{tr}(M^4) \\ &\lesssim \max\{p,q\}^4\,\mathrm{tr}(M^4),\end{aligned}$$

$$\mathbb{E}\,\left\|G^TMG\right\|_F^4 \le \mathbb{E}\,\|G\|_{op}^8\,\|M\|_F^4 \lesssim \max\{p,q\}^4\,\|M\|_F^4 ,$$

$$\mathbb{E}\,\mathrm{tr}^2(G^TMG)\,\left\|G^TMG\right\|_F^2 \le \mathbb{E}\,\|G\|_{op}^8\,\mathrm{tr}^2(M)\,\|M\|_F^2 \lesssim \max\{p,q\}^4\,\mathrm{tr}^2(M)\,\|M\|_F^2$$

and

$$\begin{aligned}&\mathbb{E}\,\mathrm{tr}(G^TMG)\,\left\|G^TMG\right\|_F \sqrt{\mathrm{tr}((G^TMG)^4)} \\ &\le \mathbb{E}\,\|G\|_{op}^2\,\mathrm{tr}(G)\cdot\|G\|_{op}^2\,\|M\|_F^2 \cdot \sqrt{\|G\|_{op}^8\,\mathrm{tr}(M^4)} \\ &= \mathbb{E}\,\|G\|_{op}^8\,\mathrm{tr}(M)\,\|M\|_F \sqrt{\mathrm{tr}(M^4)} \\ &\lesssim \max\{p,q\}^4\,\mathrm{tr}(M)\,\|M\|_F \sqrt{\mathrm{tr}(M^4)}.\end{aligned}$$

This implies that each term on the right-hand of (6) grows geometrically.

Step 1b. Next we deal with the second term in (5). We have

$$\begin{aligned}\mathop{\mathbb{E}}_G \mathrm{tr}\left((G^TMG)^2\right) = \sum_{i,j} \mathop{\mathbb{E}}_G (G^TMG)^2_{ij} &= \sum_{i,j} \mathop{\mathbb{E}}_G \left(\sum_{k,l} M_{kl} G_{ki} G_{lj}\right)^2 \\ &= \sum_{i,j} \sum_{k,l,k',l'} M_{kl} M_{k'l'} \mathop{\mathbb{E}}_G G_{ki} G_{lj} G_{k'i} G_{l'j}.\end{aligned}$$

When $i \neq j$, for non-zero contribution, it must hold that $k = l$ and $k' = l'$ and thus the nonzero contribution is

$$\sum_{i \neq j} \sum_{k,l} M_{kl}^2 = q(q-1) \|M\|_F^2 .$$

When $i = j$, the contribution is

$$\sum_{i} \sum_{k,l,k',l'} M_{kl} M_{k'l'} \mathbb{E}\, G_{ki} G_{li} G_{k'i} G_{l'i} = 2q \|M\|_F^2 + q \operatorname{tr}^2(M). \tag{7}$$

Hence

$$\mathop{\mathbb{E}}_{G} \operatorname{tr}\left((G^T M G)^2\right) = q(q+1) \|M\|_F^2 + q \operatorname{tr}^2(M)$$

and when M is random,

$$\begin{aligned}
&\operatorname{Var}\left(\mathbb{E}\operatorname{tr}((G^T M G)^2) \middle| M\right) \\
&= \operatorname{Var}\left(q(q+1)\|M\|_F^2 + q \operatorname{tr}^2(M)\right) \\
&\leq q^2(q+1)^2 \operatorname{Var}(\|M\|_F^2) + q^2 \operatorname{Var}(\operatorname{tr}^2(M)) + 2q^2(q+1)\sqrt{\operatorname{Var}(\|M\|_F^2)\operatorname{Var}(\operatorname{tr}^2(M))}.
\end{aligned} \tag{8}$$

Step 2a. Note that the $\operatorname{Var}(\operatorname{tr}^2(M))$ term on the right-hand side of (8). To bound this term, we examine the variance of $g(G)$, where $g(X) = \operatorname{tr}^2(X^T M X)$. We shall again calculate ∇g. Note that

$$\frac{\partial g}{\partial X_{rs}} = 2 \operatorname{tr}(X^T M X) \sum_i \sum_{k,l} M_{kl} \frac{\partial}{\partial X_{rs}} X_{ki} X_{li}$$

and

$$\frac{\partial}{\partial X_{rs}}(X_{ki} X_{li}) = \begin{cases} X_{li}, & (k,i) = (r,s) \text{ and } (l,i) \neq (r,s) \\ X_{ki}, & (k,i) \neq (r,s) \text{ and } (l,i) = (r,s) \\ 2X_{rs}, & (k,i) = (r,s) \text{ and } (l,i) = (r,s) \\ 0, & \text{otherwise.} \end{cases}$$

We have

$$\frac{\partial g}{\partial X_{rs}} = 4 \operatorname{tr}(X^T M X) \sum_l M_{rl} X_{ls} = 4 \sum_{\substack{1 \leq j \leq q \\ 1 \leq l,u,v \leq p}} M_{uv} M_{rl} X_{ls} X_{uj} X_{vj}$$

Next we calculate $\mathbb{E}(\partial g / \partial X_{rs})^2$ when X is i.i.d. Gaussian.

$$\left(\frac{1}{4}\frac{\partial g}{\partial X_{rs}}\right)^2 = \sum_{\substack{j,l,u,v \\ j',l',u',v'}} M_{uv} M_{u'v'} M_{rl} M_{rl'} \mathbb{E}\, X_{ls} X_{l's} X_{uj} X_{vj} X_{u'j'} X_{v'j'}$$

In order for the expectation in the summand to be non-zero, we must have one of the following cases: (1) $s \neq j \neq j'$, (2) $s = j \neq j'$, (3) $s = j' \neq j$, (4) $s \neq j = j'$, (5) $s = j = j'$. We calculate the contribution in each case below.

Case 1: it must hold that $l = l'$, $u = v$ and $u' = v'$. The contribution is $(q-1)(q-2)B^{(1)}_{rs}$, where

$$B^{(1)}_{rs} = \sum_{l,u,u'} M_{uu} M_{u'u'} M_{rl}^2 = \operatorname{tr}^2(M) \left\| M_{r,\cdot} \right\|_2^2 .$$

Case 2: it must hold that $u' = v'$. The contribution is $(q-1)B^{(2)}_{rs}$, where

$$\begin{aligned} B^{(2)}_{rs} &= \sum_{l,l',u,u',v} M_{uv} M_{u'u'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{ls} X_{l's} X_{us} X_{vs} X_{u'j'}^2 \\ &= \operatorname{tr}(M) \left(\operatorname{tr}(M) \left\| M_{r,\cdot} \right\|_2^2 + 2 \sum_{l,l'} M_{ll'} M_{rl} M_{rl'} \right) \end{aligned}$$

Case 3: this gives the same bound as Case 2.

Case 4: it must hold that $l = l'$. The contribution is $(q-1)B^{(4)}_{rs}$, where

$$B^{(4)}_{rs} = \sum_{l,u,u',v,v'} M_{uv} M_{u'v'} M_{rl}^2 \, \mathbb{E}\, X_{uj} X_{vj} X_{u'j} X_{v'j} = 3 \left\| M_{r,\cdot} \right\|_2^2 \left\| M \right\|_F^2$$

Case 5: the contribution is $B^{(5)}_{rs}$, where

$$B^{(5)}_{rs} = \sum_{\substack{l,u,v \\ l',u',v'}} M_{uv} M_{u'v'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{ls} X_{us} X_{vs} X_{l's} X_{u's} X_{v's}.$$

The only uncovered case is $l = u'$, $l' = v$, $u = v'$ and its symmetries. In such a case the contribution is at most

$$C \sum_{l,u,v} M_{uv} M_{lu} M_{rl} M_{rv} = C \sum_{u} \langle M_{r,\cdot}, M_{u,\cdot} \rangle^2.$$

Note that

$$\begin{aligned} \sum_{r,s} B^{(1)}_{rs} &= q \operatorname{tr}^2(M) \left\| M \right\|_F^2 , \\ \sum_{r,s} B^{(2)}_{rs} &= q \operatorname{tr}^2(M) \left\| M \right\|_F^2 + 2q \operatorname{tr}(M) \sum_{l,l'} M_{ll'} \langle M_{l,\cdot}, M_{l',\cdot} \rangle \\ &\le q \operatorname{tr}^2(M) \left\| M \right\|_F^2 + 2q \operatorname{tr}(M) \left\| M \right\|_F \sqrt{\operatorname{tr}(M^4)}, \\ \sum_{r,s} B^{(4)}_{rs} &= q \left\| M \right\|_F^4 , \\ \sum_{r,s} B^{(5)}_{rs} &\lesssim \sum_{r,s} B^{(1)}_{rs} + \sum_{r,s} B^{(2)}_{rs} + \operatorname{tr}(M^4). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{1}{16} \mathbb{E} \left\| \nabla g \right\|_2^2 &\le \sum_{r,s} ((q-1)(q-2) B^{(1)}_{rs} + (q-1) B^{(2)}_{rs} + (q-1) B^{(4)}_{rs} + B^{(5)}_{rs}) \\ &\lesssim q^3 \operatorname{tr}^2(M) \left\| M \right\|_F^2 + q^2 \operatorname{tr}(M) \left\| M \right\|_F \sqrt{\operatorname{tr}(M^4)} + q^2 \left\| M \right\|_F^4 + q \operatorname{tr}(M^4). \end{aligned}$$

By Poincaré's inequality,

$$\begin{aligned} &\operatorname*{Var}_G(\operatorname{tr}^2(G^T M G)) \\ &\lesssim \mathbb{E} \left\| \nabla g \right\|_2^2 \\ &\lesssim q^3 \operatorname{tr}^2(M) \left\| M \right\|_F^2 + q^2 \operatorname{tr}(M) \left\| M \right\|_F \sqrt{\operatorname{tr}(M^4)} + q^2 \left\| M \right\|_F^4 + q \operatorname{tr}(M^4). \end{aligned} \tag{9}$$

Similar to before, each term on the right-hand side grows geometrically.

Step 2b. Next we deal with $\mathrm{Var}_M(\mathbb{E}_G \operatorname{tr}^2(G^T MG)\,|M)$.

$$\mathbb{E}\operatorname{tr}^2\left(G^T MG\right) = \mathbb{E}\left(\sum_{i,k,l} M_{kl}G_{ki}G_{li}\right)^2 = \sum_{i,j}\sum_{k,l,k',l'} M_{kl}M_{k'l'}\,\mathbb{E}\,G_{ki}G_{li}G_{k'j}G_{l'j}.$$

When $i \neq j$, for non-zero contribution, it must hold that $k = l$ and $k' = l'$ and thus the nonzero contribution is

$$\sum_{i\neq j}\sum_{k,k'} M_{kk}M_{k'k'} = q(q-1)\operatorname{tr}^2(M).$$

When $i = j$, the contribution is (this is exactly the same as (7) in Step 1b.)

$$\sum_{i}\sum_{k,k',l,l'} M_{kl}M_{k'l'}\,\mathbb{E}\,G_{ki}G_{li}G_{k'i}G_{l'i} = 2q\,\|M\|_F^2 + q\operatorname{tr}^2(M).$$

Hence

$$\mathbb{E}\operatorname{tr}^2\left(G^T MG\right) = 2q\,\|M\|_F^2 + q^2\operatorname{tr}^2(M)$$

and when M is random,

$$\begin{aligned}
&\mathrm{Var}\left(\mathbb{E}\operatorname{tr}^2(G^T MG)\middle|\, M\right)\\
&= \mathrm{Var}\left(2q\,\|M\|_F^2 + q^2\operatorname{tr}^2(M)\right)\\
&\leq 4q^2\,\mathrm{Var}(\|M\|_F^2) + q^4\,\mathrm{Var}(\operatorname{tr}^2(M)) + 2q^3\sqrt{\mathrm{Var}(\|M\|_F^2)\,\mathrm{Var}(\operatorname{tr}^2(M))}.
\end{aligned} \tag{10}$$

Step 3. Let U_r denote the variance of $\operatorname{tr}((A_r^T A_r)^2)$ and V_r the variance of $\operatorname{tr}^2(A_r^T A_r)$. Combining (5), (6), (8), (9), (10), we have the following recurrence relations, where $C_1, C_2, C_3, C_4 > 0$ are absolute constants.

$$\begin{aligned}
U_{r+1} &\leq C_1 P_r + 2U_r + \frac{1}{d_r^2}V_r + \frac{3}{d_r}\sqrt{U_r V_r}\\
V_{r+1} &\leq C_2 Q_r + \frac{1}{d_r^2}U_r + V_r + \frac{2}{d_r}\sqrt{U_r V_r}\\
P_{r+1} &\leq C_3 P_r\\
Q_{r+1} &\leq C_4 Q_r\\
U_0 &= V_0 = 0
\end{aligned}$$

In the base case, set $M = I_p$ (the $p\times p$ identity matrix in (6)) and note that the second term in (5) vanishes. We see that $P_1 \lesssim (p^3 q + pq^3)/d_1^4$ after proper normalization. (Alternatively we can calculate this precisely, see Appendix D.) Similarly we have $Q_1 \lesssim p^3q^3/d_1^4$. Note that $Q_1/d_1^2 \lesssim (p^3q + pq^3)/d_1^4$. Now, we can solve that

$$U_{r+1} \leq C^r\frac{p^3q + pq^3}{d_1^4}$$

for some absolute constant $C > 0$.

References

1. Gernot Akemann and Jesper R Ipsen. Recent exact and asymptotic results for products of independent random matrices. *Acta Physica Polonica B*, pages 1747–1784, 2015.
2. Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 775–783, Cambridge, MA, USA, 2015. MIT Press.
3. Richard Bellman. Limit theorems for non-commutative operations. I. *Duke Mathematical Journal*, 21(3):491–500, 1954.
4. Giancarlo Benettin. Power-law behavior of lyapunov exponents in some conservative dynamical systems. *Physica D: Nonlinear Phenomena*, 13(1-2):211–220, 1984.
5. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
6. Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, pages 1215–1289, 2020. doi:10.1007/s00440-020-00998-3.
7. Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016. doi:10.1002/rsa.20633.
8. Sébastien Bubeck and Shirshendu Ganguly. Entropic CLT and Phase Transition in High-dimensional Wishart Matrices. *International Mathematics Research Notices*, 2018(2):588–606, 12 2016. doi:10.1093/imrn/rnw243.
9. Z. Burda, R. A. Janik, and B. Waclaw. Spectrum of the product of independent random gaussian matrices. *Physical Review E*, 81(4), Apr 2010. doi:10.1103/physreve.81.041132.
10. Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.
11. Andrea Crisanti, Giovanni Paladin, and Angelo Vulpiani. *Products of random matrices: in Statistical Physics*, volume 104. Springer Science & Business Media, 2012.
12. David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
13. S. Iida, H.A. Weidenmüller, and J.A. Zuk. Statistical scattering theory, the supersymmetry method and universal conductance fluctuations. *Annals of Physics*, 200(2):219–270, 1990.
14. Jesper R. Ipsen. *Products of independent Gaussian random matrices*. PhD thesis, Bielefeld University, 2015.
15. Hiroshi Ishitani. A central limit theorem for the subadditive process and its application to products of random matrices. *Publications of the Research Institute for Mathematical Sciences*, 12(3):565–575, 1977.
16. Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):1497–1529, Jul 2006. doi:10.1214/009117906000000205.
17. Tiefeng Jiang and Yutao Ma. Distances between random orthogonal matrices and independent normals. *Transactions of the American Mathematical Society*, 372(3):1509–1553, August 2019. doi:10.1090/tran/7470.
18. Ravindran Kannan and Santosh Vempala. *Spectral algorithms*. Now Publishers Inc, 2009.
19. Michael Kapralov, Vamsi Potluru, and David Woodruff. How to fake multiply by a gaussian matrix. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2016.
20. Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, February 2011. doi:10.1561/2200000035.
21. Satya N Majumdar and Grégory Schehr. Top eigenvalue of a random matrix: large deviations and third order phase transition. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(1):P01012, 2014.
22. Robert M May. Will a large complex system be stable? *Nature*, 238(5364):413–414, 1972.

23 P.A. Mello, P. Pereyra, and N. Kumar. Macroscopic approach to multichannel disordered conductors. *Annals of Physics*, 181(2):290–317, 1988.

24 Ralf R Muller. On the asymptotic eigenvalue distribution of concatenated vector-valued fading channels. *IEEE Transactions on Information Theory*, 48(7):2086–2091, 2002.

25 Shanmugavelayutham Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.

26 Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

27 James C Osborn. Universal results from an alternate random-matrix model for QCD with a baryon chemical potential. *Physical review letters*, 93(22):222001, 2004.

28 G Paladin and A Vulpiani. Scaling law and asymptotic distribution of lyapunov exponents in conservative dynamical systems with many degrees of freedom. *Journal of Physics A: Mathematical and General*, 19(10):1881, 1986.

29 Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for linear support vector machines. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(4):1–25, 2014.

30 Miklós Z. Rácz and Jacob Richey. A smooth transition from wishart to goe. *Journal of Theoretical Probability*, pages 898–906, 2019. `doi:10.1007/s10959-018-0808-2`.

31 Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

32 Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

33 Khalid Shebrawi and Hussein Albadawi. Trace inequalities for matrices. *Bulletin of the Australian Mathematical Society*, 87(1):139–148, 2013. `doi:10.1017/S0004972712000627`.

34 Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

35 Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012. `doi:10.1017/CBO9780511794308.006`.

36 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, October 2014. `doi:10.1561/0400000060`.

37 Fan Yang, Sifan Liu, Edgar Dobriban, and David P Woodruff. How to reduce dimension with PCA and random projections? arXiv:2005.00511 [math.ST], 2020.

A Proof of Proposition 12

Proof. We have

$$\begin{aligned}\mathbb{E}\,A_{11}^4 = \mathbb{E}\left(\sum_i B_{1i}G_{i1}\right)^4 &= \sum_{i,j,k,l}\mathbb{E}\,B_{1i}B_{1j}B_{1k}B_{1l}\,\mathbb{E}\,G_{i1}G_{j1}G_{k1}G_{l1}\\ &= 3\sum_i \mathbb{E}\,B_{1i}^4 + 3\sum_{i\neq j}\mathbb{E}\,B_{1i}^2B_{1j}^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\,A_{21}^4 = \mathbb{E}\left(\sum_i B_{2i}G_{i1}\right)^4 &= \sum_{i,j,k,l}\mathbb{E}\,B_{2i}B_{2j}B_{2k}B_{2l}\,\mathbb{E}\,G_{i1}G_{j1}G_{k1}G_{l1}\\ &= 3\sum_i \mathbb{E}\,B_{2i}^4 + 3\sum_{i\neq j}\mathbb{E}\,B_{2i}^2B_{2j}^2 = \mathbb{E}\,A_{11}^4.\end{aligned}$$ ◀

B Omitted Calculations in Section 4.1

$$S_1(p,q) = 3\sum_i \mathbb{E}\, B_{1i}^4 + 3\sum_{i\neq j} \mathbb{E}\, B_{1i}^2 B_{1j}^2 = 3dT_1(p,d) + 3d(d-1)T_4(p,d)$$

$$\begin{aligned} S_3(p,q) = \mathbb{E}\, A_{11}^2 A_{21}^2 &= \mathbb{E}\left(\sum_i B_{1i} G_{i1}\right)^2 \left(\sum_k B_{2k} G_{k1}\right)^2 \\ &= \sum_{i,j,k,l} \mathbb{E}\, B_{1i} B_{1j} B_{2k} B_{2l}\, \mathbb{E}\, G_{i1} G_{j1} G_{k1} G_{l1} \\ &= 3\sum_i \mathbb{E}\, B_{1i}^2 B_{2i}^2 + \sum_{i\neq j} \mathbb{E}\, B_{1i}^2 B_{2j}^2 + 2\sum_{i\neq j} \mathbb{E}\, B_{1i} B_{2i} B_{1j} B_{2j} \\ &= 3dT_3(p,d) + d(d-1)T_5(p,d) + 2d(d-1)T_6(p,d) \end{aligned}$$

$$\begin{aligned} S_4(p,q) = \mathbb{E}\, A_{11}^2 A_{12}^2 &= \mathbb{E}\left(\sum_i B_{1i} G_{i1}\right)^2 \left(\sum_k B_{1k} G_{k2}\right)^2 \\ &= \sum_{i,j,k,l} \mathbb{E}\, B_{1i} B_{1j} B_{1k} B_{1l}\, \mathbb{E}\, G_{i1} G_{j1} G_{k2} G_{l2} \\ &= \sum_i \mathbb{E}\, B_{1i}^4 + \sum_{i\neq j} \mathbb{E}\, B_{1i}^2 B_{1j}^2 = dT_1(p,d) + d(d-1)T_4(p,d). \end{aligned}$$

$$\begin{aligned} S_5(p,q) = \mathbb{E}\, A_{11}^2 A_{22}^2 &= \mathbb{E}\left(\sum_i B_{1i} G_{i1}\right)^2 \left(\sum_k B_{2k} G_{k2}\right)^2 \\ &= \sum_{i,j,k,l} \mathbb{E}\, B_{1i} B_{1j} B_{2k} B_{2l}\, \mathbb{E}\, G_{i1} G_{j1} G_{k2} G_{l2} \\ &= \sum_{i,j} B_{1i}^2 B_{2j}^2 = dT_3(p,d) + d(d-1)T_5(p,d) \end{aligned}$$

$$\begin{aligned} S_6(p,q) = \mathbb{E}\, A_{11} A_{12} A_{21} A_{22} &= \sum_{i,j,k,l} \mathbb{E}\, B_{1i} B_{1j} B_{2k} B_{2l}\, \mathbb{E}\, G_{i1} G_{j2} G_{k1} G_{l2} \\ &= \sum_i \mathbb{E}\, B_{1i}^2 B_{2i}^2 + \sum_{i\neq j} \mathbb{E}\, B_{1i} B_{1j} B_{2i} B_{2j} \\ &= dT_3(p,d) + d(d-1)T_6(p,d) \end{aligned}$$

C Omitted Calculations in Section 4.2

In Step 1a.

$$\begin{aligned}
B_{r,s}^{(2)} &= \sum_{l,l'} \sum_{u,v,v'} M_{uv} M_{uv'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{us}^2 X_{vj} X_{lj} X_{v'j} X_{l'j} \\
&= \underbrace{\sum_{l \neq l'} \sum_{u} M_{ul} M_{ul'} M_{rl} M_{rl'}}_{v=l\neq v'=l'} + \underbrace{\sum_{l} \sum_{\substack{u \\ v\neq l}} M_{uv}^2 M_{rl}^2}_{v=v'\neq l=l'} + \underbrace{\sum_{l\neq l'} \sum_{u} M_{ul'} M_{ul} M_{rl} M_{rl'}}_{v=l'\neq l=v'} \\
&\qquad + 3 \underbrace{\sum_{l,u} M_{ul}^2 M_{rl}^2}_{v=v'=l=l'} \\
&= \left(\sum_{u,v} M_{uv}^2\right) \left(\sum_{l} M_{rl}^2\right) + 2 \sum_{l,l',u} M_{ul'} M_{ul} M_{rl} M_{rl'} \\
&= \|M\|_F^2 \, \|M_{r,\cdot}\|_2^2 + 2 \sum_{u} \langle M_{u,\cdot}, M_{r,\cdot}\rangle^2 .
\end{aligned}$$

$$\begin{aligned}
B_{r,s}^{(3)} &= \sum_{j'\neq s} \left[\sum_{l,l'} \sum_{u,v} M_{uv} M_{u'l'} M_{rl} M_{rl'} \, \mathbb{E}\, X_{us} X_{u's} X_{vs} X_{ls} X_{l'j'}^2 \right] \\
&= \underbrace{\sum_{l,l'} \sum_{u\neq l} M_{ul} M_{ul'} M_{rl} M_{rl'}}_{u=u'\neq v=l} + \underbrace{\sum_{l,l'} \sum_{u\neq l} M_{uu} M_{ll'} M_{rl} M_{rl'}}_{u=v\neq u'=l} + \underbrace{\sum_{l,l'} \sum_{v} M_{lv} M_{vl'} M_{rl} M_{rl'}}_{u=l\neq u'=v} \\
&\qquad + 3 \underbrace{\sum_{l,l'} M_{ll} M_{ll'} M_{rl} M_{rl'}}_{u=u'=v=l} \\
&= \sum_{l,l'} \sum_{u} (2 M_{ul} M_{ul'} + M_{uu} M_{ll'}) M_{rl} M_{rl'} \\
&= \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot}\rangle + \operatorname{tr}(M) M_{ll'}) M_{rl} M_{rl'} .
\end{aligned}$$

$$\begin{aligned}
\sum_{r,s} B_{r,s}^{(3)} &= \sum_{r,s} \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot}\rangle + \operatorname{tr}(M) M_{l'l'}) M_{rl} M_{rl'} \\
&= q \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot}\rangle + \operatorname{tr}(M) M_{ll'}) \sum_{r} M_{rl} M_{rl'} \\
&= q \sum_{l,l'} (2\langle M_{l,\cdot}, M_{l',\cdot}\rangle + \operatorname{tr}(M) M_{ll'}) \langle M_{l,\cdot}, M_{l',\cdot}\rangle \\
&= q \sum_{l,l'} 2\langle M_{l,\cdot}, M_{l',\cdot}\rangle^2 + q \operatorname{tr}(M) \sum_{l,l'} M_{ll'} \langle M_{l,\cdot}, M_{l',\cdot}\rangle \\
&\leq 2q \operatorname{tr}(M^4) + q \operatorname{tr}(M) \left(\sum_{l,l'} M_{ll'}^2\right)^{\frac{1}{2}} \left(\sum_{l,l'} \langle M_{l,\cdot}, M_{l',\cdot}\rangle^2\right)^{\frac{1}{2}} \\
&\leq 2q \operatorname{tr}(M^4) + q \operatorname{tr}(M) \, \|M\|_F \, \sqrt{\operatorname{tr}(M^4)}
\end{aligned}$$

In Step 1b.

$$
\begin{aligned}
&\sum_i \sum_{\substack{k,l\\k',l'}} M_{kl} M_{k'l'}\, \mathbb{E}\, G_{ki} G_{li} G_{k'i} G_{l'i} \\
&= \sum_i \left(\underbrace{\sum_{k\neq l} M_{kl}^2}_{k=k'\neq l=l'} + \underbrace{\sum_{k\neq l} M_{kl}^2}_{k=l'\neq k'=l} + \underbrace{\sum_{k\neq l} M_{kk} M_{k'k'}}_{k=l\neq k'=l'} + 3 \underbrace{\sum_{k} M_{kk}^2}_{k=k'=l=l'} \right) \\
&= \sum_i \left(\sum_{k,l} M_{kl}^2 + \sum_{k,l} M_{kl}^2 + \sum_{k,l} M_{kk} M_{k'k'} \right) \\
&= \sum_i (2\left\|M\right\|_F^2 + \operatorname{tr}^2(M)) \\
&= 2q\left\|M\right\|_F^2 + q\operatorname{tr}^2(M).
\end{aligned}
$$

In Step 2a.

$$
\begin{aligned}
B_{rs}^{(2)} &= \sum_{l,l',u,u',v} M_{uv} M_{u'u'} M_{rl} M_{rl'}\, \mathbb{E}\, X_{ls} X_{l's} X_{us} X_{vs} X_{u'j'}^2 \\
&= \underbrace{\sum_{\substack{l\neq u\\u'}} M_{uu} M_{u'u'} M_{rl}^2}_{l=l'\neq u=v} + \underbrace{\sum_{\substack{l\neq l'\\u'}} M_{ll'} M_{u'u'} M_{rl} M_{rl'}}_{l=u\neq l'=v} + \underbrace{\sum_{\substack{l\neq l'\\u'}} M_{ll'} M_{u'u'} M_{rl} M_{rl'}}_{l=v\neq l'=u} \\
&\qquad + 3 \underbrace{\sum_{l,u'} M_{ll} M_{u'u'} M_{rl}^2}_{l=u=l'=v} \\
&= \operatorname{tr}(M) \left(\sum_{l,u} M_{uu} M_{rl}^2 + 2 \sum_{l,l'} M_{ll'} M_{rl} M_{rl'} \right) \\
&= \operatorname{tr}(M) \left(\operatorname{tr}(M)\left\|M_{r,\cdot}\right\|_2^2 + 2 \sum_{l,l'} M_{ll'} M_{rl} M_{rl'} \right)
\end{aligned}
$$

$$
\begin{aligned}
B_{rs}^{(4)} &= \sum_{l,u,u',v,v'} M_{uv} M_{u'v'} M_{rl}^2\, \mathbb{E}\, X_{uj} X_{vj} X_{u'j} X_{v'j} \\
&= \left(\sum_l M_{rl}^2 \right) \left(\underbrace{\sum_{u,v} M_{uv}^2}_{u=u'\neq v=v'} + \underbrace{\sum_{u,v} M_{uv}^2}_{u=v'\neq u=v} + \underbrace{\sum_{u,v} M_{uv}^2}_{u=v\neq u'=v'} + 3 \underbrace{\sum_{u} M_{uu}^2}_{u=v=u'=v'} \right) \\
&= 3 \left(\sum_l M_{rl}^2 \right) \sum_{u,v} M_{u,v}^2 \\
&= 3\left\|M_{r,\cdot}\right\|_2^2 \left\|M\right\|_F^2
\end{aligned}
$$

D Exact Variance when $r = 2$

Suppose that A is rotationally invariant under both left- and right-multiplication of an orthogonal matrix. Define

$$\begin{aligned}
&U_1(p,q) = \mathrm{Var}((A^TA)_{ii}^2)\\
&U_2(p,q) = \mathrm{Var}((A^TA)_{ij}^2) \quad i \neq j\\
&U_3(p,q) = \mathrm{cov}((A^TA)_{ii}^2, (A^TA)_{ik}^2) \quad i \neq k \quad \text{(same row, one entry on diagonal)}\\
&U_4(p,q) = \mathrm{cov}((A^TA)_{ij}^2, (A^TA)_{ik}^2) \quad j \neq k \quad \text{(same row, both entries off-diagonal)}\\
&U_5(p,q) = \mathrm{cov}((A^TA)_{ii}^2, (A^TA)_{jj}^2) \quad i \neq j \quad \text{(diff. rows and cols, both entries on diagonal)}\\
&U_6(p,q) = \mathrm{cov}((A^TA)_{ii}^2, (A^TA)_{jk}^2) \quad i \neq j \neq k \quad \text{(diff. rows and cols, one entry on diagonal)}\\
&U_7(p,q) = \mathrm{cov}((A^TA)_{ij}^2, (A^TA)_{kl}^2) \quad i \neq j \neq k \neq l \quad \text{(diff. rows and cols, nonsymmetric around diag.)}
\end{aligned}$$

It is clear that they are well-defined.

$$\begin{aligned}
&\mathrm{Var}(\mathrm{tr}((A^TA)^2))\\
&= \mathrm{Var}\left(\sum_{i,j}(A^TA)_{ij}^2\right)\\
&= \sum_{i,j,k,l} \mathrm{cov}((A^TA)_{ij}^2, (A^TA)_{kl}^2)\\
&= \sum_{i,j}\mathrm{Var}((A^TA)_{ij}^2) + 2\sum_i\sum_{j\neq l}\mathrm{cov}((A^TA)_{ij}^2, (A^TA)_{il}^2) + \sum_{\substack{i\neq k\\ j\neq l}}\mathrm{cov}(\mathbb{E}(A^TA)_{ij}^2, (A^TA)_{kl}^2)\\
&= q\,\mathrm{Var}((A^TA)_{11}^2) + q(q-1)\,\mathrm{Var}(\mathbb{E}(A^TA)_{12}^2)\\
&\quad + 2\left[2q(q-1)\,\mathrm{cov}((A^TA)_{11}^2, (A^TA)_{12}^2) + q(q-1)(q-2)\,\mathrm{cov}((A^TA)_{12}^2, (A^TA)_{13}^2)\right]\\
&\quad + q(q-1)\,\mathrm{cov}(\mathbb{E}(A^TA)_{11}^2, (A^TA)_{22}^2) + q(q-1)\,\mathrm{cov}(\mathbb{E}(A^TA)_{12}^2, (A^TA)_{21}^2)\\
&\qquad + 2q(q-1)(q-2)\,\mathrm{cov}((A^TA)_{11}^2, (A^TA)_{23}^2)\\
&\qquad + 2q(q-1)(q-2)\,\mathrm{cov}((A^TA)_{12}^2, (A^TA)_{31}^2)\\
&\qquad + q(q-1)(q-2)(q-3)\,\mathbb{E}(A^TA)_{12}^2(A^TA)_{34}^2\\
&= qU_1(p,q) + q(q-1)U_2(p,q) + 2q(q-1)(2U_3(p,q) + (q-2)U_4(p,q))\\
&\quad + q(q-1)(U_5(p,q) + U_2(p,q)) + 2q(q-1)(q-2)(U_6(p,q) + U_4(p,q))\\
&\quad + q(q-1)(q-2)(q-3)U_7(p,q)\\
&= qU_1(p,q) + q(q-1)(2U_2(p,q) + 4U_3(p,q) + U_5(p,q))\\
&\quad + 2q(q-1)(q-2)(2U_4(p,q) + U_6(p,q)) + q(q-1)(q-2)(q-3)U_7(p,q).
\end{aligned}$$

Let us calculate $U_1, \ldots, U_7$ for a $p \times q$ Gaussian random matrix G.

$$\begin{aligned}
U_1(p,q) = \mathbb{E}(G^TG)_{11}^4 - (\mathbb{E}(G^TG)_{11}^2)^2 &= \mathbb{E}\,\|G_1\|_2^8 - (\mathbb{E}\,\|G_1\|_2^4)^2\\
&= p(p+2)(p+4)(p+6) - (p(p+2))^2\\
&= 8p(p+2)(p+3)
\end{aligned}$$

$$
\begin{aligned}
U_2(p,q) &= \mathbb{E}(G^TG)_{12}^4 - (\mathbb{E}(G^TG)_{12}^2)^2 = \mathbb{E}\left(\sum_r G_{r1}G_{r2}\right)^4 - (\mathbb{E}\langle G_1, G_2\rangle^2)^2 \\
&= \sum_{r,s,t,u} \mathbb{E}\, G_{r1}G_{s1}G_{t1}G_{u1}G_{r2}G_{s2}G_{t2}G_{u2} - p^2 \\
&= 3\sum_{r\neq t} \mathbb{E}\, G_{r1}^2 G_{t1}^2 G_{r2}^2 G_{t2}^2 + \sum_r G_{r1}^4 G_{r2}^4 - p^2 \\
&= 3p(p-1) + 9p - p^2 = 2p(p+3). \\
U_3(p,q) &= \mathbb{E}(G^TG)_{11}^2 (G^TG)_{12}^2 - \mathbb{E}(G^TG)_{11}^2\, \mathbb{E}(G^TG)_{12}^2 \\
&= \mathbb{E}(G_1^TG_1)^2 G_1^TG_2G_2^TG_1 - \mathbb{E}\,\|G_1\|_2^4\, \mathbb{E}\langle G_1, G_2\rangle^2 \\
&= \mathbb{E}(G_1^TG_1)^2 G_1^T(\mathbb{E}\, G_2G_2^T)G_1 - p(p+2)\cdot p \\
&= \mathbb{E}(G_1^TG_1)^3 - p^2(p+2) \\
&= \mathbb{E}\,\|G_1\|_2^6 - p^2(p+2) = p(p+2)(p+4) - p^2(p+2) = 4p(p+2) \\
U_4(p,q) &= \mathbb{E}(G^TG)_{12}^2 (G^TG)_{13}^2 - \mathbb{E}(G^TG)_{12}^2\, \mathbb{E}(G^TG)_{13}^2 \\
&= \mathbb{E}\, G_1^TG_2G_2^TG_1G_1^TG_3G_3^TG_1 - p^2 \\
&= \mathbb{E}\, G_1^T\, \mathbb{E}(G_2G_2^T)G_1G_1^T\, \mathbb{E}(G_3G_3^T)G_1 - p^2 \\
&= \mathbb{E}(G_1^TG_1)^2 - p^2 = \mathbb{E}\,\|G_1\|_2^4 - p^2 = p(p+2) - p^2 = 2p \\
U_5(p,q) &= \mathbb{E}(G^TG)_{11}^2 (G^TG)_{22}^2 - \mathbb{E}(G^TG)_{11}^2\, \mathbb{E}(G^TG)_{22}^2 \\
&= \mathbb{E}\,\|G_1\|_2^4\,\|G_2\|_2^4 - \mathbb{E}\,\|G_1\|_2^4\,\|G_2\|_2^4 = 0 \\
U_6(p,q) &= \mathbb{E}(G^TG)_{11}^2 (G^TG)_{23}^2 - \mathbb{E}(G^TG)_{11}^2\, \mathbb{E}(G^TG)_{23}^2 \\
&= \mathbb{E}\,\|G_1\|_2^4\,\langle G_2, G_3\rangle^2 - \mathbb{E}\,\|G_1\|_2^4\, \mathbb{E}\langle G_2, G_3\rangle^2 = 0 \\
U_7(p,q) &= \mathbb{E}(G^TG)_{12}^2 (G^TG)_{34}^2 - \mathbb{E}(G^TG)_{12}^2\, \mathbb{E}(G^TG)_{34}^2 \\
&= \mathbb{E}\langle G_1, G_2\rangle^2 \langle G_3, G_4\rangle^2 - \mathbb{E}\langle G_1, G_2\rangle^2\, \mathbb{E}\langle G_3, G_4\rangle^2 = 0
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathrm{Var}(\mathrm{tr}((G^TG)^2)) &= qU_1 + q(q-1)(2U_2 + 4U_3 + U_5) + 2q(q-1)(q-2)(2U_4 + U_6) \\
&\quad + q(q-1)(q-2)(q-3)U_7 \\
&= qU_1 + q(q-1)(2U_2 + 4U_3) + 4q(q-1)(q-2)U_4 \\
&= 4pq(5 + 5p + 5q + 2p^2 + 5pq + 2q^2).
\end{aligned}
$$

When $r = 2$, recalling that $\mathbb{E}(A_2 - A_1) = (1 + o(1))p^2q^2/d^3$ (see (4)), we have that

$$
\frac{\sqrt{\mathrm{Var}(\mathrm{tr}((\frac{1}{\sqrt{d}}G^T \cdot \frac{1}{\sqrt{d}}G)^2))}}{p^2q^2/d^3} \le \frac{6d}{\max\{p,q\}^{\frac{1}{2}} \min\{p,q\}^{\frac{3}{2}}}.
$$

If the right-hand side above is at most a small constant c, we can distinguish A_2 from A_1 with probability at least a constant.