Elect. Comm. in Probab. **14** (2009), 176–185

ELECTRONIC COMMUNICATIONS in PROBABILITY

QUANTITATIVE ASYMPTOTICS OF GRAPHICAL PROJECTION PURSUIT

ELIZABETH S. MECKES[1]
220 Yost Hall, Department of Mathematics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106.
email: `ese3@cwru.edu`

Submitted January 12, 2009, *accepted in final form* April 14, 2009

AMS 2000 Subject classification: 60E15,62E20
Keywords: Projection pursuit, concentration inequalities, Stein's method, Lipschitz distance

Abstract

There is a result of Diaconis and Freedman which says that, in a limiting sense, for large collections of high-dimensional data most one-dimensional projections of the data are approximately Gaussian. This paper gives quantitative versions of that result. For a set of deterministic vectors $\{x_i\}_{i=1}^n$ in $\mathbb{R}^d$ with n and d fixed, let $\theta \in \mathbb{S}^{d-1}$ be a random point of the sphere and let μ_n^θ denote the random measure which puts mass $\frac{1}{n}$ at each of the points $\langle x_1, \theta \rangle, \ldots, \langle x_n, \theta \rangle$. For a fixed bounded Lipschitz test function f, Z a standard Gaussian random variable and σ^2 a suitable constant, an explicit bound is derived for the quantity $\mathbb{P}\left[\left|\int f d\mu_n^\theta - \mathbb{E}f(\sigma Z)\right| > \epsilon\right]$. A bound is also given for $\mathbb{P}\left[d_{BL}(\mu_n^\theta, \mathcal{N}(0,\sigma^2)) > \epsilon\right]$, where d_{BL} denotes the bounded-Lipschitz distance, which yields a lower bound on the waiting time to finding a non-Gaussian projection of the $\{x_i\}$ if directions are tried independently and uniformly on $\mathbb{S}^{d-1}$.

1 Introduction

A foundational tool of data analysis is the projection of high-dimensional data to a one- or two-dimensional subspace in order to visually represent the data, and, ideally, identify underlying structure. The question immediately arises: which projections are interesting? One would like to answer by saying that those projections which exhibit structure are interesting, however, identifying which projections those are is not quite as straightforward as one might think. In particular, there are several reasons that have led to the idea that one should mainly look for projections which are far from Gaussian in behavior; that Gaussian projections in fact do not generally exhibit interesting structure. One justification for this idea is the following result due to Persi Diaconis and David Freedman.

[1]RESEARCH SUPPORTED BY AN AMERICAN INSTITUTE OF MATHEMATICS FIVE-YEAR FELLOWSHIP

Theorem 1 (Diaconis-Freedman [1]). *Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that n, d and the x_i depend on a hidden index ν, so that as ν tends to infinity, so do n and d. Suppose that there is a $\sigma^2 > 0$ such that, for all $\epsilon > 0$,*

$$\frac{1}{n}\left|\left\{j \le n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\right\}\right| \xrightarrow{\nu\to\infty} 0, \tag{1}$$

and suppose that

$$\frac{1}{n^2}\left|\left\{j,k \le n : \left|\left\langle x_j, x_k\right\rangle\right| > \epsilon d\right\}\right| \xrightarrow{\nu\to\infty} 0. \tag{2}$$

Let $\theta \in \mathbb{S}^{d-1}$ be distributed uniformly on the sphere, and consider the random measure μ_ν^θ which puts mass $\frac{1}{n}$ at each of the points $\langle\theta, x_1\rangle, \ldots, \langle\theta, x_n\rangle$. Then as ν tends to infinity, the measures μ_ν^θ tend to $\mathcal{N}(0, \sigma^2)$ weakly in probability.

Heuristically, Theorem 1 can be interpreted as saying that, for a large number of high-dimensional data vectors, as long as they have nearly the same lengths and are nearly orthogonal, most one-dimensional projections are close to Gaussian regardless of the structure of the data. It is important to note that the conditions (1) and (2) are not too strong; in particular, even though only d vectors can be exactly orthogonal in $\mathbb{R}^d$, the 2^d vertices of a unit cube centered at the origin satisfy condition (2) for "rough orthogonality".

A failing of the usual interpretation of Theorem 1 is that sometimes, projections of data look nearly Gaussian for a reason; that is, it is not always due to the central-limit type effect described by the theorem. Thus the question arises: is there a way to tell whether a Gaussian projection is interesting? A possible answer lies in quantifying the theorem, and then saying that a nearly-Gaussian projection is interesting if it is "too close" to Gaussian to simply be the result of the phenomenon described by Theorem 1. By way of analogy, one has the Berry-Esséen theorem stating that the rate of convergence to normal of the sum of n independent, identically distributed random variables is of the order $\frac{1}{\sqrt{n}}$; if one has a sum of n random variables converging to Gaussian significantly faster, it must be happening for some reason other than just the usual central-limit theorem. In order to implement this idea, it is necessary (as with the Berry-Esséen theorem) to have a sharp quantitative version of the limit theorem in question.

A second motivation for proving a quantitative version of Theorem 1 is the application to waiting times for discovering an interesting direction on which to project data. If a sequence of independent random projection directions is tried until the empirical distribution of the projected data is more than some threshhold away from Gaussian (in some metric on measures), and N is the number of trials needed to find such a direction, a one can easily give a lower bound for $\mathbb{E}N$ from the type of quantitative theorem proved below.

Thus the goal of this paper is to provide a quantitative version of Theorem 1 in a fixed dimension d and for a fixed number of data vectors n. To do this, it is first necessary to replace conditions (1) and (2) with non-asymptotic conditions. The conditions we will use are the following. Let σ^2 be defined by $\frac{1}{n}\sum_{i=1}^n |x_i|^2 = \sigma^2 d$. Suppose there exist A and B, such that

$$\frac{1}{n}\sum_{i=1}^n \left|\sigma^{-2}|x_i|^2 - d\right| \le A, \tag{3}$$

and, for all $\theta \in \mathbb{S}^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^n \left\langle\theta, x_i\right\rangle^2 \le B. \tag{4}$$

For a little perspective on the restrictiveness of these conditions, note that, as for the conditions of Diaconis and Freedman, they hold for the vertices of a unit cube in $\mathbb{R}^d$ (with $A = 0$ and $B = \frac{1}{4}$). Under these assumptions, the following theorems hold.

Theorem 2. *Let $\{x_i\}_{i=1}^n$ be deterministic vectors in $\mathbb{R}^d$, subject to conditions* (3) *and* (4) *above. For a point $\theta \in \mathbb{S}^{d-1}$, let the measure μ_n^θ put equal mass at each of the points $\langle \theta, x_1 \rangle, \ldots, \langle \theta, x_n \rangle$. Fix a test function $f : \mathbb{R} \to \mathbb{R}$ with $\|f\|_{BL} := \|f\|_\infty + \sup_{x \neq y} \frac{|f(x)-f(y)|}{|x-y|} \leq 1$. Then for Z a standard Gaussian random variable, θ chosen uniformly on the sphere, σ defined as above, and $\epsilon > \max\left(\frac{2\pi\sqrt{B}}{\sqrt{d-1}}, \frac{2(A+2)}{d-1}\right)$,*

$$\mathbb{P}\left[\left|\int f(x) d\mu_n^\theta(x) - \mathbb{E} f(\sigma Z)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}} e^{-\frac{(d-1)}{2^5 B}\epsilon^2}.$$

Theorem 3. *Let $\{x_i\}_{i=1}^n$ be deterministic vectors in $\mathbb{R}^d$, subject to conditions* (3) *and* (4) *above, and again consider the measures μ_n^θ. If θ is chosen uniformly from $\mathbb{S}^{d-1}$ and $B \geq \epsilon \geq \max\left(\left[\frac{3 \cdot 2^6 \pi B}{\sqrt{d-1}}\right]^{2/5}, \frac{2(A+2)}{d-1}\right)$, then*

$$\mathbb{P}\left[d_{BL}(\mu_n^\theta, \mathcal{N}(0,\sigma^2)) > \epsilon\right] \leq \frac{c_1\sqrt{B}}{\epsilon^{3/2}} \exp\left[-\frac{c_2(d-1)\epsilon^5}{B^2}\right],$$

with $c_1 = 48\sqrt{\pi}$, $c_2 = 3^{-2}2^{-16}$, and d_{BL} denoting the bounded Lipschitz distance.

Remarks:

(i) It should be emphasized that the key difference between the results proved here and the result of Diaconis and Freedman is that Theorems 2 and 3 hold for *fixed* dimension d and number of data vectors n; there are no limits in the statements of the theorems.

(ii) It is not necessary for A and B to be absolute constants; for the the results above to be of interest as $d \to \infty$, it is easy to see from the statements that it is only necessary that $A = o(d)$ and $B = o(d)$ for Theorem 2 while $B = o(\sqrt{d})$ for Theorem 3. The reader may also be wondering where the dependence on n is in the statements above; it is built into the definition of B. Note that, by definition, $B \geq \frac{|x_i|^2}{n}$ for each i; in particular, $B \geq \frac{\sigma^2 d}{n}$. It is thus necessary that $n \to \infty$ as $d \to \infty$ for Theorem 2 and $n \gg \sqrt{d}$ for Theorem 3.

(iii) For Theorem 2, consider the special case that $\epsilon^2 = \frac{C^2 \cdot 2^5 B}{d-1}$ for a large constant C. Then the statement becomes

$$\mathbb{P}\left[\left|\int f(x) d\mu_n^\theta(x) - \mathbb{E} f(\sigma Z)\right| > \frac{C'}{\sqrt{d-1}}\right] \leq \sqrt{\frac{\pi}{2}} e^{-C^2},$$

with $C' = C \cdot 4\sqrt{2B}$. That is, roughly speaking, $\left|\int f(x) d\mu_n^\theta(x) - \mathbb{E} f(\sigma Z)\right|$ is likely to be on the order of $\frac{1}{\sqrt{d}}$ or smaller.

(iv) It is similarly useful to consider the following special case for Theorem 3. Let $C > \frac{3}{10}$, and consider the case $\epsilon^5 = C\left(\frac{9 \cdot 2^{16} B^2}{d-1}\right) \log(d-1)$. Then the bound above becomes:

$$\mathbb{P}\left[d_{BL}(\mu_n^\theta, \mathcal{N}(0,\sigma^2)) > \left(C' \frac{\log(d-1)}{d-1}\right)^{1/5}\right] \leq \frac{C'' B}{(d-1)^{C - \frac{3}{10}}},$$

where $C' = 9 \cdot 2^{16} C B^2$ and $C'' = 48\sqrt{\pi} C^{-3/10}$. Thus, roughly speaking, the bounded Lipschitz distance from the random measure μ_n^θ to the Gaussian measure with mean zero and variance σ^2 is unlikely to be more than a large multiple of $\left(\frac{\log(d-1)}{d-1}\right)^{1/5}$. We make no claims of the sharpness of this result.

Theorem 3 can easily be used to give an estimate on the waiting time until a non-Gaussian direction is found, if directions are tried randomly and independently. Specifically, we have the following corollary.

Corollary 4. *Let $\theta_1, \theta_2, \theta_3, \dots$ be a sequence of independent, uniformly distributed random points on $\mathbb{S}^{d-1}$. Let $T_\epsilon := \min\{j : d_{BL}(\mu_n^{\theta_j}, \mathcal{N}(0,\sigma^2)) > \epsilon\}$. Then there are constants c, c' such that*

$$\mathbb{E} T_\epsilon \geq \frac{c\epsilon^{3/2}}{\sqrt{B}} \exp\left(\frac{c'(d-1)\epsilon^5}{B^2}\right).$$

2 Proofs

This section is mainly devoted to the proofs of Theorems 2 and 3, with some additional remarks following the proofs. For the proof of Theorem 2, several auxiliary results are needed. The first is an abstract normal approximation for bounding the distance of a random variable to a Gaussian random variable in the presence of a continuous family of exchangeable pairs. The theorem is an abstraction of an idea used by Stein in [6] to bound the distance to Gaussian of the trace of a power of a random orthogonal matrix.

Theorem 5 (Meckes [4]). *Suppose that (W, W_ϵ) is a family of exchangeable pairs defined on a common probability space, such that $\mathbb{E}W = 0$ and $\mathbb{E}W^2 = \sigma^2$. Let $\mathscr{F}$ be a σ-algebra on this space with $\sigma(W) \subseteq \mathscr{F}$. Suppose there is a function $\lambda(\epsilon)$ and random variables E, E' measurable with respect to $\mathscr{F}$, such that*

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

(ii) $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| \mathscr{F}\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E$.

(iii) $\frac{1}{\lambda(\epsilon)}\mathbb{E}|W_\epsilon - W|^3 \xrightarrow{\epsilon \to 0} 0$.

Then if Z is a standard normal random variable,

$$d_{TV}(W, \sigma Z) \leq \mathbb{E}\left|E\right| + \sqrt{\frac{\pi}{2}}\mathbb{E}\left|E'\right|.$$

The next result gives expressions for some mixed moments of entries of a Haar-distributed orthogonal matrix. See [3], Lemma 3.3 and Theorem 1.6 for a detailed proof.

Lemma 6. *If $U = \left[u_{ij}\right]_{i,j=1}^d$ is an orthogonal matrix distributed according to Haar measure, then $\mathbb{E}\left[\prod u_{ij}^{r_{ij}}\right]$ is non-zero if and only if $r_{i\bullet} := \sum_{j=1}^d r_{ij}$ and $r_{\bullet j} := \sum_{i=1}^d r_{ij}$ are even for each i and j. Second and fourth-degree moments are as follows:*

(i) For all i, j,

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

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

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

(iii) For the matrix $Q = \left[q_{ij}\right]_{i,j=1}^d$ defined by $q_{ij} := u_{i1}u_{j2} - u_{i2}u_{j1}$, and for all i, j, ℓ, p,

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

Finally, we will need to make use of the concentration of measure on the sphere, in the form of the following lemma.

Lemma 7 (Lévy, see [5])**.** *For a function $F : \mathbb{S}^{d-1} \to \mathbb{R}$, let M_F denote its median with respect to the uniform measure (that is, for θ distributed uniformly on $\mathbb{S}^{d-1}$, $\mathbb{P}\left[F(\theta) \le M_F\right] \ge \frac{1}{2}$ and $\mathbb{P}\left[F(\theta) \ge M_F\right] \ge \frac{1}{2}$) and let L denote its Lipschitz constant. Then*

$$\mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon\right] \le \sqrt{\frac{\pi}{2}}\exp\left[-\frac{(d-1)\epsilon^2}{2L^2}\right].$$

With these results, it is now possible to give the proof of Theorem 2.

Proof of Theorem 2. The proof divides into two parts. First, an "annealed" version of the theorem is proved using the infinitesimal version of Stein's method given by Theorem 5. Then, for a fixed test function f and Z a standard Gaussian random variable, the quantity $\mathbb{P}\left[\left|\int f\,d\mu_\nu^\theta - \mathbb{E}f(\sigma Z)\right| > \epsilon\right]$ is bounded using the annealed theorem together with the concentration of measure phenomenon.

Let θ be a uniformly distributed random point of $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$, and let I be a uniformly distributed element of $\{1, \ldots, n\}$, independent of θ. Consider the random variable $W := \langle \theta, x_I \rangle$. Then $\mathbb{E}W = 0$ by symmetry and $\mathbb{E}W^2 = \sigma^2$ by the condition $\frac{1}{n}\sum_{i=1}^n |x_i|^2 = \sigma^2 d$. Theorem 5 will be used to bound the total variation distance from W to σZ, where Z is a standard Gaussian random variable. The family of exchangeable pairs needed to apply the theorem is constructed as follows. For $\epsilon > 0$ fixed, let

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

where $\delta = O(\epsilon^4)$. Let U be a Haar-distributed $d \times d$ random orthogonal matrix, independent of θ and I, and let $W_\epsilon = \langle UA_\epsilon U^T\theta, x_I \rangle$; the pair (W, W_ϵ) is exchangeable for each $\epsilon > 0$.

Let K be the $d \times 2$ matrix made of the first two columns of U and $C_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. Define $Q := KC_2K^T$ (note that this is the same Q as in part (iii) of Theorem 6). Then by the construction of W_ϵ,

$$W_\epsilon - W = -\left(\frac{\epsilon^2}{2} + \delta\right)\left\langle KK^T\theta, x_I\right\rangle + \epsilon\left\langle Q\theta, x_I\right\rangle. \tag{5}$$

The conditions of Theorem 5 can be verified using the expressions in Lemma 6 as follows. By the lemma, $\mathbb{E}\left[KK^T\right] = \frac{2}{d}I$ and $\mathbb{E}\left[Q\right] = 0$, and so it follows from (5) that

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \left(-\frac{\epsilon^2}{d} + \frac{2\delta}{n}\right)W.$$

Condition (i) of Theorem 5 is thus satisfied for $\lambda(\epsilon) = \frac{\epsilon^2}{d}$ and $E' = 0$.
For the condition (ii), taking $\mathscr{F} = \sigma(\theta, I)$, Lemma 6, part (iii) yields

$$\begin{aligned}
\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| \mathscr{F}\right] &= \frac{d}{2\sigma^2}\mathbb{E}\left[\left\langle Q\theta, x_I\right\rangle^2 \middle| \mathscr{F}\right] + O(\epsilon) \\
&= \frac{d}{2\sigma^2}\sum_{i,j,r,s=1}^{d}\mathbb{E}\left[q_{ij}q_{rs}\theta_j\theta_s x_{I,i}x_{I,r}\middle|\mathscr{F}\right] + O(\epsilon) \\
&= \frac{1}{\sigma^2(d-1)}\left[\sum_{i,j=1}^{d}\theta_j^2 x_{I,i}^2 - \sum_{i,j=1}^{d}\theta_i\theta_j x_{I,i}x_{I,j}\right] + O(\epsilon) \\
&= \frac{1}{\sigma^2(d-1)}\left[|x_I|^2 - W^2\right] + O(\epsilon) \\
&= 1 + \frac{1}{d-1}\left[\frac{|x_I|^2}{\sigma^2} - d + 1 - \frac{W^2}{\sigma^2}\right] + O(\epsilon).
\end{aligned}$$

Condition (ii) of Theorem 5 is thus satisfied with $E = \frac{1}{d-1}\left[\frac{|x_I|^2}{\sigma^2} - d + 1 - \frac{W^2}{\sigma^2}\right]$. Condition (iii) of the theorem is trivial by (5); it follows that

$$d_{TV}(W, \sigma Z) \le \frac{1}{d-1}\mathbb{E}\left|\frac{|x_I|^2}{\sigma^2} - d + 1 - \frac{W^2}{\sigma^2}\right| \le \frac{1}{d-1}\left[\frac{1}{n}\sum_{i-1}^{n}\left|\frac{|x_i|^2}{\sigma^2} - d\right| + 2\right] \le \frac{A+2}{d-1}. \tag{6}$$

This is the annealed statement referred to at the beginning of the proof.

We next use the concentration of measure on the sphere to show that, for a large measure of $\theta \in \mathbb{S}^{d-1}$, the random measure μ_n^θ which puts mass $\frac{1}{n}$ at each of the $\left\langle \theta, x_i\right\rangle$ is close to the average behavior. To do this, we make use of Lévy's Lemma (Lemma 7). Let $f : \mathbb{R} \to \mathbb{R}$ be such that $\|f\|_{BL} := \|f\|_\infty + \sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|} \le 1$. Consider the function F defined on the sphere by

$$F(\theta) := \int f(x)d\mu_n^\theta(x) = \frac{1}{n}\sum_{i=1}^{n} f(\left\langle \theta, x_i\right\rangle).$$

In order to apply Lemma 7, it is necessary to determine the Lipschitz constant of F. Let $\theta, \theta' \in$

$\mathbb{S}^{d-1}$. Then, using $\|f\|_{BL} \le 1$ together with equation (4),

$$\begin{aligned}\left|F(\theta') - F(\theta)\right| &= \frac{1}{n}\left|\sum_{i=1}^{n} f(\langle \theta', x_i\rangle) - f(\langle \theta, x_i\rangle)\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n} |\langle \theta' - \theta, x_i\rangle| \\ &\le \left[\frac{1}{n}\sum_{i=1}^{n}\langle \theta' - \theta, x_i\rangle^2\right]^{1/2} \\ &\le |\theta' - \theta|\sqrt{B},\end{aligned}$$

thus the Lipschitz constant of F is bounded by $\sqrt{B}$. It follows from Lemma 7 that

$$\mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon\right] \le \sqrt{\frac{\pi}{2}}e^{-\frac{(d-1)\epsilon^2}{2B}},$$

where M_F is the median of the function F.
Now, if θ is a random point of $\mathbb{S}^{d-1}$, then

$$\begin{aligned}\left|\mathbb{E}F(\theta) - M_F\right| &\le \mathbb{E}\left|F(\theta) - M_F\right| \\ &= \int_0^\infty \mathbb{P}\left[\left|F(\theta) - M_F\right| > t\right] dt \\ &\le \int_0^\infty \sqrt{\frac{\pi}{2}}e^{-\frac{(d-1)t^2}{2B}} dt \\ &= \frac{\pi\sqrt{B}}{2\sqrt{d-1}},\end{aligned} \tag{7}$$

thus if $\epsilon > \frac{\pi\sqrt{B}}{\sqrt{d-1}}$, we may use concentration about the median of F to obtain concentration about the mean, with only a loss in constants.
Note that

$$\mathbb{E}F(\theta) = \mathbb{E}\int f\, d\mu_n^\theta = \mathbb{E}f(W)$$

for $W = \langle \theta, x_I\rangle$ as above, and so by the bound (6),

$$\left|\mathbb{E}F(\theta) - \mathbb{E}f(\sigma Z)\right| \le \frac{A+2}{d-1}.$$

Putting these pieces together, if $\epsilon > \max\left(\frac{2\pi\sqrt{B}}{\sqrt{d-1}}, \frac{2(A+2)}{d-1}\right)$, then

$$\begin{aligned}\mathbb{P}\left[\left|\int f\, d\mu_n^\theta - \mathbb{E}f(\sigma Z)\right| > \epsilon\right] &\le \mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon - |M_F - \mathbb{E}F(\theta)| - |\mathbb{E}F(\theta) - \mathbb{E}f(\sigma Z)|\right] \\ &\le \mathbb{P}\left[\left|F(\theta) - M_F\right| > \frac{\epsilon}{4}\right] \\ &\le \sqrt{\frac{\pi}{2}}e^{-\frac{(d-1)}{2^5B}\epsilon^2}.\end{aligned}$$

□

Proof of Theorem 3. The first two steps of the proof of Theorem 3 were essentially done already in the proof of Theorem 2. From that proof, we have that if $W = \langle \theta, x_I \rangle$ for θ distributed uniformly on $\mathbb{S}^{d-1}$ and I independent of θ and uniformly distributed in $\{1, \ldots, n\}$, then

$$d_{TV}(W, \sigma Z) \le \frac{A+2}{d-1}, \tag{8}$$

for A as in equation (3). Furthermore, it follows from equation (7) in the proof of Theorem 2 that for $F(\theta) := \int f \, d\mu_n^\theta$ and $\epsilon > \frac{\pi\sqrt{B}}{\sqrt{d-1}}$, then

$$\begin{aligned} \mathbb{P}\left[|F(\theta) - \mathbb{E}F(\theta)| > \epsilon\right] &\le \mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon - \left|M_f - \mathbb{E}F(\theta)\right|\right] \\ &\le \mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon - \frac{\pi\sqrt{B}}{2\sqrt{d-1}}\right] \\ &\le \sqrt{\frac{\pi}{2}} e^{-\frac{(d-1)}{8B}\epsilon^2}. \end{aligned} \tag{9}$$

In this proof, this last statement is used together with a series of successive approximations of arbitrary bounded Lipschitz functions as used by Guionnet and Zeitouni [2] to obtain a bound for $\mathbb{P}\left[d_{BL}(\mu_n^\theta, \mathcal{N}(0,\sigma^2)) > \epsilon\right]$.
By definition,

$$\mathbb{P}\left[d_{BL}(\mu_n^\theta, \mathbb{E}\mu_n^\theta) > \epsilon\right] = \mathbb{P}\left[\sup_{\|f\|_{BL} \le 1} \left|\int f \, d\mu_n^\theta - \mathbb{E}\int f \, d\mu_n^\theta\right| > \epsilon\right].$$

First consider the subclass $\mathscr{F}_{BL,K} = \{f : \|f\|_{BL} \le 1, supp(f) \subseteq K\}$ for a compact set $K \subseteq \mathbb{R}$. Let $\Delta = \frac{\epsilon}{4}$; for $f \in \mathscr{F}_{BL,K}$, define the approximation f_Δ as in Guionnet and Zeitouni [2] as follows. Let $x_o = \inf K$ and let

$$g(x) = \begin{cases} 0 & x \le 0; \\ x & 0 \le x \le \Delta; \\ \Delta & x \ge \Delta. \end{cases}$$

For $x \in K$, define f_Δ recursively by $f_\Delta(x_o) = 0$ and

$$f_\Delta(x) = \sum_{i=0}^{\lceil \frac{x - x_o}{\Delta} \rceil} \Big(2\mathbb{I}\left[f(x_o + (i+1)\Delta) \ge f_\Delta(x_o + i\Delta)\right] - 1\Big) g(x - x_o - i\Delta).$$

That is, the function f_Δ is just an approximation of f by a function which is piecewise linear and has slope 1 or -1 on each of the intervals $[x_o + i\Delta, x_o + (i+1)\Delta]$. Note that, because $\|f\|_{BL} \le 1$, it follows that $\|f - f_\Delta\|_\infty \le \Delta$ and the number of distinct functions whose linear span is used to approximate f in this way is bounded by $\frac{|K|}{\Delta}$, where $|K|$ is the diameter of K. If $\{h_k\}_{k=1}^N$ denotes the set of functions used in the approximation f_Δ and ϵ_k their coefficients, then for $\epsilon^2 > 8\pi|K|\sqrt{\frac{B}{d-1}}$,

$$\begin{aligned}\mathbb{P}\left[\sup_{f\in\mathscr{F}_{BL,K}}\left|\int f\,d\mu_n^\theta - \mathbb{E}\int f\,d\mu_n^\theta\right| > \epsilon\right] &\le \mathbb{P}\left[\sup_{f\in\mathscr{F}_{BL,K}}\left|\int f_\Delta d\mu_n^\theta - \mathbb{E}\int f_\Delta d\mu_n^\theta\right| > \epsilon - 2\Delta\right]\\ &= \mathbb{P}\left[\sup_{f\in\mathscr{F}_{BL,K}}\left|\sum_{k=1}^N \epsilon_k\left(\int h_k d\mu_n^\theta - \mathbb{E}\int h_k d\mu_n^\theta\right)\right| > \frac{\epsilon}{2}\right]\\ &\le \mathbb{P}\left[\sum_{k=1}^N\left|\int h_k d\mu_n^\theta - \mathbb{E}\int h_k d\mu_n^\theta\right| > \frac{\epsilon}{2}\right]\\ &\le \sum_{k=1}^N \mathbb{P}\left[\left|\int h_k d\mu_n^\theta - \mathbb{E}\int h_k d\mu_n^\theta\right| > \frac{\epsilon}{2N}\right]\\ &\le \sqrt{\frac{\pi}{2}}Ne^{-\frac{(d-1)}{8B}\left(\frac{\epsilon}{2N}\right)^2}\\ &\le \frac{2\sqrt{2\pi}|K|}{\epsilon}e^{-\frac{(d-1)}{8B}\left(\frac{\epsilon^2}{8|K|}\right)^2}.\end{aligned}$$

The second-last line follows from equation (9) above, and the last line from the bound $N \le \frac{4|K|}{\epsilon}$. To move to the full set $\mathscr{F}_{BL} := \{f : \|f\|_{BL} \le 1\}$, we make a truncation argument. Given $f \in \mathscr{F}_{BL}$ and $M > 0$, define f_M by

$$f_M(x) = \begin{cases} 0 & x \le -M - |f(-M)|;\\ sgn(f(-M))\big[x + M + |f(-M)|\big] & -M - |f(-M)| < x \le -M;\\ f(x) & -M < x \le M;\\ sgn(f(M))\big[|f(M)| + M - x\big] & M < x \le M + |f(M)|;\\ 0 & x > M + |f(M)|;\end{cases}$$

that is, f_M is equal to f on $[-M,M]$ and is drops off to zero linearly with slope 1 outside $[-M,M]$. Then, since $f(x) = f_M(x)$ for $x \in [-M,M]$ and $|f(x) - f_M(x)| \le 1$ for $x \notin [-M,M]$,

$$\left|\int [f - f_M]\,d\mu_n^\theta\right| \le \mathbb{P}\big[|\langle x_I,\theta\rangle| > M\big] \le \frac{1}{M^2}\mathbb{E}\big[\langle x_I,\theta\rangle^2\big] \le \frac{B}{M^2}.$$

Choosing M such that $\frac{B}{M^2} = \frac{\epsilon}{4}$, it follows that for $\epsilon^{5/2} > \frac{3\cdot 2^6\pi B}{\sqrt{d-1}}$,

$$\begin{aligned}\mathbb{P}\left[\sup_{f\in\mathscr{F}_{BL}}\left|\int f\,d\mu_n^\theta - \mathbb{E}\int f\,d\mu_n^\theta\right| > \epsilon\right] &\le \mathbb{P}\left[\sup_{f\in\mathscr{F}_{BL}}\left|\int f_M d\mu_n^\theta - \mathbb{E}\int f_M d\mu_n^\theta\right| > \epsilon - \frac{2B}{M^2}\right]\\ &\le \mathbb{P}\left[\sup_{g\in\mathscr{F}_{BL,[-M-1,M+1]}}\left|\int g\,d\mu_n^\theta - \mathbb{E}\int g\,d\mu_n^\theta\right| > \frac{\epsilon}{2}\right]\\ &\le \frac{4\sqrt{2\pi}(M+1)}{\epsilon}e^{-\frac{(d-1)}{8B}\left(\frac{\epsilon^2}{16(M+1)}\right)^2}\\ &\le \frac{12\sqrt{2\pi B}}{\epsilon^{3/2}}e^{-\frac{(d-1)\epsilon^5}{9\cdot 2^{11}B^2}},\end{aligned}$$

assuming that $B \geq \epsilon$.
Recall that $\mathbb{E}\int f\,d\mu_n^\theta = \mathbb{E}f(W)$ for $W = \langle \theta, x_I \rangle$, and so by the bound (8),

$$\sup_{f\in\mathscr{F}_{BL}} \left|\mathbb{E}\int f\,d\mu_n^\theta - \mathbb{E}f(\sigma Z)\right| \leq \frac{A+2}{d-1},$$

thus for ϵ bounded below as above and also satisfying $\epsilon > \frac{2(A+2)}{d-1}$,

$$\mathbb{P}\left[d_{BL}(W,\sigma Z) > \epsilon\right] \leq \frac{48\sqrt{\pi B}}{\epsilon^{3/2}}\exp\left[-\frac{(d-1)\epsilon^5}{9\cdot 2^{16}B^2}\right].$$

□

Proof of Corollary 4. The proof is essentially trivial. Note that

$$\mathbb{P}[T_\epsilon > m] \geq \left[1 - \frac{c_1\sqrt{B}}{\epsilon^{3/2}}\exp\left(\frac{c_2(d-1)\epsilon^5}{B^2}\right)\right]^m$$

by independence of the θ_j and Theorem 3, since $T_\epsilon > m$ if and only if $d_{BL}(\mu_n^{\theta_j}, \mathcal{N}(0,\sigma^2)) \leq \epsilon$ for all $1 \leq j \leq m$. This bound can be used in the identity $\mathbb{E}T_\epsilon = \sum_{m=0}^\infty \mathbb{P}[T_\epsilon > m]$ to obtain the bound in the corollary. □

Remark: One of the features of the proofs given above is that they can be generalized to the case of k-dimensional projections of the d-dimensional data vectors $\{x_i\}$, with k fixed or even growing with d. The proof of the higher-dimensional analog of Theorem 2 goes through essentially the same way. However, the analog of the proof of Theorem 3 from Theorem 2 is rather more involved in the multivariate setting and will be the subject of a future paper.

References

[1] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 1984. MR0751274

[2] A. Guionnet and O. Zeitouni. Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.*, 5:119–136 (electronic), 2000. MR1781846

[3] Elizabeth Meckes. An infinitesimal version of Stein's method of exchangeable pairs. Doctoral dissertation, Stanford University, 2006.

[4] Elizabeth Meckes. Linear functions on the classical matrix groups. *Trans. Amer. Math. Soc.*, 360(10):5355–5366, 2008. MR2415077

[5] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov. MR0856576

[6] C. Stein. The accuracy of the normal approximation to the distribution of the traces of powers of random orthogonal matrices. 1995. Technical Report No. 470, Stanford University Department of Statistics.