Alea 4, 257–283 (2008)

Multivariate normal approximation using exchangeable pairs

Sourav Chatterjee and Elizabeth Meckes

367 Evans Hall #3860, Department of Statistics, University of California at Berkeley, Berkeley, CA 94720-3860.
E-mail address: sourav@stat.berkeley.edu
URL: http://www.stat.berkeley.edu/~sourav/

220 Yost Hall, Department of Mathematics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44122.
E-mail address: ese3@cwru.edu
URL: http://case.edu/artsci/math/esmeckes/

Abstract. Since the introduction of Stein's method in the early 1970s, much research has been done in extending and strengthening it; however, there does not exist a version of Stein's original method of exchangeable pairs for multivariate normal approximation. The aim of this article is to fill this void. We present three abstract normal approximation theorems using exchangeable pairs in multivariate contexts, one for situations in which the underlying symmetries are discrete, and real and complex versions of a theorem for situations involving continuous symmetry groups. Our main applications are proofs of the approximate normality of rank k projections of Haar measure on the orthogonal and unitary groups, when $k = o(n)$.

1. Introduction

Stein's method was introduced by Charles Stein (1972) as a tool for proving central limit theorems for sums of dependent random variables. Stein's version of his method, best known as the "method of exchangeable pairs", is described in detail in his later work (Stein, 1986). The method of exchangeable pairs is a general technique whose applicability is not restricted to sums of random variables; for some recent examples, one can look at the work of Jason Fulman (2005) on central

Received by the editors January 15, 2008; accepted July 28, 2008.
2000 *Mathematics Subject Classification.* 60F05; 60B15.
Key words and phrases. Stein's method, exchangeable pairs, normal approximation, random matrices.
S. Chatterjee's research partially supported by NSF grant DMS-0707054 and a Sloan Research Fellowship.
E. Meckes's research partially supported by an ARCS Fellowship and an American Institute of Mathematics Five-year Fellowship.

limit theorems for complicated objects arising from the representation theory of permutation groups, and the work of the second-named author (Meckes, 2007) on the distribution of eigenfunctions of the Laplacian on Riemannian manifolds.

One of the significant advantages of the method is that it automatically gives concrete error bounds. Although Stein's original theorem does not generally give Kolmogorov distance bounds of the correct order, there has been substantial research on modifications of Stein's result to obtain rate-optimal Berry-Esséen type bounds (see e.g. the works of Rinott and Rotar, 1997 and Shao and Su, 2006). The "infinitesimal" version of the method described in Meckes (2006) and in our Theorems 2.4 and 2.5 below frequently does produce bounds of the correct order, in total variation distance in the univariate case and in Wasserstein distance in the multivariate case.

Heuristically, the method of exchangeable pairs for univariate normal approximation goes as follows. Suppose that a random variable W is conjectured to be approximately a standard Gaussian. The first step in the method is to construct a second random variable W' on the same probability space such that (W, W') is an exchangeable pair, i.e. (W, W') has the same distribution as (W', W). The random variable W' is generally constructed by making a small random change in W, so that W and W' are close.

Let $\Delta = W - W'$. The next step is to verify the existence of a small number λ such that

$$\mathbb{E}(\Delta \mid W) = \lambda W + r_1, \tag{1.1}$$

$$\mathbb{E}(\Delta^2 \mid W) = 2\lambda + r_2, \quad \text{and} \tag{1.2}$$

$$\mathbb{E}|\Delta|^3 = r_3, \tag{1.3}$$

where the random quantities r_1, r_2, and r_3 are all negligible compared to λ. If the above relations hold, then, depending on the sizes of λ and the r_i's, one can conclude that W is approximately Gaussian. The exact statement of Stein's abstract normal approximation theorem for piecewise differentiable test functions is the following:

Theorem 1.1 (Stein (1986), page 35)**.** *Let (W, W') be an exchangeable pair of real random variables such that $\mathbb{E}W^2 = 1$ and $\mathbb{E}\big[W - W' \mid W\big] = \lambda W$ for some $0 < \lambda < 1$. Let $\Delta = W - W'$. Let $h : \mathbb{R} \to \mathbb{R}$ be bounded with piecewise continuous derivative h'. Then for Z a standard normal random variable,*

$$\big|\mathbb{E}h(W) - \mathbb{E}h(Z)\big| \le \frac{\|h - \mathbb{E}h(Z)\|_\infty}{\lambda}\sqrt{\mathrm{Var}(\mathbb{E}\big[\Delta^2 \mid W\big])} + \frac{\|h'\|_\infty}{4\lambda}\mathbb{E}|\Delta|^3.$$

Observe that the condition $\mathbb{E}\big[W - W'\big|W\big] = \lambda W$ implies that $\mathbb{E}\Delta^2 = 2\lambda$, and thus the bound in Stein's theorem above can also be stated as:

$$\big|\mathbb{E}h(W) - \mathbb{E}h(Z)\big| \le 2\|h - \mathbb{E}h(Z)\|_\infty\sqrt{\mathbb{E}\left[\frac{1}{2\lambda}\mathbb{E}\left[\Delta^2\big|W\right] - 1\right]} + \frac{\|h'\|_\infty}{4\lambda}\mathbb{E}|\Delta|^3.$$

Powerful as it is, the above theorem and all its existing modifications cater only to *univariate* normal approximation. There has been some previous work in proving multivariate central limit theorems using Stein's method, though none of these approaches have used exchangeable pairs. Rinott and Rotar (1996) proved multivariate central limit theorems for sums of dependent random vectors using the dependency graph version of Stein's method. Around the same time, Goldstein and Rinott (1996a) developed the size-bias coupling version of Stein's method for multivariate normal approximation. Both of these techniques are well-known and in regular use. More recently, Raič (2004) proved a new multivariate central limit theorem for sums of dependent random vectors with the dependency graph approach which removed the need for finite third moments. However, as in the univariate case, there are many problems which are more amenable to analysis via exchangeable pairs (particularly the adaptation to the case of continuous symmetries) which necessitates the creation of a multivariate version of this method. The present authors introduced, for the first time, a multivariate version of Theorem 1.1 in an earlier draft of this manuscript that was posted on arXiv. Subsequently, an extension of one of our main results (Theorem 2.3) to the case of multivariate normal approximation with non-identity covariance was formulated by Reinert and Röllin (2008). Our current draft is mainly a reorganization of the original manuscript, with better error bounds in several examples. We refer to the Reinert-Röllin paper for many other interesting applications.

The contents of this paper are as follows. In Section 2, we prove three abstract normal approximation theorems which give a framework for using the method of exchangeable pairs in a multivariate context. The first is for situations in which the symmetry used in constructing the exchangeable pair is discrete, and is a fairly direct analog of Theorem 1.1 above. An an example, the theorem is applied in Section 3 to prove a basic central limit theorem for a sum of independent, identically distributed random vectors.

The second abstract theorem of Section 2 includes an additional modification, making it useful in situations in which continuous symmetries are present. The idea for the modification was introducted in the technical report by Stein (1995) and further developed in Meckes (2006). Section 4 contains two applications of this theorem. First, for Y a random vector in $\mathbb{R}^n$ with spherically symmetric distribution, sufficient conditions are given under which the first k coordinates are approximately distributed as a standard normal random vector in $\mathbb{R}^k$. We then give a treatment of projections of Haar measure on the orthogonal group. Specifically, for M a random $n \times n$ orthogonal matrix and $A_1, \ldots, A_k$ fixed matrices over $\mathbb{R}$, we give an explicit bound on the Wasserstein distance between $(\mathrm{Tr}\,(A_1 M), \ldots, \mathrm{Tr}\,(A_k M))$ and a Gaussian random vector.

As a corollary to the theorem discussed above, we state a theorem for bounding the distance between a complex random vector and a complex Gaussian random vector, in the context of continuous groups of symmetries. The main application of this version of the theorem in given in Section 4, where for M a random $n \times n$ unitary matrix and $A_1, \ldots, A_n$ fixed matrices over $\mathbb{C}$, we derive an explicit bound on the Wasserstein distance between $(\mathrm{Tr}\,(A_1 M), \ldots, \mathrm{Tr}\,(A_k M))$ and a complex Gaussian random vector.

Before moving into Section 2, we give the following very brief outline of the literature around the various other versions of Stein's method.

Other versions of Stein's method. The three most notable variants of Stein's method are (i) the dependency graph approach introduced by Baldi and Rinott (1989) and further developed by Arratia et al. (1990) and Barbour et al. (1989),

(ii) the size-biased coupling method of Goldstein and Rinott (1996b) (see also Barbour et al. (1992)), and (iii) the zero-biased coupling technique due to Goldstein and Reinert (1997). In addition to these three basic approaches, an important contribution was made by Andrew Barbour (1990), who noticed the connection between Stein's method and diffusion approximation. This connection has subsequently been widely exploited by practitioners of Stein's method, and is a mainstay of some of our proofs.

Besides normal approximation, Stein's method has been successfully used for proving convergence to several other distributions as well. Shortly after the method was introduced for normal approximation by Stein, Poisson approximation by Stein's method was introduced by Chen (1975) and became popular after the publication of Arratia et al. (1989, 1990). The method has also been developed for gamma approximation by Luk (1994); for chi-square approximation by Pickett (2004); for the uniform distribution on the discrete circle by Diaconis (2004); for the semi-circle law by Götze and Tikhomirov (2005); for the binomial and multinomial distributions by Holmes (2004) and Loh (1992); and the hypergeometric distribution, also by Holmes (2004).

The method of exchangeable pairs was extended to Poisson approximation by Chatterjee, Diaconis and Meckes in the survey paper Chatterjee et al. (2005), and to a general method of normal approximation for arbitrary functions of independent random variables in Chatterjee (2008).

For further references and exposition (particularly to the method of exchangeable pairs), we refer to the recent monograph edited by Diaconis and Holmes (2004).

1.1. *Notation and conventions.* The total variation distance $d_{TV}(\mu,\nu)$ between the measures μ and ν on $\mathbb{R}$ is defined by

$$d_{TV}(\mu,\nu) = \sup_A \big|\mu(A) - \nu(A)\big|,$$

where the supremum is over measurable sets A. This is equivalent to

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sup_f \left|\int f(t)d\mu(t) - \int f(t)d\nu(t)\right|,$$

where the supremum is taken over continuous functions which are bounded by 1 and vanish at infinity; this is the definition most commonly used in what follows. The total variation distance between two random variables X and Y is defined to be the total variation distance between their distributions:

$$d_{TV}(X,Y) = \sup_A \big|\mathbb{P}(X\in A) - \mathbb{P}(Y\in A)\big| = \frac{1}{2}\sup_f \big|\mathbb{E}f(X) - \mathbb{E}f(Y)\big|.$$

If the Banach space of signed measures on $\mathbb{R}$ is viewed as dual to the space of continuous functions on $\mathbb{R}$ vanishing at infinity, then the total variation distance is (up to the factor of $\frac{1}{2}$) the norm distance on that Banach space.

The Wasserstein distance $d_W(X,Y)$ between the random variables X and Y is defined by

$$d_W(X,Y) = \sup_{M_1(g)\leq 1} \big|\mathbb{E}g(X) - \mathbb{E}g(Y)\big|,$$

where $M_1(g) = \sup_{x\neq y}\frac{|g(x)-g(y)|}{|x-y|}$ is the Lipschitz constant of g. Note that Wasserstein distance is not directly comparable to total variation distance, since the class of functions considered is required to be Lipschitz but not required to be bounded.

In particular, total variation distance is always bounded by 1, whereas the statement that the Wasserstein distance between two distributions is bounded by 1 has content. On the space of probability distributions with finite absolute first moment, Wasserstein distance induces a stronger topology than the usual one described by weak convergence, but not as strong as the topology induced by the total variation distance. See Dudley (1989) for detailed discussion of the various notions of distance between probability distributions.

We will use $\mathfrak{N}(\mu, \sigma^2)$ to denote the normal distribution on $\mathbb{R}$ with mean μ and variance σ^2; unless otherwise stated, the random variable $Z = (Z_1, \dots, Z_k)$ is understood to be a standard Gaussian random vector on $\mathbb{R}^k$.

In $\mathbb{R}^n$, the Euclidean inner product is denoted $\langle \cdot, \cdot \rangle$ and the Euclidean norm is denoted $|\cdot|$. On the space of real (resp. complex) $n \times n$ matrices, the Hilbert-Schmidt inner product is defined by

$$\langle A, B \rangle_{H.S.} = \mathrm{Tr}\,(AB^T), \qquad \big(\text{resp. } \langle A, B \rangle_{H.S.} = \mathrm{Tr}\,(AB^*)\big)$$

with corresponding norms

$$\|A\|_{H.S.} = \sqrt{\mathrm{Tr}\,(AA^T)}, \qquad \left(\text{resp. } \|A\|_{H.S.} = \sqrt{\mathrm{Tr}\,(AA^*)}\right).$$

The operator norm of a matrix A over $\mathbb{R}$ is defined by

$$\|A\|_{op} = \sup_{|v|=1, |w|=1} |\langle Av, w \rangle|.$$

The $n \times n$ identity matrix is denoted I_n, the $n \times n$ matrix of all zeros is denoted 0_n, and $A \oplus B$ is the block direct sum of A and B.

For Ω a domain in $\mathbb{R}^k$, the notation $C^k(\Omega)$ will be used for the space of k-times continuously differentiable real-valued functions on Ω, and $C_o^k(\Omega) \subseteq C^k(\Omega)$ are those C^k functions on Ω with compact support. For $g : \mathbb{R}^k \to \mathbb{R}$, let

$$M_1(g) := \sup_{x \neq y} \frac{|g(x) - g(y)|}{|x - y|};$$

if $g \in \mathbb{C}^1(\mathbb{R}^k)$ also, then let

$$M_2(g) := \sup_{x \neq y} \frac{|\nabla g(x) - \nabla g(y)|}{|x - y|};$$

if $g \in \mathbb{C}^2(\mathbb{R}^k)$ as well, then

$$M_3(g) := \sup_{x \neq y} \frac{\|\mathrm{Hess}\, g(x) - \mathrm{Hess}\, g(y)\|_{op}}{|x - y|}.$$

The last definition differs from the one in Raič (2004), where M_3 is defined in terms of the Hilbert-Schmidt norm as opposed to the operator norm. Note that if $g \in C^1(\mathbb{R}^k)$, then $M_1(g) = \sup_x |\nabla g(x)|$, and if $g \in C^2(\mathbb{R}^k)$, then $M_2(g) = \sup_x \|\mathrm{Hess}\, g(x)\|_{op}$.

Acknowledgement. The authors wish to thank an anonymous referee, whose helpful comments and suggestions significantly improved both the results and the exposition of this paper.

2. Three abstract normal approximation theorems

In this section we develop the general machine that will be applied in the examples in Sections 3 and 4. In the following, we use the notation $\mathcal{L}(X)$ to denote the law of a random vector or variable X. The following lemma gives a second-order characterizing operator for the Gaussian distribution on $\mathbb{R}^k$.

Lemma 2.1. *Let $Z \in \mathbb{R}^k$ be a random vector with $\{Z_i\}_{i=1}^k$ independent, identically distributed standard Gaussians.*

(i) *If $f : \mathbb{R}^k \to \mathbb{R}$ is two times continuously differentiable and compactly supported, then*

$$\mathbb{E}\big[\Delta f(Z) - \langle Z, \nabla f(Z)\rangle\big] = 0.$$

(ii) *If $Y \in \mathbb{R}^k$ is a random vector such that*

$$\mathbb{E}\big[\Delta f(Y) - \langle Y, \nabla f(Y)\rangle\big] = 0$$

for every $f \in C^2(\mathbb{R}^k)$ with $\mathbb{E}\big|\Delta f(Y) - \langle Y, \nabla f(Y)\rangle\big| < \infty$, then $\mathcal{L}(Y) = \mathcal{L}(Z)$.

(iii) *If $g \in C_o^\infty(\mathbb{R}^k)$, then the function*

$$U_o g(x) := \int_0^1 \frac{1}{2t}\big[\mathbb{E}g(\sqrt{t}x + \sqrt{1-t}Z) - \mathbb{E}g(Z)\big]dt$$

is a solution to the differential equation

$$\Delta h(x) - \langle x, \nabla h(x)\rangle = g(x) - \mathbb{E}g(Z). \tag{2.1}$$

Remark. The form of $U_o g$ is a direct rewriting of the inverse of the Ornstein-Uhlenbeck generator (see Barbour, 1990).

Proof. Part (i) is just integration by parts.

Part (ii) follows easily from part (iii): note that if

$$\mathbb{E}\big[\Delta f(Y) - \langle Y, \nabla f(Y)\rangle\big] = 0$$

for every $f \in C^2(\mathbb{R}^k)$ with $\mathbb{E}\big|\Delta f(Y) - \langle Y, \nabla f(Y)\rangle\big| < \infty$, then for $g \in C_o^\infty$ given,

$$\mathbb{E}g(Y) - \mathbb{E}g(Z) = \mathbb{E}\big[\Delta(U_o g)(Y) - \langle Y, \nabla(U_o g)(Y)\rangle\big] = 0,$$

and so $\mathcal{L}(Y) = \mathcal{L}(Z)$ since C_o^∞ is dense in the class of bounded continuous functions vanishing at infinity, with respect to the supremum norm.

A proof of part (iii) is given in Barbour (1990), Götze (1991) and Raič (2004), all using results about Markov semi-groups. For a direct proof, see Meckes (2006). □

The next lemma gives useful bounds on $U_o g$ and its derivatives in terms of g and its derivatives. As in Raič (2004), bounds are most naturally given in terms of the quantities $M_i(g)$ defined in the introduction.

Lemma 2.2. *For $g : \mathbb{R}^k \to \mathbb{R}$ given, $U_o g$ satisfies the following bounds:*

(i)

$$\sup_{x \in R^k} \|\mathrm{Hess}\, U_o g(x)\|_{H.S.} \le M_1(g).$$

(ii)

$$M_3(U_o g) \le \frac{\sqrt{2\pi}}{4} M_2(g).$$

Proof. Write $h(x) = U_o g(x)$ and $Z_{x,t} = \sqrt{t}x + \sqrt{1-t}Z$. Note that by the formula for $U_o g$,

$$\frac{\partial^r h}{\partial x_{i_1} \cdots \partial x_{i_r}}(x) = \int_0^1 (2t)^{-1} t^{r/2} \mathbb{E}\left[\frac{\partial^r g}{\partial x_{i_1} \cdots \partial x_{i_r}}(Z_{x,t})\right] dt. \tag{2.2}$$

It follows by integration by parts on the Gaussian expectation that

$$\begin{aligned} \frac{\partial^2 h}{\partial x_i \partial x_j}(x) &= \int_0^1 \frac{1}{2}\mathbb{E}\left[\frac{\partial^2 g}{\partial x_i \partial x_j}(\sqrt{t}x + \sqrt{1-t}Z)\right] dt \\ &= \int_0^1 \frac{1}{2\sqrt{1-t}}\mathbb{E}\left[Z_i \frac{\partial g}{\partial x_j}(Z_{x,t})\right] dt, \end{aligned} \tag{2.3}$$

and so

$$\operatorname{Hess} h(x) = \int_0^1 \frac{1}{2\sqrt{1-t}}\mathbb{E}\left[Z\left(\nabla g(Z_{x,t})\right)^T\right] dt. \tag{2.4}$$

Fix a $k \times k$ matrix A. Then

$$\langle \operatorname{Hess} h(x), A\rangle_{H.S.} = \int_0^1 \frac{1}{2\sqrt{1-t}}\mathbb{E}\left[\left\langle A^T Z, \nabla g(Z_{x,t})\right\rangle\right] dt,$$

thus

$$|\langle \operatorname{Hess} h(x), A\rangle_{H.S.}| \le M_1(g)\mathbb{E}|A^T Z| \int_0^1 \frac{1}{2\sqrt{1-t}} dt = M_1(g)\mathbb{E}|A^T Z|.$$

If $A = [a_{ij}]_{i,j=1}^k$, then

$$\mathbb{E}|AZ| \le \sqrt{\mathbb{E}|AZ|^2} = \sqrt{\mathbb{E}\sum_{i=1}^k\left(\sum_{j=1}^k a_{ji}Z_j\right)^2} = \sqrt{\sum_{i,j=1}^k a_{ij}^2} = \|A\|_{H.S.},$$

and thus

$$\|\operatorname{Hess} h(x)\|_{H.S.} \le M_1(g)$$

for all $x \in \mathbb{R}^k$, hence part (i).

For part (ii), let u and v be fixed vectors in $\mathbb{R}^k$ with $|u| = |v| = 1$. Then it follows from (2.4) that

$$\langle (\operatorname{Hess} h(x) - \operatorname{Hess} h(y))\, u, v\rangle = \int_0^1 \frac{1}{2\sqrt{1-t}}\mathbb{E}\left[\langle Z, v\rangle \langle \nabla g(Z_{x,t}) - \nabla g(Z_{y,t}), u\rangle\right] dt,$$

and so

$$\begin{aligned} |\langle (\operatorname{Hess} h(x) - \operatorname{Hess} h(y))u, v\rangle| &\le |x-y|\, M_2(g)\, \mathbb{E}|\langle Z, v\rangle| \int_0^1 \frac{\sqrt{t}}{2\sqrt{1-t}} dt \\ &= |x-y|\, M_2(g)\frac{\sqrt{2\pi}}{4}, \end{aligned}$$

since $\langle Z, v\rangle$ is just a standard Gaussian random variable. This completes the proof of part (ii). □

There is an important difference in the behavior of solutions to the Stein equation (iii) in the context of multivariate approximation versus univariate approximation. In the univariate case, one can replace the expression on the left-hand side of (iii) with the first-order expression $h'(x) - xh(x)$; the function $g(x) = U_o h(x)$ which solves the differential equation

$$h'(x) - xh(x) = g(x) - \mathbb{E}g(Z)$$

satisfies the bounds (see Stein, 1986)

$$\|g\|_\infty \leq \sqrt{\frac{\pi}{2}}\|h - \mathbb{E}h(Z)\|_\infty \qquad M_1(g) \leq 2\|h - \mathbb{E}h(Z)\|_\infty \qquad M_2(g) \leq 2M_1(h),$$

and the fact that the differential equation is first order rather than second then allows for reducing the degree of smoothness needed by one, over what is required in the multivariate case. Alternatively, one can use the same expression as in (iii) above; in this case, $M_3(g) \leq 2M_1(g)$ (see Raič, 2004), also decreasing by one the degree of smoothenss needed. This improvement allowed the univariate version proved in Meckes (2008) of Theorem 4.5 below, on the approximation of projections of Haar measure on the orthogonal group by Gaussian measure, to be proved in total variation distance as opposed to Wasserstein distance.

This improvement is not possible in the multivariate case; it can be shown, for example (see Raič, 2004), that if

$$f(x, y) = \max\{\min\{x, y\}, 0\},$$

then $U_o f$ defined as in Lemma 2.1 is twice differentiable but $\frac{\partial^2(U_o f)}{\partial x^2}$ is not Lipschitz.

Theorem 2.3. *Let X and X' be two random vectors in $\mathbb{R}^k$ such that $\mathcal{L}(X) = \mathcal{L}(X')$, and let $Z = (Z_1, \ldots, Z_k) \in \mathbb{R}^k$ be a standard Gaussian random vector. Suppose there is a constant λ such that*

$$\frac{1}{\lambda}\mathbb{E}\left[X' - X\middle|X\right] = -X. \tag{2.5}$$

Define the random matrix E by

$$\frac{1}{2\lambda}\mathbb{E}\left[(X' - X)(X' - X)^T\middle|X\right] = \sigma^2 I_k + \mathbb{E}\left[E\middle|X\right]. \tag{2.6}$$

Then if $g \in C^2(\mathbb{R}^k)$ with $M_1(g) < \infty$ and $M_2(g) < \infty$,

$$\left|\mathbb{E}g(X) - \mathbb{E}g(\sigma Z)\right| \leq \frac{1}{\sigma}M_1(g)\mathbb{E}\|E\|_{H.S.} + \left(\frac{\sqrt{2\pi}}{24\sigma}\right)\frac{M_2(g)}{\lambda}\mathbb{E}|X' - X|^3. \tag{2.7}$$

Proof. Fix g, and let $U_o g$ be as in Lemma 2.1. Note that it suffices to assume that $g \in C^\infty(\mathbb{R}^k)$: let $h : \mathbb{R}^k \to \mathbb{R}$ be a centered Gaussian density with covariance matrix $\epsilon^2 I_k$. Approximate g by $g * h$; clearly $\|g * h - g\|_\infty \to 0$ as $\epsilon \to 0$, and by Young's inequality, $M_1(g * h) \leq M_1(g)$ and $M_2(g * h) \leq M_2(g)$.

Note also that if $f(x) = g(\sigma x)$, then $\left|\mathbb{E}g(X) - \mathbb{E}g(\sigma Z)\right| = \left|\mathbb{E}f(\sigma^{-1}X) - \mathbb{E}f(Z)\right|$. It is easy to see that $M_1(f) = \sigma M_1(g)$ and $M_2(f) = \sigma^2 M_2(g)$. It thus follows from the theorem for $\sigma = 1$ that

$$\begin{aligned}\left|\mathbb{E}g(X) - \mathbb{E}g(\sigma Z)\right| &\leq \sigma M_1(g)\mathbb{E}\left\|\sigma^{-2}E\right\|_{H.S.} + \left(\frac{\sqrt{2\pi}}{24}\right)\frac{\sigma^2 M_2(g)}{\lambda}\mathbb{E}\left|\sigma^{-3}(X' - X)\right|^3 \\ &= \frac{M_1(g)}{\sigma}\mathbb{E}\|E\|_{H.S.} + \frac{\sqrt{2\pi}\,M_2(g)}{24\sigma\lambda}\mathbb{E}|X' - X|^3;\end{aligned}$$

we therefore restrict our attention to the case $\sigma = 1$.

For notational convenience, write $h(x) = U_o g(x)$. Then

$$\begin{aligned}
0 &= \frac{1}{\lambda}\mathbb{E}\left[h(X') - h(X)\right] \\
&= \frac{1}{\lambda}\mathbb{E}\left[\langle X' - X, \nabla h(X)\rangle + \frac{1}{2}(X' - X)^T(\operatorname{Hess} h(X))(X' - X) + R\right] \\
&= \mathbb{E}\left[-\langle X, \nabla h(X)\rangle + \Delta h(X) + \langle E, \operatorname{Hess} h(X)\rangle_{H.S.} + \frac{R}{\lambda}\right] \\
&= \mathbb{E}g(X) - \mathbb{E}g(Z) + \mathbb{E}\left[\langle E, \operatorname{Hess} h(X)\rangle_{H.S.} + \frac{R}{\lambda}\right], \qquad (2.8)
\end{aligned}$$

where R is the error in the second-order expansion. By an alternate form of Taylor's theorem (see Yomdin, 1983),

$$\mathbb{E}|R| \leq \frac{M_3(h)}{6}\mathbb{E}|X' - X|^3 \leq \frac{\sqrt{2\pi}M_2(g)}{24}\mathbb{E}|X' - X|^3.$$

Furthermore,

$$\mathbb{E}\left|\langle E, \operatorname{Hess} h(X)\rangle\right| \leq \left(\sup_{y\in\mathbb{R}^k}\|\operatorname{Hess} h(y)\|_{H.S.}\right)\mathbb{E}\|E\|_{H.S.} \leq M_1(g)\mathbb{E}\|E\|_{H.S.}.$$

This completes the proof.

□

Remarks.

(i) Usually the X and X' of the theorem will make an exchangeable pair, but this is not required for the proof.

(ii) The coupling assumed in (2.5) implies that $\mathbb{E}X = 0$. It is not required that X have a scalar covariance matrix, however, it follows from (2.5) and (2.6) that

$$\mathbb{E}\big[E\big] = \mathbb{E}\big[XX^T\big] - \sigma^2 I_k.$$

It should therefore be the case that the covariance matrix of X is not too far from $\sigma^2 I_k$.

The following is a continuous analog of Theorem 2.3. A univariate version which gives approximation in total variation distance was proved in Meckes (2008). As was noted following the proof of Lemma 2.2, a bound on total variation distance in the multivariate context is not possible with the method used here because of the difference in the behavior of solutions to the Stein equation in the multivariate context.

Theorem 2.4. *Let X be a random vector in $\mathbb{R}^k$ and for each $\epsilon > 0$ let X_ϵ be a random vector such that $\mathcal{L}(X) = \mathcal{L}(X_\epsilon)$, with the property that $\lim_{\epsilon\to 0} X_\epsilon = X$ almost surely. Let Z be a normal random vector in $\mathbb{R}^k$ with mean zero and covariance matrix $\sigma^2 I_k$. Suppose there is a function $\lambda(\epsilon)$ and a random matrix F such that the following conditions hold.*

(i)

$$\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[(X_\epsilon - X)_i \middle| X\right] \xrightarrow[\epsilon\to 0]{L_1} -X.$$

(ii)

$$\frac{1}{2\lambda(\epsilon)}\mathbb{E}\left[(X_\epsilon - X)(X_\epsilon - X)^T \middle| X\right] \xrightarrow[\epsilon\to 0]{L_1} \sigma^2 I_k + \mathbb{E}\left[F\middle|X\right].$$

(iii) *For each* $\rho > 0$,

$$\lim_{\epsilon\to 0}\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[\left|X_\epsilon - X\right|^2 \mathbb{I}(|X_\epsilon - X|^2 > \rho)\right] = 0.$$

Then

$$d_W(X,Z) \le \frac{1}{\sigma}\mathbb{E}\|F\|_{H.S.} \tag{2.9}$$

Proof. Fix a test function g; as in the proof of Theorem 2.3, it suffices to assume that $g \in C^\infty(\mathbb{R}^k)$ and to consider only the case $\sigma = 1$; the general result follows exactly as before.

Let $U_o g$ be as in Lemma 2.1, and as before, write $h(x) = U_o g(x)$. Observe

$$\begin{aligned} 0 &= \frac{1}{\lambda(\epsilon)}\mathbb{E}\left[h(X_\epsilon) - h(X)\right] \\ &= \frac{1}{\lambda(\epsilon)}\mathbb{E}\left[\langle X_\epsilon - X, \nabla h(X)\rangle + \frac{1}{2}(X_\epsilon - X)^T(\operatorname{Hess} h(X))(X_\epsilon - X) + R\right], \end{aligned} \tag{2.10}$$

where R is the error in the second-order approximation of $h(X_\epsilon) - h(X)$. By Taylor's theorem, there is a constant K (depending on h) and a function δ with $\delta(x) \le K\min\{x^2, x^3\}$, such that $\left|R\right| \le \delta(|X' - X|)$. Fix $\rho > 0$. Then by breaking up the integrand over the sets $\left\{|X_\epsilon - X| \le \rho\right\}$ and $\left\{|X_\epsilon - X| > \rho\right\}$,

$$\begin{aligned} \frac{1}{\lambda(\epsilon)}\mathbb{E}\left|R\right| &\le \frac{K}{\lambda(\epsilon)}\mathbb{E}\Big[|X_\epsilon - X|^3\mathbb{I}(|X_\epsilon - X| \le \rho) + |X_\epsilon - X|^2\mathbb{I}(|X_\epsilon - X| > \rho)\Big] \\ &\le \frac{K\rho\mathbb{E}\left|X_\epsilon - X\right|^2}{\lambda(\epsilon)} + \frac{K}{\lambda(\epsilon)}\mathbb{E}\Big[|X_\epsilon - X|^2\mathbb{I}(|X' - X| > \rho)\Big]. \end{aligned}$$

The second term tends to zero as $\epsilon \to 0$ by condition (iii); condition (ii) implies that the first is bounded by $CK\rho$ for a constant C depending on k and on the distribution of X. It follows that

$$\lim_{\epsilon\to 0}\frac{1}{\lambda(\epsilon)}\mathbb{E}\left|R\right| = 0.$$

For the first two terms of (2.10),

$$\begin{aligned} &\lim_{\epsilon\to 0}\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[\langle X_\epsilon - X, \nabla h(X)\rangle + \frac{1}{2}(X_\epsilon - X)^T(\operatorname{Hess} h(X))(X_\epsilon - X)\right] \\ &\quad= \lim_{\epsilon\to 0}\frac{1}{\lambda(\epsilon)}\mathbb{E}\Bigg[\left\langle \mathbb{E}\left[(X_\epsilon - X)\middle|X\right], \nabla h(X)\right\rangle \\ &\qquad\qquad + \frac{1}{2}\left\langle \mathbb{E}\left[\left(X_\epsilon - X\right)\left(X_\epsilon - X\right)^T\middle|X\right], \operatorname{Hess} h(X)\right\rangle_{H.S.}\Bigg] \\ &\quad= \mathbb{E}\left[-\langle X, \nabla h(X)\rangle + \Delta h(X) + \left\langle \mathbb{E}\left[F\middle|X\right], \operatorname{Hess} h(X)\right\rangle_{H.S.}\right] \\ &\quad= \mathbb{E}g(X) - \mathbb{E}g(Z) + \mathbb{E}\left[\left\langle \mathbb{E}\left[F\middle|X\right], \operatorname{Hess} h(X)\right\rangle_{H.S.}\right], \end{aligned} \tag{2.11}$$

where conditions (i) and (ii) together with the boundedness of ∇h and $\operatorname{Hess} h$ are used to get the third line and the definition of $h = U_o g$ is used to get the fourth line. We have thus shown that

$$\mathbb{E}\big[g(X) - g(Z)\big] = -\mathbb{E}\left\langle F, \operatorname{Hess} h(X)\right\rangle_{H.S.}. \tag{2.12}$$

The result now follows immediately by applying the Cauchy-Schwarz inequality to (2.12) and then the bound $\|\operatorname{Hess} h(x)\|_{H.S.} \le M_1(g)$ from Lemma 2.2 (i). □

Remarks.

(i) It is easy to see that if

(iii′) $\lim_{\epsilon\to 0} \frac{1}{\lambda(\epsilon)}\mathbb{E}\big|X_\epsilon - X\big|^3 = 0,$

then condition (iii) of the theorem holds. This is what is done in the applications below.

(ii) As in Theorem 2.3, the condition (i) implies that $\mathbb{E}X = 0$ and it follows from (i) and (ii) that

$$\mathbb{E}F = \mathbb{E}XX^T - \sigma^2 I;$$

the covariance matrix of X should thus not be far from $\sigma^2 I$.

Theorem 2.4 has the following corollary for complex random vectors.

Theorem 2.5. *Let W be a random vector in $\mathbb{C}^k$ and for each $\epsilon > 0$ let W_ϵ be a random vector such that $\mathcal{L}(W) = \mathcal{L}(W_\epsilon)$, with the property that $\lim_{\epsilon\to 0} W_\epsilon = W$ almost surely. Let $Z = (Z_1, \dots, Z_k)$ be a standard complex Gaussian random vector; i.e., with covariance matrix of the corresponding random vector in $\mathbb{R}^{2k}$ given by $\frac{1}{2}I_{2k}$. Suppose there is a function $\lambda(\epsilon)$ and complex $k\times k$ random matrices $\Gamma = \big[\gamma_{ij}\big]$ and $\Lambda = \big[\lambda_{ij}\big]$ such that*

(i) $\dfrac{1}{\lambda(\epsilon)}\mathbb{E}\left[(W_\epsilon - W)\middle|W\right] \xrightarrow[\epsilon\to 0]{L_1} -W.$

(ii) $\dfrac{1}{2\lambda(\epsilon)}\mathbb{E}\left[(W_\epsilon - W)(W_\epsilon - W)^*|W\right] \xrightarrow[\epsilon\to 0]{L_1} I_k + \mathbb{E}\left[\Gamma\middle|W\right].$

(iii) $\dfrac{1}{2\lambda(\epsilon)}\mathbb{E}\left[(W_\epsilon - W)(W_\epsilon - W)^T|W\right] \xrightarrow[\epsilon\to 0]{L_1} \mathbb{E}\left[\Lambda\middle|W\right].$

(iv) *For each $\rho > 0$,* $\lim\limits_{\epsilon\to 0}\dfrac{1}{\lambda(\epsilon)}\mathbb{E}\left[\big|W_\epsilon - W\big|^2\mathbb{I}(|W_\epsilon - W|^2 > \rho)\right] = 0.$

Then

$$d_W(W, Z) \le \mathbb{E}\|\Gamma\|_{H.S.} + \mathbb{E}\|\Lambda\|_{H.S.}.$$

Proof. Identifying $\mathbb{C}^k$ with $\mathbb{R}^{2k}$, W satisfies the conditions of Theorem 2.4 with $\sigma^2 = \frac{1}{2}$ and F given as a $k\times k$ matrix of 2×2 blocks, with the i-jth block equal to

$$\frac{1}{2}\begin{bmatrix}\operatorname{Re}(\gamma_{ij} + \lambda_{ij}) & \operatorname{Im}(\lambda_{ij} - \gamma_{ij})\\ \operatorname{Im}(\lambda_{ij} + \gamma_{ij}) & \operatorname{Re}(\gamma_{ij} - \lambda_{ij})\end{bmatrix}.$$

Thus $\|F\|_{H.S.}^2 = \frac{1}{2}(\|\Gamma\|_{H.S.}^2 + \|\Lambda\|_{H.S.}^2)$ and

$$\mathbb{E}\|F\|_{H.S.} \le \frac{1}{\sqrt{2}}\Big[\mathbb{E}\|\Gamma\|_{H.S.} + \mathbb{E}\|\Lambda\|_{H.S.}\Big].$$

□

3. Examples using Theorem 2.3

3.1. *A basic central limit theorem.* As a simple illustration of the use of Theorem 2.3, we derive error bounds in the classical multivariate CLT for sums of independent random vectors. While the question of error bounds in the univariate CLT was settled long ago, the optimal bounds in the multivariate case are still unknown and much work has been done in this direction. One important contribution was made by Götze (1991), who used Stein's method in conjunction with induction. To the best of our knowledge, the most recent results are due to V. Bentkus (2003), where one can also find extensive pointers to the literature.

Suppose Y is a random vector in $\mathbb{R}^k$ with mean zero and identity covariance. Let W be the normalized sum of n i.i.d. copies of Y. Götze (1991) and Bentkus (2003) both give bounds on quantities like $\Delta_n = \sup_{f\in\mathcal{A}} |\mathbb{E}f(W) - \mathbb{E}f(Z)|$, where $Z = (Z_1, \dots, Z_k)$ is a standard k-dimensional normal random vector and $\mathcal{A}$ is any collection of functions satisfying certain properties. For example, when $\mathcal{A}$ is the class of indicator functions of convex sets, Bentkus gets $\Delta_n \le 400k^{1/4}n^{-1/2}\mathbb{E}|Y|^3$, improving on Götze's earlier bound which has a coefficient of $k^{1/2}$ rather than $k^{1/4}$. Note that $\mathbb{E}|Y|^3 = O(k^{3/2})$.

Theorem 2.3 allows us to easily obtain uniform bounds on $|\mathbb{E}g(S_n) - \mathbb{E}g(Z)|$ for large classes of smooth functions.

Theorem 3.1. *Let $\{Y_i\}_{i=1}^n$ be a set of independent, identically distributed random vectors in $\mathbb{R}^k$. Assume that the Y_i are such that*

$$\mathbb{E}(Y_i) = 0, \qquad \mathbb{E}(Y_iY_i^T) = I_k.$$

Let $W = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$. Then for any $g \in C_o^2$,

$$\big|\mathbb{E}g(W) - \mathbb{E}g(Z)\big| \le \frac{M_1(g)}{2\sqrt{n}}\sqrt{\mathbb{E}|Y_1|^4 - k} + \frac{\sqrt{2\pi}}{3\sqrt{n}}M_2(g)\mathbb{E}|Y_1|^3.$$

Proof. To apply Theorem 2.3, make an exchangeable pair (W, W') as follows. For each i, let X_i be an independent copy of Y_i, and let I be a uniform random variable in $\{1, \dots, n\}$, independent of everything. Define W' by

$$W' = W - \frac{Y_I}{\sqrt{n}} + \frac{X_I}{\sqrt{n}}.$$

Then

$$\begin{aligned}\mathbb{E}\left[W' - W\middle|W\right] &= \frac{1}{\sqrt{n}}\mathbb{E}\left[X_I - Y_I\middle|W\right]\\ &= \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[X_i - Y_i\middle|W\right] = -\frac{1}{n}W,\end{aligned}$$

where the independence of X_i and W has been used in the last line. Thus condition 2.5 of Theorem 2.3 holds with $\lambda = \frac{1}{n}$.

It remains to check condition 2 and bound the E_{ij}. Write $Y_i = (Y_i^1, \dots, Y_i^k)$. For $1 \le j, \ell \le k$,

$$\begin{aligned} E_{j\ell} &= \frac{n}{2}\mathbb{E}\left[(W_j' - W_j)(W_\ell' - W_\ell)\big|W\right] - \delta_{j\ell} \\ &= \frac{1}{2}\mathbb{E}\left[(X_I^j - Y_I^j)(X_I^\ell - Y_I^\ell)\big|W\right] - \delta_{j\ell} \\ &= \frac{1}{2n}\sum_{i=1}^n \mathbb{E}\left[X_i^j X_i^\ell - X_i^j Y_i^\ell - X_i^\ell Y_i^j + Y_i^j Y_i^\ell\big|W\right] - \delta_{j\ell} \\ &= \frac{1}{2n}\sum_{i=1}^n \mathbb{E}\left[Y_i^j Y_i^\ell - \delta_{j\ell}\big|W\right], \end{aligned}$$

by the independence of the X_i and the Y_i. Thus

$$\begin{aligned} \mathbb{E}E_{j\ell}^2 &= \frac{1}{4n^2}\mathbb{E}\left(\mathbb{E}\left[\sum_{i=1}^n (Y_i^j Y_i^\ell - \delta_{j\ell})\bigg|W\right]\right)^2 \\ &\le \frac{1}{4n^2}\mathbb{E}\left[\left(\sum_{i=1}^n (Y_i^j Y_i^\ell - \delta_{j\ell})\right)^2\right] \\ &= \frac{1}{4n^2}\mathbb{E}\sum_{i=1}^n (Y_i^j Y_i^\ell - \delta_{j\ell})^2 \\ &= \frac{1}{4n}\mathbb{E}\left[Y_1^j Y_1^\ell - \delta_{j\ell}\right]^2 \\ &= \frac{1}{4n}\left[\mathbb{E}\left(Y_1^j Y_1^\ell\right)^2 - \delta_{j\ell}\right], \end{aligned}$$

where the independence of the Y_i has been used to get the third line. It follows that

$$\mathbb{E}\|E\|_{H.S.} \le \sqrt{\mathbb{E}\|E\|_{H.S.}^2} \le \frac{1}{2\sqrt{n}}\sqrt{\sum_{j,\ell}\left(\mathbb{E}(Y_1^j Y_1^\ell)^2 - \delta_{j\ell}\right)} = \frac{1}{2\sqrt{n}}\sqrt{\mathbb{E}|Y_1|^4 - k}.$$

It remains to bound the second term of Theorem 2.3.

$$\begin{aligned} \frac{1}{\lambda}\mathbb{E}|W' - W|^3 &= \frac{1}{\sqrt{n}}\mathbb{E}\big|X_I - Y_I\big|^3 \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\big|X_1 - Y_1\big|^3 \\ &\le \frac{1}{\sqrt{n}}\mathbb{E}\left(|X_1|^3 + 3|X_1|^2|Y_1| + 3|Y_1|^2|X_1| + |Y_1|^3\right). \end{aligned}$$

Applying Hölder's inequality with $p = \frac{3}{2}$ and $q = 3$,

$$\mathbb{E}|X_1|^2|Y_1| \le \left(\mathbb{E}|X_1|^3\right)^{2/3}\left(\mathbb{E}|Y_1|^3\right)^{1/3} = \mathbb{E}|Y_1|^3.$$

It follows that

$$\frac{1}{\lambda}\mathbb{E}|W' - W|^3 \le \frac{8\mathbb{E}|Y_1|^3}{\sqrt{n}}.$$

Together with Theorem 2.3, this finishes the proof. □

4. Examples using Theorem 2.4

4.1. *Rank k projections of spherically symmetric measures on* $\mathbb{R}^n$. Consider a random vector $Y \in \mathbb{R}^n$ whose distribution is spherically symmetric; i.e., if U is a fixed orthogonal matrix, then the distribution of Y is the same as the distribution of UY. Assume that Y is normalized such that $\mathbb{E}Y_1^2 = 1$. Note that the spherical symmetry then implies that $\mathbb{E}YY^T = I_n$. Assume further that there is a constant a (independent of n) so that

$$\mathrm{Var}(|Y|^2) \leq a. \tag{4.1}$$

For k fixed, let P_k denote the orthogonal projection of $\mathbb{R}^n$ onto the span of the first k standard basis vectors. In this section, Theorem 2.4 is applied to show that $P_k(Y) = (Y_, \ldots, Y_k)$ is approximately distributed as a standard k-dimensional Gaussian random vector if $k = o(n)$. That $\mathbb{E}P_k(Y)P_k(Y)^T = I_k$ is immediate from the symmetry and normalization, as above.

This example is closely related to the following result from Diaconis and Freedman (1987).

Theorem 4.1 (Diaconis-Freedman). *Let* $Z_1,\ldots,Z_n$ *be independent standard Gaussian random variables and let* $\mathbb{P}_\sigma^k$ *be the law of* $(\sigma Z_1, \ldots, \sigma Z_k)$. *For a probability* μ *on* $[0,\infty)$, *define* $\mathbb{P}_{\mu,k}$ *by*

$$\mathbb{P}_{\mu,k} = \int \mathbb{P}_\sigma^k d\mu(\sigma).$$

Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ *be a spherically symmetric random vector, and let* $\mathbb{P}_k$ *be the law of* $(Y_1, \ldots, Y_k)$. *Then there is a probability measure* μ *on* $[0,\infty)$ *such that for* $1 \leq k \leq n-4$,

$$d_{TV}(\mathbb{P}_k, \mathbb{P}_{\mu,k}) \leq \frac{2(k+3)}{n-k-3}.$$

Furthermore, the mixing measure μ *can be taken to be the law of* $\frac{1}{\sqrt{n}}|Y|$.

In some cases, the explicit form given in Theorem 4.1 for the mixing measure has allowed the theorem to be used to prove central limit theorems of interest in convex geometry; see Brehm and Voigt (2000) and Klartag (2007). Theorem 4.3 below says that the variance bound (4.1) is sufficient to show that the mixing measure of Theorem 4.1 can be taken to be a point mass. In fact, it is not too difficult to obtain the total variation analog of Theorem 4.3 directly from the Diaconis-Freedman result and (4.1); however, the Stein's method proof given below is considerably simpler than the direct proof given in Diaconis and Freedman (1987). The rates obtained are of the same order, though the rate obtained by Diaconis and Freedman is in the total variation distance, whereas the rate below is in the Wasserstein distance.

To apply Theorem 2.4, construct a family of exchangeable pairs as follows. For $\epsilon > 0$ fixed, let

$$\begin{aligned} A_\epsilon &= \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2} \\ &= I_n + \begin{bmatrix} -\frac{\epsilon^2}{2} + \delta & \epsilon \\ -\epsilon & -\frac{\epsilon^2}{2} + \delta \end{bmatrix} \oplus 0_{n-2}, \end{aligned}$$

where δ is a deterministic constant and $\delta = O(\epsilon^4)$. Let U be a Haar-distributed $n \times n$ random orthogonal matrix, independent of Y, and let $Y_\epsilon = \left(UA_\epsilon U^T\right) Y$. Thus Y_ϵ is a small random rotation of Y. In what follows, Theorem 2.4 is applied to the exchangeable pair $(P_k(Y), P_k(Y_\epsilon))$.

Let K be the $k \times 2$ matrix made of the first two columns of U and $C_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. Define $Q := KC_2K^T$. Then by the construction of Y_ϵ,

$$P_k(Y_\epsilon) - P_k(Y) = \epsilon \left[-\left(\frac{\epsilon}{2} + \epsilon^{-1}\delta \right) P_k KK^T + P_k Q \right] Y, \tag{4.2}$$

and $\epsilon^{-1}\delta = O(\epsilon^3)$.

To check the conditions of Theorem 2.4, the following lemma is needed; see Meckes (2006), Lemma 3.3 and Theorem 1.6 for a detailed proof.

Lemma 4.2. *If $U = [u_{ij}]_{i,j=1}^n$ is an orthogonal matrix distributed according to Haar measure, then $\mathbb{E}\left[\prod u_{ij}^{k_{ij}}\right]$ is non-zero if and only if the number of entries from each row and from each column is even. Second and fourth-degree moments are as follows:*

(i) *For all i, j,*

$$\mathbb{E}\left[u_{ij}^2\right] = \frac{1}{n}.$$

(ii) *For all $i, j, r, s, \alpha, \beta, \lambda, \mu$,*

$$\begin{aligned} &\mathbb{E}\big[u_{ij}u_{rs}u_{\alpha\beta}u_{\lambda\mu}\big] \\ &\quad = -\frac{1}{(n-1)n(n+2)}\Big[\delta_{ir}\delta_{\alpha\lambda}\delta_{j\beta}\delta_{s\mu} + \delta_{ir}\delta_{\alpha\lambda}\delta_{j\mu}\delta_{s\beta} + \delta_{i\alpha}\delta_{r\lambda}\delta_{js}\delta_{\beta\mu} \\ &\qquad\qquad + \delta_{i\alpha}\delta_{r\lambda}\delta_{j\mu}\delta_{\beta s} + \delta_{i\lambda}\delta_{r\alpha}\delta_{js}\delta_{\beta\mu} + \delta_{i\lambda}\delta_{r\alpha}\delta_{j\beta}\delta_{s\mu}\Big] \\ &\qquad + \frac{n+1}{(n-1)n(n+2)}\Big[\delta_{ir}\delta_{\alpha\lambda}\delta_{js}\delta_{\beta\mu} + \delta_{i\alpha}\delta_{r\lambda}\delta_{j\beta}\delta_{s\mu} + \delta_{i\lambda}\delta_{r\alpha}\delta_{j\mu}\delta_{s\beta}\Big]. \end{aligned}$$

(iii) *For the matrix $Q = \left[q_{ij}\right]_{i,j=1}^n$ defined as above, $q_{ij} = u_{i1}u_{j2} - u_{i2}u_{j1}$. For all i, j, ℓ, p,*

$$\mathbb{E}\left[q_{ij}q_{\ell p}\right] = \frac{2}{n(n-1)}\left[\delta_{i\ell}\delta_{jp} - \delta_{ip}\delta_{j\ell}\right].$$

By the lemma, $\mathbb{E}\big[KK^T\big] = \frac{2}{n}I$ and $\mathbb{E}\big[Q\big] = 0$, and so

$$\lim_{\epsilon \to 0} \frac{n}{\epsilon^2}\mathbb{E}\Big[(P_k(Y_\epsilon) - P_k(Y))\Big| P_k(Y)\Big] = -P_k(Y);$$

condition (i) of Theorem 2.4 thus holds with $\lambda(\epsilon) = \frac{\epsilon^2}{n}$.

Fix $i, j \le k$. By (4.2),

$$\lim_{\epsilon \to 0} \frac{n}{2\epsilon^2} \mathbb{E}\Big[(P_k(Y_\epsilon) - P_k(Y))_i (P_k(Y_\epsilon) - P_k(Y))_j \Big| Y\Big]$$
$$= \frac{n}{2}\mathbb{E}\left[(P_kQY)_i(P_kQY)_j \middle| Y\right]$$
$$= \frac{n}{2}\mathbb{E}\left[\sum_{\ell,m} Y_\ell Y_m q_{i\ell} q_{jm} \middle| Y\right]$$
$$= \frac{1}{(n-1)}\mathbb{E}\left[\sum_{\ell,m} Y_\ell Y_m \left(\delta_{ij}\delta_{\ell m} - \delta_{im}\delta_{\ell j}\right)\right]$$
$$= \frac{1}{(n-1)}\left[\delta_{ij}|Y|^2 - Y_iY_j\right].$$

Thus

$$F = \frac{1}{(n-1)}\left[\left(\mathbb{E}\left[|Y|^2 - (n-1)\big|P_k(Y)\right]\right) \cdot I_k - P_k(Y)P_k(Y)^T\right].$$

Now,

$$\mathbb{E}\|P_k(Y)P_k(Y)^T\|_{H.S.} = \mathbb{E}\,|P_k(Y)|_2^2 = k$$

by assumption, and

$$\mathbb{E}\left|\mathbb{E}\left[|Y|^2 - (n-1)\big|P_k(Y)\right]\right| \le \sqrt{a} + 1,$$

so applying Theorem 2.4 gives:

Theorem 4.3. *With notation as above,*

$$d_W(P_k(Y), Z) \le \frac{k(\sqrt{a}+2)}{n-1}.$$

4.2. *Rank k projections of Haar measure on $\mathcal{O}_n$.*

A theme in studying random matrices from the compact classical matrix groups is that these matrices are in many ways (though not all ways) similar to Gaussian random matrices. For example, it was shown in D'Aristotile et al. (2003) that if M is a random matrix in the orthogonal group $\mathcal{O}_n$ distributed according to Haar measure, then

$$\sup_{\substack{A:\,\mathrm{Tr}\,(AA^T)=n \\ -\infty<x<\infty}} \left|\mathbb{P}(\mathrm{Tr}\,(AM) \le x) - \Phi(x)\right| \to 0$$

as $n \to \infty$. In Meckes (2008), this result was refined to include a rate of convergence (in total variation) of $W = \mathrm{Tr}\,(AM)$ to a standard Gaussian random variable, depending only on the value of $\mathrm{Tr}\,(AA^T)$. That is, rank one projections of Haar measure on $\mathcal{O}_n$ are uniformly close to Gaussian, and rank one projections of Gaussian random matrices are exactly Gaussian.

A natural question is whether rank k projections of Haar measure on $\mathcal{O}_n$ are close, in some sense, to multivariate Gaussian distributions, and if so, how large k can be. This is a more refined comparison of the type mentioned above, since the distributions of all projections of any rank of Gaussian matrices are Gaussian. In remarkable recent work, Jiang (2006) has shown that the entries of any $p_n \times q_n$ submatrix of an $n \times n$ random orthogonal matrix are close to i.i.d. Gaussians in total variation distance whenever $p_n = o(\sqrt{n})$ and $q_n = o(\sqrt{n})$, and that these orders of p_n and q_n are best possible. This improved an earlier result of Diaconis et al. (1992), which proved the result in the case of $p_n = o(n^{1/3})$ and $q_n = o(n^{1/3})$. As this article was in preparation, Collins and Stolz (2008) proved that for r fixed, $A_1, \ldots, A_r$ deterministic parameter matrices, and M a uniformly distributed element of a classical compact symmetric space (represented as a space of matrices), the random vector $(\mathrm{Tr}\,(A_1 M), \ldots, \mathrm{Tr}\,(A_r M))$ converges weakly to a Gaussian random vector, as the dimension of the space tends to infinity. Their work in particular covers the cases of M a Haar-distributed random orthogonal or unitary matrix, but goes farther to consider more general homogeneous spaces.

In this section, it is shown that rank k projections of Haar measure on $\mathcal{O}_n$ are close in Wasserstein distance to Gaussian for $k = o(n)$. This in particular recovers Jiang's result (in Wasserstein distance), but is more general in that it is uniform over all rank k projections, and not just those having the special form of truncation to a sub-matrix. The theorem also strengthens the result of Collins and Stolz, in the case that M is a random element of $\mathcal{O}_n$.

Theorem 4.4. *Let $B_1, \ldots, B_k$ be linearly independent $n \times n$ matrices (i.e. the only linear combination of them which is equal to the zero matrix has all coefficients equal to zero) over $\mathbb{R}$ such that $\mathrm{Tr}\,(B_i B_i^T) = n$ for each i. Let $b_{ij} = \mathrm{Tr}\,(B_i B_j^T)$. Let M be a random orthogonal matrix and let*

$$X = (\mathrm{Tr}\,(B_1 M), \mathrm{Tr}\,(B_2 M), \ldots, \mathrm{Tr}\,(B_k M)) \in \mathbb{R}^k.$$

Let $Y = (Y_1, \ldots, Y_k)$ be a random vector whose components have the standard Gaussian distribution, with covariance matrix $C := \frac{1}{n}(b_{ij})_{i,j=1}^k$. Then for $n \geq 2$,

$$d_W(X, Y) \leq \frac{k\sqrt{2\|C\|_{op}}}{n-1}.$$

Remark. Lemma 4.2 and an easy computation show that for all i, j,

$$\mathbb{E}\big[\mathrm{Tr}\,(B_i M)\mathrm{Tr}\,(B_j M)\big] = \frac{1}{n}\langle B_i, B_j \rangle,$$

thus the matrix C above is also the covariance matrix of X.

It is shown below that Theorem 4.4 follows fairly easily from the following special case.

Theorem 4.5. *Let $A_1, \ldots, A_k$ be $n \times n$ matrices over $\mathbb{R}$ satisfying $\mathrm{Tr}\,(A_i A_j^T) = n\delta_{ij}$; for $i \neq j$, A_i and A_j are orthogonal with respect to the Hilbert-Schmidt inner product. Let M be a random orthogonal matrix, and consider the vector $X = (\mathrm{Tr}\,(A_1 M), \mathrm{Tr}\,(A_2 M), \ldots, \mathrm{Tr}\,(A_k M)) \in \mathbb{R}^k$. Let $Z = (Z_1, \ldots, Z_k)$ be a random vector whose components are independent standard normal random variables. Then for $n \geq 2$,*

$$\big|\mathbb{E}f(X) - \mathbb{E}f(Z)\big| \leq \frac{\sqrt{2}M_1(f)k}{n-1}$$

where $M_1(f)$ is the Lipschitz constant of f.

Example. Let M be a random $n \times n$ orthogonal matrix, and let $0 < a_1 < a_2 < \ldots < a_k = n$. For each $1 \leq i \leq n$, let

$$B_i = \sqrt{\frac{n}{a_i}} I_{a_i} \oplus \mathbf{0}_{n-a_i};$$

B_i has $\sqrt{\frac{n}{a_i}}$ in the first a_i diagonal entries and zeros everywhere else. If $i \leq j$, then $\langle B_i, B_j \rangle_{HS} = n\sqrt{\frac{a_i}{a_j}}$; in particular, $\langle B_i, B_i \rangle_{HS} = n$. The B_i are linearly independent w.r.t. the Hilbert-Schmidt inner product since the a_i are all distinct, so to apply Theorem 4.4, we have only to bound the eigenvalues of the matrix $\left(\sqrt{\frac{a_{\min(i,j)}}{a_{\max(i,j)}}}\right)_{i,j=1}^k$. But this is easy, since $|\lambda| \leq \sqrt{\sum_{i,j=1}^k \frac{a_{\min(i,j)}}{a_{\max(i,j)}}} \leq k$ for all eigenvalues λ (see, e.g., Bhatia, 1997). It now follows from Theorem 4.4 that if Y is a vector of standard normals with covariance matrix $\left(\sqrt{\frac{a_{\min(i,j)}}{a_{\max(i,j)}}}\right)_{i,j=1}^k$ and $X = (\mathrm{Tr}\,(B_1 M), \ldots, \mathrm{Tr}\,(B_k M))$, then

$$\sup_{|f|_L \leq 1} \left|\mathbb{E}f(X) - \mathbb{E}f(Y)\right| \leq \frac{\sqrt{2}k^{3/2}}{n-1}.$$

Proofs

Proof of Theorem 4.4 from Theorem 4.5. Perform the Gram-Schmidt algorithm on the matrices $\{B_1, \ldots, B_k\}$ with respect to the Hilbert-Schmidt inner product $\langle C, D \rangle = \mathrm{Tr}\,(CD^T)$ to get matrices $\{A_1, \ldots, A_k\}$ which are mutually orthogonal and have H-S norm $\sqrt{n}$. Denote the matrix which takes the B's to the A's by D^{-1} for $D = \left[d_{ij}\right]_{i,j=1}^n$; the matrix is invertible since the B's are linearly independent. Now by assumption,

$$\begin{aligned} b_{ij} &= \langle B_i, B_j \rangle \\ &= \left\langle \sum_l d_{il} A_l, \sum_p d_{jp} A_p \right\rangle \\ &= n \sum_l d_{il} d_{jl}. \end{aligned}$$

Thus $DD^T = C = \frac{1}{n} \left(b_{ij}\right)_{i,j=1}^k$.

Now, let $f : \mathbb{R}^k \to \mathbb{R}$ with $M_1(f) \leq 1$. Define $h : \mathbb{R}^k \to \mathbb{R}$ by $h(x) = f(Dx)$. Then $M_1(h) \leq \|D\|_{op} \leq \sqrt{\|DD^T\|_{op}}$. By Theorem 4.5,

$$\left|\mathbb{E}h(\mathrm{Tr}\,(A_1 M), \ldots, \mathrm{Tr}\,(A_k M)) - \mathbb{E}h(Z)\right| \leq \frac{k\sqrt{2\|C\|_{op}}}{n-1}$$

for Z a standard Gaussian random vector in $\mathbb{R}^k$. But $D\big(\mathrm{Tr}\,(A_1 M), \ldots, \mathrm{Tr}\,(A_k M)\big) = \big(\mathrm{Tr}\,(B_1 M), \ldots, \mathrm{Tr}\,(B_k M)\big)$ and DZ has standard normal components with covariance matrix $C = \frac{1}{n} \left(b_{ij}\right)_{i,j=1}^k$. □

Proof of Theorem 4.5. Make an exchangeable pair (M, M_ϵ) as before; let A_ϵ be the rotation

$$A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{n-2} = I_n + \begin{bmatrix} \sqrt{1-\epsilon^2}-1 & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2}-1 \end{bmatrix} \oplus \mathbf{0}_{n-2},$$

let U be a Haar-distributed random orthogonal matrix, independent of M, and let

$$M_\epsilon = UAU^TM.$$

Let $X_\epsilon = (\mathrm{Tr}\,(A_1M_\epsilon), \ldots, \mathrm{Tr}\,(A_kM_\epsilon))$.

As in section 4.1, define K to be the first two columns of U and $C_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, and let $Q = KC_2K^T$. Then

$$M_\epsilon - M = \epsilon\left[\left(\frac{-\epsilon}{2} + O(\epsilon^3)\right)KK^T + Q\right]M. \tag{4.3}$$

It follows from Lemma 4.2 that $\mathbb{E}\big[KK^T\big] = \frac{2}{n}I$ and $\mathbb{E}\big[Q\big] = 0$, thus

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{n}{\epsilon^2}\mathbb{E}\big[(X_\epsilon - X)_i\big|M\big] \\
&\qquad = \lim_{\epsilon\to 0}\frac{n}{\epsilon^2}\mathbb{E}\left[\mathrm{Tr}\,[A_i(M_\epsilon - M)]\big|M\right] \\
&\qquad = \lim_{\epsilon\to 0}\frac{n}{\epsilon^2}\left[\left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right)\mathbb{E}\left[\mathrm{Tr}\,(A_iKK^TM)\big|M\right] + \epsilon\mathbb{E}\left[\mathrm{Tr}\,(A_iQM)\big|M\right]\right] \\
&\qquad = \lim_{\epsilon\to 0}\frac{n}{\epsilon^2}\left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right)\frac{2}{n}X_i \\
&\qquad = -X_i.
\end{aligned}$$

Condition (i) of Theorem 2.4 is thus satisfied with $\lambda(\epsilon) = \frac{\epsilon^2}{n}$. The random matrix F is computed as follows. For notational convenience, write $A_i = A = (a_{pq})$ and $A_j = B = (b_{\alpha\beta})$. By (4.3),

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{n}{2\epsilon^2}\mathbb{E}\Big[(X_\epsilon - X)_i(X_\epsilon - X)_j\Big|M\Big] \\
&\qquad = \frac{n}{2}\mathbb{E}\left[\mathrm{Tr}\,(AQM)\mathrm{Tr}\,(BQM)\big|M\right] \\
&\qquad = \frac{n}{2}\mathbb{E}\left[\sum_{p,q,r,\alpha,\beta,\gamma} a_{pq}b_{\alpha\beta}m_{rp}m_{\gamma\alpha}q_{qr}q_{\beta\gamma}\middle|M\right] \\
&\qquad = \frac{n}{2}\mathbb{E}\left[\sum_{p,q,r,\alpha,\beta\gamma} a_{pq}b_{\alpha\beta}m_{rp}m_{\gamma\alpha}\left(\frac{2}{n(n-1)}\right)(\delta_{q\beta}\delta_{r\gamma} - \delta_{q\gamma}\delta_{r\beta})\right] \qquad (4.4) \\
&\qquad = \frac{1}{(n-1)}\mathbb{E}\left[\langle MA, MB\rangle_{H.S.} - \mathrm{Tr}\,(AMBM)\right] \\
&\qquad = \frac{1}{(n-1)}\mathbb{E}\left[\langle A, B\rangle_{H.S.} - \mathrm{Tr}\,(MAMB)\right] \\
&\qquad = \frac{1}{(n-1)}\left[n\delta_{ij} - \mathrm{Tr}\,(MAMB)\right].
\end{aligned}$$

Thus

$$F = \frac{1}{(n-1)}\mathbb{E}\left[\left[\delta_{ij} - \mathrm{Tr}\,(A_iMA_jM)\right]_{i,j=1}^k\middle|X\right].$$

Claim: If $n \geq 2$, then $\mathbb{E}\left[\mathrm{Tr}\,(A_iMA_jM) - \delta_{ij}\right]^2 \leq 2$ for all i and j.

The claim gives that, for $n \geq 2$,

$$\mathbb{E}\|F\|_{H.S.} \leq \sqrt{\mathbb{E}\|F\|_{H.S.}^2} \leq \frac{\sqrt{2}k}{n-1},$$

thus completing the proof.

To prove the claim, first observe that Lemma 4.2 implies

$$\mathbb{E}\big[\mathrm{Tr}\,(A_iMA_jM)\big] = \frac{1}{n}\langle A_i, A_j\rangle = \delta_{ij}.$$

Again writing $A_i = A$ and $A_j = B$, applying Lemma 4.2 gives, (ii),

$$\begin{aligned}
&\mathbb{E}\Big[\mathrm{Tr}\,(AMBM)\Big]^2 \\
&\quad = \mathbb{E}\left[\sum_{\substack{p,q,r,s\\ \alpha,\beta,\mu,\lambda}} a_{sp}a_{\mu\alpha}b_{qr}b_{\beta\lambda}m_{pq}m_{rs}m_{\alpha\beta}m_{\lambda\mu}\right] \\
&\quad = -\frac{2}{(n-1)n(n+2)}\left[\mathrm{Tr}\,(A^TAB^TB) + \mathrm{Tr}\,(AB^TAB^T) + \mathrm{Tr}\,(AA^TBB^T)\right] \\
&\qquad + \frac{n+1}{(n-1)n(n+2)}\left[2\langle A, B\rangle_{H.S.} + \|A\|_{H.S.}^2\|B\|_{H.S.}^2\right]
\end{aligned}$$

Now, as the Hilbert-Schmidt norm is submultiplicative (see Bhatia, 1997, page 94),

$$\mathrm{Tr}\,(A^TAB^TB) \leq \|A^TA\|_{H.S.}\|B^TB\|_{H.S.} \leq \|A\|_{H.S.}^2\|B\|_{H.S.}^2 = n^2,$$

and the other two summands of the first line are bounded by n^2 in the same way. Also,

$$2\langle A, B\rangle_{H.S.} + \|A\|_{H.S.}^2\|B\|_{H.S.}^2 = n^2(1 + 2\delta_{ij}),$$

Thus

$$\mathbb{E}\left[\mathrm{Tr}\,(A_iMA_jM) - \delta_{ij}\right]^2 \leq \frac{-6n^2 + (n+1)n^2(1+2\delta_{ij}) - (n-1)n(n+2)\delta_{ij}}{(n-1)n(n+2)} \leq 2.$$

□

4.3. *Complex-linear functions of random unitary matrices.*

In this section, we consider Haar-distributed random matrices in $\mathcal{U}_n$. As discussed in the previous section, a general theme in studying random matrices from the classical compact matrix groups has been to compare to the corresponding Gaussian distribution. In particular, it was shown in D'Aristotile et al. (2003) that if $M = \Gamma + i\Lambda$ is a random $n \times n$ unitary matrix and A and B are fixed real diagonal matrices with $\mathrm{Tr}\,(AA^T) = \mathrm{Tr}\,(BB^T) = n$, then $\mathrm{Tr}\,(A\Gamma) + i\mathrm{Tr}\,(B\Lambda)$ converges in distribution to standard complex normal. This implies in particular that $\mathrm{Re}(\mathrm{Tr}\,(AM))$ converges in distribution to $\mathfrak{N}\left(0, \frac{1}{2}\right)$. A total variation rate of convergence for this last statement was obtained in Meckes (2008), giving as an easy consequence the weak-star convergence of the random variable $W = \mathrm{Tr}\,(AM)$ to standard complex normal, for A an $n \times n$ matrix over $\mathbb{C}$ with $\mathrm{Tr}\,(AA^*) = n$. The approaches used in D'Aristotile et al. (2003) and Meckes (2008) are somewhat awkward, partly due to the fact that the limiting behavior of W is a multivariate question. In this section, Theorem 2.5 is applied to prove the analogous result to Theorem 4.5 for complex-rank k projections of Haar measure on the space of random unitary matrices. As in the previous section, this result recovers and strengthens the result of Collins and Stolz (2008), in the case that M is a Haar-distributed unitary matrix.

Theorem 4.6. *Let $M \in \mathcal{U}_n$ be distributed according to Haar measure, and let $\{A_i\}_{i=1}^k$ be fixed $n \times n$ matrices over $\mathbb{C}$ such that $\operatorname{Tr}(A_i A_j^*) = n\delta_{ij}$. Let $W(M) = (\operatorname{Tr}(A_1 M), \ldots, \operatorname{Tr}(A_k M))$ and let Z be a standard complex Gaussian random vector in $\mathbb{C}^k$. Then there is a universal constant c such that*

$$d_W(W, Z) \le \frac{ck}{n}.$$

Remark: The constant c given by the proof is asymptotically equal to $\sqrt{2}$; for $n \ge 4$, c can be taken to be 3.

For the proof, the following lemma is needed. See Meckes (2006), Lemma 3.5 for a detailed proof.

Lemma 4.7. *Let $H = \big[h_{ij}\big]_{i,j} \in \mathcal{U}_n$ be distributed according to Haar measure. Then the expected value of a product of entries of H and their conjugates is non-zero only when there are the same number of entries as conjugates of entries from each row and from each column. Second- and fourth-degree moments are as follows.*

(i) *For all i, j,*

$$\mathbb{E}\big[|h_{ij}|^2\big] = \frac{1}{n},$$

(ii) *For all $i, j, r, s, \alpha, \beta, \lambda, \mu$,*

$$\begin{aligned}\mathbb{E}\big[h_{ij} h_{rs} \overline{h}_{\alpha\beta} \overline{h}_{\lambda\mu}\big] &= \frac{1}{(n-1)(n+1)}\Big[\delta_{i\alpha}\delta_{r\lambda}\delta_{j\beta}\delta_{s\mu} + \delta_{i\lambda}\delta_{r\alpha}\delta_{j\mu}\delta_{s\beta}\Big] \\ &\quad - \frac{1}{(n-1)n(n+1)}\Big[\delta_{i\alpha}\delta_{r\lambda}\delta_{j\mu}\delta_{s\beta} + \delta_{i\lambda}\delta_{r\alpha}\delta_{j\beta}\delta_{s\mu}\Big],\end{aligned}$$

(iii)

$$\begin{aligned}&\mathbb{E}\big[(h_{i1}\overline{h}_{j2} - h_{i2}\overline{h}_{j1})(h_{r1}\overline{h}_{s2} - h_{r2}\overline{h}_{s1})\big] \\ &\qquad = -\frac{2}{(n-1)(n+1)}\,\delta_{is}\delta_{jr} + \frac{2}{(n-1)n(n+1)}\,\delta_{ij}\delta_{rs}.\end{aligned}$$

Proof of Theorem 4.6. The theorem is proved as an application of Theorem 2.5, similarly to the proof of Theorem 4.5 via Theorem 2.4. Construct a family of pairs (W, W_ϵ) analogously to what was done in the orthogonal case: let $U \in \mathcal{U}_n$ be a random unitary matrix, independent of M, and let $M_\epsilon = U A_\epsilon U^* M$, where as before

$$A_\epsilon = \begin{bmatrix}\sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2}\end{bmatrix} \oplus I_{n-2},$$

thus M_ϵ is a small random rotation of M. Let $W_\epsilon = W(M_\epsilon)$; (W, W_ϵ) is exchangeable by construction.

As in the previous sections, let I_2 be the 2×2 identity matrix, K the $n \times 2$ matrix made from the first two columns of $U = \big[u_{ij}\big]_{i,j}$, and let

$$C_2 = \begin{bmatrix}0 & 1 \\ -1 & 0\end{bmatrix}.$$

Define the matrix $Q = KC_2K^*$. Then

$$M_\epsilon = M + K\big[(\sqrt{1-\epsilon^2}-1)I_2 + \epsilon C_2\big]K^*M,$$

and

$$\mathrm{Tr}\,(A_iM_\epsilon) - \mathrm{Tr}\,(A_iM) = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right)\mathrm{Tr}\,(A_iKK^*M) + \epsilon\mathrm{Tr}\,(A_iQM). \quad (4.5)$$

It follows from Lemma 4.7 that $\mathbb{E}\big[KK^*\big] = \frac{2}{n}I$ and $\mathbb{E}\big[Q\big] = 0$, thus

$$\lim_{\epsilon\to 0}\frac{n}{\epsilon^2}\mathbb{E}\left[\mathrm{Tr}\,(A_iM_\epsilon) - \mathrm{Tr}\,(A_iM)\big|M\right] = -\mathrm{Tr}\,(A_iM), \quad (4.6)$$

and the first condition of Theorem 2.5 holds with $\lambda(\epsilon) = \frac{\epsilon^2}{n}$.

Let $A_i =: A = \big[a_{pq}\big]$ and $A_j =: B = \big[b_{\alpha\beta}\big]$; by (4.5) and Lemma 4.7,

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{n}{2\epsilon^2}\mathbb{E}\left[(W_\epsilon - W)_i(W_\epsilon - W)_j\big|W\right]\\
&\quad= \frac{n}{2}\mathbb{E}\left[(\mathrm{Tr}\,(AQM))\,(\mathrm{Tr}\,(BQM))\,\big|W\right]\\
&\quad= \frac{n}{2}\mathbb{E}\left[\sum_{p,q,r,\alpha,\beta,\mu} a_{pq}m_{rp}b_{\alpha\beta}m_{\gamma\alpha}(u_{q1}\overline{u}_{r2} - u_{q2}\overline{u}_{r1})(u_{\beta 1}\overline{u}_{\gamma 2} - u_{\beta 2}\overline{u}_{\gamma 1})\right|W\Bigg]\\
&\quad= \frac{1}{(n-1)(n+1)}\left[\sum_{p,q,\alpha,\beta} a_{pq}m_{qp}b_{\alpha\beta}m_{\beta\alpha} - n\sum_{p,q,\alpha\beta} a_{pq}b_{\alpha\beta}m_{\beta p}m_{q\alpha}\right]\\
&\quad= \frac{1}{(n-1)(n+1)}\left[\mathrm{Tr}\,(AM)\mathrm{Tr}\,(BM) - n\mathrm{Tr}\,(AMBM)\right].
\end{aligned} \quad (4.7)$$

Similarly, one can use part (iii) of Lemma 4.7 with the roles of r and s reversed to get

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{n}{2\epsilon^2}\mathbb{E}\big[(W_\epsilon - W)_i\overline{(W_\epsilon - W)}_j\big|W\big] = \frac{n}{2}\mathbb{E}\left[\mathrm{Tr}\,(AQM)\overline{\mathrm{Tr}\,(BQM)}\big|W\right]\\
&\quad= \frac{n}{2}\mathbb{E}\left[\sum_{p,q,r,\alpha,\beta,\gamma} a_{pq}m_{rp}\overline{b}_{\alpha\beta}\overline{m}_{\gamma\alpha}(u_{q1}\overline{u}_{r2} - u_{q2}\overline{u}_{r1})(\overline{u}_{\beta 1}u_{\gamma 2} - \overline{u}_{\beta 2}u_{\gamma 1})\right|W\Bigg]\\
&\quad= \frac{1}{(n-1)(n+1)}\left[n\sum_{p,\alpha}\left(\sum_q a_{pq}\overline{b}_{\alpha q}\right)\left(\sum_\gamma m_{\gamma p}\overline{m}_{\gamma\alpha}\right) - \sum_{p,q,\alpha,\beta} a_{pq}\overline{b}_{\alpha\beta}m_{qp}\overline{m}_{\beta\alpha}\right]\\
&\quad= \frac{1}{(n-1)(n+1)}\left[n^2\delta_{ij} - \mathrm{Tr}\,(AM)\overline{\mathrm{Tr}\,(BM)}\right]\\
&\quad= \delta_{ij} + \frac{1}{(n-1)(n+1)}\Big[\delta_{ij} - \mathrm{Tr}\,(AM)\overline{\mathrm{Tr}\,(BM)}\Big],
\end{aligned} \quad (4.8)$$

where the fact that M is unitary and the assumption $\mathrm{Tr}\,(A_iA_j^*) = n\delta_{ij}$ have been used to get the second to last line.

One can thus take

$$\gamma_{ij} = \frac{\delta_{ij} - \mathrm{Tr}\,(A_iM)\overline{\mathrm{Tr}\,(A_jM)}}{(n-1)(n+1)} \qquad \lambda_{ij} = \frac{\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM) - n\mathrm{Tr}\,(A_iMA_jM)}{(n-1)(n+1)}.$$

By the Cauchy-Schwarz inequality,

$$\mathbb{E}\|\Gamma\|_{H.S.} \leq \sqrt{\mathbb{E}\sum_{i,j}|\gamma_{ij}|^2}$$

and

$$\mathbb{E}|\gamma_{ij}|^2 = \frac{1}{(n-1)^2(n+1)^2}\mathbb{E}\Big[\delta_{ij} - 2\,\mathrm{Re}(\mathrm{Tr}\,(A_iM)\overline{\mathrm{Tr}\,(A_jM)}) + |\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM)|^2\Big].$$

Now,

$$\mathbb{E}|\mathrm{Tr}\,(A_iM)\overline{\mathrm{Tr}\,(A_jM)}| \leq \sqrt{\mathbb{E}|\mathrm{Tr}\,(A_iM)|^2\mathbb{E}|\mathrm{Tr}\,(A_jM)|^2} = 1$$

by the normalization of the matrices A_i. Again writing $A = A_i$ and $B = A_j$,

$$\begin{aligned}
&\mathbb{E}|\mathrm{Tr}\,(AM)\mathrm{Tr}\,(BM)|^2 \\
&= \sum_{\substack{p,q,r,s\\ \alpha,\beta,\lambda,\mu}} a_{pq}\overline{a}_{rs}b_{\alpha\beta}\overline{b}_{\lambda\mu}\mathbb{E}\big[m_{qp}m_{\beta\alpha}\overline{m}_{sr}\overline{m}_{\mu\lambda}\big] \\
&= \frac{1}{(n-1)(n+1)}\Bigg[\mathrm{Tr}\,(AA^*)\mathrm{Tr}\,(BB^*) + (\mathrm{Tr}\,(AB^*))^2 \\
&\qquad -\frac{1}{n}\mathrm{Tr}\,(AA^*BB^*) - \frac{1}{n}\mathrm{Tr}\,(A^*AB^*B)\Bigg] \\
&= \frac{1}{(n-1)(n+1)}\left[n^2(1+\delta_{ij}) - \frac{1}{n}\mathrm{Tr}\,(AA^*BB^*) - \frac{1}{n}\mathrm{Tr}\,(A^*AB^*B)\right],
\end{aligned}$$

where Lemma 4.7 has been used to get the third line and the normalization and orthogonality conditions on the A_i have been used to get the last line. Now,

$$\begin{aligned}
|\mathrm{Tr}\,(AA^*BB^*)| &\leq \|AA^*\|_{H.S.}\|BB^*\|_{H.S.} \\
&\leq \|A\|_{H.S.}\|A^*\|_{H.S.}\|B\|_{H.S.}\|B^*\|_{H.S.} = n^2;
\end{aligned}$$

the first inequality is just the Cauchy-Schwarz inequality for the Hilbert-Schmidt inner product and the second is due to the submultiplicativity of the Hilbert-Schmidt norm (see Bhatia, 1997, page 94). It now follows that

$$\begin{aligned}
\mathbb{E}|\gamma_{ij}|^2 &\leq \frac{1}{(n-1)^2(n+1)^2}\left[\delta_{ij} + 2 + \frac{n^2(1+\delta_{ij}) + 2n}{(n-1)(n+1)}\right] \\
&\leq \frac{1}{(n-1)^2(n+1)^2}\left[5 + \frac{2}{n-1}\right],
\end{aligned}$$

and thus

$$\mathbb{E}\|\Gamma\|_{H.S.} \leq \frac{k}{(n-1)(n+1)}\sqrt{5 + \frac{2}{n-1}}. \tag{4.9}$$

Taking a similar approach to bounding $\mathbb{E}\|\Lambda\|_{H.S.}$,

$$\begin{aligned}
\mathbb{E}|\lambda_{ij}|^2 = \frac{1}{(n-1)^2(n+1)^2}\mathbb{E}\Big[&|\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM)|^2 + n^2|\mathrm{Tr}\,(A_iMA_jM)|^2 \\
&- 2n\,\mathrm{Re}(\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM)\overline{\mathrm{Tr}\,(A_iMA_jM)})\Big].
\end{aligned} \tag{4.10}$$

It has already been shown that

$$\mathbb{E}|\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM)|^2 \le \frac{n^2(1+\delta_{ij})+2n}{(n-1)(n+1)} \le 2+\frac{2}{n-1}.$$

One can use Lemma 4.7 to compute the other two terms similarly:

$$\begin{aligned}
&\mathbb{E}\Big[\mathrm{Tr}\,(AM)\mathrm{Tr}\,(BM)\overline{\mathrm{Tr}\,(AMBM)}\Big] \\
&\quad = \sum_{\substack{p,q,r,s\\ \alpha,\beta,\lambda,\mu}} a_{pq}\overline{a}_{\alpha\beta}b_{rs}\overline{b}_{\lambda\mu}\mathbb{E}\big[m_{qp}m_{sr}\overline{m}_{\beta\lambda}\overline{m}_{\mu\alpha}\big] \\
&\quad = \frac{1}{(n-1)(n+1)}\Bigg[\mathrm{Tr}\,(AA^*BB^*)+\mathrm{Tr}\,(A^*AB^*B) \\
&\qquad\qquad -\frac{1}{n}\mathrm{Tr}\,(AA^*)\mathrm{Tr}\,(BB^*)-\frac{1}{n}(\mathrm{Tr}\,(AB^*))^2\Bigg] \\
&\quad = \frac{1}{(n-1)(n+1)}\left[\mathrm{Tr}\,(AA^*BB^*)+\mathrm{Tr}\,(A^*AB^*B)-n(1+\delta_{ij})\right],
\end{aligned}$$

thus

$$\left|\mathbb{E}\left[2n\,\mathrm{Re}\Big(\mathrm{Tr}\,(A_iM)\mathrm{Tr}\,(A_jM)\mathrm{Tr}\,(A_iMA_jM)\Big)\right]\right| \le \frac{4n^3+2n(1+\delta_{ij})}{(n-1)(n+1)} \le 4n+\frac{8}{n-1};$$

and

$$\begin{aligned}
&\mathbb{E}|\mathrm{Tr}\,(AMBM)|^2 \\
&\quad = \sum_{\substack{p,q,r,s\\ \alpha,\beta,\lambda,\mu}} a_{pq}\overline{a}_{\alpha\beta}b_{rs}\overline{b}_{\lambda\mu}\mathbb{E}\big[m_{qr}m_{sp}\overline{m}_{\beta\lambda}\overline{m}_{\mu\alpha}\big] \\
&\quad = \frac{1}{(n-1)(n+1)}\Bigg[\mathrm{Tr}\,(AA^*)\mathrm{Tr}\,(BB^*)+(\mathrm{Tr}\,(AB^*))^2 \\
&\qquad\qquad -\frac{1}{n}\mathrm{Tr}\,(AA^*BB^*)-\frac{1}{n}\mathrm{Tr}\,(A^*AB^*B)\Bigg] \\
&\quad = \frac{1}{(n-1)(n+1)}\left[n^2(1+\delta_{ij})-\frac{1}{n}\mathrm{Tr}\,(AA^*BB^*)-\frac{1}{n}\mathrm{Tr}\,(A^*AB^*B)\right],
\end{aligned}$$

thus

$$n^2\mathbb{E}|\mathrm{Tr}\,(A_iMA_jM)|^2 \le \frac{n^4(1+\delta_{ij})+2n^3}{(n-1)(n+1)} \le 2n^2+\frac{2n^2}{n-1}.$$

Using these three bounds in (4.10) yields

$$\mathbb{E}\|\Lambda\|_{H.S.} \le \sqrt{\sum_{i,j}\mathbb{E}|\lambda_{ij}|^2} \le \frac{k}{(n-1)}\sqrt{2+\frac{2(n^2+5)}{(n-1)(n+1)^2}}.$$

□

References

R. Arratia, L. Goldstein and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* **17** (1), 9–25 (1989). ISSN 0091-1798.

Richard Arratia, Larry Goldstein and Louis Gordon. Poisson approximation and the Chen-Stein method. *Statist. Sci.* **5** (4), 403–434 (1990). ISSN 0883-4237. With comments and a rejoinder by the authors.

Pierre Baldi and Yosef Rinott. On normal approximations of distributions in terms of dependency graphs. *Ann. Probab.* **17** (4), 1646–1650 (1989). ISSN 0091-1798.

A. D. Barbour. Stein's method for diffusion approximations. *Probab. Theory Related Fields* **84** (3), 297–322 (1990). ISSN 0178-8051.

A. D. Barbour, Lars Holst and Svante Janson. *Poisson approximation*, volume 2 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York (1992). ISBN 0-19-852235-5. Oxford Science Publications.

A. D. Barbour, Michał Karoński and Andrzej Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *J. Combin. Theory Ser. B* **47** (2), 125–145 (1989). ISSN 0095-8956.

V. Bentkus. On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Inference* **113** (2), 385–402 (2003). ISSN 0378-3758.

Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York (1997). ISBN 0-387-94846-5.

Ulrich Brehm and Jürgen Voigt. Asymptotics of cross sections for convex bodies. *Beiträge Algebra Geom.* **41** (2), 437–454 (2000). ISSN 0138-4821.

Sourav Chatterjee. A new method of normal approximation. *Ann. Probability* **36** (4), 1584–1610 (2008).

Sourav Chatterjee, Persi Diaconis and Elizabeth Meckes. Exchangeable pairs and Poisson approximation. *Probab. Surv.* **2**, 64–106 (electronic) (2005). ISSN 1549-5787.

Louis H. Y. Chen. Poisson approximation for dependent trials. *Ann. Probability* **3** (3), 534–545 (1975).

Benoît Collins and Michael Stolz. Borel theorems for random matrices from the classical compact symmetric spaces. *Ann. Probab.* **36** (3), 876–895 (2008). ISSN 0091-1798.

Anthony D'Aristotile, Persi Diaconis and Charles M. Newman. Brownian motion and the classical groups. In *Probability, statistics and their applications: papers in honor of Rabi Bhattacharya*, volume 41 of *IMS Lecture Notes Monogr. Ser.*, pages 97–116. Inst. Math. Statist., Beachwood, OH (2003).

Persi Diaconis. Stein's method for Markov chains: first examples. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 27–43. Inst. Math. Statist., Beachwood, OH (2004).

Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theory. *Ann. Inst. H. Poincaré Probab. Statist.* **23** (2, suppl.), 397–423 (1987). ISSN 0246-0203.

Persi Diaconis and Susan Holmes, editors. *Stein's method: expository lectures and applications*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 46. Institute of Mathematical Statistics, Beachwood, OH (2004). ISBN 0-940600-62-5. Papers from the Workshop on Stein's Method held at Stanford University, Stanford, CA, 1998.

Persi W. Diaconis, Morris L. Eaton and Steffen L. Lauritzen. Finite de Finetti theorems in linear models and multivariate analysis. *Scand. J. Statist.* **19** (4), 289–315 (1992). ISSN 0303-6898.

Richard M. Dudley. *Real analysis and probability.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA (1989). ISBN 0-534-10050-3.

Jason Fulman. Stein's method and Plancherel measure of the symmetric group. *Trans. Amer. Math. Soc.* **357** (2), 555–570 (electronic) (2005). ISSN 0002-9947.

Larry Goldstein and Gesine Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* **7** (4), 935–952 (1997). ISSN 1050-5164.

Larry Goldstein and Yosef Rinott. Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33** (1), 1–17 (1996a). ISSN 0021-9002.

Larry Goldstein and Yosef Rinott. Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33** (1), 1–17 (1996b). ISSN 0021-9002.

F. Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** (2), 724–739 (1991). ISSN 0091-1798.

F. Götze and A. N. Tikhomirov. Limit theorems for spectra of random matrices with martingale structure. In *Stein's method and applications*, volume 5 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 181–193. Singapore Univ. Press, Singapore (2005).

Susan Holmes. Stein's method for birth and death chains. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 45–67. Inst. Math. Statist., Beachwood, OH (2004).

Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *Ann. Probab.* **34** (4), 1497–1529 (2006). ISSN 0091-1798.

B. Klartag. A central limit theorem for convex sets. *Invent. Math.* **168** (1), 91–131 (2007). ISSN 0020-9910.

Wei-Liem Loh. Stein's method and multinomial approximation. *Ann. Appl. Probab.* **2** (3), 536–554 (1992). ISSN 1050-5164.

H.M. Luk. *Stein's method for the gamma distribution and related statistical applications.* Ph.D. thesis, University of Southern California (1994).

Elizabeth Meckes. *An infinitesimal version of Stein's method of exchangeable pairs.* Ph.D. thesis, Stanford University (2006).

Elizabeth Meckes. On the approximate normality of eigenfunctions of the laplacian (2007). To appear in *Trans. Amer. Math. Soc.*. Available at `http://arxiv.org/abs/0705.1342`.

Elizabeth Meckes. Linear functions on the classical matrix groups. *Trans. Amer. Math. Soc.* **360** (10), 5355–5366 (2008).

A. Pickett. *Rates of Convergence of χ^2 Approximations via Stein's Method.* Ph.D. thesis, University of Oxford (2004).

M. Raič. A multivariate CLT for decomposable random vectors with finite second moments. *J. Theoret. Probab.* **17** (3), 573–603 (2004). ISSN 0894-9840.

G. Reinert and A. Röllin. Multivariate normal approximation with stein's method of exchangeable pairs under a general linearity condition (2008). Available at `http://arxiv.org/abs/0711.1082`.

Yosef Rinott and Vladimir Rotar. A multivariate CLT for local dependence with $n^{-1/2} \log n$ rate and applications to multivariate graph related statistics. *J. Multivariate Anal.* **56** (2), 333–350 (1996). ISSN 0047-259X.

Yosef Rinott and Vladimir Rotar. On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics. *Ann. Appl. Probab.* **7** (4), 1080–1105 (1997). ISSN 1050-5164.

Qi-Man Shao and Zhong-Gen Su. The Berry-Esseen bound for character ratios. *Proc. Amer. Math. Soc.* **134** (7), 2153–2159 (electronic) (2006). ISSN 0002-9939.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602. Univ. California Press, Berkeley, Calif. (1972).

Charles Stein. *Approximate computation of expectations*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7. Institute of Mathematical Statistics, Hayward, CA (1986). ISBN 0-940600-08-0.

Charles Stein. The accuracy of the normal approximation to the distribution of the traces of powers of random orthogonal matrices. Technical report, Stanford University Department of Statistics (1995). Technical Report No. 470.

Y. Yomdin. The geometry of critical and near-critical values of differentiable mappings. *Math. Ann.* **264** (4), 495–515 (1983). ISSN 0025-5831.