Chapter 4
L^1 Bounds

In this chapter we focus on normal approximation using smooth functions, and the L^1 norm in particular. We begin with a discussion of distances induced by function classes. Any class of functions $\mathcal{H}$ mapping $\mathbb{R}$ to $\mathbb{R}$ induces a measure of the separation between the distributions $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ of the random variables X and Y by

$$\big\|\mathcal{L}(X) - \mathcal{L}(Y)\big\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} \big|Eh(X) - Eh(Y)\big|. \tag{4.1}$$

Certain choices of $\mathcal{H}$ lead to classical distances, for instance, taking

$$\mathcal{H} = \big\{\mathbf{1}(x \le z),\ z \in \mathbb{R}\big\} \tag{4.2}$$

leads to the Kolmogorov, L^∞, or supremum norm distance, while the class of measurable functions

$$\mathcal{H} = \big\{h\colon\ 0 \le h(x) \le 1,\ \forall x \in \mathbb{R}\big\} \tag{4.3}$$

leads to the total variation distance.

Calculations with smooth functions are typically simpler than those with functions such as the discontinuous indicators in (4.2), or the bounded measurable functions in (4.3). Our main focus in this chapter is the L^1 distance, given by (4.1) with $\mathcal{H} = \mathfrak{L}$, the collection of Lipschitz functions in (4.7). In Sect. 4.8 we move to the distance $\|\mathcal{L}(W) - \mathcal{L}(Z)\|_{\mathcal{H}_{m,\infty}}$, produced by taking $\mathcal{H}$ to be the collection of functions $\mathcal{H}_{m,\infty}$ defined in (4.183), a class including functions allowed to posses some small number of additional higher order derivatives.

Our L^1 examples include: the sums of independent random variables and an associated contraction principle, hierarchical structures, cone measure on the sphere, combinatorial central limit theorems, simple random sampling, coverage processes, and locally dependent random variables. To illustrate our approach for the smooth functions $\mathcal{H}_{m,\infty}$ we show how fast convergence rates may result under a vanishing third moment assumption. The use of Stein's method for L^1 approximation was pioneered by Erickson (1974).

We begin now by recalling that the L^1 distance between distribution functions F and G is defined by

L.H.Y. Chen et al., *Normal Approximation by Stein's Method*,
Probability and Its Applications,
DOI 10.1007/978-3-642-15007-4_4,

$$\|F - G\|_1 = \int_{-\infty}^{\infty} \big|F(t) - G(t)\big| dt. \tag{4.4}$$

This distance has a number of equivalent forms, and, perhaps for that reason, is known by many names, including Gini's measure of discrepancy, the Kantarovich metric (see Rachev 1984), and the Wasserstein, Dudley, and the Fortet–Mourier distance (see e.g., Barbour et al. 1992). In addition to writing the L^1 distance as in (4.4), we will also let $\|\mathcal{L}(X) - \mathcal{L}(Y)\|_1$ denote the L^1 distance between the distributions of random variables X and Y.

That zero biasing seems to be particularly suited to produce L^1 bounds is evidenced in the following theorem from Goldstein (2004).

Theorem 4.1 *Let W be a mean zero, variance 1 random variable with distribution function F and let W^* have the W-zero biased distribution and be defined on the same space as W. Then, with Φ the cumulative distribution function of the standard normal,*

$$\|F - \Phi\|_1 \le 2E|W^* - W|. \tag{4.5}$$

As there may exist many couplings of W and W^* on a joint space, the challenge in producing good L^1 bounds is to find one in which the variables are close.

Before proving Theorem 4.1, we recall some facts about the L^1 norm which can be found in Rachev (1984). First, the 'dual form' of the L^1 distance is given by

$$\|F - G\|_1 = \inf E|X - Y|, \tag{4.6}$$

where the infimum is over all couplings of X and Y on a joint space with marginal distributions F and G, respectively. As $\mathbb{R}$ is a Polish space, this infimum is achieved. A yet equivalent form of the L^1 distance is given by (4.1) with $\mathfrak{L}$ the collection of Lipschitz functions

$$\mathfrak{L} = \big\{h : \mathbb{R} \to \mathbb{R}\colon\ \big|h(y) - h(x)\big| \le |y - x|\big\}, \tag{4.7}$$

that is,

$$\big\|\mathcal{L}(Y) - \mathcal{L}(X)\big\|_1 = \sup_{h \in \mathfrak{L}} \big|Eh(Y) - Eh(X)\big|. \tag{4.8}$$

We will also make use of the fact that the elements in $\mathfrak{L}$ are exactly those absolutely continuous functions whose derivatives are (a.e.) bounded by 1 in absolute value. Though the L^1 distance is, therefore, just one example of a metric induced by a collection of smooth functions such as those we will study in Sect. 4.8, its many equivalent forms lead to a rich theory which accommodates numerous examples.

Part (ii) of Proposition 2.4 leads directly to the following proof of Theorem 4.1.

Proof First, let (W, W^*) achieve the infimum $\|W - W^*\|_1$ in (4.6). As (2.77) holds with $\Delta = W^* - W$, (2.79) yields

$$\big|Eh(W) - Nh\big| \le 2\|h'\|E|W - W^*| = 2\|W - W^*\|_1.$$

Taking supremum over $h \in \mathfrak{L}$ and using (4.8) shows

$$\|F - \Phi\|_1 \leq 2\|W - W^*\|_1. \tag{4.9}$$

Now for (W, W^*) any coupling of W to a variable W^* with the W-zero bias distribution, inequality (4.6) shows that the right hand side of (4.9) can be no greater than that of (4.5), and the result follows. □

The majority of this chapter is devoted to the exploration of various consequences of this bound, starting with sums of independent random variables.

4.1 Sums of Independent Variables

4.1.1 L^1 Berry–Esseen Bounds

Continuing the discussion in Sect. 3.1, and Theorem 3.1 in particular, in this section we elaborate on the theme of L^1 bounds for sums of independent random variables. In particular, we demonstrate the application of Theorem 4.1 and the construction (2.61) in Lemma 2.8 to produce L^1 bounds with small, explicit, and distributionally specific constants for the distance between the distribution of a sum of independent variables and the normal. The utility of Theorem 4.2 below is reflected by the fact that the L^1 distance on the left hand side of (4.13) is that of a convolution to the normal, but is bounded on the right by terms which require only the calculation of integrals of the form (4.4) involving marginal distributions.

The proof of Theorem 4.2 requires the following simple proposition. For H a distribution function on $\mathbb{R}$ let

$$H^{-1}(u) = \sup\{x\colon\ H(x) < u\} \quad \text{for } u \in (0, 1)$$

and let $\mathcal{U}(a, b)$ denote the uniform distribution on (a, b). It is well known that when $U \sim \mathcal{U}[0, 1]$ then $H^{-1}(U)$ has distribution function H.

Proposition 4.1 *For F and G distribution functions and $U \sim \mathcal{U}(0, 1)$,*

$$\|F - G\|_1 = E\left|F^{-1}(U) - G^{-1}(U)\right|.$$

Further, for any $a \geq 0$ and $b \in \mathbb{R}$, with $F_{a,b}$ and $G_{a,b}$ the distribution functions of $aX + b$ and $aY + b$, respectively, we have

$$\|F_{a,b} - G_{a,b}\|_1 = a\|F - G\|_1. \tag{4.10}$$

Proof The first claim is stated in (iii), Sect. 2.3 of Rachev (1984); the second follows immediately from the dual form (4.6) of the L^1 distance. □

Note that one consequence of the proposition is a representation of a pair of variables which achieve the infimum in (4.6).

For X a random variable with finite third absolute moment let

$$B(X) = \frac{2\operatorname{Var}(X)\|\mathcal{L}(X^*) - \mathcal{L}(X)\|_1}{E|X|^3}. \tag{4.11}$$

Applying (4.10) we have

$$B(aX) = B(X) \quad \text{for } a \neq 0. \tag{4.12}$$

Theorem 4.2 *Let $\xi_i, i = 1, \ldots, n$ be independent mean zero random variables with variances $\sigma_i^2 = \operatorname{Var}(\xi_i)$ satisfying $\sum_{i=1}^n \sigma_i^2 = 1$. Then for F the distribution function of*

$$W = \sum_{i=1}^n \xi_i$$

and Φ that of the standard normal,

$$\|F - \Phi\|_1 \leq \sum_{i=1}^n B(\xi_i) E|\xi_i|^3. \tag{4.13}$$

Additionally, when $W = \sum_{i=1}^n X_i/(\sigma\sqrt{n})$ with $X, X_1, \ldots, X_n$ i.i.d. mean zero, variance σ^2 random variables, then

$$\|F - \Phi\|_1 \leq \frac{1}{\sigma^3\sqrt{n}} B(X) E|X|^3. \tag{4.14}$$

Proof Let $U_1, \ldots, U_n$ be mutually independent $\mathcal{U}(0,1)$ variables and set

$$(\xi_i, \xi_i^*) = \left(G_i^{-1}(U_i), \left(G_i^*\right)^{-1}(U_i)\right), \quad i = 1, \ldots, n,$$

where $G_1^*, \ldots, G_n^*$ are the distribution functions of $\xi_1^*, \ldots, \xi_n^*$, respectively. Then ξ_i and ξ_i^* have distribution functions G_i and G_i^*, respectively, and by Proposition 4.1,

$$E\left|\xi_i^* - \xi_i\right| = \left\|G_i^* - G_i\right\|_1.$$

Constructing W^* as in Lemma 2.8 yields $W^* - W = \xi_I^* - \xi_I$, with I having distribution $P(I = i) = \sigma_i^2$, so applying Theorem 4.1 we have

$$\begin{aligned}
\|F - \Phi\|_1 &\leq 2E|W^* - W| \\
&= 2E\left|\xi_I^* - \xi_I\right| \\
&= 2\sum_{i=1}^n \sigma_i^2 E\left|\xi_i^* - \xi_i\right| \\
&= 2\sum_{i=1}^n \sigma_i^2 \left\|G_i^* - G_i\right\|_1 \\
&= \sum_{i=1}^n B(\xi_i) E|\xi_i|^3,
\end{aligned}$$

thus proving (4.13).

If $X, X_1, \ldots, X_n$ are i.i.d. with mean zero and variance σ^2 then applying (4.13) with $\xi_i = X_i/(\sigma\sqrt{n})$, and (4.12), yields the bound

$$\|F - \Phi\|_1 \le \frac{1}{\sigma^3 n^{3/2}} \sum_{i=1}^{n} B\left(\frac{X_i}{\sigma\sqrt{n}}\right) E|X_i|^3 = \frac{B(X)E|X|^3}{\sigma^3\sqrt{n}},$$

proving (4.14). □

Specializing (4.14) to particular cases leads to the following corollary.

Corollary 4.1 *When $X = (\xi - p)/\sqrt{pq}$ where ξ has the Bernoulli distribution with success probability $1 - q = p \in (0, 1)$,*

$$B(X) = 1 \quad \textit{and} \quad \|F - \Phi\|_1 \le \frac{E|X|^3}{\sqrt{n}} = \frac{p^2 + q^2}{\sqrt{npq}} \quad \textit{for all } n = 1, 2, \ldots.$$

When X has the uniform distribution $\mathcal{U}[-\sqrt{3}, \sqrt{3}]$, then

$$B(X) = 1/3 \quad \textit{and} \quad \|F - \Phi\|_1 \le \frac{E|X|^3}{3\sqrt{n}} = \frac{\sqrt{3}}{4\sqrt{n}} \quad \textit{for all } n = 1, 2, \ldots.$$

Proof In the Bernoulli case, by (2.55), X^* has the uniform distribution function

$$G^*(x) = \sqrt{pq}x + p \quad \text{for } x \in \left[\frac{-p}{\sqrt{pq}}, \frac{q}{\sqrt{pq}}\right],$$

that is, $X^* =_d (U - p)/\sqrt{pq}$, where $U \sim \mathcal{U}[0, 1]$. Hence, by Proposition 4.1,

$$\|G^* - G\|_1 = \left\| \frac{U - p}{\sqrt{pq}} - \frac{\xi - p}{\sqrt{pq}} \right\|_1 = \frac{1}{\sqrt{pq}} \|U - \xi\|_1 = \frac{p^2 + q^2}{2\sqrt{pq}}.$$

Calculating $E|X|^3 = (p^2 + q^2)/\sqrt{pq}$ and using $\mathrm{Var}(X) = 1$ gives $B(X) = 1$, and the claimed bound.

For the uniform distribution $\mathcal{U}[-\sqrt{3}, \sqrt{3}]$, (2.55) yields

$$G^*(x) = -\frac{\sqrt{3}x^3}{36} + \frac{\sqrt{3}x}{4} + \frac{1}{2} \quad \text{for } x \in [-\sqrt{3}, \sqrt{3}]$$

and from (4.4) we obtain

$$\|G^* - G\|_1 = \frac{\sqrt{3}}{8}.$$

Calculating $E|X|^3 = 3\sqrt{3}/4$ now gives $B(X) = 1/3$, and the claimed bound. □

Constants $B(X)$ and bounds for other distributions may be calculated in a similar fashion. A universal L^1 constant over a class of distributions $\mathcal{F}$, by Theorem 4.2, is given by

$$B(\mathcal{F}) = \sup_{\mathcal{L}(X) \in \mathcal{F}} B(X).$$

The following result, by Goldstein (2010a) and Tyurin (2010), shows that the Bernoulli distribution achieves the worst case $B(X)$.

Theorem 4.3 *For $\sigma > 0$ let $\mathcal{F}_\sigma$ be the collection of all mean zero distributions with variance σ^2 and finite absolute third moment. Then*

$$B(\mathcal{F}) = 1 \quad \text{where } \mathcal{F} = \bigcup_{\sigma>0} \mathcal{F}_\sigma.$$

Theorems 4.3 and 4.2 immediately give

Corollary 4.2 *If $\xi_i, i = 1, \ldots, n$ are independent mean zero random variables with variances $\sigma_i^2 = \mathrm{Var}(\xi_i)$ satisfying $\sum_{i=1}^n \sigma_i^2 = 1$ and $W = \xi_1 + \cdots + \xi_n$, then*

$$\|F - \Phi\|_1 \le \sum_{i=1}^n E|\xi_i|^3.$$

In particular, if $W = n^{-1/2} \sum X_i$ with $X, X_1, \ldots, X_n$ i.i.d. variables with mean zero and variance 1, then

$$\|F - \Phi\|_1 \le \frac{E|X|^3}{\sqrt{n}}.$$

Though it may be difficult to achieve the optimal L^1 coupling between X and X^* in particular applications, especially those involving dependence, the following proposition shows how to construct a coupling which results in a constant bounded by 1 when X is symmetric. Proposition 4.2 is applied in Theorem 4.7 to improve the leading constant in Goldstein (2007) for projections of cone measure.

Proposition 4.2 *Let χ be a random variable with a symmetric distribution, variance $\sigma^2 \in (0, \infty)$ and finite third absolute moment. Let $\overline{X}$ and $\overline{Y}$ be constructed on a joint space with $0 \le \overline{X} \le \overline{Y}$ a.s. having marginal distributions given by $\overline{X} =_d |\chi|$ and $\overline{Y} =_d |\chi^\square|$, where $\chi^\square$ is as defined in Proposition 2.3. Let $V \sim \mathcal{U}[0, 1]$ and ϵ take the values 1 and -1 with equal probability, and be independent of each other and of $\overline{X}$ and $\overline{Y}$. Then $X = \epsilon \overline{X}$ has distribution χ, the variable*

$$X^* = \epsilon V \overline{Y}$$

has the χ-zero biased distribution, and

$$\frac{2\sigma^2 E|X^* - X|}{E|X|^3} \le 1. \tag{4.15}$$

Proof That $X =_d \chi$ follows by the symmetry of χ. Again, by the symmetry of χ,

$$\sigma^2 Ef(\chi^\square) = E\big[\chi^2 f(\chi)\big] = E\big[(-\chi)^2 f(-\chi)\big] = E\big[\chi^2 f(-\chi)\big] = \sigma^2 Ef(-\chi^\square).$$

Hence $\chi^\square$ is symmetric, and as $\epsilon V \sim \mathcal{U}[-1, 1]$ and is independent of $\overline{Y}$, by Proposition 2.3,

$$X^* = \epsilon V \overline{Y} =_d \epsilon V \chi^\square =_d \chi^*.$$

Now,

$$E|X^* - X| = E|\epsilon V\overline{Y} - \epsilon\overline{X}| = E|V\overline{Y} - \overline{X}| = \int_{x\geq 0, y>0}\int_0^1 |vy - x| dv dF(x, y)$$

where $dF(x, y)$ is the joint distribution of $(\overline{X}, \overline{Y})$. Since $dF(x, y)$ is zero on sets where $x > y$, we may decompose the integral above as

$$\begin{aligned}
&\int_{x\geq 0, y>0}\int_{x/y<v\leq 1} (vy - x) dv dF(x, y) + \int_{x\geq 0, y>0}\int_{0<v<x/y} (x - vy) dv dF(x, y) \\
&\quad = \int_{x\geq 0, y>0}\left(\frac{1}{2}y\left(1 - \left(\frac{x}{y}\right)^2\right) - x\left(1 - \frac{x}{y}\right)\right) dF(x, y) \\
&\qquad + \int_{x\geq 0, y>0}\left(x\left(\frac{x}{y}\right) - \frac{1}{2}y\left(\frac{x}{y}\right)^2\right) dF(x, y) \\
&\quad = E\left(\frac{1}{2}\overline{Y} - \overline{X} + \frac{\overline{X}^2}{\overline{Y}}\right).
\end{aligned}$$

As $\overline{X}/\overline{Y} \leq 1$, we have $\overline{X}^2/\overline{Y} \leq \overline{X}$, and therefore

$$E|X^* - X| \leq \frac{1}{2}E\overline{Y} = \frac{1}{2}E\left|\chi^{\square}\right| = \frac{1}{2\sigma^2}E|X|^3.$$

Substituting into (4.15) yields the desired inequality. □

Let X be any random variable with mean zero and variance σ^2, and let ϕ be an increasing function on $[0, \infty)$. Since x^2 is an increasing function on $[0, \infty)$, X^2 will be positively correlated with $\phi(|X|)$, that is,

$$\sigma^2 E\phi\left(\left|X^{\square}\right|\right) = EX^2\phi\left(|X|\right) \geq EX^2 E\phi\left(|X|\right) = \sigma^2 E\phi\left(|X|\right),$$

showing $|X^{\square}|$ is stochastically larger than $|X|$. Hence there always exists a coupling where $|X^{\square}| \geq |X|$ a.s., even when X is not symmetric. Though an optimal L^1 coupling is similarly assured, in principle, by Proposition 4.1, couplings constructed by following Proposition 4.2 seem to be of more practical use; see in particular where this proposition is applied for cone measure in item 3 of Proposition 4.5.

4.1.2 Contraction Principle

In this section we show that the distribution of a standardized sum of i.i.d. variables is closer in L^1 to the normal, in a zero bias sense, than the distribution of the summands themselves. This result leads to a type of L^1 contraction principle for the CLT. For some additional generality we will consider weighted averages of i.i.d. random variable.

Let $\|\boldsymbol{\alpha}\|$ denote the Euclidean norm of a vector $\boldsymbol{\alpha} \in \mathbb{R}^k$, and when $\boldsymbol{\alpha}$ is nonzero let

$$\varphi(\boldsymbol{\alpha}) = \frac{\sum_{i=1}^k |\alpha_i|^3}{(\sum_{i=1}^k \alpha_i^2)^{3/2}}. \tag{4.16}$$

Inequality (4.17) of Lemma 4.1 says that taking weighted averages of i.i.d. variables is a contraction in the L^1 distance to normal in a zero biased sense.

Lemma 4.1 *For $\boldsymbol{\alpha} \in \mathbb{R}^k$ with $\lambda = \|\boldsymbol{\alpha}\| \neq 0$, let*

$$Y = \sum_{i=1}^k \frac{\alpha_i}{\lambda} W_i,$$

where W_i are mean zero, variance one, independent random variables distributed as W. Then

$$\big\|\mathcal{L}(Y^*) - \mathcal{L}(Y)\big\|_1 \leq \varphi \big\|\mathcal{L}(W^*) - \mathcal{L}(W)\big\|_1 \tag{4.17}$$

with $\varphi = \varphi(\boldsymbol{\alpha})$ as in (4.16), and $\varphi < 1$ if and only if $\boldsymbol{\alpha}$ is not a multiple of a standard basis vector.

If W_0 is any mean zero, variance one random variable with finite absolute third moment, $\boldsymbol{\alpha}_n$, $n = 0, 1, \ldots$ a sequence of nonzero vectors in $\mathbb{R}^k$, $\lambda_n = \|\boldsymbol{\alpha}_n\|$, $\varphi_n = \varphi(\boldsymbol{\alpha}_n)$, and

$$W_{n+1} = \sum_{i=1}^k \frac{\alpha_{n,i}}{\lambda_n} W_{n,i} \quad \text{for } n = 0, 1, \ldots \tag{4.18}$$

where $W_{n,i}$ are i.i.d. copies of W_n, then

$$\big\|\mathcal{L}\big(W_n^*\big) - \mathcal{L}(W_n)\big\|_1 \leq \left(\prod_{j=0}^{n-1} \varphi_j\right). \tag{4.19}$$

If $\limsup_n \varphi_n = \varphi < 1$, then for any $\gamma \in (\varphi, 1)$ there exists C such that

$$\big\|\mathcal{L}(W_n) - \mathcal{L}(Z)\big\|_1 \leq C\gamma^n \quad \text{for all } n, \tag{4.20}$$

while if $\boldsymbol{\alpha}_n = \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha}$ and all n, then

$$\big\|\mathcal{L}(W_n) - \mathcal{L}(Z)\big\|_1 \leq 2\varphi^n \quad \text{for all } n, \tag{4.21}$$

with $\varphi = \varphi(\boldsymbol{\alpha})$.

We begin the proof of the lemma by studying how φ behaves in terms of $\boldsymbol{\alpha}$, and prove a bit more than we need now, saving the additional results for use in Sect. 4.2.

Lemma 4.2 *For $\boldsymbol{\alpha} \in \mathbb{R}^k$ with $\lambda = \|\boldsymbol{\alpha}\| \neq 0$,*

$$\sum_{i=1}^k \frac{|\alpha_i|^p}{\lambda^p} \leq 1 \quad \text{for all } p > 2, \tag{4.22}$$

with equality if and only if $\boldsymbol{\alpha}$ is a multiple of a standard basis vector. With φ as in (4.16),

$$\frac{1}{\sqrt{k}} \le \varphi \le 1, \tag{4.23}$$

with equality to the upper bound if and only if $\boldsymbol{\alpha}$ *is a multiple of a standard basis vector, and equality to the lower bound if and only if* $|\alpha_i| = |\alpha_j|$ *for all* i, j.

In addition, when $\alpha_i \ge 0$ *and* $\sum_{i=1}^n \alpha_i = 1$ *then*

$$\lambda \le \varphi, \tag{4.24}$$

with equality if and only if $\boldsymbol{\alpha}$ *is equal to a standard basis vector.*

Proof Since $|\alpha_i|/\lambda \le 1$ we have $|\alpha_i|^{p-2}/\lambda^{p-2} \le 1$, yielding

$$\sum_{i=1}^k \frac{|\alpha_i|^p}{\lambda^p} = \sum_{i=1}^k \left(\frac{|\alpha_i|^{p-2}}{\lambda^{p-2}} \right) \frac{\alpha_i^2}{\lambda^2} \le \sum_{i=1}^k \frac{\alpha_i^2}{\lambda^2} = 1,$$

with equality if and only if $|\alpha_i| = \lambda$ for some i and $\alpha_j = 0$ for all $j \ne i$. Specializing to the case $p = 3$ yields the claims about the upper bound in (4.23).

By Hölder's inequality with $p = 3, q = 3/2$, we have

$$\left(\sum_{i=1}^k \alpha_i^2 \right)^{3/2} = \left(\sum_{i=1}^k 1 \cdot \alpha_i^2 \right)^{3/2} \le \sqrt{k} \sum_{i=1}^k |\alpha_i|^3,$$

giving the lower bound (4.23), with equality if and only if α_i^2 is proportional to 1 for all i.

The claim (4.24) follows from the inequality

$$(EY)^2 \le EY^2 \quad \text{when } P(Y = \alpha_i) = \alpha_i,$$

which is an equality if and only if the variable Y is constant. □

We may now proceed to the proof of the lemma.

Proof of Lemma 4.1 Let F_{W^*} and F_W be the distribution functions of W^* and W, respectively, and with $U_1, \ldots, U_n$ independent $\mathcal{U}[0, 1]$ variables let

$$\left(W_i^*, W_i\right) = \left(F_{W^*}^{-1}(U_i), F_W^{-1}(U_i)\right) \quad i = 1, \ldots, n.$$

By Proposition 4.1, $E|W_i^* - W_i| = \|\mathcal{L}(W^*) - \mathcal{L}(W)\|_1$ for all $i = 1, \ldots, n$.

By Lemma 2.8 and (2.59), with I a random index independent of all other variables with distribution

$$P(I = i) = \frac{\alpha_i^2}{\lambda^2},$$

the variable

$$Y^* = Y - \frac{\alpha_I}{\lambda}\left(W_I - W_I^*\right) \tag{4.25}$$

has the Y-zero biased distribution. Using (4.6) for the first inequality, we now obtain (4.17) by

$$\begin{aligned}\big\|\mathcal{L}(Y^*) - \mathcal{L}(Y)\big\|_1 &\le E|Y^* - Y| \\ &= E\sum_{i=1}^{k} \frac{|\alpha_i|}{\lambda}\big|W_i^* - W_i\big|\mathbf{1}(I=i) \\ &= \sum_{i=1}^{k} \frac{|\alpha_i|^3}{\lambda^3} E\big|W_i^* - W_i\big| \\ &= \varphi\big\|\mathcal{L}(W) - \mathcal{L}(W^*)\big\|_1.\end{aligned}$$

That $\varphi < 1$ if and only if $\boldsymbol{\alpha}$ is not a multiple of a standard basis vector was shown in Lemma 4.2.

To obtain (4.19), note that induction and (4.17) yield

$$\big\|\mathcal{L}\big(W_n^*\big) - \mathcal{L}(W_n)\big\|_1 \le \left(\prod_{j=0}^{n-1} \varphi_j\right)\big\|\mathcal{L}\big(W_0^*\big) - \mathcal{L}(W_0)\big\|_1,$$

and $\|\mathcal{L}(W_0^*) - \mathcal{L}(W_0)\|_1 \le 1$ by Theorem 4.3.

When $\limsup_n \varphi_n = \varphi < \gamma < 1$ there exists n_0 such that

$$\varphi_j \le \gamma \quad \text{for all } j \ge n_0.$$

Hence, for all $n \ge n_0$

$$\prod_{j=0}^{n-1} \varphi_j = \left(\prod_{j=0}^{n_0-1} \frac{\varphi_j}{\gamma}\right)\gamma^{n_0} \prod_{j=n_0}^{n-1} \varphi_j \le \left(\prod_{j=0}^{n_0-1} \frac{\varphi_j}{\gamma}\right)\gamma^{n}.$$

The bound (4.20) now follows from this inequality and Theorem 4.1.

The last claim (4.21) is immediate from (4.19) and Theorem 4.1. □

We note that the standardized, classical case (4.14) is recovered from (4.17) and Theorem 4.1 when $\alpha_i = 1/\sqrt{n}$. In Sect. 4.2 we study nonlinear versions of recursion (4.18) with applications to physical models.

4.2 Hierarchical Structures

For $k \ge 2$ an integer, $\mathcal{D} \subset \mathbb{R}$, and $F : \mathcal{D}^k \to \mathcal{D}$ a given function, every distribution for a random variable X_0 with $P(X_0 \in \mathcal{D}) = 1$ generates the sequence of 'hierarchical' distributions through the recursion

$$X_{n+1} = F(\mathbf{X}_n), \quad n \ge 0, \tag{4.26}$$

where $\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,k})^\intercal$ with $X_{n,i}$ independent, each with distribution X_n. Such hierarchical variables have been considered extensively in the physics literature (see Li and Rogers 1999 and the references therein), in particular to model conductivity of random media.

The special case where the function F is determined by the conductivity properties of the diamond lattice has been considered in Griffiths and Kaufman (1982) and Schlösser and Spohn (1992). Figure 4.1 shows the progression of the diamond lattice from large to small scale. At the large scale (a), the conductivity of the system can be measured along the bond connecting its top and bottom nodes. Inspection of the lattice on a finer scale reveals that this bond is actually comprised of four smaller bonds, each similar to (a), connected as shown in (b). Inspection on an even finer scaler reveals that each of the four bonds in (b) are constructed in a self-similar way from bonds at a smaller level, giving the successive diagram (c), and so on.

To determine the conductivity function F associated with a given lattice, first recall that conductances add in parallel, that is, if two components with conductances x_1 and x_2 are placed in parallel, then the net conductance of the system is

$$L_1(x_1, x_2) = x_1 + x_2. \tag{4.27}$$

Similarly, resistances add for components placed in series. Hence, for these same two components in series, as resistance and conductance are inverses, the resulting conductance of the system is

$$L_{-1}(x_1, x_2) = \left(x_1^{-1} + x_2^{-1}\right)^{-1}. \tag{4.28}$$

For the diamond lattice in particular, assume that each bond has a fixed 'baseline' conductivity characteristic $w \geq 0$ such that when a component with conductivity $x \geq 0$ is present along the bond its net conductivity is wx. For bonds in the diamond lattice as in (b), we associate conductivities characteristics $\mathbf{w} = (w_1, w_2, w_3, w_4)^\intercal$, numbering bonds from the top and proceeding counter-clockwise. Hence, if $\mathbf{x} = (x_1, x_2, x_3, x_4)^\intercal$ are the conductances of four elements each as in (a) which are present along the bonds in (b), then the two components in series on the left side have conductance $L_{-1}(w_1x_1, w_2x_2)$, and similarly, the conductance for the two components in series on the right is $L_{-1}(w_3x_3, w_4x_4)$. Combining these two subsystems in parallel gives

$$F(\mathbf{x}) = L_1\big(L_{-1}(w_1x_1, w_2x_2), L_{-1}(w_3x_3, w_4x_4)\big), \tag{4.29}$$

that is,

$$F(\mathbf{x}) = \left(\frac{1}{w_1x_1} + \frac{1}{w_2x_2}\right)^{-1} + \left(\frac{1}{w_3x_3} + \frac{1}{w_4x_4}\right)^{-1}. \tag{4.30}$$

Returning to the sequence of distributions generated by the recursion (4.26), conditions on F which imply the weak law

$$X_n \to_p c \tag{4.31}$$

for some constant c have been considered by various authors. Recall that we say F is homogeneous, or positively homogeneous, if

$$F(ax_1, \ldots, ax_k) = a^k F(x_1, \ldots, x_k)$$

hold for all $a \in \mathbb{R}$, or all $a > 0$, respectively. Shneiberg (1986) proves that (4.31) holds if $\mathcal{D} = [a, b]$ and F is continuous, monotonically increasing, positively homogeneous, convex and satisfies the normalization condition $F(\mathbf{1}_k) = 1$ where $\mathbf{1}_k$

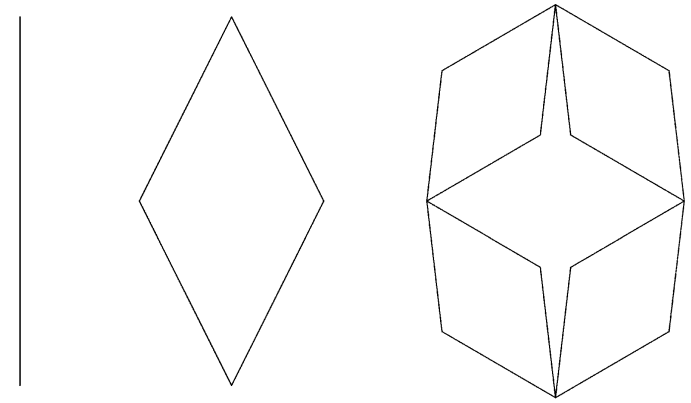

Fig. 4.1 The diamond lattice

is the vector of all ones in $\mathbb{R}^k$. Li and Rogers (1999) provide rather weak conditions under which (4.31) holds for closed $\mathcal{D} \subset (-\infty, \infty)$. See also Wehr (1997) and Wehr (2001), and Jordan (2002) for an extension of the model to random F and applications of hierarchical structures to computer science.

Letting X_0 have mean c and variance σ^2, the classical central limit theorem can be set in the framework of hierarchical sequences by letting

$$F(x_1, x_2) = \frac{1}{2}(x_1 + x_2), \tag{4.32}$$

which gives

$$X_n =_d \frac{X_{0,1} + \cdots + X_{0,2^n}}{2^n} \tag{4.33}$$

where $X_{0,m}, m = 1, \ldots, 2^n$ are independent and identically distributed as X_0. Hence, $X_n \to_p c$ by the weak law of large numbers, and since X_n is the average of $N = 2^n$ i.i.d. variables with finite variance, we have additionally that

$$W_n = \sqrt{N}\left(\frac{X_n - c}{\sigma}\right) \to_d \mathcal{N}(0, 1).$$

Moreover, when X_0 has a bounded absolute third moment (4.21) yields

$$\left\| \mathcal{L}(W_n) - \mathcal{L}(Z) \right\|_1 \le C\gamma^n \tag{4.34}$$

with $C = 2$ and $\gamma = 1/\sqrt{2}$.

The function F in (4.32) is a simple average, and one would, therefore, expect normal limiting behavior more generally when the function F averages its inputs in some sense.

Definition 4.1 We say that $F : \mathcal{D}^k \to \mathcal{D}$ is an averaging function when it satisfies the following three properties on its domain:

1. $\min_i x_i \le F(\mathbf{x}) \le \max_i x_i$.
2. $F(\mathbf{x}) \le F(\mathbf{y})$ whenever $x_i \le y_i$.

3. For all $x < y$ and for any two distinct indices $i_1 \neq i_2$, there exists $x_i \in \{x, y\}$, $i = 1, \ldots, k$ such that $x_{i_1} = x, x_{i_2} = y$ and $x < F(\mathbf{x}) < y$.

We say F is strictly averaging if F satisfies Properties 1 and 2 with strict inequality when $\min_i x_i < \max_i x_i$, and when $x_i < y_i$ for some i, respectively.

Properties 1 and 2 say that the 'average' returned by F should lie inbetween the values being 'averaged' and that that 'average' increases with those values. Note that Property 1 says that for F to be an averaging function it is necessary that $F(\mathbf{1}_k) = 1$. Property 3 says that F is sensitive, that is, depends on, all of its coordinates. We note that if F is strictly averaging then F satisfies Property 3 thusly: if $x < y$ and $x_{i_i} = x, x_{i_2} = y$, then any assignment of the values x, y to the remaining coordinates gives $x < F(\mathbf{x}) < y$ by the strict form of Property 1. Hence all strictly averaging functions are averaging. We note that the function $F(\mathbf{x}) = \min_i x_i$ satisfies the first two properties but not the third, and it gives rise to extreme value, rather than normal, limiting behavior.

Normal limits are proved by Wehr and Woo (2001) for sequences X_n, $n = 0, 1, \ldots$ determined by the recursion (4.26) when the function $F(\mathbf{x})$ is averaging by showing that such recursions can be treated as the approximate linear recursion around the mean $c_n = EX_n$ with small perturbation Z_n,

$$X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n + Z_n, \quad n \geq 0, \tag{4.35}$$

where $\boldsymbol{\alpha}_n = F'(\mathbf{c}_n)$, the gradient of F at $\mathbf{c}_n$ where $\mathbf{c}_n = (c_n, \ldots, c_n)^\intercal \in \mathbb{R}^k$. In Sect. 4.2.1 we prove Theorem 4.6, which gives the bound (4.34) for the L^1 distance to the normal for sequences generated by the approximate linear recursion (4.35) under Conditions 4.1 and 4.2, which guarantee that Z_n is small relative to X_n.

In Sect. 4.2.2 we prove Theorem 4.4 which shows that the normal convergence of the hierarchical sequence $X_n, n = 0, 1, \ldots$ holds with bound (4.34) under mild conditions, and specifies the exponential rate γ in an explicit range. Theorem 4.4 is proved by invoking Theorem 4.6 after showing that the required moment conditions are satisfied for a linearization of $X_{n+1} = F(\mathbf{X}_n)$.

Theorem 4.4 *For some $a < b$ let X_0 be a non constant random variable with $P(X_0 \in [a, b]) = 1$ and let*

$$X_{n+1} = F(\mathbf{X}_n), \quad n \geq 0,$$

where $\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,k})^\intercal$ with $X_{n,i}$ independent, each with distribution X_n and $F : [a, b]^k \to [a, b]$, twice continuously differentiable. Suppose F is averaging and that $X_n \to_p c$, with $\boldsymbol{\alpha} = F'(\mathbf{1}_k c)$ not a scalar multiple of a standard basis vector. Then with $W_n = (X_n - c_n)/\sqrt{\mathrm{Var}(X_n)}$ and Z a standard normal variable, for all $\gamma \in (\varphi, 1)$ there exists C such that

$$\left\| \mathcal{L}(W_n) - \mathcal{L}(Z) \right\|_1 \leq C\gamma^n \quad \text{for all } n \geq 0,$$

where φ, given by $\boldsymbol{\alpha}$ through (4.16), is a positive number strictly less than 1. The value φ achieves a minimum of $1/\sqrt{k}$ if and only if the components of $\boldsymbol{\alpha}$ are equal.

As in (4.33), the variable X_n is a function of $N = k^n$ variables, so achieving the rate φ^n exactly corresponds to a 'classical rate' of $N^{-\theta}$ where

$$\varphi^n = N^{-\theta} = k^{-n\theta} \quad \text{or} \quad \theta = -\log_k \varphi. \tag{4.36}$$

Hence when φ achieves its minimum value $1/\sqrt{k}$ we have $\theta = -1/2$ and the rate $N^{-1/2}$, and achieving this rate for all $\gamma > \varphi$ therefore corresponds to the rate $N^{-1/2+\epsilon}$ for every $\epsilon > 0$. Further, when $\boldsymbol{\alpha}$ is close to a standard basis vector, φ is close to 1, so the bound can have rate $N^{-\theta}$ for θ arbitrarily close to zero. This behavior is anticipated: for the simple hierarchical sequence generated by the function $F(x_1, x_2) = (1-\epsilon)x_1 + \epsilon x_2$, convergence to the normal will be slow indeed for small $\epsilon > 0$. The condition in Theorem 4.4 that the gradient $\boldsymbol{\alpha} = F'(\mathbf{c})$ of F at the limiting value $\mathbf{c}$ not be a scalar multiple of a standard basis vector rules out cases which behave in the limit degenerately as $F(x_1, x_2) = x_1$.

The function (4.32), and (4.30) when $F(\mathbf{1}_4) = 1$, are examples of averaging functions. To handle multiples, we say that $G(\mathbf{y})$ with $G(\mathbf{1}_k) \neq 0$ is a scaled averaging function if $G(\mathbf{y})/G(\mathbf{1}_k)$ is averaging. Now suppose that $G(\mathbf{y})$ is scaled averaging and homogeneous, and that

$$Y_{n+1} = G(\mathbf{Y}_n) \quad \text{for } n \geq 0,$$

where Y_0 is a given random variable and $\mathbf{Y}_n \in \mathbb{R}^k$ is a vector of independent copies of Y_n. Then letting

$$a_{n+1} = ka_n + 1 \quad \text{for all } n \geq 0, \text{ and } a_0 = 0,$$

and setting $F(\mathbf{y}) = G(\mathbf{y})/G(\mathbf{1}_k)$, which is an averaging function, and $\mathbf{X}_n = \mathbf{Y}_n/G(\mathbf{1}_k)^{a_n}$ and likewise for X_n, we have

$$\begin{aligned} X_{n+1} &= Y_{n+1}/G(\mathbf{1}_k)^{a_{n+1}} \\ &= G(\mathbf{Y}_n)/G(\mathbf{1}_k)^{a_{n+1}} = F(\mathbf{Y}_n)/G(\mathbf{1}_k)^{ka_n} = F(\mathbf{X}_n). \end{aligned}$$

As the scaled and centered and variables $(X_n - EX_n)/\sqrt{\mathrm{Var}(X_n)}$ and $(Y_n - EY_n)/\sqrt{\mathrm{Var}(Y_n)}$ are equal, the conclusion of Theorem 4.4 holds for Y_n when it holds for X_n.

Theorem 4.4 is applied in Sect. 4.2.3 to the specific hierarchical variables generated by the diamond lattice conductivity function (4.30), and, in (4.67), the value φ determining the range of γ is given as an explicit function of the weights $\mathbf{w}$; for the diamond lattice all rates $N^{-\theta}$ for $\theta \in (0, 1/2)$ are exhibited. Interestingly, there appears to be no such formula, simple or otherwise, for the limiting mean or variance of the sequence X_n.

To proceed we introduce another equivalent formulation of the L^1 distance. With $\mathfrak{L}$ as in (4.7), let

$$\mathcal{F} = \big\{ f\colon\ f \text{ absolutely continuous } f(0) = f'(0) = 0,\ f' \in \mathfrak{L} \big\}. \tag{4.37}$$

Clearly, if $f \in \mathcal{F}$ then $h \in \mathfrak{L}$ for $h = f'$. On the other hand, if $h \in \mathfrak{L}$ then

$$f \in \mathcal{F} \quad \text{and} \quad f'(y) - f'(x) = h(y) - h(x) \quad \text{for } f(x) = \int_0^x \big[h(u) - h(0)\big] du.$$

Then, from (4.8),

$$\left\|\mathcal{L}(Y) - \mathcal{L}(X)\right\|_1 = \sup_{f \in \mathcal{F}} \left|E\left(f'(Y) - f'(X)\right)\right|. \tag{4.38}$$

For the application of Theorem 4.4, it is necessary to verify that the function $F(\mathbf{x})$ in (4.26) is averaging. Proposition 3 of Wehr and Woo (2001) shows that the effective conductance of a resistor network is an averaging function of the conductances of its individual components. Theorem 4.5, which shows that strict averaging is preserved under certain compositions, yields an independent proof that, for instance, (4.30) is strictly averaging under natural scaling and positivity conditions on the weights. In addition, Theorem 4.5 provides an additional source of averaging functions to which Theorem 4.4 may be applied.

Theorem 4.5 *Let $k \geq 1$ and set $I_0 = \{1, \ldots, k\}$. Suppose subsets $I_i \subset I_0$, $i \in I_0$ satisfy $\bigcup_{i \in I_0} I_i = I_0$. For $\mathbf{x} \in \mathbb{R}^k$ and $i \in I_0$ let $\mathbf{x}_i = (x_{j_1}, \ldots, x_{j_{|I_i|}})$ where $\{j_1, \ldots, j_{|I_i|}\} = I_i$ with $j_1 < \cdots < j_{|I_i|}$. Let $F_i : \mathbb{R}^{|I_i|} \to \mathbb{R}$ (or $F_i : [0, \infty)^{|I_i|} \to [0, \infty)$), $i = 0, \ldots, k$. If $F_0, F_1, \ldots, F_k$ are strictly averaging and F_0 is (positively) homogeneous, then the composition*

$$F_{\mathbf{s}}(\mathbf{x}) = F_0\left(s_1 F_1(\mathbf{x}_1), \ldots, s_k F_k(\mathbf{x}_k)\right)$$

is strictly averaging for any $\mathbf{s}$ which satisfies $F_0(\mathbf{s}) = 1$ and $s_i > 0$ for all i. If $F_0, F_1, \ldots, F_k$ are scaled, strictly averaging and F_0 is (positively) homogeneous, then

$$F_{\mathbf{1}}(\mathbf{x}) = F_0\left(F_1(\mathbf{x}_1), \ldots, F_k(\mathbf{x}_k)\right)$$

is a scaled strictly averaging function.

Note that the parallel and series combination rules (4.27) and (4.28) are the $p = 1$ and $p = -1$ special cases, respectively, with $w_i = 1$, of the weighted L^p norm functions

$$L_p^{\mathbf{w}}(\mathbf{x}) = \left(\sum_{i=1}^{k} (w_i x_i)^p\right)^{1/p}, \quad \mathbf{w} = (w_1, \ldots, w_k)^\intercal, \quad w_i \in (0, \infty),$$

which are scaled, strictly averaging, and positively homogeneous on $[0, \infty)^k$ for $p > 0$ and on $(0, \infty)$ for $p < 0$. Since $F(\mathbf{x})$ in (4.30) is represented by the composition (4.29), Theorem 4.5 obtains to show that F is a scaled, strictly averaging function on $(0, \infty)^4$ for any choice of positive weights. In particular, for positive weights such that $F(\mathbf{1}) = 1$, the function F is strictly averaging on $(0, \infty)^4$. Theorem 4.4 requires F to have domain $[a, b]^k$. However, if F is an averaging function on, say, $(0, \infty)^4$, then Property 1 implies that $F : [a, b]^k \to [a, b]$ for all $[a, b] \subset (0, \infty)$, and hence F will be averaging on this smaller domain. Note lastly that Theorem 4.5 shows the same conclusion holds when the resistor parallel L_1 and series L_{-1} combination rules in this network are replaced by, say, L_2 and L_{-2} respectively.

4.2.1 Bounds to the Normal for Approximately Linear Recursions

In this section we study sequences $\{X_n\}_{n\geq 0}$ generated by the approximate linear recursion

$$X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n + Z_n, \quad n \geq 0, \tag{4.39}$$

where X_0 is a given nontrivial random variable and the components $X_{n,1}, \ldots, X_{n,k}$ of $\mathbf{X}_n$ are independent copies of X_n. We present Theorem 4.6 which shows the exponential bound (4.34) holds when the perturbation term Z_n, which measures the departure from linearity, is small. The effective size of Z_n is measured by the quantity β_n of (4.42), which will be small when the moment bounds in Conditions 4.1 and 4.2 are satisfied. When the recursion is nearly linear, X_{n+1} will be approximately equal to $\boldsymbol{\alpha}_n \cdot \mathbf{X}_n$, and therefore its variance σ_{n+1}^2 will be close to $\sigma_n^2 \lambda_n^2$ where $\lambda_n = \|\boldsymbol{\alpha}_n\|$. Iterating, the variance of X_n will grow like a some constant C^2 times $\lambda_{n-1}^2 \cdots \lambda_0^2$, so when $\boldsymbol{\alpha}_n \to \boldsymbol{\alpha}$, like $C^2 \lambda^{2n}$. Condition 4.1 assures that Z_n is small relative to X_n in that its variance grows at a slower rate. This condition was assumed in Wehr and Woo (2001) for deriving a normal limiting law for the standardized sequence generated by (4.39).

Condition 4.1 *The nonzero sequence of vectors $\boldsymbol{\alpha}_n \in \mathbb{R}^k$, $k \geq 2$, converges to $\boldsymbol{\alpha}$, not equal to any multiple of a standard basis vector. With $\lambda = \|\boldsymbol{\alpha}\|$, there exist $0 < \delta_1 < \delta_2 < 1$ and positive constants $C_{X,2}, C_{Z,2}$ such that for all n,*

$$\begin{aligned} \mathrm{Var}(X_n) &\geq C_{X,2}^2 \lambda^{2n} (1-\delta_1)^{2n}, \\ \mathrm{Var}(Z_n) &\leq C_{Z,2}^2 \lambda^{2n} (1-\delta_2)^{2n}. \end{aligned}$$

Bounds on the distance between X_n and the normal can be provided under the following additional conditions on the fourth order moments of X_n and Z_n. Condition 4.2 on the higher order moments is satisfied under the same averaging assumption on F used in Wehr and Woo (2001) to guarantee Condition 4.1 for weak convergence to the normal.

Condition 4.2 *With δ_1 and δ_2 as in Condition 4.1, there exists $\delta_3 \geq 0$ and $\delta_4 \geq 0$ such that*

$$\phi_1 = \frac{(1-\delta_2)(1+\delta_3)^3}{(1-\delta_1)^4} < 1 \quad \text{and} \quad \phi_2 = \left(\frac{1-\delta_4}{1-\delta_1}\right)^2 < 1,$$

and constants $C_{X,4}, C_{Z,4}$ such that

$$\begin{aligned} E(X_n - EX_n)^4 &\leq C_{X,4}^4 \lambda^{4n} (1+\delta_3)^{4n}, \\ E(Z_n - EZ_n)^4 &\leq C_{Z,4}^4 \lambda^{4n} (1-\delta_4)^{4n}. \end{aligned}$$

The following is our main result on L^1 bounds for approximately linear recursions.

Theorem 4.6 *Let X_0 be a random variable with variance $\sigma_0^2 \in (0,\infty)$ and*

$$X_{n+1} = \boldsymbol{\alpha}_n \cdot \mathbf{X}_n + Z_n \quad \text{for } n \geq 0 \tag{4.40}$$

with $\boldsymbol{\alpha}_n \in \mathbb{R}^k$, $\lambda_n = \|\boldsymbol{\alpha}_n\| \neq 0$ and $\mathbf{X}_n$ a vector in $\mathbb{R}^k$ with independent components distributed as X_n with mean c_n and finite, non-zero variance σ_n^2. Set $Y_0 = 0$ and $\mathbf{W}_n = (\mathbf{X}_n - E\mathbf{X}_n)/\sigma_n$, and for $n \geq 0$ let

$$W_n = \frac{X_n - c_n}{\sigma_n}, \qquad Y_{n+1} = \frac{\boldsymbol{\alpha}_n}{\lambda_n} \cdot \mathbf{W}_n, \tag{4.41}$$

and

$$\beta_n = E|W_n - Y_n| + \frac{1}{2} E\left|W_n^3 - Y_n^3\right|. \tag{4.42}$$

If there exist $(\beta, \varphi) \in (0,1)^2$ such that

$$\limsup_{n\to\infty} \frac{\beta_n}{\beta^n} < \infty \tag{4.43}$$

and $\varphi_n = \varphi(\boldsymbol{\alpha}_n)$ in (4.16) satisfies

$$\limsup_{n\to\infty} \varphi_n = \varphi, \tag{4.44}$$

then with $\gamma = \beta$ when $\beta > \varphi$, and for any $\gamma \in (\varphi, 1)$ when $\beta \leq \varphi$, there exists C such that

$$\left\|\mathcal{L}(W_n) - \mathcal{L}(Z)\right\|_1 \leq C\gamma^n \quad \text{for all } n \geq 0. \tag{4.45}$$

Under Conditions 4.1 and 4.2, the bound (4.45) holds for all $\gamma \in (\max(\beta,\varphi), 1)$ with $\beta = \max\{\phi_1, \phi_2\} < 1$ and $\varphi = \sum_{i=1}^k |\alpha_i|^3/\lambda^3 < 1$ where $\boldsymbol{\alpha}$ and λ are the limiting values of $\boldsymbol{\alpha}_n$ and λ_n, respectively.

Proof Let $f \in \mathcal{F}$ with $\mathcal{F}$ given by (4.37). Then f' is absolutely continuous with

$$\left|f''(w)\right| \leq 1, \quad \text{and in addition} \quad \left|f'(w)\right| \leq |w| \quad \text{and} \quad \left|f(w)\right| \leq w^2/2.$$

Letting h be given by

$$h(w) = f'(w) - wf(w) \tag{4.46}$$

we have $Nh = 0$ by Lemma 2.1. Differentiation yields

$$h'(w) = f''(w) - wf'(w) - f(w),$$

and therefore

$$\left|h'(w)\right| \leq \left(1 + \frac{3}{2}w^2\right). \tag{4.47}$$

Letting

$$r_n = \frac{\lambda_n \sigma_n}{\sigma_{n+1}} \quad \text{and} \quad T_n = \frac{\sigma_n}{\sigma_{n+1}}\left(\frac{Z_n - EZ_n}{\sigma_n}\right) \tag{4.48}$$

and using (4.41), write the recursion (4.40) as

$$\begin{aligned} W_{n+1} &= \frac{X_{n+1} - EX_{n+1}}{\sigma_{n+1}} \\ &= \frac{\sigma_n}{\sigma_{n+1}} \left(\boldsymbol{\alpha}_n \cdot \frac{\mathbf{X}_n - E\mathbf{X}_n}{\sigma_n} + \frac{Z_n - EZ_n}{\sigma_n} \right) \\ &= \frac{\sigma_n}{\sigma_{n+1}} \left(\boldsymbol{\alpha}_n \cdot \mathbf{W}_n + \frac{Z_n - EZ_n}{\sigma_n} \right) \\ &= r_n Y_{n+1} + T_n. \end{aligned} \tag{4.49}$$

Now by (4.47) and the definition of β_n in (4.42),

$$E\left|h(W_n) - h(Y_n)\right| = E\left| \int_{Y_n}^{W_n} h'(u)du \right| \le \beta_n.$$

Now by (2.51), that $\mathrm{Var}(W_{n+1}) = 1$, (4.46) and $Nh = 0$, we have

$$\begin{aligned} \left|Ef'(W_{n+1}) - Ef'\left(W_{n+1}^*\right)\right| &= \left|Ef'(W_{n+1}) - W_{n+1}f(W_{n+1})\right| \\ &= \left|Eh(W_{n+1}) - Nh\right| \\ &= \left|E\left(h(W_{n+1}) - h(Y_{n+1}) + h(Y_{n+1}) - Nh\right)\right| \\ &\le \beta_{n+1} + \left|Eh(Y_{n+1}) - Nh\right| \\ &= \beta_{n+1} + \left|E\left(f'\left(Y_{n+1}^*\right) - f'(Y_{n+1})\right)\right| \\ &\le \beta_{n+1} + \left\|Y_{n+1}^* - Y_{n+1}\right\|_1 \quad \text{by (4.38)} \\ &\le \beta_{n+1} + \varphi_n \left\|W_n^* - W_n\right\|_1 \quad \text{by Lemma 4.1.} \end{aligned}$$

Taking supremum over $f \in \mathcal{F}$ on the left hand side, using (4.38) again and letting $d_n = \|W_n^* - W_n\|_1$ we obtain, for all $n \ge 0$,

$$d_{n+1} \le \beta_{n+1} + \varphi_n d_n.$$

Iteration yields that for all $n, n_0 \ge 0$,

$$d_{n_0+n} \le \sum_{j=n_0+1}^{n_0+n} \left(\prod_{i=j}^{n_0+n-1} \varphi_i \right) \beta_j + \left(\prod_{i=n_0}^{n_0+n-1} \varphi_i \right) d_{n_0}. \tag{4.50}$$

Now suppose the bounds (4.43) and (4.44) hold on β_n and φ_n, respectively, and recall the choice of γ. When $\beta > \varphi$ take $\overline{\varphi} \in (\varphi, \beta)$ so that $\varphi < \overline{\varphi} < \beta = \gamma$; when $\beta \le \varphi$ take $\overline{\varphi} \in (\varphi, \gamma)$ so that $\beta \le \varphi < \overline{\varphi} < \gamma$. Then for any $\overline{B} > \limsup_n \beta_n/\beta^n$ there exists n_0 such that for all $n \ge n_0$

$$\beta_n \le \overline{B}\beta^n \quad \text{and} \quad \varphi_n \le \overline{\varphi}.$$

Applying these inequalities in (4.50) and summing yields, for all $n \ge 0$,

$$d_{n+n_0} \le \overline{B}\beta^{n_0+1} \left(\frac{\beta^n - \overline{\varphi}^n}{\beta - \overline{\varphi}} \right) + \overline{\varphi}^n d_{n_0}.$$

Since $\max(\beta, \overline{\varphi}) \le \gamma$, for some C we have that $d_n \le C\gamma^n$ for all $n \ge n_0$, and by enlarging C if necessary, for all $n \ge 0$. Now (4.45) follows from Theorem 4.1.

To prove the final claim under Conditions 4.1 and 4.2 it suffices to show that (4.43) and (4.44) hold with $\beta = \max\{\phi_1, \phi_2\}$ and $\varphi = \sum_{i=1}^{k} |\alpha_i|^3/\lambda^3 < 1$ where $\boldsymbol{\alpha}$ is the limiting value of $\boldsymbol{\alpha}_n$. Lemma 6 of Wehr and Woo (2001) gives that the limit as $n \to \infty$ of $\sigma_n/(\lambda_0 \cdots \lambda_{n-1})$ exists in $(0, \infty)$, and therefore that

$$\lim_{n\to\infty} r_n = 1 \quad \text{and} \quad \lim_{n\to\infty} \frac{\sigma_{n+1}}{\sigma_n} = \lambda. \tag{4.51}$$

Referring to the definition of T_n in (4.48) and using (4.51) and Conditions 4.1 and 4.2, there exist positive constants $C_{T,2}$, $C_{T,4}$ such that

$$\big(E|T_n|\big)^2 \le ET_n^2 = \mathrm{Var}(T_n) = \left(\frac{\sigma_n}{\sigma_{n+1}}\right)^2 \frac{\mathrm{Var}(Z_n)}{\mathrm{Var}(X_n)} \le C_{T,2}^2 \left(\frac{1-\delta_2}{1-\delta_1}\right)^{2n},$$
$$\text{and} \quad ET_n^4 = \left(\frac{\sigma_n}{\sigma_{n+1}}\right)^4 E\left(\frac{Z_n - EZ_n}{\sigma_n}\right)^4 \le C_{T,4}^4 \left(\frac{1-\delta_4}{1-\delta_1}\right)^{4n}.$$

By independence, a simple bound and Condition 4.2 for the second inequality we have

$$\big(E|Y_n|\big)^2 \le EY_n^2 = \mathrm{Var}(Y_n) = 1, \quad \text{and}$$
$$EY_{n+1}^4 \le 6E\left(\frac{X_n - c_n}{\sigma_n}\right)^4 \le 6C_{X,4}^4 \left(\frac{1+\delta_3}{1-\delta_1}\right)^{4n}.$$

Using the recursion (4.39) and writing $\sigma_{Z_n}^2 = \mathrm{Var}(Z_n)$, we have $\sigma_{n+1} \le \lambda_n \sigma_n + \sigma_{Z_n}$ and $\lambda_n \sigma_n \le \sigma_{n+1} + \sigma_{Z_n}$, hence with $C_{r,1} = C_{T,2}$ we have

$$|\lambda_n \sigma_n - \sigma_{n+1}| \le \sigma_{Z_n} \quad \text{so} \quad |r_n - 1| \le C_{r,1} \left(\frac{1-\delta_2}{1-\delta_1}\right)^n.$$

Now, since $|r_n^p - 1| = |(r_n - 1 + 1)^p - 1| \le \sum_{j=1}^{p} \binom{p}{j} |r_n - 1|^j$ and $0 < \delta_1 < \delta_2 < 1$, there are constants $C_{r,p}$ such that

$$\big|r_n^p - 1\big| \le C_{r,p} \left(\frac{1-\delta_2}{1-\delta_1}\right)^n, \quad p = 1, 2, \ldots.$$

Now considering the first term of β_n in (4.42), recalling (4.49),

$$E|W_{n+1} - Y_{n+1}| = E|(r_n - 1)Y_{n+1} + T_n|$$
$$\le |r_n - 1| E|Y_{n+1}| + E|T_n| \le (C_{r,1} + C_{T,2}) \left(\frac{1-\delta_2}{1-\delta_1}\right)^n,$$

which is upper bounded by a constant times ϕ_1^{n+1}.

For the second term of (4.42) we have

$$E\big|W_{n+1}^3 - Y_{n+1}^3\big| = E\big|\big(r_n^3 - 1\big)Y_{n+1}^3 + 3r_n^2 Y_{n+1}^2 T_n + 3r_n Y_{n+1} T_n^2 + T_n^3\big|.$$

Applying the triangle inequality, the first term which results may be bounded as

$$\big|r_n^3 - 1\big| E\big|Y_{n+1}^3\big| \le \big|r_n^3 - 1\big| \big(EY_{n+1}^4\big)^{3/4}$$
$$\le 6^{3/4} C_{r,3} C_{X,4}^3 \left(\frac{(1-\delta_2)(1+\delta_3)^3}{(1-\delta_1)^4}\right)^n,$$

which is smaller than some constant times ϕ_1^{n+1}.

Since $r_n \to 1$ by (4.51), it suffices to bound the next two terms without the factor of r_n. Thus,

$$E\left|Y_{n+1}^2 T_n\right| \le \sqrt{EY_{n+1}^4 ET_n^2} \le 6^{1/2} C_{X,4}^2 C_{T,2} \left(\frac{(1-\delta_2)(1+\delta_3)^2}{(1-\delta_1)^3}\right)^n,$$

which is less than a constant times ϕ_1^{n+1}. Lastly,

$$E\left|Y_{n+1} T_n^2\right| \le \sqrt{EY_{n+1}^2 ET_n^4} \le C_{T,4}^2 \left(\frac{1-\delta_4}{1-\delta_1}\right)^{2n} = C_{T,4}^2 \phi_2^n \quad \text{and}$$

$$E\left|T_n^3\right| \le \left(ET_n^4\right)^{3/4} \le C_{T,4}^3 \left(\frac{1-\delta_4}{1-\delta_1}\right)^{3n} \le C_{T,4}^3 \phi_2^{3n/2}.$$

Hence (4.43) holds with the given β.

Since $\boldsymbol{\alpha}_n \to \boldsymbol{\alpha}$, we have $\varphi_n \to \varphi$, verifying (4.44). Under Condition 4.1, $\boldsymbol{\alpha}$ is not a scalar multiple of a standard basis vector and hence $\varphi < 1$ by Lemma 4.1. As the first part of the theorem shows that (4.43) and (4.44) imply that (4.45) holds for all $\gamma \in (\max(\beta, \varphi), 1)$, the last claim is shown. □

We note that this proof reverses the way in which the Stein equation is typically applied, where h is given and the properties of f are dependent on those assumed for h. In particular, in the proof of Theorem 4.6 the function $f \in \mathcal{F}$ is taken as given, and the function h, whose properties are determined by f through (2.4), plays only an auxiliary role.

4.2.2 *Normal Bounds for Hierarchical Sequences*

The following result, extending Proposition 9 of Wehr and Woo (2001) to higher orders, is used to show that the moment bounds of Conditions 4.1 and 4.2 are satisfied under the hypotheses of Theorem 4.4, allowing Theorem 4.6 to be invoked. The dependence of the constants in (4.53) and (4.54) on ϵ is suppressed for notational simplicity.

Lemma 4.3 *Let the hypotheses of Theorem 4.4 be satisfied for the recursion*

$$X_{n+1} = F(\mathbf{X}_n) \quad \text{for } n \ge 0.$$

With $c_n = EX_n$ and $\boldsymbol{\alpha}_n = F'(c_n)$, define

$$Z_n = F(\mathbf{X}_n) - \boldsymbol{\alpha}_n \cdot \mathbf{X}_n. \tag{4.52}$$

Then with $\boldsymbol{\alpha}$ the limit of $\boldsymbol{\alpha}_n$ and $\lambda = \|\boldsymbol{\alpha}\|$, for any integer $p \ge 1$ and $\epsilon > 0$, there exists constants $C_{X,p}$, $C_{Z,p}$ such that

$$E|Z_n - EZ_n|^p \le C_{Z,p}^p (\lambda + \epsilon)^{2pn} \quad \text{for all } n \ge 0, \tag{4.53}$$

and

$$E|X_n - c_n|^p \le C_{X,p}^p (\lambda + \epsilon)^{pn} \quad \text{for all } n \ge 0. \tag{4.54}$$

Proof Expanding $F(\mathbf{X}_n)$ around the mean $\mathbf{c}_n = \mathbf{1}_k c_n$ of $\mathbf{X}_n$ yields

$$F(\mathbf{X}_n) = F(\mathbf{c}_n) + \sum_{i=1}^{k} \alpha_{n,i}(X_{n,i} - c_n) + R_2(\mathbf{c}_n, \mathbf{X}_n), \tag{4.55}$$

where

$$R_2(\mathbf{c}_n, \mathbf{X}_n) = \sum_{i,j=1}^{k} \int_0^1 (1-t) \frac{\partial^2 F}{\partial x_i \partial x_j}\big(\mathbf{c}_n + t(\mathbf{X}_n - \mathbf{c}_n)\big)(X_{n,i} - c_n)(X_{n,j} - c_n)dt.$$

Since the second partials of F are continuous on the compact set $\mathcal{D} = [a, b]^k$, with $\|\cdot\|$ the supremum norm on $\mathcal{D}$ we have

$$B = \frac{1}{2} \max_{i,j} \left\| \frac{\partial^2 F}{\partial x_i \partial x_j} \right\| < \infty,$$

and therefore

$$\big|R_2(\mathbf{c}_n, \mathbf{X}_n)\big| \le B \sum_{i,j=1}^{k} \big|(X_{n,i} - c_n)(X_{n,j} - c_n)\big|. \tag{4.56}$$

Using (4.52), (4.55) and (4.56), we have for all $p \ge 1$

$$\begin{aligned}
&E|Z_n - EZ_n|^p \\
&\quad = E\left|F(\mathbf{X}_n) - EF(\mathbf{X}_n) - \sum_{i=1}^{k} \alpha_{n,i}(X_{n,i} - c_n)\right|^p \\
&\quad = E\big|F(\mathbf{c}_n) - EF(\mathbf{X}_n) + R_2(\mathbf{c}_n, \mathbf{X}_n)\big|^p \\
&\quad \le 2^{p-1}\left(\big|F(\mathbf{c_n}) - EF(\mathbf{X}_n)\big|^p + B^p E\left(\sum_{i,j}\big|(X_{n,i} - c_n)(X_{n,j} - c_n)\big|\right)^p\right).
\end{aligned} \tag{4.57}$$

For the first term of (4.57), again using (4.56),

$$\begin{aligned}
\big|F(\mathbf{c_n}) - EF(\mathbf{X_n})\big|^p &= \big|ER_2(\mathbf{c}_n, \mathbf{X}_n)\big|^p \\
&\le B^p\left(E\sum_{i,j}\big|(X_{n,i} - c_n)(X_{n,j} - c_n)\big|\right)^p \\
&\le B^p\left(\sum_{i,j} E(X_n - c_n)^2\right)^p \\
&= B^p k^{2p}\big[E(X_n - c_n)^2\big]^p \\
&\le B^p k^{2p} E(X_n - c_n)^{2p},
\end{aligned} \tag{4.58}$$

using Jensen's inequality for the final step.

Similarly, for the second term in (4.57),

$$E\left(\sum_{i,j}\left|(X_{n,i}-c_n)(X_{n,j}-c_n)\right|\right)^p \le k^p E\left(\sum_{i=1}^k (X_{n,i}-c_n)^2\right)^p$$
$$\le k^{2p-1} E\left(\sum_{i=1}^k (X_{n,i}-c_n)^{2p}\right)$$
$$= k^{2p} E(X_n - c_n)^{2p}. \tag{4.59}$$

Applying the bounds (4.58) and (4.59) in (4.57) we obtain for all $p \ge 1$, with $C_p = 2^p B^p k^{2p}$,

$$E|Z_n - EZ_n|^p \le C_p E(X_n - c_n)^{2p}. \tag{4.60}$$

To demonstrate the proposition it therefore suffices to prove (4.54).

Note that since $X_n \to c$ for $X_n \in [a,b]$ the bounded convergence theorem implies that $c_n = EX_n \to c$. Lemma 8 of Wehr and Woo (2001) shows that if $F : [a,b]^k \to [a,b]$ is an averaging function and there exists $c \in [a,b]$ such that $X_n \to_p c$, then

$$\forall \epsilon \in (0,1) \exists M \quad \text{such that} \quad \text{for all } n \ge 0, \quad P\big(|X_n - c| > \epsilon\big) \le M\epsilon^n. \tag{4.61}$$

In particular the large deviation estimate (4.61) holds under the given assumptions, and therefore also with c replaced by c_n.

We now show that if a_n, $n = 0, 1, \ldots$ is a sequence such that for every $\epsilon > 0$ there exists M and $n_0 \ge 0$ such that

$$a_{n+1} \le (\lambda + \epsilon)^p a_n + M(\lambda+\epsilon)^{p(n+1)} \quad \text{for all } n \ge n_0, \tag{4.62}$$

then for all $\epsilon > 0$ there exists C such that

$$a_n \le C(\lambda+\epsilon)^{pn} \quad \text{for all } n \ge 0. \tag{4.63}$$

Let $\epsilon > 0$ be given, and let M and n_0 be such that (4.62) holds with ϵ replaced by $\epsilon/2$. Setting

$$C = \max\left\{\frac{a_{n_0}}{(\lambda+\epsilon)^{n_0}}, \frac{M}{1-(\frac{\lambda+\epsilon/2}{\lambda+\epsilon})^p}\right\},$$

it is trivial that (4.63) holds for $n = n_0$, and a direct induction shows (4.63) holds for all $n \ge n_0$. By increasing C if necessary, we have that (4.63) holds for all $n \ge 0$.

Unqualified statements in the remainder of the proof below involving ϵ and M are to be read to mean that for every $\epsilon > 0$ there exists M such that the statement holds for all n; the values of ϵ and M are not necessarily the same at each occurrence, even from line to line. By (4.61) and that $X_n \in [a,b]$ we have

$$E(X_n - c_n)^{2p} = E\big[(X_n - c_n)^{2p}; |X_n - c_n| \le \epsilon\big] + E\big[(X_n - c_n)^{2p}; |X_n - c_n| > \epsilon\big]$$
$$\le \epsilon^p E|X_n - c_n|^p + M\epsilon^n.$$

From (4.60), this inequality gives that

$$E|Z_n - EZ_n|^p \le \epsilon^p E|X_n - c_n|^p + M\epsilon^n. \tag{4.64}$$

Since for all $\epsilon > 0$ we have

$$\lim_{x\to\infty} (x+1)^p - (1+\epsilon)x^p = -\infty \quad \text{and therefore}$$
$$\sup_{x\ge 0} (x+1)^p - (1+\epsilon)x^p < \infty,$$

substituting $x = |w|/|z|$ when $z \neq 0$ we see that there exists M such that for all w, z we have

$$|w+z|^p \le (1+\epsilon)|w|^p + M|z|^p,$$

noting that the inequality holds trivially with $M = 1$ for $z = 0$. Now applying definition (4.52),

$$E|X_{n+1} - c_{n+1}|^p \le (1+\epsilon)E\left|\sum_{i=1}^k \alpha_{n,i}(X_{n,i} - c_n)\right|^p + ME|Z_n - EZ_n|^p. \tag{4.65}$$

Specializing (4.65) to the case $p = 2$ gives

$$E(X_{n+1} - c_{n+1})^2 \le (\lambda+\epsilon)^2 E(X_n - c_n)^2 + ME(Z_n - EZ_n)^2.$$

Applying (4.64) with $p = 2$ to this inequality yields

$$\begin{aligned} E(X_{n+1} - c_{n+1})^2 &\le (\lambda+\epsilon)^2 E(X_n - c_n)^2 + M\epsilon^{2n+2} \\ &\le (\lambda+\epsilon)^2 E(X_n - c_n)^2 + M(\lambda+\epsilon)^{2(n+1)}. \end{aligned}$$

Hence inequality (4.62), and therefore (4.63), are true for $a_n = E(X_n - c_n)^2$ and $p = 2$, yielding (4.54) for $p = 2$. Now Hölder's inequality shows that (4.54) is also true for $p = 1$.

Now let $p > 2$ be an integer and suppose that (4.54) is true for all integers q, $1 \le q < p$. In expanding the first term in (4.65) we let $\mathbf{p} = (p_1, \ldots, p_k)$ denote a multi-index and $|\mathbf{p}| = \sum_i p_i$. Use the induction hypotheses, and (4.22) of Lemma 4.2 in (4.66), to obtain, with $A_{X,p} = \max_{q<p} C_{X,q}$ and $B^p_{X,p} = k^{p-1}A^p_{X,p}$, that

$$\begin{aligned} &E\left|\sum_{i=1}^k \alpha_{n,i}(X_{n,i} - c_n)\right|^p \\ &\le \sum_{i=1}^k |\alpha_{n,i}|^p E|X_{n,i} - c_n|^p + \sum_{|\mathbf{p}|=p, 0\le p_i<p} \binom{p}{\mathbf{p}} E\prod_{i=1}^k |\alpha_{n,i}|^{p_i}|X_{n,i} - c_n|^{p_i} \\ &\le E|X_n - c_n|^p \sum_{i=1}^k |\alpha_{n,i}|^p + \sum_{|\mathbf{p}|=p, 0\le p_i<p} \binom{p}{\mathbf{p}} \prod_{i=1}^k |\alpha_{n,i}|^{p_i} C^{p_i}_{X,p_i}(\lambda+\epsilon)^{p_i n} \\ &\le E|X_n - c_n|^p \sum_{i=1}^k |\alpha_{n,i}|^p + A^p_{X,p}(\lambda+\epsilon)^{pn} \sum_{|\mathbf{p}|=p} \binom{p}{\mathbf{p}} \prod_{i=1}^k |\alpha_{n,i}|^{p_i} \end{aligned}$$

$$= E|X_n - c_n|^p \sum_{i=1}^{k} |\alpha_{n,i}|^p + A^p_{X,p}(\lambda+\epsilon)^{pn}\left(\sum_{i=1}^{k}|\alpha_{n,i}|\right)^p$$

$$\le \sum_{i=1}^{k} |\alpha_{n,i}|^p \left(E|X_n - c_n|^p + B^p_{X,p}(\lambda+\epsilon)^{pn}\right)$$

$$\le (\lambda+\epsilon)^p E|X_n - c_n|^p + B^p_{X,p}(\lambda+\epsilon)^{p(n+1)}. \tag{4.66}$$

Applying (4.64) and (4.66) in (4.65) gives

$$E|X_{n+1} - c_{n+1}|^p \le (\lambda+\epsilon)^p E|X_n - c_n|^p + M(\lambda+\epsilon)^{p(n+1)},$$

from which we can conclude that (4.63) holds for $a_n = E|X_n - c_n|^p$, completing the induction on p. □

Proof of Theorem 4.4 By Theorem 4.6 it suffices to show that Conditions 4.1 and 4.2 are satisfied for some δ_i, $i = 1, 2, 3, 4$ satisfying $\beta < \varphi$.

By Property 1 of averaging functions, $F(\mathbf{1}_k c) = c$, and differentiation with respect to c yields $\sum_{i=1}^{n} \alpha_i = 1$. By Property 2, monotonicity, $\alpha_i \ge 0$, and (4.24) of Lemma 4.2 yields $0 < \lambda < \varphi < 1$, using that $\boldsymbol{\alpha}$ is not a multiple of a standard basis vector.

Let $\delta_4 \in (1-\varphi, 1-\lambda)$. Since $\delta_4 < 1-\lambda$ we have $\lambda^2 < \lambda(1-\delta_4)$, and therefore there exists $\epsilon > 0$ such that $(\lambda+\epsilon)^2 < \lambda(1-\delta_4)$. By Lemma 4.3, for $p = 2$ and $p = 4$, for this ϵ there exists $C^p_{Z,p}$ such that

$$E(Z_n - EZ_n)^p \le C^p_{Z,p}(\lambda+\epsilon)^{2pn} \le C^p_{Z,p}\lambda^{pn}(1-\delta_4)^{pn}.$$

Hence the fourth and second moment bounds in Conditions 4.1 and 4.2 on Z_n are satisfied with δ_4 and $\delta_2 = \delta_4$, respectively.

Since $1 - \delta_4 < \varphi$ there $\delta_1 \in (0, \delta_2)$ and $\delta_3 > 0$ such that $\eta < \varphi$ where

$$\eta = \frac{(1-\delta_4)(1+\delta_3)^3}{(1-\delta_1)^4}.$$

Proposition 10 of Wehr and Woo (2001) shows that under the assumptions of Theorem 4.4, for every $\epsilon > 0$ there exists $C^2_{X,2}$ such that

$$\mathrm{Var}(X_n) \ge C^2_{X,2}(\lambda-\epsilon)^{2n}.$$

Taking $\epsilon = \lambda\delta_1$, we have $\mathrm{Var}(X_n)$ satisfies the lower bound in Condition 4.1. Applying Lemma 4.3 with $p = 4$ and $\epsilon = \lambda\delta_3$ we see the fourth moment bound on X_n in Condition 4.2 is satisfied.

With these choices for δ_i, $i = 1, \ldots, 4$, as $\eta < \varphi < 1$, we have $\phi_2 < \eta < 1$ and $\phi_1 = \eta < 1$, hence Conditions 4.1 and 4.2 are satisfied. Noting that $\beta = \max\{\phi_1, \phi_2\} = \eta < \varphi$ now completes the proof. □

4.2.3 Convergence Rates for the Diamond Lattice

We now apply Theorem 4.4 to hierarchical sequences generated by the diamond lattice conductivity function $F(\mathbf{x})$ in (4.30). We have already argued that Theorem 4.5 implies that $F(\mathbf{x})$ is strictly averaging on, say $[a, b]^4$, for any $0 < a < b$ and choice of positive weights satisfying $F(\mathbf{1}_4) = 1$, and on this domain such an $F(\mathbf{x})$ is easily seen to be twice continuously differentiable. For all such $F(\mathbf{x})$ the result of Shneiberg (1986) quoted in Sect. 4.2 shows that X_n satisfies a weak law.

We now study the quantity φ which determines the exponential decay rate of the upper bound of Theorem 4.4 to zero. The first partial derivative $\partial F(\mathbf{x})/\partial x_1$ has the form

$$\frac{\partial F(\mathbf{x})}{\partial x_1} = \frac{(w_1 x_1^2)^{-1}}{((w_1 x_1)^{-1} + (w_2 x_2)^{-1})^2},$$

and similarly for the other partials. Hence $F'(t\mathbf{x}) = F'(\mathbf{x})$ for all $t \neq 0$. As X_n is a random variable on $[a, b]$ we have $c_n = EX_n \neq 0$, and therefore

$$\boldsymbol{\alpha}_n = F'(c_n \mathbf{1}_4) = F'(\mathbf{1}_4) \quad \text{for all } n \geq 0.$$

In particular, $\boldsymbol{\alpha} = \lim_{n\to\infty} \boldsymbol{\alpha}_n$ is given by

$$\boldsymbol{\alpha} = \left[\frac{w_1^{-1}}{(w_1^{-1} + w_2^{-1})^2}, \frac{w_2^{-1}}{(w_1^{-1} + w_2^{-1})^2}, \frac{w_3^{-1}}{(w_3^{-1} + w_4^{-1})^2}, \frac{w_4^{-1}}{(w_3^{-1} + w_4^{-1})^2} \right]^{\mathsf{T}}.$$

Since we are considering the case where all the weights are positive, the vector $\boldsymbol{\alpha}$ is not a scalar multiple of a standard basis vector. Now from (4.16) we compute

$$\varphi = \lambda^{-3} \left(\frac{w_1^{-3} + w_2^{-3}}{(w_1^{-1} + w_2^{-1})^6} + \frac{w_3^{-3} + w_4^{-3}}{(w_3^{-1} + w_4^{-1})^6} \right), \tag{4.67}$$

where

$$\lambda = \left(\frac{w_1^{-2} + w_2^{-2}}{(w_1^{-1} + w_2^{-1})^4} + \frac{w_3^{-2} + w_4^{-2}}{(w_3^{-1} + w_4^{-1})^4} \right)^{1/2}.$$

As an illustration of the bounds provided by Theorem 4.4, first consider the 'side equally weighted network', the one with $\mathbf{w} = (w, w, 2-w, 2-w)^{\mathsf{T}}$ for $w \in [1, 2)$; we recall the weights $\mathbf{w}$ refer to the bonds in the lattice traversed counterclockwise from the top in Fig. 4.1(c). The vector of weights for w in this range are positive and satisfy $F(\mathbf{1}_4) = 1$. For $w = 1$ all weights are equal and $\boldsymbol{\alpha} = 4^{-1}\mathbf{1}_4$, so φ achieves its minimum value $1/2 = 1/\sqrt{k}$ with $k = 4$. By Theorem 4.4, for all $\gamma \in (1/2, 1)$ there exists a constant C such that $\|W_n - Z\|_1 \leq C\gamma^n$. The values of γ just above $1/2$ correspond, in view of (4.36), to the rate $N^{-\theta}$ for θ just below $-\log_4 1/2 = 1/2$, that is, $N^{-1/2+\epsilon}$ for small $\epsilon > 0$, where $N = 4^n$, the number of variables at stage n. As w increases from 1 to 2, φ increases continuously from $1/2$ to $1/\sqrt{2}$, with w approaching 2 from below corresponding to the least favorable rate for the side equally weighted network of θ just under $-\log_4 1/\sqrt{2} = 1/4$, that is, of $N^{-1/4+\epsilon}$ for any $\epsilon > 0$.

With only the restriction that the weights are positive and satisfy $F(\mathbf{1}_4) = 1$ consider

$$\mathbf{w} = (1 + 1/t, s, t, 1/t)^\mathsf{T}$$
$$\text{where } s = \left[\left(1 - (1/t + t)^{-1}\right)^{-1} - (1 + 1/t)^{-1}\right]^{-1}, \quad t > 0.$$

When $t = 1$ we have $s = 2/3$ and $\varphi = 11\sqrt{2}/27$. As $t \to \infty$, $s/t \to 1/2$ and $\boldsymbol{\alpha}$ tends to the standard basis vector $(1, 0, 0, 0)$, so $\varphi \to 1$. Since $11\sqrt{2}/27 \in (1/2, 1/\sqrt{2})$, the above two examples show that the value of γ given by Theorem 4.4 for the diamond lattice can take any value in the range $(1/2, 1)$, corresponding to $N^{-\theta}$ for any $\theta \in (0, 1/2)$.

4.3 Cone Measure Projections

In this section we use Stein's method to obtain L^1 bounds for the normal approximation of one dimensional projections of the form

$$Y = \boldsymbol{\theta} \cdot \mathbf{X}, \tag{4.68}$$

where for some $p > 0$, the vector $\mathbf{X} \in \mathbb{R}^n$ has the cone measure distribution $\mathcal{C}_p^n$ given in (4.71) below, and $\boldsymbol{\theta} \in \mathbb{R}^n$ is of unit length. The normal approximation of projections of random vectors in lesser and greater generality has been studied by many authors, and under a variety of metrics. In the case $p = 2$, when cone measure is uniform on the surface of the unit Euclidean sphere in $\mathbb{R}^n$, Diaconis and Freedman (1987) show that the low dimensional projections of $\mathbf{X}$ are close to normal in total variation. It is particularly easy to see in this case, and true in general, that cone measure $\mathcal{C}_p^n$ is *coordinate symmetric*, that is,

$$(X_1, \ldots, X_n) =_d (e_1 X_1, \ldots, e_n X_n) \quad \text{for all } (e_1, \ldots, e_n) \in \{-1, 1\}^n. \tag{4.69}$$

Meckes and Meckes (2007) derive bounds using Stein's method for the normal approximation of random vectors with symmetries in general, including coordinate-symmetry, considering the supremum and total variation norm. Goldstein and Shao (2009) give L^∞ bounds on the projections of coordinate symmetric random vectors of order $1/\sqrt{n}$ without applying Stein's method. Klartag (2009) proves bounds of order $1/n$ on the L^∞ distance under additional conditions on the distribution of $\mathbf{X}$, including that its density be log concave. One special case of note where $\mathbf{X}$ is coordinate symmetric is when its distribution is uniform over a convex set which has symmetry with respect to all coordinate planes. For general results on the projections of vectors sampled uniformly from convex sets, see Klartag (2007) and references therein. Studying here the specific instance of the projections of cone measure allows, naturally, for the sharpening of general results about projections of coordinate symmetric vectors to this particular case.

To define cone measure let

$$\begin{aligned} S(\ell_p^n) &= \left\{\mathbf{x} \in \mathbb{R}^n \colon \sum_{i=1}^n |x_i|^p = 1\right\} \quad \text{and} \\ B(\ell_p^n) &= \left\{\mathbf{x} \in \mathbb{R}^n \colon \sum_{i=1}^n |x_i|^p \le 1\right\}. \end{aligned} \tag{4.70}$$

Then with μ^n Lebesgue measure in $\mathbb{R}^n$, the cone measure of $A \subset S(\ell_p^n)$ is given by

$$\mathcal{C}_p^n(A) = \frac{\mu^n([0,1]A)}{\mu^n(B(\ell_p^n))} \quad \text{where } [0,1]A = \{ta\colon\ a \in A,\ 0 \le t \le 1\}. \tag{4.71}$$

The main result in the this section on the projections of $\mathcal{C}_p^n$ is the following.

Theorem 4.7 *Let* $\mathbf{X}$ *have cone measure* $\mathcal{C}_p^n$ *on the sphere* $S(\ell_p^n)$ *for some* $p > 0$ *and let*

$$Y = \sum_{i=1}^{n} \theta_i X_i$$

be the one-dimensional projection of $\mathbf{X}$ *along the direction* $\boldsymbol{\theta} \in \mathbb{R}^n$ *with* $\|\boldsymbol{\theta}\| = 1$. *Then with* $\sigma_{n,p}^2 = \mathrm{Var}(X_1)$ *and* $m_{n,p} = E|X_1|^3/\sigma_{n,p}^2$, *given in* (4.84) *and* (4.87), *respectively, and* F *the distribution function of the normalized sum* $W = Y/\sigma_{n,p}$, *we have*

$$\|F - \Phi\|_1 \le \left(\frac{m_{n,p}}{\sigma_{n,p}}\right) \sum_{i=1}^{n} |\theta_i|^3 + \left(\frac{1}{p} \vee 1\right) \frac{4}{n+2}, \tag{4.72}$$

where Φ *is the cumulative distribution function of the standard normal.*

We note that by the limits in (4.84) and (4.88), the constant $m_{n,p}/\sigma_{n,p}$ that multiplies the sum in the bound (4.72) is of the order of a constant with asymptotic value

$$\lim_{n\to\infty} \frac{m_{n,p}}{\sigma_{n,p}} = \frac{\Gamma(4/p)\sqrt{\Gamma(1/p)}}{\Gamma(3/p)^{3/2}}.$$

Since, for $\boldsymbol{\theta} \in \mathbb{R}^n$ with $\|\boldsymbol{\theta}\| = 1$, we have

$$\sum |\theta_i|^3 \ge \frac{1}{\sqrt{n}},$$

the second term in (4.72) is always of smaller order than the first, so the decay rate of the bound to zero is determined by $\sum_i |\theta_i|^3$. The minimal rate $1/\sqrt{n}$ is achieved when $\theta_i = 1/\sqrt{n}$.

In the special cases $p = 1$ and $p = 2$, $\mathcal{C}_p^n$ is uniform on the simplex $\sum_{i=1}^n |x_i| = 1$ and the unit Euclidean sphere $\sum_{i=1}^n x_i^2 = 1$, respectively. By (4.84) and (4.87) for $p = 1$,

$$\sigma_{n,1}^2 = \frac{2}{n(n+1)} \quad \text{and} \quad m_{n,1} = \frac{3}{n+2},$$

and, using also (4.88) for $p = 2$,

$$\sigma_{n,2}^2 = \frac{1}{n} \quad \text{and} \quad m_{n,2} \le \sqrt{\frac{3}{n+2}};$$

these relations yield

$$\frac{m_{n,1}}{\sigma_{n,1}} = 3\sqrt{\frac{n(n+1)}{2(n+2)^2}} \le \frac{3}{\sqrt{2}} \quad \text{and} \quad \frac{m_{n,2}}{\sigma_{n,2}} \le \sqrt{\frac{3n}{n+2}} \le \sqrt{3}.$$

Substituting into (4.72) now gives

$$\|F - \Phi\|_1 \le \frac{3}{\sqrt{p+1}} \sum_{i=1}^{n} |\theta_i|^3 + \frac{4}{n+2} \quad \text{for } p \in \{1, 2\}. \tag{4.73}$$

4.3.1 Coupling Constructions for Coordinate Symmetric Variables and Their Projections

We generalize the construction in Proposition 2.3 to coordinate symmetric vectors, beginning by generalizing the notion of square biasing, given there, to square biasing in coordinates.

To begin, note that if $\mathbf{Y}$ is a coordinate symmetric random vector in $\mathbb{R}^n$ and $EY_i^2 < \infty$ for $i = 1, \ldots, n$, then the symmetry condition (4.69) implies

$$EY_i = -EY_i \quad \text{and} \quad EY_iY_j = -EY_iY_j \quad \text{for all } i \ne j,$$

and hence

$$EY_i = 0 \quad \text{and} \quad EY_iY_j = \sigma_i^2 \delta_{ij} \quad \text{for all } i, j, \tag{4.74}$$

where $\sigma_i^2 = \mathrm{Var}(Y_i) = EY_i^2$. By removing any component which has zero variance, and lowering the dimension accordingly, we may assume without loss of generality that $\sigma_i^2 > 0$ for all $i = 1, \ldots, n$. For such $\mathbf{Y}$, for all $i = 1, \ldots, n$, we claim there exists a distribution $\mathbf{Y}^i$ such that for all functions $f : \mathbb{R}^n \to \mathbb{R}$ for which the expectation of the left hand side below exists,

$$EY_i^2 f(\mathbf{Y}) = \sigma_i^2 Ef\left(\mathbf{Y}^i\right), \tag{4.75}$$

and say that $\mathbf{Y}^i$ has the $\mathbf{Y}$-square bias distribution in direction i. In particular, the distribution of $\mathbf{Y}^i$ is absolutely continuous with respect to $\mathbf{Y}$ with

$$dF^i(\mathbf{y}) = \frac{y_i^2}{\sigma_i^2} dF(\mathbf{y}). \tag{4.76}$$

By specializing (4.75) to the case where f depends only on Y_i, we see, in the language of Proposition 2.3, that $Y_i^i =_d Y_i^\square$, that is, that Y_i^i has the Y_i-square bias distribution.

Proposition 4.3 shows how to construct the zero bias distribution Y^* for the sum Y of the components of a coordinate-symmetric vector in terms of $\mathbf{Y}^i$ and a random index in a way that parallels the construction for size biasing given in Proposition 2.2. Again we let $\mathcal{U}[a, b]$ denote the uniform distribution on $[a, b]$.

Proposition 4.3 *Let* $\mathbf{Y} \in \mathbb{R}^n$ *be a coordinate-symmetric random vector with* $\mathrm{Var}(Y_i) = \sigma_i^2 \in (0, \infty)$ *for all* $i = 1, 2, \ldots, n$, *and*

$$Y = \sum_{i=1}^{n} Y_i.$$

Let $\mathbf{Y}^i, i = 1, \ldots, n$, *have the square bias distribution given in* (4.75), I *a random index with distribution*

$$P(I = i) = \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2} \tag{4.77}$$

and $U_i \sim \mathcal{U}[-1, 1]$, *with* $\mathbf{Y}^i$, I *and* U_i *mutually independent for all* $i = 1, \ldots, n$. *Then*

$$Y^* = U_I Y_I^I + \sum_{j \neq I} Y_j^I \tag{4.78}$$

has the Y*-zero bias distribution.*

Proof Let f be an absolutely continuous function with $E|Yf(Y)| < \infty$. Staring with the given form of Y^* then averaging over the index I, integrating out the uniform variable U_i and applying (4.75) and (4.69) we obtain

$$\begin{aligned}
\sigma^2 E f'(Y^*) &= \sigma^2 E f'\left(U_I Y_I^I + \sum_{j \neq I} Y_j^I \right) \\
&= \sigma^2 \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2} E f'\left(U_i Y_i^i + \sum_{j \neq i} Y_j^i \right) \\
&= \sum_{i=1}^{n} \sigma_i^2 E\left(\frac{f(Y_i^i + \sum_{j \neq i} Y_j^i) - f(-Y_i^i + \sum_{j \neq i} Y_j^i)}{2Y_i^i} \right) \\
&= \sum_{i=1}^{n} E Y_i \left(\frac{f(Y_i + \sum_{j \neq i} Y_j) - f(-Y_i + \sum_{j \neq i} Y_j)}{2} \right) \\
&= \sum_{i=1}^{n} E Y_i f\left(Y_i + \sum_{j \neq i} Y_j \right) \\
&= E Y f(Y).
\end{aligned}$$

Thus, Y^* has the Y-zero bias distribution. □

Factoring (4.76) as

$$\begin{aligned}
dF^i(\mathbf{y}) &= dF_i^i(y_i) dF(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n | y_i) \\
\text{where } dF_i^i(y_i) &= \frac{y_i^2 dF_i(y_i)}{\sigma_i^2}
\end{aligned} \tag{4.79}$$

provides an alternate way of seeing that $Y_i^i =_d Y_i^\square$. Moreover, it suggests a coupling between Y and Y^* where, given $\mathbf{Y}$, an index $I = i$ is chosen with weight proportional to the variance σ_i^2, the summand Y_i is replaced by Y_i^i having that summand's 'square bias' distribution and then multiplied by U, and, finally, the remaining variables of $\mathbf{Y}$ are perturbed, so that they achieve their original distribution conditional on the ith variable now taking on the value Y_i^i. Typically the remaining variables are changed as little as possible in order to make the coupling between Y and Y^* close.

Now let $\mathbf{X} \in \mathbb{R}^n$ be an exchangeable coordinate-symmetric random vector with components having finite second moments and let $\boldsymbol{\theta} \in \mathbb{R}^n$ have unit length. Then, by (4.74), the projection Y of $\mathbf{X}$ along the direction $\boldsymbol{\theta}$,

$$Y = \sum_{i=1}^n \theta_i X_i$$

has mean zero and variance σ^2 equal to the common variance of the components of $\mathbf{X}$. To form Y^* using the construction just outlined, in view of (4.79) in particular, requires a vector of random variables to be 'adjusted' according to their original distribution, conditional on one coordinate taking on a newly chosen, biased, value. Random vectors which have the 'scaling-conditional' property in Definition 4.2 can easily be so adjusted. Let $\mathcal{L}(V)$ and $\mathcal{L}(V|X = x)$ denote the distribution of V, and the conditional distribution of V given $X = x$, respectively.

Definition 4.2 Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an exchangeable random vector and $\mathcal{D} \subset \mathbb{R}$ the support of the distribution of X_1. If there exists a function $g : \mathcal{D} \to \mathbb{R}$ such that $P(g(X_1) = 0) = 0$ and

$$\mathcal{L}(X_2, \ldots, X_n | X_1 = a) = \mathcal{L}\left(\frac{g(a)}{g(X_1)} (X_2, \ldots, X_n) \right) \quad \text{for all } a \in \mathcal{D}, \tag{4.80}$$

we say that $\mathbf{X}$ is scaling g-conditional, or simply scaling-conditional.

Proposition 4.4 is an application of Theorem 4.1 and Proposition 4.3 to projections of coordinate symmetric, scaling-conditional vectors.

Proposition 4.4 *Let $\mathbf{X} \in \mathbb{R}^n$ be an exchangeable, coordinate symmetric and scaling g-conditional random vector with finite second moment. For $\boldsymbol{\theta} \in \mathbb{R}^n$ of unit length set*

$$Y = \sum_{i=1}^n \theta_i X_i, \quad \sigma^2 = \mathrm{Var}(Y), \quad \textit{and} \quad F(x) = P(Y/\sigma \le x).$$

Then any construction of $(\mathbf{X}, X_i^i)$ on a joint space for each $i = 1, \ldots, n$ with X_i^i having the X_i-square biased distribution provides the upper bound

$$\|F - \Phi\|_1 \le \frac{2}{\sigma} E \left| \theta_I \left(U_I X_I^I - X_I \right) + \left(\frac{g(X_I^I)}{g(X_I)} - 1 \right) \sum_{j \ne I} \theta_j X_j \right|, \tag{4.81}$$

where $P(I = i) = \theta_i^2$ and $U_i \sim \mathcal{U}[-1, 1]$ with $\{X_i^i, X_j, j \ne i\}$, I and U_i mutually independent for $i = 1, 2, \ldots, n$.

Proof For all $i = 1, \ldots, n$, since $\mathbf{X}$ is scaling g-conditional, given $\mathbf{X}$ and X_i^i with the X_i-square bias distribution, by (4.79) and (4.80) the vector

$$\mathbf{X}^i = \left(\frac{g(X_i^i)}{g(X_i)} X_1, \ldots, \frac{g(X_i^i)}{g(X_i)} X_{i-1}, X_i^i, \frac{g(X_i^i)}{g(X_i)} X_{i+1}, \ldots, \frac{g(X_i^i)}{g(X_i)} X_n \right)$$

has the $\mathbf{X}$-square bias distribution in direction i as given in (4.75), that is, for every h for which the expectation on the left-hand side below exists,

$$EX_i^2 h(\mathbf{X}) = EX_i^2 Eh\big(\mathbf{X}^i\big). \tag{4.82}$$

We now apply Proposition 4.3 to $\mathbf{Y} = (\theta_1 X_1, \ldots, \theta_n X_n)$. First, the coordinate symmetry of $\mathbf{Y}$ follows from that of $\mathbf{X}$. Next, we claim

$$\mathbf{Y}^i = \big(\theta_1 X_1^i, \ldots, \theta_n X_n^i\big)$$

has the $\mathbf{Y}$-square bias distribution in direction i. Given f, let

$$h(\mathbf{X}) = f(\theta_1 X_1, \ldots, \theta_n X_n).$$

Applying (4.82) we obtain

$$\begin{aligned} EY_i^2 f(\mathbf{Y}) &= E\theta_i^2 X_i^2 f(\mathbf{Y}) \\ &= \theta_i^2 EX_i^2 h(\mathbf{X}) \\ &= \theta_i^2 EX_i^2 Eh\big(\mathbf{X}^i\big) \\ &= E\theta_i^2 X_i^2 Ef\big(\mathbf{Y}^i\big) \\ &= EY_i^2 Ef\big(\mathbf{Y}^i\big). \end{aligned}$$

Since $\mathbf{X}$ is exchangeable, the variance of Y_i is proportional to θ_i^2 and the distribution of I in (4.77) specializes to the one claimed. Lastly, as $\mathbf{Y}^i$, I and U_i are mutually independent for $i = 1, \ldots, n$, Proposition 4.3 yields that

$$Y^* = U_I Y_I^I + \sum_{j \neq I} Y_j^I$$

has the Y-zero bias distribution.

The difference $Y^* - Y$ is given by

$$\begin{aligned} Y^* - Y &= U_I Y_I^I + \sum_{j \neq I} Y_j^I - \sum_{i=1}^n Y_i \\ &= U_I \theta_I X_I^I + \sum_{j \neq I} \theta_j X_j^I - \sum_{j=1}^n \theta_j X_j \\ &= \theta_I \big(U_I X_I^I - X_I\big) + \sum_{j \neq I} \theta_j \big(X_j^I - X_j\big) \\ &= \theta_I \big(U_I X_I^I - X_I\big) + \sum_{j \neq I} \theta_j \left(\frac{g(X_I^I)}{g(X_I)} - 1 \right) X_j \\ &= \theta_I \big(U_I X_I^I - X_I\big) + \left(\frac{g(X_I^I)}{g(X_I)} - 1 \right) \sum_{j \neq I} \theta_j X_j. \end{aligned}$$

The proof is completed by dividing both sides by σ, applying (2.59) to yield $Y^*/\sigma = (Y/\sigma)^*$, and invoking Theorem 4.1. □

4.3.2 Construction and Bounds for Cone Measure

Proposition 4.5 below shows that Proposition 4.4 can be applied to cone measure. We denote the Gamma and Beta distributions with parameters α, β as $\Gamma(\alpha, \beta)$ and $B(\alpha, \beta)$, respectively. That is, with the Gamma function at $\alpha > 0$ given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx,$$

with $\beta > 0$, the density of the $\Gamma(\alpha, \beta)$ distribution is

$$\frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \mathbf{1}_{\{x>0\}};$$

the density of the Beta distribution $B(\alpha, \beta)$ is given in (4.90).

Proposition 4.5 *Let C_p^n denote cone measure as given in (4.71) for some $n \in \mathbb{N}$ and $p > 0$.*

1. *Cone measure C_p^n is exchangeable and coordinate-symmetric. For $\{G_j, \epsilon_j, j = 1, \ldots, n\}$ independent variables with $G_j \sim \Gamma(1/p, 1)$ and ϵ_j taking values -1 and $+1$ with equal probability, setting $G_{a,b} = \sum_{i=a}^b G_i$ we have*

$$\mathbf{X} = \left(\epsilon_1 \left(\frac{G_1}{G_{1,n}} \right)^{1/p}, \ldots, \epsilon_n \left(\frac{G_n}{G_{1,n}} \right)^{1/p} \right) \sim C_p^n. \tag{4.83}$$

2. *The common marginal distribution X_i of cone measure is characterized by*

$$X_i =_d -X_i \quad \text{and} \quad |X_i|^p \sim B\big(1/p, (n-1)/p\big),$$

and the variance $\sigma_{n,p}^2 = \mathrm{Var}(X_i)$ is given by

$$\sigma_{n,p}^2 = \frac{\Gamma(3/p)\Gamma(n/p)}{\Gamma(1/p)\Gamma((n+2)/p)} \tag{4.84}$$

and satisfies

$$\lim_{n\to\infty} n^{2/p} \sigma_{n,p}^2 = \frac{p^{2/p}\Gamma(3/p)}{\Gamma(1/p)}.$$

3. *The square bias distribution X_i^i of X_i is characterized by*

$$X_i^i =_d -X_i^i \quad \text{and} \quad \big|X_i^i\big|^p \sim B\big(3/p, (n-1)/p\big). \tag{4.85}$$

Letting $\{G_j, G_j', \epsilon_j, j = 1, \ldots, n\}$ be independent variables with $G_j \sim \Gamma(1/p, 1)$, $G_j' \sim \Gamma(2/p, 1)$ and ϵ_j taking values -1 and $+1$ with equal probability, for each $i = 1, \ldots, n$, a construction of $(\mathbf{X}, X_i^i)$ on a joint space is given by the representation of $\mathbf{X}$ in (4.83) along with

$$X_i^i = \epsilon_i \left(\frac{G_i + G_i'}{G_{1,n} + G_i'} \right)^{1/p}. \tag{4.86}$$

The mean $m_{n,p} = E|X_i^i| = E|X_i^3|/\sigma_{n,p}^2$ *for all* $i = 1, \ldots, n$ *is given by*

$$m_{n,p} = \frac{\Gamma(4/p)\Gamma((n+2)/p)}{\Gamma(3/p)\Gamma((n+3)/p)} \tag{4.87}$$

and satisfies

$$\lim_{n\to\infty} n^{1/p} m_{n,p} = \frac{p^{1/p}\Gamma(4/p)}{\Gamma(3/p)} \quad \textit{and} \quad m_{n,p} \le \left(\frac{3}{n+2} \right)^{1/(p\vee 1)}. \tag{4.88}$$

4. *Cone measure* $\mathcal{C}_p^n$ *is scaling* $(1 - |x|^p)^{1/p}$ *conditional.*

The proof of Proposition 4.5 is deferred to the end of this section. Before proceeding to Theorem 4.7, we remind the reader of the following known facts about the Gamma and Beta distributions; see Bickel and Doksum (1977), Theorem 1.2.3 for the case $n = 2$ of the first claim, the extension to general n and the following claim being straightforward. For $\gamma_i \sim \Gamma(\alpha_i, \beta), i = 1, \ldots, n$, independent with $\alpha_i > 0$ and $\beta > 0$,

$$\gamma_1 + \gamma_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta), \quad \frac{\gamma_1}{\gamma_1 + \gamma_2} \sim B(\alpha_1, \alpha_2), \tag{4.89}$$

$$\text{and} \quad \left(\frac{\gamma_1}{\sum_{i=1}^n \gamma_i}, \ldots, \frac{\gamma_n}{\sum_{i=1}^n \gamma_i} \right) \quad \text{and} \quad \sum_{i=1}^n \gamma_i \quad \text{are independent;}$$

the Beta distribution $B(\alpha, \beta)$ has density

$$p_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1} \mathbf{1}_{u\in[0,1]}$$

$$\text{and } \kappa > 0 \text{ moment} \frac{\Gamma(\alpha+\kappa)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+\kappa)\Gamma(\alpha)}. \tag{4.90}$$

Proof of Theorem 4.7 Using Proposition 4.5, we apply Proposition 4.4 for $\mathbf{X}$ with $g(x) = (1 - |x|^p)^{1/p}$ and the joint construction of $(\mathbf{X}, X_i^i)$ given in item 3. Note that Proposition 4.2 applies, using the notation there, with $V \sim \mathcal{U}[0, 1]$, independent of all other variables, $U_i = \epsilon_i V$, and

$$\overline{X}_i = \left(\frac{G_i}{G_{1,n}} \right)^{1/p} \quad \text{and} \quad \overline{Y}_i = \left(\frac{G_i + G_i'}{G_{1,n} + G_i'} \right)^{1/p}.$$

Applying the triangle inequality on (4.81) yields the bound on the L^1 norm $\|F - \Phi\|_1$ of

$$\frac{2}{\sigma_{n,p}} \left(E\left|\theta_I \left(U_I X_I^I - X_I \right)\right| + E\left| \left(\frac{g(X_I^I)}{g(X_I)} - 1 \right) \sum_{j\neq I} \theta_j X_j \right| \right). \tag{4.91}$$

We begin by averaging the first term over I. Note that

$$|X_1| = \left(\frac{G_1}{G_{1,n}}\right)^{1/p} \le \left(\frac{G_1 + G_1'}{G_{1,n} + G_1'}\right)^{1/p} = |X_1^1|,$$

and therefore, recalling $P(I = i) = \theta_i^2$, we may invoke Proposition 4.2 to conclude

$$\begin{aligned} E|\theta_I(U_I X_I^I - X_I)| &= \sum_{i=1}^n |\theta_i|^3 E|U_i X_i^i - X_i| \\ &= E|U_1 X_1^1 - X_1| \sum_{i=1}^n |\theta_i|^3 \\ &\le \frac{E|X_1|^3}{2\sigma_{n,p}^2} \sum_{i=1}^n |\theta_i|^3 = \frac{m_{n,p}}{2} \sum_{i=1}^n |\theta_i|^3. \end{aligned} \tag{4.92}$$

Now, averaging the second term in (4.91) over the distribution of I yields

$$E\left|\left(\frac{g(X_I^I)}{g(X_I)} - 1\right) \sum_{j \ne I} \theta_j X_j\right| = \sum_{i=1}^n E\left|\left(\frac{g(X_i^i)}{g(X_i)} - 1\right) \sum_{j \ne i} \theta_j X_j\right| \theta_i^2. \tag{4.93}$$

Using (4.83), (4.86) and $g(x) = (1 - |x|^p)^{1/p}$, we have

$$\frac{g(X_i^i)}{g(X_i)} - 1 = \left(\frac{G_{1,n}}{G_{1,n} + G_i'}\right)^{1/p} - 1. \tag{4.94}$$

Applying (4.89) we have that $\{G_{1,n}, G_i'\}$ are independent of $X_1, \ldots, X_n$; hence, the term (4.94) is independent of the sum it multiplies in (4.93) and therefore (4.93) equals

$$\sum_{i=1}^n E\left|\frac{g(X_i^i)}{g(X_i)} - 1\right| E\left|\sum_{j \ne i} \theta_j X_j\right| \theta_i^2. \tag{4.95}$$

To bound the first expectation in (4.95), since $G_{1,n}/(G_{1,n} + G_i') \sim B(n/p, 2/p)$, we have

$$E\left|\frac{g(X_i^i)}{g(X_i)} - 1\right| = E\left(1 - \left(\frac{G_{1,n}}{G_{1,n} + G_i'}\right)^{1/p}\right) \le \left(\frac{1}{p} \vee 1\right) \frac{2}{n+2} \tag{4.96}$$

since for $p \ge 1$, using (4.90) with $\kappa = 1$,

$$\begin{aligned} &E\left(1 - \left(\frac{G_{1,n}}{G_{1,n} + G_i'}\right)^{1/p}\right) \\ &\quad \le E\left(1 - \left(\frac{G_{1,n}}{G_{1,n} + G_i'}\right)\right) = 1 - \frac{n/p}{(n+2)/p} = \frac{2}{n+2}, \end{aligned}$$

while for $0 < p < 1$, using Jensen's inequality and the fact that

$$(1 - x)^{1/p} \ge 1 - x/p \quad \text{for } x \le 1,$$

we have

$$E\left(1-\left(\frac{G_{1,n}}{G_{1,n}+G_i'}\right)^{1/p}\right)$$
$$\le 1-\left(E\left(\frac{G_{1,n}}{G_{1,n}+G_i'}\right)\right)^{1/p}=1-\left(\frac{n}{n+2}\right)^{1/p}\le\frac{2}{p(n+2)}.$$

We may bound the second expectation in (4.95) by $\sigma_{n,p}$ since

$$\left(E\left|\sum_{j\neq i}\theta_j X_j\right|\right)^2$$
$$\le E\left(\sum_{j\neq i}\theta_j X_j\right)^2=\mathrm{Var}\left(\sum_{j\neq i}\theta_j X_j\right)=\sigma_{n,p}^2\sum_{j\neq i}\theta_j^2\le\sigma_{n,p}^2.$$

Neither this bound nor the bound (4.96) depends on i, so substituting them into (4.95) and summing over i, again using $\sum_i\theta_i^2=1$, yields

$$\sum_{i=1}^n E\left|\frac{g(X_i^i)}{g(X_i)}-1\right|E\left|\sum_{j\neq i}\theta_j X_j\right|\theta_i^2\le\sigma_{n,p}\left(\frac{1}{p}\vee 1\right)\frac{2}{n+2}. \tag{4.97}$$

Adding (4.92) and (4.97) and multiplying by $2/\sigma_{n,p}$ in accordance with (4.81) yields (4.72). □

Proof of Proposition 4.5

1. For $A\subset S(\ell_p^n)$, $\mathbf{e}=(e_1,\dots,e_n)\in\{-1,1\}^n$ and a permutation $\pi\in\mathcal{S}_n$, let

$$A_{\mathbf{e}}=\{\mathbf{x}\colon\ (e_1x_1,\dots,e_nx_n)\in A\}$$
$$\text{and}\quad A_\pi=\{\mathbf{x}\colon\ (x_{\pi(1)},\dots,x_{\pi(n)})\in A\}.$$

By the properties of Lebesgue measure, $\mu^n([0,1]A_{\mathbf{e}})=\mu^n([0,1]A_\pi)=\mu^n([0,1]A)$, so by (4.71), cone measure is coordinate symmetric and exchangeable.

The coordinate symmetry of $\mathbf{X}$ implies that

$$P(\mathbf{X}\in A)=P(\mathbf{X}\in A_{\mathbf{e}})\quad\text{for all }\mathbf{e}\in\{-1,1\}^n,$$

so with ϵ_i, $i=1,\dots,n$, i.i.d. variables taking the values 1 and -1 with probability $1/2$ and independent of $\mathbf{X}$,

$$\begin{aligned}P\big((\epsilon_1X_1,\dots,\epsilon_nX_n)\in A\big)&=P(\mathbf{X}\in A_\epsilon)\\&=\frac{1}{2^n}\sum_{\mathbf{e}\in\{-1,1\}^n}P(\mathbf{X}\in A_{\mathbf{e}})\\&=P(\mathbf{X}\in A),\end{aligned}$$

and hence $(\epsilon_1X_1,\dots,\epsilon_nX_n)=_d(X_1,\dots,X_n)$. Note that for any $(s_1,\dots,s_n)\in\{-1,1\}^n$ that

$$(\epsilon_1s_1,\dots,\epsilon_ns_n)=_d(\epsilon_1,\dots,\epsilon_n),\quad\text{and is independent of }\mathbf{X}.$$

Hence, since $P(X_i = 0) = 0$, with $s_i = X_i/|X_i|$, the sign of X_i, we have

$$\begin{aligned} P\big((\epsilon_1|X_1|, \ldots, \epsilon_n|X_n|) \in A\big) &= P\big((\epsilon_1 s_1 X_1, \ldots, \epsilon_n s_n X_n) \in A\big) \\ &= P\big((\epsilon_1 X_1, \ldots, \epsilon_n X_n) \in A\big) \\ &= P\big((X_1, \ldots, X_n) \in A\big). \end{aligned}$$

We thus obtain (4.83) applying that $\mathbf{X} \sim \mathcal{C}_p^n$ satisfies

$$\big(|X_1|, \ldots, |X_n|\big) =_d \left(\left(\frac{G_1}{G_{1,n}}\right)^{1/p}, \ldots, \left(\frac{G_n}{G_{1,n}}\right)^{1/p}\right) \tag{4.98}$$

shown, for instance, by Schechtman and Zinn (1990).

2. Applying the coordinate symmetry of $\mathbf{X}$ coordinatewise gives $X_i =_d -X_i$ and (4.98) yields $|X_i|^p = G_i/G_{1,n}$, which has the claimed Beta distribution, by (4.89). As $EX_i = 0$, we have

$$\mathrm{Var}(X_i) = EX_i^2 = E\big(|X_i|^p\big)^{2/p} \tag{4.99}$$

and the variance claim in (4.84) follows from (4.90) for $\alpha = 1/p$, $\beta = (n-1)/p$ and $\kappa = 2/p$.

From Stirlings formula, for all $x > 0$,

$$\lim_{m\to\infty} \frac{m^x \Gamma(m)}{\Gamma(m+x)} = 1,$$

so letting $m = n/p$ and $x = k/p$,

$$\lim_{n\to\infty} \frac{n^{k/p}\Gamma(n/p)}{\Gamma((n+k)/p)} = p^{k/p}. \tag{4.100}$$

The limit (4.84) now follows.

3. If X is symmetric with variance $\sigma^2 > 0$ and $X^\square$ has the X-square bias distribution, then for all bounded continuous functions f

$$\begin{aligned} &\sigma^2 Ef\big(X^\square\big) \\ &\quad = EX^2 f(X) = E\big[(-X)^2 f(-X)\big] = EX^2 f(-X) = \sigma^2 Ef\big(-X^\square\big), \end{aligned}$$

showing $X^\square$ is symmetric.

From (4.90) and a change of variable, a random variable X satisfies

$$|X|^p \sim B(\alpha/p, \beta/p)$$

if and only if the density $p_{|X|}(u)$ of $|X|$ is

$$p_{|X|}(u) = \frac{p\Gamma((\alpha+\beta)/p)}{\Gamma(\alpha/p)\Gamma(\beta/p)} u^{\alpha-1}\big(1-u^p\big)^{\beta/p-1}\mathbf{1}_{u\in[0,1]}. \tag{4.101}$$

Hence, since $|X_i|^p \sim B(1/p, (n-1)/p)$ by item 2, the density $p_{|X_i|}(u)$ of $|X_i|$ is

$$p_{|X_i|}(u) = \frac{p\Gamma(n/p)}{\Gamma(1/p)\Gamma((n-1)/p)}\big(1-u^p\big)^{(n-1)/p-1}\mathbf{1}_{u\in[0,1]}.$$

Multiplying by u^2 and renormalizing produces the $|X_i^i|$ density

$$p_{|X_i^i|}(u) = \frac{u^2 p_{|X_i|}(u)}{E X_i^2}$$
$$= \frac{p\Gamma((n+2)/p)}{\Gamma(3/p)\Gamma((n-1)/p)} u^2 \left(1-u^p\right)^{(n-1)/p-1} \mathbf{1}_{u\in[0,1]}, \qquad (4.102)$$

and comparing (4.102) to (4.101) shows the second claim in (4.85). The representation (4.86) now follows from (4.89) and the symmetry of X_i^i. The moment formula (4.87) for $m_{n,p}$ follows from (4.90) for $\alpha = 3/p$, $\beta = (n-1)/p$ and $\kappa = 1/p$, and the limit in (4.88) follows from (4.100).

Regarding the last claim in (4.88), for $p \geq 1$ Hölder's inequality gives

$$m_{n,p} = E\left|X^1\right| \leq \left(E\left|X^1\right|^p\right)^{1/p} = \left(\frac{3}{n+2}\right)^{1/p},$$

while for $0 < p < 1$, we have

$$m_{n,p} = E\left|X^1\right| = E\left(\frac{G_i + G_i'}{G_{1,n} + G_i'}\right)^{1/p} \leq E\left(\frac{G_i + G_i'}{G_{1,n} + G_i'}\right) = \frac{3}{n+2}.$$

4. We consider the conditional distribution on the left-hand side of (4.80), and use the representation, and notation $G_{a,b}$, given in (4.83). The second equality below follows from the coordinate-symmetry of $\mathbf{X}$, and the fourth follows since we may replace $G_{1,n}$ by $G_{2,n}/(1-|a|^p)$ on the conditioning event. Using the notation $a\mathcal{L}(V)$ for the distribution of aV, we have

$$\begin{aligned}
&\mathcal{L}(X_2, \ldots, X_n | X_1 = a) \\
&= \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{1,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{1,n}}\right)^{1/p} \middle| \epsilon_1\left(\frac{G_1}{G_{1,n}}\right)^{1/p} = a\right) \\
&= \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{1,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{1,n}}\right)^{1/p} \middle| \left(\frac{G_1}{G_{1,n}}\right)^{1/p} = |a|\right) \\
&= \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{1,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{1,n}}\right)^{1/p} \middle| \frac{G_{2,n}}{G_{1,n}} = 1-|a|^p\right) \\
&= \left(1-|a|^p\right)^{1/p} \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{2,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{2,n}}\right)^{1/p} | \frac{G_{2,n}}{G_{1,n}} = 1-|a|^p\right) \\
&= \left(1-|a|^p\right)^{1/p} \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{2,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{2,n}}\right)^{1/p} \middle| \frac{G_1}{G_{1,n}} = |a|^p\right) \\
&= \left(1-|a|^p\right)^{1/p} \mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{2,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{2,n}}\right)^{1/p}\right) \\
&= g(a)\mathcal{L}\left(\epsilon_2\left(\frac{G_2}{G_{2,n}}\right)^{1/p}, \ldots, \epsilon_n\left(\frac{G_n}{G_{2,n}}\right)^{1/p}\right). \qquad (4.103)
\end{aligned}$$

In the penultimate step may we remove the conditioning on $G_1/G_{1,n}$ since (4.89) and the independence of G_1 from all other variables gives that

$$\left(\frac{G_2}{G_{2,n}}, \ldots, \frac{G_n}{G_{2,n}}\right) \quad \text{is independent of } (G_1, G_{2,n})$$

and therefore independent of $G_1/(G_1+G_{2,n}) = G_1/G_{1,n}$.

Regarding the right-hand side of (4.80), using $1-|X_1|^p = \sum_{i=2}^n |X_i|^p$ and the representation (4.83), we obtain

$$\begin{aligned} g(a)(X_2,\dots,X_n)/g(X_1) &= g(a)\left(\frac{(X_2,\dots,X_n)}{(|X_2|^p+\cdots+|X_n|^p)^{1/p}}\right) \\ &= g(a)\left(\frac{(\epsilon_2(\frac{G_2}{G_{1,n}})^{1/p},\dots,\epsilon_n(\frac{G_n}{G_{1,n}})^{1/p})}{((\frac{G_2}{G_{1,n}})+\cdots+(\frac{G_n}{G_{1,n}}))^{1/p}}\right) \\ &= g(a)\left(\frac{(\epsilon_2 G_2^{1/p},\dots,\epsilon_n G_n^{1/p})}{(G_2+\cdots+G_n)^{1/p}}\right) \\ &= g(a)\left(\epsilon_2\left(\frac{G_2}{G_{2,n}}\right)^{1/p},\dots,\epsilon_n\left(\frac{G_n}{G_{2,n}}\right)^{1/p}\right) \end{aligned}$$

matching the distribution (4.103). □

In principle, Proposition 4.3 and Theorem 4.1 may be applied to compute bounds to the normal for projections of other coordinate-symmetric vectors when the required couplings, and conditioning, are as tractable as here.

4.4 Combinatorial Central Limit Theorems

In this section we apply Theorem 4.1 to derive L^1 bounds in the combinatorial central limit theorem, that is, for random variables Y of the form

$$Y = \sum_{i=1}^n a_{i,\pi(i)}, \tag{4.104}$$

where π is a permutation distributed uniformly over the symmetric group $\mathcal{S}_n$, and $\{a_{ij}\}_{1\le i,j\le n}$ are the components of a matrix $A \in \mathbb{R}^{n\times n}$.

Random variables of this form are of interest in permutation tests. In particular, given a function $d(x,y)$ which in some sense measures the closeness of two observations x and y, given values $x_1,\dots,x_n$ and $y_1,\dots,y_n$ and a putative 'matching' permutation τ that associates x_i to $y_{\tau(i)}$, one can test whether the level of matching given by τ, as measured by

$$y_\tau = \sum_{i=1}^n a_{i\tau(i)} \quad \text{where } a_{ij} = d(x_i,y_j),$$

is unusually high by seeing how large the matching level y_τ is relative to that provided by a random matching, that is, by seeing whether $P(Y \ge y_\tau)$ is significantly small.

Motivated by these considerations, Wald and Wolfowitz (1944) proved the central limit theorem as $n\to\infty$ when the factorization $a_{ij} = b_i c_j$ holds; Hoeffding (1951) later generalized this result to arrays $\{a_{ij}\}_{1\le i,j\le n}$. Motoo (1957) gave

Lindeberg-type sufficient conditions for the normal limit to hold. In Sect. 6.1 the L^∞ distance to the normal is considered for the case where π is uniformly distributed, and also when its distribution is constant on conjugacy classes of $\mathcal{S}_n$.

Letting

$$a_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}, \quad a_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij} \quad \text{and} \quad a_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij},$$

straightforward calculations show that when π is uniform over $\mathcal{S}_n$ the mean μ_A and variance σ_A^2 of Y are given by

$$\begin{aligned} \mu_A &= n a_{\cdot\cdot} \quad \text{and} \\ \sigma_A^2 &= \frac{1}{n-1}\sum_{i,j}\left(a_{ij}^2 - a_{i\cdot}^2 - a_{\cdot j}^2 + a_{\cdot\cdot}^2\right) \\ &= \frac{1}{n-1}\sum_{i,j}(a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot})^2. \end{aligned} \tag{4.105}$$

For simplicity, writing μ and σ^2 for μ_A and σ_A^2, respectively, we prove in (4.124) the following equivalent representation for σ^2,

$$\sigma^2 = \frac{1}{4n^2(n-1)}\sum_{i,j,k,l}\left[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})\right]^2, \tag{4.106}$$

and assume in what follows that $\sigma^2 > 0$ to rule out trivial cases. By (4.106), $\sigma^2 = 0$ if and only if $a_{il} - a_{i\cdot}$ does not depend on i, that is, if and only if the difference between any two rows $\mathbf{a}_i$ and $\mathbf{a}_j$ of A satisfy $\mathbf{a}_i - \mathbf{a}_j = (a_{i\cdot} - a_{j\cdot})(1, \ldots, 1)$.

For each $n \geq 3$, Theorem 4.8 provides an L^1 bound between the standardized version of the variable Y given in (4.104) and the normal, with an explicit constant depending on the third-moment-type quantity

$$\gamma = \gamma_A, \quad \text{where } \gamma_A = \sum_{i,j=1}^{n} |a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}|^3. \tag{4.107}$$

When the elements of A are all of comparable order, σ^2 is of order n and γ of order n^2, resulting in a bound of order $n^{-1/2}$.

Theorem 4.8 *For $n \geq 3$, let $\{a_{ij}\}_{i,j=1}^n$ be the components of a matrix $A \in \mathbb{R}^{n\times n}$, let π be a random permutation uniformly distributed over $\mathcal{S}_n$, and let Y be given by* (4.104). *Then, with μ, σ^2 given in* (4.105), *and γ given in* (4.107), *F the distribution function of $W = (Y - \mu)/\sigma$ and Φ that of the standard normal,*

$$\|F - \Phi\|_1 \leq \frac{\gamma}{(n-1)\sigma^3}\left(16 + \frac{56}{(n-1)} + \frac{8}{(n-1)^2}\right).$$

The proof of this theorem depends on a construction of the zero bias variable using an exchangeable pair, which we now describe.

4.4.1 Use of the Exchangeable Pair

We recall that the exchangeable variables Y', Y'' form a λ-Stein pair if

$$E(Y''|Y') = (1-\lambda)Y' \tag{4.108}$$

for some $0 < \lambda < 1$. When $\mathrm{Var}(Y') = \sigma^2 \in (0, \infty)$, Lemma 2.7 yields

$$EY' = 0 \quad \text{and} \quad E(Y'-Y'')^2 = 2\lambda\sigma^2. \tag{4.109}$$

The following proposition is in some sense a two variable version of Proposition 2.3.

Proposition 4.6 *Let Y', Y'' be a λ-Stein pair with $\mathrm{Var}(Y') = \sigma^2 \in (0,\infty)$ and distribution $F(y', y'')$. Then when $Y^\dagger$, $Y^\ddagger$ have distribution*

$$dF^\dagger(y', y'') = \frac{(y'-y'')^2}{2\lambda\sigma^2} dF(y', y''), \tag{4.110}$$

and $U \sim \mathcal{U}[0,1]$ is independent of $Y^\dagger$, $Y^\ddagger$, the variable

$$Y^* = UY^\dagger + (1-U)Y^\ddagger \quad \textit{has the } Y'\textit{-zero biased distribution.} \tag{4.111}$$

Proof For all absolutely continuous functions f for which the expectations below exist,

$$\begin{aligned}
\sigma^2 Ef'(Y^*) &= \sigma^2 Ef'\big(UY^\dagger + (1-U)Y^\ddagger\big) \\
&= \sigma^2 E\left(\frac{f(Y^\dagger) - f(Y^\ddagger)}{Y^\dagger - Y^\ddagger}\right) \\
&= \frac{1}{2\lambda} E\left(\left(\frac{f(Y'') - f(Y')}{Y''-Y'}\right)(Y''-Y')^2\right) \\
&= \frac{1}{2\lambda} E\big(f(Y'') - f(Y')(Y''-Y')\big) \\
&= \frac{1}{\lambda} E\big(Y'f(Y') - Y''f(Y')\big) \\
&= \frac{1}{\lambda} E\big(Y'f(Y') - (1-\lambda)Y'f(Y')\big) \\
&= EY'f(Y').
\end{aligned}$$

□

The following lemma, leading toward the construction of zero bias variables, is motivated by generalizing the framework of Example 2.3, where the Stein pair is a function of some underlying random variables $\xi_\alpha, \alpha \in \chi$ and a random index $\mathbf{I}$.

Lemma 4.4 *Let $F(y', y'')$ be the distribution of a Stein pair and suppose there exists a distribution*

$$F(\mathbf{i}, \xi_\alpha, \alpha \in \chi) \tag{4.112}$$

and an $\mathbb{R}^2$ *valued function* $(y', y'') = \psi(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$ *such that when* $\mathbf{I}$ *and* $\{\Xi_\alpha, \alpha \in \mathcal{X}\}$ *have distribution* (4.112) *then*

$$(Y', Y'') = \psi(\mathbf{I}, \Xi_\alpha, \alpha \in \mathcal{X})$$

has distribution $F(y', y'')$. *If* $\mathbf{I}^\dagger, \{\Xi_\alpha^\dagger, \alpha \in \chi\}$ *have distribution*

$$dF^\dagger(\mathbf{i}, \xi_\alpha, \alpha \in \mathcal{X}) = \frac{(y' - y'')^2}{E(Y' - Y'')^2} dF(\mathbf{i}, \xi_\alpha, \alpha \in \mathcal{X}) \tag{4.113}$$

then the pair

$$(Y^\dagger, Y^\ddagger) = \psi\left(\mathbf{I}^\dagger, \Xi_\alpha^\dagger, \alpha \in \mathcal{X}\right)$$

has distribution $F^\dagger(y^\dagger, y^\ddagger)$ *satisfying*

$$dF^\dagger(y', y'') = \frac{(y' - y'')^2}{2\lambda\sigma^2} dF(y', y'').$$

Proof For any bounded measurable function f

$$\begin{aligned} Ef(Y^\dagger, Y^\ddagger) &= Ef\left(\psi\left(\mathbf{I}^\dagger, \Xi_\alpha^\dagger, \alpha \in \mathcal{X}\right)\right) \\ &= \int f\left(\psi(\mathbf{i}, \xi_\alpha, \alpha \in \chi)\right) dF^\dagger(\mathbf{i}, \xi_\alpha, \alpha \in \chi) \\ &= \int f(y', y'') \frac{(y' - y'')^2}{2\lambda\sigma^2} dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi) \\ &= E\left(\frac{(Y' - Y'')^2}{2\lambda\sigma^2} f(Y', Y'')\right), \end{aligned}$$

where (Y', Y'') has distribution $F(y', y'')$. □

We continue building a general framework around Example 2.3, where the random index is chosen independently of the permutation, so their joint distribution factors, leading to

$$dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi) = P(\mathbf{I} = \mathbf{i}) dF(\xi_\alpha, \alpha \in \chi). \tag{4.114}$$

Moreover, in view of (2.47), that is, that

$$Y'' - Y' = b\left(i, j, \pi(i), \pi(j)\right) \quad \text{where } b(i, j, k, l) = a_{il} + a_{jk} - (a_{ik} + a_{jl}),$$

we will pay special attention to situations where

$$Y'' - Y' = b(\mathbf{I}, \Xi_\alpha, \alpha \in \chi_\mathbf{I}) \tag{4.115}$$

where $\mathbf{I}$ and $\chi_\mathbf{I}$ are vectors of small dimensions with components in $\mathcal{I}$ and χ, respectively. In other words, we consider situations where the difference between Y'' and Y' depends on only a few variables. In such cases, it will be convenient to further decompose $dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$ as

$$dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi) = P(\mathbf{I} = \mathbf{i}) dF_\mathbf{i}(\xi_\alpha, \alpha \in \chi_\mathbf{i}) dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_\mathbf{i} | \xi_\alpha, \alpha \in \chi_\mathbf{i}), \tag{4.116}$$

where $dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}})$ is the marginal distribution of ξ_α for $\alpha \in \chi_{\mathbf{i}}$, and $dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_{\mathbf{i}}|\xi_\alpha, \alpha \in \chi_{\mathbf{i}})$ the conditional distribution of ξ_α for $\alpha \notin \chi_{\mathbf{i}}$ given ξ_α for $\alpha \in \chi_{\mathbf{i}}$. One notes, however, that the factorization (4.114) guarantees that the marginal distributions of any ξ_α does not depend on $\mathbf{i}$. In terms of generating variables having the specified distributions for the purposes of coupling, the decomposition (4.116) corresponds to first generating $\mathbf{I}$, then $\{\xi_\alpha, \alpha \in \chi_{\mathbf{I}}\}$, and lastly $\{\xi_\alpha, \alpha \notin \chi_{\mathbf{I}}\}$ conditional on $\{\xi_\alpha, \alpha \in \chi_{\mathbf{I}}\}$. In what follows we will continue the slight abuse notation of letting $\{\alpha\colon\ \alpha \in \chi_{\mathbf{i}}\}$ denote the set of components of the vector $\chi_{\mathbf{i}}$.

We now consider the square bias distribution $F^\dagger$ in (4.113) when the factorization (4.116) of F holds. Letting $\mathbf{I}$ and $\{\Xi_\alpha\colon\ \alpha \in \chi\}$ have distribution (4.114), by (4.109), (4.115) and independence we obtain

$$2\lambda\sigma^2 = E(Y' - Y'')^2 = Eb^2(\mathbf{I}, \Xi_\alpha, \alpha \in \chi_{\mathbf{I}}) = \sum_{\mathbf{i}\subset\mathcal{I}} P(\mathbf{I} = \mathbf{i}) Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_{\mathbf{i}}).$$

In particular, we may define a distribution for a vector of indices $\mathbf{I}^\dagger$ with components in $\mathcal{I}$ by

$$P(\mathbf{I}^\dagger = \mathbf{i}) = \frac{r_{\mathbf{i}}}{2\lambda\sigma^2} \quad \text{with } r_{\mathbf{i}} = P(\mathbf{I} = \mathbf{i}) Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_{\mathbf{i}}). \tag{4.117}$$

Hence, substituting (4.115) and (4.116) into (4.113),

$$\begin{aligned}
&dF^\dagger(\mathbf{i}, \xi_\alpha, \alpha \in \chi) \\
&\quad = \frac{P(\mathbf{I} = \mathbf{i}) b^2(\mathbf{i}, \xi_\alpha, \alpha \in \chi_{\mathbf{i}})}{2\lambda\sigma^2} dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_{\mathbf{i}}|\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) \\
&\quad = \frac{r_{\mathbf{i}}}{2\lambda\sigma^2} \frac{b^2(\mathbf{i}, \xi_\alpha, \alpha \in \chi_{\mathbf{i}})}{Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_{\mathbf{i}})} dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_{\mathbf{i}}|\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) \\
&\quad = P(\mathbf{I}^\dagger = \mathbf{i}) dF_{\mathbf{i}}^\dagger(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_{\mathbf{i}}|\xi_\alpha, \alpha \in \chi_{\mathbf{i}}),
\end{aligned} \tag{4.118}$$

where

$$dF_{\mathbf{i}}^\dagger(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) = \frac{b^2(\mathbf{i}, \xi_\alpha, \alpha \in \chi_{\mathbf{i}})}{Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_{\mathbf{i}})} dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}). \tag{4.119}$$

Definition (4.119) represents $dF^\dagger(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$ in a manner parallel to (4.116) for $dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$. This representation gives the parallel construction of variables $\mathbf{I}^\dagger$, $\{\Xi_\alpha^\dagger,\ \alpha \in \chi\}$ with distribution $dF^\dagger(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$ as follows. First generate $\mathbf{I}^\dagger$ according to the distribution $P(\mathbf{I}^\dagger = \mathbf{i})$. Then, when $\mathbf{I}^\dagger = \mathbf{i}$, generate $\{\Xi_\alpha^\dagger, \alpha \in \chi_{\mathbf{i}}\}$ according to $dF_{\mathbf{i}}^\dagger(\xi_\alpha, \alpha \in \chi_{\mathbf{i}})$ and then $\{\Xi_\alpha^\dagger,\ \alpha \notin \chi_{\mathbf{i}}\}$ according to $dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \chi_{\mathbf{i}}|\xi_\alpha, \alpha \in \chi_{\mathbf{i}})$. As this last factor is the same as the last factor in (4.116) an opportunity for coupling is presented. In particular, it may be possible to set $\Xi_\alpha^\dagger$ equal to Ξ_α for many $\alpha \notin \chi_{\mathbf{i}}$, thus making the pair $Y^\dagger$, $Y^\ddagger$ close to Y', Y''.

4.4.2 Construction and Bounds for the Combinatorial Central Limit Theorem

In this section we prove Theorem 4.8 by specializing the construction given in Sect. 4.4.1 to handle the combinatorial central limit theorem, and then applying Theorem 4.1. Recall that by (2.45) we may, without loss of generality, replace a_{ij} by $a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}$, and assume

$$a_{i\cdot} = a_{\cdot j} = a_{\cdot\cdot} = 0, \tag{4.120}$$

noting that by doing so we may now write

$$W = Y/\sigma, \tag{4.121}$$

and that (4.107) becomes $\gamma = \sum_{ij} |a_{ij}|^3$.

Now, denoting Y and π by Y' and π', respectively, when convenient, the construction given in Example 2.3 applies. That is, given π, uniform over $\mathcal{S}_n$, take (I, J) independent of π with a uniform distribution over all distinct pairs in $\{1, \ldots, n\}$, in other words, with distribution

$$p_1(i, j) = \frac{1}{(n)_2} \mathbf{1}(i \neq j). \tag{4.122}$$

Letting τ_{ij} be the permutation which transposes i and j, set $\pi'' = \pi\tau_{I,J}$ and let Y'' be given by (4.104) with π'' replacing π. Example 2.3 shows that (Y, Y'') is a $2/(n-1)$-Stein pair, and (2.48) gives

$$Y - Y'' = (a_{I,\pi(I)} + a_{J,\pi(J)}) - (a_{I,\pi(J)} + a_{J,\pi(I)}). \tag{4.123}$$

In particular, averaging over $I, J, \pi(I)$ and $\pi(J)$ we now obtain (4.106) as follows, using (4.109) for the second equality,

$$\begin{aligned} \frac{1}{n^2(n-1)^2} \sum_{i,j,k,l} \big[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})\big]^2 &= E(Y' - Y'')^2 \\ &= 2\lambda\sigma^2 \\ &= \frac{4\sigma^2}{n-1}. \end{aligned} \tag{4.124}$$

We first demonstrate an intermediate result before presenting a coupling construction of Y', Y'' to $Y^\dagger, Y^\ddagger$, leading to a coupling of Y' and Y^*.

Lemma 4.5 *Let π be chosen uniformly from $\mathcal{S}_n$ and suppose $i \neq j$ and $k \neq l$ are elements of $\{1, \ldots, n\}$. Then*

$$\pi^\dagger = \begin{cases} \pi\tau_{\pi^{-1}(k),j} & \text{if } l = \pi(i), k \neq \pi(j), \\ \pi\tau_{\pi^{-1}(l),i} & \text{if } l \neq \pi(i), k = \pi(j), \\ \pi\tau_{\pi^{-1}(k),i}\tau_{\pi^{-1}(l),j} & \text{otherwise}, \end{cases} \tag{4.125}$$

is a permutation that satisfies

$$\pi^\dagger(m) = \pi(m) \quad \textit{for all } m \notin \left\{i, j, \pi^{-1}(k), \pi^{-1}(l)\right\}, \tag{4.126}$$

$$\left\{\pi^\dagger(i), \pi^\dagger(j)\right\} = \{k, l\}, \tag{4.127}$$

and

$$P\left(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\}\right) = \frac{1}{(n-2)!} \tag{4.128}$$

for all distinct $\xi_m^\dagger, m \notin \{i, j\}$ *with* $\xi_m^\dagger \notin \{k, l\}$.

Proof That $\pi^\dagger$ satisfies (4.126) is clear from its definition. To show (4.127) and that $\pi^\dagger$ is a permutation, let A_1, A_2 and A_3 denote the three cases of (4.125) in their respective order. Clearly under A_1 we have

$$\pi^\dagger(t) = \pi(t) \quad \text{for all } t \notin \left\{j, \pi^{-1}(k)\right\}.$$

Hence, as $i \neq j$ and $i = \pi^{-1}(l) \neq \pi^{-1}(k)$, we have $\pi^\dagger(i) = \pi(i) = l$. Also,

$$\pi^\dagger(j) = \pi\tau_{\pi^{-1}(k),j}(j) = \pi\left(\pi^{-1}(k)\right) = k,$$

showing (4.127) holds on A_1. As $\pi^\dagger(\pi^{-1}(k)) = \pi(j)$, both π and $\pi^\dagger$ map the set $\{j, \pi^{-1}(k)\}$ to $\{\pi(j), k\}$, and, as their images agree on $\{j, \pi^{-1}(k)\}^c$, we conclude that $\pi^\dagger$ is a permutation on A_1. As A_2 becomes A_1 upon interchanging i with j and k with l, these conclusions hold also on A_2.

Under A_3, either $l = \pi(i)$, $k = \pi(j)$ or $l \neq \pi(i)$, $k \neq \pi(j)$. In the first instance $\pi^\dagger = \pi$, so $\pi^\dagger$ is a permutation, and (4.127) is immediate. Otherwise, as $i \neq j$ and $i \neq \pi^{-1}(l)$, we have

$$\pi^\dagger(i) = \pi\tau_{\pi^{-1}(k),i}\tau_{\pi^{-1}(l),j}(i) = \pi\tau_{\pi^{-1}(k),i}(i) = \pi\left(\pi^{-1}(k)\right) = k$$

and similarly, as $j \neq i$ and $j \neq \pi^{-1}(k)$,

$$\pi^\dagger(j) = \pi\tau_{\pi^{-1}(k),i}\tau_{\pi^{-1}(l),j}(j) = \pi\tau_{\pi^{-1}(k),i}\left(\pi^{-1}(l)\right), \tag{4.129}$$

and now, as $l \neq k$ and $l \neq \pi(i)$,

$$\pi\tau_{\pi^{-1}(k),i}\left(\pi^{-1}(l)\right) = \pi\left(\pi^{-1}(l)\right) = l,$$

so (4.127) holds under A_3. As both π and $\pi^\dagger$ map $\{i, j, \pi^{-1}(k), \pi^{-1}(l)\}$ to $\{\pi(i), \pi(j), k, l\}$, and agree on $\{i, j, \pi^{-1}(k), \pi^{-1}(l)\}^c$, we conclude that $\pi^\dagger$ is a permutation on A_3.

We now turn our attention to (4.128). Let $\xi_m^\dagger, m \notin \{i, j\}$ be distinct and satisfy $\xi_m^\dagger \notin \{k, l\}$. Under A_1 we have $k \neq \pi(j)$, and have shown that $i \neq \pi^{-1}(k)$. Hence $\pi^{-1}(k) \notin \{i, j\}$ and therefore $\xi^\dagger_{\pi^{-1}(k)} \notin \{k, l\}$. Setting $\xi_i^\dagger = l$, we have

$$\begin{aligned}
&P\left(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\}, A_1\right) \\
&\quad = P\left(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\},\ \pi(i) = l,\ \pi(j) \neq k\right) \\
&\quad = P\left(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{j\},\ \pi(j) \neq k\right) \\
&\quad = P\left(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \left\{j, \pi^{-1}(k)\right\},\ \pi(j) \neq k,\ \pi^\dagger\left(\pi^{-1}(k)\right) = \xi^\dagger_{\pi^{-1}(k)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= P\big(\pi(m) = \xi_m^\dagger,\ m \notin \big\{j, \pi^{-1}(k)\big\},\ \pi(j) \neq k,\ \pi(j) = \xi_{\pi^{-1}(k)}^\dagger\big) \\
&= P\big(\pi(m) = \xi_m^\dagger,\ m \notin \big\{j, \pi^{-1}(k)\big\},\ \pi(j) = \xi_{\pi^{-1}(k)}^\dagger\big) \\
&= \sum_{q \notin \{i,j\}} P\big(\pi(m) = \xi_m^\dagger,\ m \notin \{j, q\},\ \pi(j) = \xi_q^\dagger,\ \pi(q) = k\big) \\
&= \frac{(n-2)}{n!}.
\end{aligned}
$$

Case A_2 being the same upon interchanging i with j and k with l, we obtain

$$
P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\},\ A_1 \cup A_2\big) = \frac{2(n-2)}{n!}. \tag{4.130}
$$

Under A_3 there are subcases depending on

$$
R = \big|\big\{\pi(i), \pi(j)\big\} \cap \{k, l\}\big|,
$$

and we let $A_{3,r} = A_3 \cap \{R = r\}$ for $r = 0, 1, 2$. When $R = 0$ the elements $\pi(i), \pi(j), k, l$ are distinct, and so $A_{3,0} = \{R = 0\}$. Additionally $R = 0$ if and only if the inverse images $i, j, \pi^{-1}(k), \pi^{-1}(l)$ under π are also distinct, and so

$$
\begin{aligned}
&P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\}, A_{3,0}\big) \\
&\quad = P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \big\{i, j, \pi^{-1}(k),\ \pi^{-1}(l)\big\}, \\
&\qquad \pi^\dagger\big(\pi^{-1}(k)\big) = \xi_{\pi^{-1}(k)}^\dagger,\ \pi^\dagger\big(\pi^{-1}(l)\big) = \xi_{\pi^{-1}(l)}^\dagger, A_{3,0}\big) \\
&\quad = P\big(\pi(m) = \xi_m^\dagger,\ m \notin \big\{i, j, \pi^{-1}(k), \pi^{-1}(l)\big\}, \\
&\qquad \pi(i) = \xi_{\pi^{-1}(k)}^\dagger, \pi(j) = \xi_{\pi^{-1}(l)}^\dagger, A_{3,0}\big) \\
&\quad = \sum_{\{q,r\}:\ |\{q,r,i,j\}|=4} P\big(\pi(m) = \xi_m^\dagger,\ k \notin \{i, j, q, r\}, \\
&\qquad\qquad \pi(i) = \xi_q^\dagger,\ \pi(j) = \xi_r^\dagger,\ \pi(q) = k,\ \pi(r) = l\big) \\
&\quad = \frac{(n-2)(n-3)}{n!}.
\end{aligned} \tag{4.131}
$$

Considering the case $R = 1$, in view of (4.125) we find

$$
A_{3,1} = A_3 \cap \{R = 1\} = A_{3,1a} \cup A_{3,1b},
$$

where

$$
A_{3,1a} = \big\{\pi(i) = k, \pi(j) \neq l\big\}, \quad \text{and} \quad A_{3,1b} = \big\{\pi(i) \neq k, \pi(j) = l\big\}.
$$

Since by appropriate relabeling each of these cases becomes A_1, we have

$$
P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\}, A_{3,1}\big) = \frac{2(n-2)}{n!}. \tag{4.132}
$$

For $R = 2$ we have $A_{3,2} = A_{3,2a} \cup A_{3,2b}$ where

$$
A_{3,2a} = \big\{\pi(i) = l,\ \pi(j) = k\big\} \quad \text{and} \quad A_{3,2b} = \big\{\pi(j) = l,\ \pi(i) = k\big\}.
$$

Under $A_{3,2a}$,

$$\begin{aligned} &P\big(\pi^\dagger(m) = \xi_m^\dagger, m \notin \{i, j\}, A_{3,2a}\big) \\ &\quad = P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\},\ \pi(i) = l,\ \pi(j) = k\big) = \frac{1}{n!}, \end{aligned}$$

and the same holding for $A_{3,2b}$, by symmetry, yields

$$P\big(\pi^\dagger(m) = \xi_m^\dagger,\ m \notin \{i, j\}, A_{3,2}\big) = \frac{2}{n!}. \tag{4.133}$$

Summing the contributions from (4.130), (4.131), (4.132) and (4.133) we obtain

$$P\big(\pi^\dagger(m) = \xi_m^\dagger,\ k \notin \{i, j\}\big) = \frac{4(n-2)}{n!} + \frac{(n-2)(n-3)}{n!} + \frac{2}{n!} = \frac{1}{(n-2)!}$$

as claimed. □

The following lemma shows how to choose the 'special' indices in Lemma 4.5 to form the square bias, and hence, zero bias, distributions. In addition, as values of the $\pi^\dagger$ permutation can be made to coincide with those of a given π using (4.125), a coupling of these variables on the same space is achieved. Before stating the lemma we note that (4.134) is a distribution by virtue of (4.106).

Lemma 4.6 *Let*

$$Y = \sum_{i=1}^n a_{i,\pi(i)}$$

with π chosen uniformly from $\mathcal{S}_n$, and let $(I^\dagger, J^\dagger, K^\dagger, L^\dagger)$ be independent of π with distribution

$$p_2(i, j, k, l) = \frac{[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})]^2}{4n^2(n-1)\sigma^2}. \tag{4.134}$$

Further, let $\pi^\dagger$ be constructed from π as in (4.125) with $I^\dagger$, $J^\dagger$, $K^\dagger$ and $L^\dagger$ replacing i, j, k and l, respectively and $\pi^\ddagger = \pi^\dagger \tau_{I^\dagger, J^\dagger}$. Then

$$\pi(i) = \pi^\dagger(i) = \pi^\ddagger(i) \quad \text{for all } i \notin \mathcal{I} \tag{4.135}$$

where $\mathcal{I} = \{I^\dagger, J^\dagger, \pi^{-1}(K^\dagger), \pi^{-1}(L^\dagger)\}$, the variables

$$Y^\dagger = \sum_{i=1}^n a_{i,\pi^\dagger(i)} \quad \text{and} \quad Y^\ddagger = \sum_{i=1}^n a_{i,\pi^\ddagger(i)} \tag{4.136}$$

have the square bias distribution (4.113), and with U an uniform variable on $[0, 1]$, independent of all other variables

$$Y^* = UY^\dagger + (1-U)Y^\dagger$$

has the Y-zero bias distribution.

Proof The claim (4.135) follows from (4.126) and the definition of $\pi^{\ddagger}$. When $\mathbf{I} = (I, J)$ is independent of π with distribution (4.122), $\chi = \{1, \ldots, n\}$ and $\Xi_\alpha = \pi(\alpha)$ for $\alpha \in \chi$, let ψ be the $\mathbb{R}^2$ valued function of $\{\mathbf{I}, \Xi_\alpha, \alpha \in \chi\}$ which yields the exchangeable pair Y', Y'' in Example 2.3. In view of Lemma 4.6, to prove the remainder of the claims it suffices to verify the hypotheses of Lemma 4.4, that is, with $\mathbf{I}^\dagger = (I^\dagger, J^\dagger)$ that $\{\mathbf{I}^\dagger, \Xi^\dagger_\alpha, \alpha \in \chi\}$, or equivalently $\{\mathbf{I}^\dagger, \pi^\dagger(\alpha), \alpha \in \chi\}$, has distribution (4.113). Relying on the discussion following Lemma 4.4, we prove this latter claim by considering the factorization (4.116) of $dF(\mathbf{i}, \xi_\alpha, \alpha \in \chi)$ and show that $\{\mathbf{I}^\dagger, \pi^\dagger(\alpha), \alpha \in \chi\}$ follows the corresponding square bias distribution (4.118).

With $\mathbf{i} = (i, j)$ and $P(\mathbf{I} = \mathbf{i})$ already specified by (4.122), we identify the remaining parts of the factorization (4.116) by noting that the distribution $dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) = dF_{\mathbf{i}}(\xi_i, \xi_j)$ of the images of i and j under π is uniform over all $\xi_i \neq \xi_j$, and, for such ξ_i, ξ_j, $dF_{\mathbf{i}^c|\mathbf{i}}(\xi_\alpha, \alpha \notin \{i, j\}|\xi_i, \xi_j)$ is uniform over all distinct elements $\xi_\alpha, \alpha \in \chi$ that do not intersect $\{\xi_i, \xi_j\}$, that is, for such values

$$dF_{\mathbf{i}^c|\mathbf{i}}\big(\xi_\alpha, \alpha \notin \{i, j\}|\xi_i, \xi_j\big) = \frac{1}{(n-2)!}. \tag{4.137}$$

Now consider the corresponding factorization (4.118). First, this expression specifies the joint distribution of the values $\mathbf{I}^\dagger$ and their images $\Xi^\dagger_\alpha$, $\alpha \in \mathbf{I}^\dagger$ under $\pi^\dagger$ by

$$\begin{aligned} &P(\mathbf{I}^\dagger = \mathbf{i}) dF^\dagger_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}) \\ &\quad = \frac{P(\mathbf{I} = \mathbf{i})}{2\lambda\sigma^2} b^2(\mathbf{i}, \xi_\alpha, \alpha \in \chi_{\mathbf{i}}) dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}}), \end{aligned} \tag{4.138}$$

where from (2.47) for the difference $Y' - Y''$ we have

$$b(i, j, \xi_i, \xi_j) = (a_{i,\xi_i} + a_{j,\xi_j}) - (a_{i,\xi_j} + a_{j,\xi_i}). \tag{4.139}$$

Since the distribution (4.122) of $\mathbf{I}$ is uniform over the range where $i \neq j$, and for such distinct i and j, the distribution $dF_{\mathbf{i}}(\xi_\alpha,\ \alpha \in \chi_{\mathbf{i}})$ is uniform over all distinct choices of images ξ_i and ξ_j, we conclude that the joint distribution (4.138) of $\mathbf{I}^\dagger$ and their 'biased permutation images' $(\Xi^\dagger_{I^\dagger}, \Xi^\dagger_{J^\dagger})$ is proportional to $\mathbf{1}_{i \neq j, k \neq l} b^2(i, j, k, l)$. This is exactly the distribution $p_2(i, j, k, l)$ from which $I^\dagger, J^\dagger, K^\dagger, L^\dagger$ is chosen. In addition, the values $\{K^\dagger, L^\dagger\}$ are the images of $\{I^\dagger, J^\dagger\}$ under the permutation $\pi^\dagger$ constructed as specified in the statement of the lemma, as follows. By (4.134) $I^\dagger \neq J^\dagger$ and $K^\dagger \neq L^\dagger$ with probability one. As $\{I^\dagger, J^\dagger, K^\dagger, L^\dagger\}$ and π are independent, the construction and conclusions of Lemma 4.5 apply, conditional on these indices. Invoking Lemma 4.5, $\pi^\dagger$ is a permutation that maps $\{I^\dagger, J^\dagger\}$ to $\{K^\dagger, L^\dagger\}$.

To show that the remaining values are distributed according to $dF_{\mathbf{i}}(\xi_\alpha, \alpha \in \chi_{\mathbf{i}})$, again by Lemma 4.5, if $\xi^\dagger_m, m \notin \{I^\dagger, J^\dagger\}$ are distinct values not lying in $\{K^\dagger, L^\dagger\}$, then

$$P\big(\pi^\dagger(m) = \xi^\dagger_m, m \notin \{I^\dagger, J^\dagger\}|I^\dagger, J^\dagger, K^\dagger, L^\dagger\big) = \frac{1}{(n-2)!}. \tag{4.140}$$

As (4.140) agrees with (4.137), the proof of the lemma is complete. □

Note that in general even when $\mathbf{I}$ is uniformly distributed, the index $\mathbf{I}^\dagger$ need not be. In fact, from (4.117) it is clear that when $\mathbf{I}$ is uniform the distribution of $\mathbf{I}^\dagger$ is given by $P(\mathbf{I}^\dagger = \mathbf{i}) = 0$ for all $\mathbf{i}$ such that $P(\mathbf{I} = \mathbf{i}) = 0$, and otherwise

$$P(\mathbf{I}^\dagger = \mathbf{i}) = \frac{Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_\mathbf{i})}{\sum_\mathbf{i} Eb^2(\mathbf{i}, \Xi_\alpha, \alpha \in \chi_\mathbf{i})}. \tag{4.141}$$

In particular, the distribution (4.134) selects the indices $\mathbf{I}^\dagger = (I^\dagger, J^\dagger)$ jointly with their 'biased permutation' images $(K^\dagger, L^\dagger)$ with probability that preferentially makes the squared difference large. One can see this effect directly by calculating the marginal distribution of $I^\dagger, J^\dagger$, which, by (4.141), is proportional to $[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})]^2$, by expanding and applying (4.120), yielding

$$\begin{aligned}
&\sum_{k,l} \big[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})\big]^2 \\
&\quad = 2 \sum_{k,l} \big(a_{ik}^2 + a_{jl}^2 - a_{ik}a_{jk} - a_{jl}a_{il}\big) \\
&\quad = 2n \sum_{k=1}^{n} (a_{ik} - a_{jk})^2,
\end{aligned}$$

and hence the generally nonuniform distribution

$$P(I^\dagger = i, J^\dagger = j) = \frac{\sum_{k=1}^n (a_{ik} - a_{jk})^2}{2n(n-1)\sigma^2}.$$

With the construction of the zero bias variable now in hand, Theorem 4.8 follows from Lemma 4.6, Theorem 4.1, (4.10) of Proposition 4.1, and the following lemma.

Lemma 4.7 *For Y and Y^* constructed as in Lemma 4.6*

$$\big\| \mathcal{L}(Y^*) - \mathcal{L}(Y) \big\|_1 \le \frac{\gamma}{(n-1)\sigma^2} \left(8 + \frac{28}{(n-1)} + \frac{4}{(n-1)^2} \right).$$

With π and the indices $\{I^\dagger, J^\dagger, K^\dagger, L^\dagger\}$ constructed as in Lemma 4.6 the calculation of the bound proceeds by decomposing

$$V = Y^* - Y \quad \text{as } V = V\mathbf{1}_2 + V\mathbf{1}_1 + V\mathbf{1}_0$$

where

$$\mathbf{1}_k = \mathbf{1}(R = k) \quad \text{with } R = \big| \big\{ \pi(I^\dagger), \pi(J^\dagger) \big\} \cap \{K^\dagger, L^\dagger\} \big|.$$

The three factors give rise to the three terms of the bound. The proof of the lemma, though not difficult, requires some attention to detail, and can be found in the Appendix to this chapter.

4.5 Simple Random Sampling

Theorem 4.9 gives an L^1 bound for the exchangeable pair coupling. After proving the theorem, we will record a corollary and use it to prove an L^1 bound for simple random sampling. Recall that (Y, Y') is a λ-Stein pair for $\lambda \in (0, 1)$ if (Y, Y') are exchangeable and satisfy the linear regression condition

$$E(Y'|Y) = (1 - \lambda)Y. \tag{4.142}$$

Theorem 4.9 *Let W, W' be a mean zero, variance 1, λ-Stein pair. Then if F is the distribution function of W,*

$$\|F - \Phi\|_1 \le \sqrt{\frac{2}{\pi}} E\left| E\left(1 - \frac{(W' - W)^2}{2\lambda}\middle| W\right)\right| + \frac{1}{2\lambda} E|W' - W|^3.$$

Proof Letting $\Delta = W - W'$, the result follows directly from Proposition 2.4 and Lemma 2.7, the latter which shows that identity (2.76) is satisfied with $R = 0$, $\hat{K}(t)$ given by (2.38), $\hat{K}_1 = E(\Delta^2|W)/2\lambda$ by (2.39), and

$$\begin{aligned}\hat{K}_2 &= \frac{|\Delta|}{2\lambda}\left(\mathbf{1}_{\{-\Delta\le 0\}}\int_{-\Delta}^{0}(-t)dt + \mathbf{1}_{\{-\Delta>0\}}\int_{0}^{-\Delta} t\,dt\right)\\ &= \frac{|\Delta|}{2\lambda}\left(\mathbf{1}_{\{-\Delta\le 0\}}\frac{\Delta^2}{2} + \mathbf{1}_{\{-\Delta>0\}}\frac{\Delta^2}{2}\right) = \frac{|\Delta^3|}{4\lambda}.\end{aligned}$$

□

In many applications calculation of the expectation of the absolute value of the conditional expectation may be difficult. However, by (2.34) we have

$$E\left(\frac{(W' - W)^2}{2\lambda}\right) = 1 \quad \text{so that} \quad E\left(E\left(1 - \frac{(W' - W)^2}{2\lambda}\middle| W\right)\right) = 0.$$

Hence, by the Cauchy–Schwarz inequality,

$$\begin{aligned}E\left|E\left(1 - \frac{(W' - W)^2}{2\lambda}\middle| W\right)\right| &\le \sqrt{\mathrm{Var}\left(E\left(1 - \frac{(W' - W)^2}{2\lambda}\middle| W\right)\right)}\\ &= \frac{1}{2\lambda}\sqrt{\mathrm{Var}\big(E\big((W' - W)^2|W\big)\big)}.\end{aligned}$$

Though the variance of the conditional expectation $E((W' - W)^2|W)$ may still be troublesome, the inequality

$$\mathrm{Var}\big(E(Y|W)\big) \le \mathrm{Var}\big(E(Y|\mathcal{F})\big) \quad \text{when } \sigma\{W\} \subset \mathcal{F} \tag{4.143}$$

often leads to the computation of a tractable bound, and provides estimates which result in the optimal rate. To show (4.143), first note that the conditional variance formula, for any X, yields

$$\mathrm{Var}\big[E(X|W)\big] \le E\big[\mathrm{Var}(X|W)\big] + \mathrm{Var}\big[E(X|W)\big] = \mathrm{Var}(X).$$

However, for $X = E(Y|\mathcal{F})$ we have

$$E(X|W) = E\big(E(Y|\mathcal{F})|W\big) = E(Y|W),$$

and substituting yields (4.143). Hence we arrive at the following corollary to Theorem 4.9.

Corollary 4.3 *Under the assumptions of Theorem 4.9, when $\mathcal{F}$ is any σ-algebra containing $\sigma\{W\}$,*

$$\|F - \Phi\|_1 \le \frac{1}{\lambda}\left(\frac{1}{\sqrt{2\pi}}\Theta + \frac{1}{2}E|W' - W|^3\right),$$

where

$$\Theta = \sqrt{\mathrm{Var}\big(E\big((W' - W)^2|\mathcal{F}\big)\big)}. \tag{4.144}$$

We use Corollary 4.3 to prove an L^1 bound for the sum of numerical characteristics of a simple random sample, that is, for a sample of a population $\{1, \ldots, N\}$ drawn so that all subsets of size n, with $0 < n < N$, are equally likely. The limiting normal distribution for simple random sampling was obtained by Wald and Wolfowitz (1944) (see also Madow 1948; Erdös and Rényi 1959a; and Hájek 1960).

Let $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, N$ denote the characteristic of interest associated with individual i, and let Y be the sum of the characteristics $\{X_1, \ldots, X_n\}$ of the sampled individuals. One can easily verify that the mean μ and variance σ^2 of Y are given by

$$\mu = n\bar{a} \quad \text{and} \quad \sigma^2 = \frac{n(N-n)}{N(N-1)}\sum_{i=1}^{N}(a_i - \bar{a})^2 \quad \text{where } \bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i. \tag{4.145}$$

As we are interested in bounds to the normal for the standardized variable $(Y - \mu)/\sigma$, by replacing a by $(a - \bar{a})/\sqrt{\sum_{b\in\mathcal{A}}(b - \bar{a})^2}$ we may assume in what follows without loss of generality that

$$\bar{a} = 0 \quad \text{and} \quad \sum_{i=1}^{N} a_i^2 = 1. \tag{4.146}$$

For $m = 1, \ldots, n$ let $(n)_m = n(n-1)\cdots(n-m+1)$, the falling factorial of n, and

$$f_m = \frac{(n)_m}{(N)_m}. \tag{4.147}$$

Theorem 4.10 *Let the numerical characteristics $\mathcal{A} = \{a_i,\ i = 1, 2, \ldots, N\}$ of a population of size N satisfy (4.146), and let Y be the sum of characteristics in a simple random sample of size n from $\mathcal{A}$ with $1 < n < N$. Let*

$$\begin{aligned} \sigma^2 &= \frac{n(N-n)}{N(N-1)}, \\ \lambda &= \frac{N}{n(N-n)}, \quad A_4 = \sum_{a\in\mathcal{A}} a^4, \quad \textit{and} \quad \gamma = \sum_{a\in\mathcal{A}} |a|^3. \end{aligned} \tag{4.148}$$

Then with F the distribution function of Y/σ,

$$\|F - \Phi\|_1 \le \frac{1}{\lambda}\left(\frac{R_1}{\sqrt{2\pi}} + \frac{R_2}{2}\right),$$

where

$$R_1 = \frac{1}{n}\sqrt{\frac{2}{\sigma^2}S_1 + \frac{8}{\sigma^4(N-n)^2}S_2}$$

with

$$\begin{aligned} S_1 &= A_4 - \frac{1}{N}, \\ S_2 &= A_4(f_1 - 7f_2 + 6f_3 - 6f_4) + 3(f_2 - f_3 + f_4) - \sigma^4 \quad \textit{and} \\ R_2 &= 8f_1\gamma/\sigma^3. \end{aligned}$$

In the usual asymptotic n and N tend to infinity together with the sampling fraction $f_1 = n/N$ bounded away from zero and one; in such cases $\lambda = O(1/n)$ and $f_m = O(1)$. Additionally, if $a \in \mathcal{A}$ satisfy $\sum_{a\in\mathcal{A}} a^2 = 1$ and are of comparable size then $a = O(1/\sqrt{N})$ which implies $A_4 = O(1/n)$ and $\gamma = O(1/\sqrt{n})$. Overall then the bound provided by the theorem in such an asymptotic, which has main contribution from R_2, is $O(1/\sqrt{n})$.

Since distinct labels may be appended to a_i, $i = 1, \ldots, N$, say as a second coordinate which is neglected when taking sums, we may assume in what follows that elements of $\mathcal{A} = \{a_i,\ i = 1, \ldots, N\}$ are distinct. The first main point of attention is the construction of a Stein pair, which can be achieved as follows. Let $X_1, X_2, \ldots, X_{n+1}$ be a simple random sample of size $n+1$ from the population and let I and I' be two distinct indices drawn uniformly from $\{1, \ldots, n+1\}$. Now set

$$Y = X_I + T \quad \text{and} \quad Y' = X_{I'} + T \quad \text{where } T = \sum_{i\in\{1,\ldots,n+1\}\setminus\{I,I'\}} X_i.$$

As $(X_I, X_{I'}, T) =_d (X_{I'}, X_I, T)$ the variables Y and Y' are exchangeable. By exchangeability and the first condition in (4.146) we have

$$E(X_I|Y) = \frac{1}{n}Y \quad \text{and} \quad E(X_{I'}|Y) = -\frac{1}{N-n}Y,$$

and therefore

$$E(Y'|Y) = E(Y - X_I + X_{I'}|Y) = (1-\lambda)Y$$

where $\lambda \in (0, 1)$ is given by (4.148); the linearity condition (4.142) is satisfied.

Before starting the proof we pause to simplify the required moment calculations for $\mathcal{X} = \{X_1, \ldots, X_n\}$, a simple random sample of $\mathcal{A}$. For $m \in \mathbb{N}$, $\{k_1, \ldots, k_m\} \subset \mathbb{N}$ and $\mathbf{k} = (k_1, \ldots, k_m)$ let

$$[\mathbf{k}] = E\left(\sum_{\{a,b,\ldots,c\} \subset \mathcal{X},\, |\{a,b,\ldots,c\}|=m} a^{k_1} b^{k_2} \cdots c^{k_m} \right)$$

and

$$\langle \mathbf{k} \rangle = \sum_{\{y_1,\ldots,y_m\} \subset \mathcal{A},\, |\{y_1,\ldots,y_m\}|=m} y_1^{k_1} y_2^{k_2} \ldots y_m^{k_m}.$$

Now observe that, with f_m given in (4.147),

$$[\mathbf{k}] = f_m \langle \mathbf{k} \rangle. \tag{4.149}$$

As $[\mathbf{k}]$ and $\langle \mathbf{k} \rangle$ are invariant under any permutation of its components we may always use the canonical representation where $k_1 \geq \cdots \geq k_m$.

Let e_j^m be the jth unit vector in $\mathbb{R}^m$. When the population characteristics satisfy (4.146) we have

$$\langle k_1, \ldots, k_{m-1}, 1 \rangle = -\sum_{j=1}^{m-1} \left\langle (k_1, \ldots, k_{m-1}) + e_j^{m-1} \right\rangle \quad \text{and}$$

$$\langle k_1, \ldots, k_{m-1}, 2 \rangle = \langle k_1, \ldots, k_{m-1} \rangle - \sum_{j=1}^{m-1} \left\langle (k_1, \ldots, k_{m-1}) + 2e_j^{m-1} \right\rangle.$$

Note then that

$$\begin{aligned} \langle 2 \rangle &= 1 \\ \langle 3, 1 \rangle &= -\langle 4 \rangle \\ \langle 2, 2 \rangle &= \langle 2 \rangle - \langle 4 \rangle \\ \langle 2, 1, 1 \rangle &= -\langle 3, 1 \rangle - \langle 2, 2 \rangle = \langle 4 \rangle - \langle 2 \rangle + \langle 4 \rangle = 2\langle 4 \rangle - \langle 2 \rangle \\ \langle 1, 1, 1, 1 \rangle &= -3\langle 2, 1, 1 \rangle = -6\langle 4 \rangle + 3\langle 2 \rangle. \end{aligned} \tag{4.150}$$

Proof of Theorem 4.10 We may assume $n \leq N/2$, as otherwise we may replace Y, a sample of size n from $\mathcal{A}$, by $-Y$, a sample of size $N - n$; this assumption is used in (4.151).

We apply Corollary 4.3, beginning with the first term in the bound. Letting $\mathcal{X} = \{X_j,\ j \neq I'\}$ and $\mathcal{F} = \sigma(\mathcal{X})$, applying inequality (4.143) yields

$$\begin{aligned} \mathrm{Var}\big(E\big((Y' - Y)^2 | Y\big)\big) &\leq \mathrm{Var}\big(E\big((Y' - Y)^2 | \mathcal{F}\big)\big) \\ &= \mathrm{Var}\big(E\big((X_{I'} - X_I)^2 | \mathcal{F}\big)\big) \\ &= \mathrm{Var}\big(E\big(X_{I'}^2 - 2X_{I'}X_I + X_I^2 | \mathcal{F}\big)\big). \end{aligned}$$

For these three conditional expectations,

$$E\big(X_{I'}^2|\mathcal{F}\big) = \frac{1}{N-n}\sum_{b\notin\mathcal{X}} b^2,$$

$$E(X_{I'}X_I|\mathcal{F}) = \frac{1}{n(N-n)}\sum_{a\in\mathcal{X},b\notin\mathcal{X}} ab \quad \text{and} \quad E\big(X_I^2|\mathcal{F}\big) = \frac{1}{n}\sum_{a\in\mathcal{X}} a^2.$$

By the standardization (4.146) we have,

$$\frac{1}{N-n}\sum_{b\notin\mathcal{X}} b^2 = \frac{1}{N-n}\Big(1-\sum_{a\in\mathcal{X}} a^2\Big)$$

$$\text{and} \quad \frac{1}{n(N-n)}\sum_{a\in\mathcal{X}}\sum_{b\notin\mathcal{X}} ab = -\frac{1}{n(N-n)}\Big(\sum_{a\in\mathcal{X}} a\Big)^2.$$

Hence, using $\mathrm{Var}(U+V) \le 2(\mathrm{Var}(U)+\mathrm{Var}(V))$,

$$\begin{aligned}
&\mathrm{Var}\big(E\big((Y'-Y)^2|Y\big)\big)\\
&\le \mathrm{Var}\Big(\frac{N-2n}{n(N-n)}\sum_{a\in\mathcal{X}} a^2 + \frac{2}{n(N-n)}\Big(\sum_{a\in\mathcal{X}} a\Big)^2\Big)\\
&\le 2\Big(\frac{1}{n^2}\mathrm{Var}\Big(\sum_{a\in\mathcal{X}} a^2\Big) + \Big(\frac{2}{n(N-n)}\Big)^2\mathrm{Var}\Big(\sum_{a\in\mathcal{X}} a\Big)^2\Big). \qquad (4.151)
\end{aligned}$$

Calculating the first variance in (4.151), using (4.149), we begin with

$$\Big(E\sum_{a\in\mathcal{X}} a^2\Big)^2 = [2]^2 = \big(f_1\langle 2\rangle\big)^2 = f_1^2.$$

Next, note

$$\begin{aligned}
E\Big(\sum_{a\in\mathcal{X}} a^2\Big)^2 &= [4]+[2,2] = f_1\langle 4\rangle + f_2\langle 2,2\rangle\\
&= f_1\langle 4\rangle + f_2\big(\langle 2\rangle - \langle 4\rangle\big) = \frac{n(N-n)}{N(N-1)}\langle 4\rangle + f_2,
\end{aligned}$$

and therefore

$$\mathrm{Var}\Big(\sum_{a\in\mathcal{X}} a^2\Big) = \frac{n(N-n)}{N(N-1)}\Big(\langle 4\rangle - \frac{1}{N}\Big) = \sigma^2 S_1.$$

For the second variance in (4.151), using (4.149) and (4.150) we first obtain the expectation

$$E\Big(\sum_{a\in\mathcal{X}} a\Big)^2 = [2]+[1,1] = f_1 - f_2 = \sigma^2. \qquad (4.152)$$

Similarly, for the second moment we compute

$$E\left(\sum_{a\in\mathcal{X}} a\right)^4 = [4] + 4[3,1] + 3[2,2] + 3[2,1,1] + [1,1,1,1]$$
$$= f_1\langle 4\rangle + f_2\big(4\langle 3,1\rangle + 3\langle 2,2\rangle\big) + f_3 3\langle 2,1,1\rangle + f_4\langle 1,1,1,1\rangle$$
$$= \langle 4\rangle(f_1 - 7f_2 + 6f_3 - 6f_4) + 3(f_2 - f_3 + f_4).$$

The variance of this term is now obtained by subtracting the square of the expectation (4.152), resulting in the quantity S_2.

Hence, from (4.151),

$$\mathrm{Var}\big(E\big((Y'-Y)^2|Y\big)\big) \le \frac{1}{n^2}\left(2\sigma^2 S_1 + \frac{8}{(N-n)^2}S_2\right),$$

and therefore, with $W = Y/\sigma$ and $W' = Y'/\sigma$, we have

$$\sqrt{\mathrm{Var}\big(E\big((W'-W)^2|W\big)\big)} = \sqrt{\mathrm{Var}\big(E\big((Y'-Y)^2|Y\big)\big)/\sigma^4} = R_1.$$

Regarding the second term in Corollary 4.3, as

$$E|Y'-Y|^3 = E|X_{I'} - X_I|^3 \le 8E|X_I|^3 = 8\frac{n}{N}\sum_{a\in\mathcal{A}}|a|^3 = 8f_1\gamma,$$

we obtain

$$E|W'-W|^3 = 8f_1\gamma/\sigma^3 = R_2.$$

□

4.6 Chatterjee's L^1 Theorem

The basis of all normal Stein identities is that $Z \sim \mathcal{N}(0,1)$ if and only if

$$E\big[Zf(Z)\big] = E\big[f'(Z)\big] \tag{4.153}$$

for all absolutely continuous functions f for which these expectations exist. For a mean zero, variance one random variable W which may be close to normal, (4.153) may hold approximately, and there may therefore be a related identity which holds exactly for W. One way the identity (4.153) may be altered to hold exactly for some given W is to no longer insist that the same variable, W, appear on the right hand side as on the left, thus leading to the zero bias identity (2.51)

$$E\big[Wf(W)\big] = E\big[f'(W^*)\big], \tag{4.154}$$

as discussed in Sect. 2.3.3. Insisting that W appear on both sides, one may be lead instead to consider identities of the form

$$E\big[Wf(W)\big] = E\big[f'(W)T\big], \tag{4.155}$$

for some random variable T, defined on the same space as W. When such a T exists, by conditioning we obtain

$$E\big[f'(W^*)\big] = E\big[Wf(W)\big] = E\big[f'(W)T\big] = E\big[f'(W)E(T|W)\big],$$

which reveals that

$$E(T|W=w) = \frac{dF^*(w)}{dF(w)}$$

is the Radon–Nikodym derivative of the zero bias distribution of W with respect to the distribution of W. In particular, as W^* always has an absolutely continuous distribution, for there to exist a T such that (4.155) holds it is necessary for W to be absolutely continuous; naturally, in other cases, considering approximations allows the equality to become relaxed. Identities of the form (4.155), in some generality, were considered in Cacoullos and Papathanasiou (1992), but T was constrained to be a function of W. As we will see, much more flexibility is provided by removing this restriction.

Theorem 4.11, of Chatterjee (2008), gives bounds to the normal, in the L^1 norm, for a mean zero function $\psi(\mathbf{X})$ of a vector of independent random variables $\mathbf{X} = (X_1, \ldots, X_n)$ taking values in some space $\mathcal{X}$. For the identity (4.155), or an approximate form thereof, to be useful, a viable T must be produced. Towards this goal, with $\mathbf{X}'$ an independent copy of $\mathbf{X}$, and $A \subset \{1, \ldots, n\}$, let $\mathbf{X}^A$ be the random vector with components

$$X_j^A = \begin{cases} X_j' & j \in A, \\ X_j & j \notin A. \end{cases} \tag{4.156}$$

For $i \in \{1, \ldots, n\}$, writing i for $\{i\}$ when notationally convenient, let

$$\Delta_i \psi(\mathbf{X}) = \psi(\mathbf{X}) - \psi\big(\mathbf{X}^i\big), \tag{4.157}$$

which measures the sensitivity of the function ψ to the values in its ith coordinate. Now, for any $A \subset \{1, \ldots, n\}$, let

$$T_A = \sum_{i \notin A} \Delta_i \psi(\mathbf{X}) \Delta_i \psi\big(\mathbf{X}^A\big) \quad \text{and} \quad T = \frac{1}{2} \sum_{\substack{A \subset \{1,\ldots,n\} \\ |A| \neq n}} \frac{T_A}{\binom{n}{|A|}(n - |A|)}. \tag{4.158}$$

Theorem 4.11 *Let $W = \psi(\mathbf{X})$ be a function of a vector of independent random variables $\mathbf{X} = (X_1, \ldots, X_n)$, and have mean zero and variance 1. Then, with Δ_i as defined in (4.157) and T given in (4.158) we have that $ET = 1$ and*

$$\big\|\mathcal{L}(W) - \mathcal{L}(Z)\big\|_1 \leq \sqrt{2/\pi}\sqrt{\mathrm{Var}\big(E(T|W)\big)} + \frac{1}{2}\sum_{i=1}^n E\big|\Delta_i \psi(\mathbf{X})\big|^3.$$

We present the proof, from Chatterjee (2008), at the end of this section.

To explore a simple application, let $\psi(\mathbf{X}) = \sum_{i=1}^n X_i$ where $X_1, \ldots, X_n$ are independent with mean zero, variances $\sigma_1^2, \ldots, \sigma_n^2$ summing to one, and fourth moments $\tau_1, \ldots, \tau_n$. For $A \subset \{1, \ldots, n\}$ and $i \notin A$,

$$\begin{aligned} \Delta_i \psi\big(\mathbf{X}^A\big) &= \psi\big(\mathbf{X}^A\big) - \psi\big(\mathbf{X}^{A \cup i}\big) \\ &= \sum_{j \notin A} X_j + \sum_{j \in A} X_j' - \Bigg(\sum_{j \notin A \cup i} X_j + \sum_{j \in A \cup i} X_j' \Bigg) = X_i - X_i'. \end{aligned} \tag{4.159}$$

Hence,

$$T_A = \sum_{i \notin A} \Delta_i \psi(\mathbf{X}) \Delta_i \psi(\mathbf{X}^A) = \sum_{i \notin A} (X_i - X_i')^2,$$

and

$$\begin{aligned} T &= \frac{1}{2} \sum_{A \subset \{1,\ldots,n\},\, |A| \neq n} \frac{T_A}{\binom{n}{|A|}(n - |A|)} \\ &= \frac{1}{2} \sum_{a=0}^{n-1} \frac{1}{\binom{n}{a}(n-a)} \sum_{A \subset \{1,\ldots,n\},\, |A|=a} T_A \\ &= \frac{1}{2} \sum_{a=0}^{n-1} \frac{1}{\binom{n}{a}(n-a)} \sum_{A \subset \{1,\ldots,n\}, |A|=a} \sum_{i \notin A} (X_i - X_i')^2 \\ &= \frac{1}{2} \sum_{a=0}^{n-1} \frac{1}{\binom{n}{a}(n-a)} \sum_{i=1}^{n} \sum_{A \subset \{1,\ldots,n\},\, |A|=a, A \not\ni i} (X_i - X_i')^2. \end{aligned}$$

As for each $i \in \{1, \ldots, n\}$ there are $\binom{n-1}{a}$ subsets of A of size a that do not contain i, we obtain

$$\begin{aligned} T &= \frac{1}{2} \sum_{a=0}^{n-1} \frac{1}{\binom{n}{a}(n-a)} \sum_{i=1}^{n} (X_i - X_i')^2 \sum_{A \subset \{1,\ldots,n\},\, |A|=a, A \not\ni i} 1 \\ &= \left(\frac{1}{2} \sum_{i=1}^{n} (X_i - X_i')^2 \right) \left(\sum_{a=0}^{n-1} \frac{1}{\binom{n}{a}(n-a)} \binom{n-1}{a} \right) \\ &= \frac{1}{2} \sum_{i=1}^{n} (X_i - X_i')^2. \end{aligned}$$

For the first term in the theorem, applying the bound (4.143) with $\mathcal{F}$ the σ-algebra generated by $\mathbf{X}$ we obtain

$$\operatorname{Var}\big(E(T|W)\big) \leq \operatorname{Var}(T) = \frac{1}{4} \sum_{i=1}^{n} \operatorname{Var}\big((X_i - X_i')^2\big) = \frac{1}{2} \sum_{i=1}^{n} (\tau_i + 3\sigma_i^4).$$

From (4.159),

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^{n} E\big|\Delta_i \psi(\mathbf{X})\big|^3 = \frac{1}{2} \sum_{i=1}^{n} E\big|X_i - X_i'\big|^3 &\leq \frac{1}{2} \sum_{i=1}^{n} \big(E(X_i - X_i')^4\big)^{3/4} \\ &= \frac{1}{2^{1/4}} \sum_{i=1}^{n} (\tau_i + 3\sigma_i^4)^{3/4}. \end{aligned}$$

Invoking Theorem 4.11 yields,

$$\big\|\mathcal{L}(W)-\mathcal{L}(Z)\big\|_1 \le \sqrt{\frac{1}{\pi}\sum_{i=1}^n\big(\tau_i+3\sigma_i^4\big)}+\frac{1}{2^{1/4}}\sum_{i=1}^n\big(\tau_i+3\sigma_i^4\big)^{3/4}.$$

When $X_1,\dots,X_n$ are independent, mean zero variables having common second and fourth moments, say, σ^2 and τ, respectively, then applying this result to $W=(X_1+\cdots+X_n)/\sqrt{n}$ yields

$$\big\|\mathcal{L}(W)-\mathcal{L}(Z)\big\|_1 \le n^{-1/2}\left(\sqrt{\frac{1}{\pi}\big(\tau+3\sigma^4\big)}+\frac{1}{2^{1/4}}\big(\tau+3\sigma^4\big)^{3/4}\right).$$

For a different application of Theorem 4.11 we consider normal approximation of quadratic forms. Let $\mathrm{Tr}(A)$ denote the trace of A.

Proposition 4.7 *Let $\mathbf{X}=(X_1,\dots,X_n)$ be a vector of independent variables taking the values $+1,-1$ with equal probability, A a real symmetric matrix and $Y=\sum_{i\le j}a_{ij}X_iX_j$. Then the mean μ and variance σ^2 of Y are given by*

$$\mu=\mathrm{Tr}(A)\quad\text{and}\quad\sigma^2=\frac{1}{2}\mathrm{Tr}\big(A^2\big),\tag{4.160}$$

and $W=(Y-\mu)/\sigma$ satisfies

$$\big\|\mathcal{L}(W)-\mathcal{L}(Z)\big\|_1 \le \left(\frac{1}{\pi\sigma^4}\mathrm{Tr}\big(A^4\big)\right)^{1/2}+\frac{7}{2\sigma^3}\sum_{i=1}^n\left(\sum_{j=1}^n a_{ij}^2\right)^{3/2}.$$

Proof The mean and variance formulas (4.160) can be obtained by specializing Theorems 1.5 and 1.6 of Seber and Lee (2003) to $\mathbf{X}$ with the given distribution. By subtracting the mean and then replacing a_{ij} by a_{ij}/σ it suffices to prove the result when $a_{ii}=0$ and $\sigma^2=1$. Letting

$$\psi(\mathbf{x})=\sum_{i<j}a_{ij}x_ix_j$$

for $\mathbf{x}\in\mathbb{R}^n$, with $\mathbf{x}^i$ the vector $\mathbf{x}$ with x_i' replacing x_i and using the symmetry of A we have

$$\begin{aligned}\Delta_i\psi(\mathbf{x})&=\psi(\mathbf{x})-\psi\big(\mathbf{x}^i\big)\\&=\sum_{j:\,i<j}a_{ij}x_ix_j+\sum_{j:\,j<i}a_{ji}x_jx_i-\sum_{j:\,i<j}a_{ij}x_i'x_j-\sum_{j:\,j<i}a_{ji}x_jx_i'\\&=\big(x_i-x_i'\big)\sum_{j=1}^n a_{ij}x_j.\end{aligned}$$

By replacing $\mathbf{x}$ above by $\mathbf{X}^A$, for $i\notin A$ we have

$$\Delta_i\psi\big(\mathbf{X}^A\big)=\big(X_i-X_i'\big)\left(\sum_{j\notin A}a_{ij}X_j+\sum_{j\in A}a_{ij}X_j'\right).$$

We apply the bound $\mathrm{Var}(E(T|W))\le\mathrm{Var}(E(T|\mathbf{X}))$, from (4.143). For the calculation of $E(T|\mathbf{X})$, with $A\subset\{1,\dots,n\}$ and $i\notin A$, using that X_i, X_i' are in -1, 1, we

have

$$
\begin{aligned}
&E\left(\Delta_i\psi(\mathbf{X})\Delta_i\psi\left(\mathbf{X}^A\right)|\mathbf{X}\right)\\
&\quad= E\left(\left(X_i - X_i'\right)^2\left(\sum_{j=1}^n a_{ij}X_j\right)\left(\sum_{j\notin A} a_{ij}X_j + \sum_{j\in A} a_{ij}X_j'\right)\middle|\mathbf{X}\right)\\
&= \left(\sum_{j=1}^n a_{ij}X_j\right)E\left(\left(X_i - X_i'\right)^2\left(\sum_{j\notin A} a_{ij}X_j + \sum_{j\in A} a_{ij}X_j'\right)\middle|\mathbf{X}\right)\\
&\quad= 2\left(\sum_{j=1}^n a_{ij}X_j\right)E\left(\left(1 - X_iX_i'\right)\left(\sum_{j\notin A} a_{ij}X_j + \sum_{j\in A} a_{ij}X_j'\right)\middle|\mathbf{X}\right)\\
&\quad= 2\left(\sum_{j=1}^n a_{ij}X_j\right)\left(\sum_{j\notin A} a_{ij}X_j\right),
\end{aligned}
$$

where, since $i\notin A$, all the remaining terms have conditional mean zero. Hence we may write

$$
E\left(\Delta_i\psi(\mathbf{X})\Delta_i\psi\left(\mathbf{X}^A\right)|\mathbf{X}\right) = 2\sum_{j\in\{1,\ldots,n\},\, k\notin A} a_{ij}a_{ik}X_jX_k.
$$

Summing over all $i\notin A$, (4.158) yields

$$
E(T_A|\mathbf{X}) = 2\sum_{i\notin A}\sum_{j\in\{1,\ldots,n\},\, k\notin A} a_{ij}a_{ik}X_jX_k.
$$

From the definition of T, again from (4.158),

$$
\begin{aligned}
E(T|\mathbf{X}) &= \frac{1}{2}\sum_{\substack{A\subset\{1,\ldots,n\}\\ |A|\neq n}} \frac{T_A}{\binom{n}{|A|}(n-|A|)}\\
&= \sum_{\substack{A\subset\{1,\ldots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{i\notin A}\sum_{j\in\{1,\ldots,n\},\, k\notin A} a_{ij}a_{ik}X_jX_k\\
&= \sum_{j=1}^n \sum_{\substack{A\subset\{1,\ldots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{i\notin A}\sum_{k\notin A} a_{ij}a_{ik}X_jX_k\\
&= \sum_{1\le i,j,k\le n} a_{ij}a_{ik}X_jX_k \sum_{A\cap\{i,k\}=\emptyset} \frac{1}{\binom{n}{|A|}(n-|A|)}\\
&= \sum_{1\le i,j,k\le n} a_{ij}a_{ik}X_jX_k \sum_{a=0}^{n-2}\sum_{A\cap\{i,k\}=\emptyset,\,|A|=a} \frac{1}{\binom{n}{a}(n-a)}\\
&= \sum_{1\le i,j,k\le n} a_{ij}a_{ik}X_jX_k \sum_{a=0}^{n-2} \frac{\binom{n-2}{a}}{\binom{n}{a}(n-a)}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{1\le i,j,k\le n} a_{ij}a_{ik}X_jX_k \sum_{a=0}^{n-2} \frac{n-a-1}{n(n-1)} \\
&= \sum_{1\le i,j,k\le n} a_{ij}a_{ik}X_jX_k \left(\frac{1}{n(n-1)} \sum_{a=1}^{n-1} a \right) \\
&= \frac{1}{2} \sum_{1\le i,j,k\le n} a_{ji}a_{ik}X_jX_k \\
&= \frac{1}{2}\mathbf{X}^\intercal A^2 \mathbf{X}.
\end{aligned}
$$

Letting $b_{jk} = \sum_{1\le i,j\le n} a_{ji}a_{ik}$, the jkth element of A^2, again using $X_i^2 = 1$,

$$
\operatorname{Var}\big(E(T|\mathbf{X})\big) = \operatorname{Var}\left(\sum_{j<k} b_{jk}X_jX_k \right) = \sum_{j<k} b_{jk}^2 \le \frac{1}{2}\operatorname{Tr}\big(A^4\big).
$$

To bound the final term in Theorem 4.11, we apply Khintchine's inequality, see Haagerup (1982), which yields

$$
E\left| \sum_{j=1}^n a_jX_j \right|^p \le B_p^p \left(\sum_{j=1}^n a_j^2 \right)^{p/2}
$$

$$
\text{where } B_p = \begin{cases} 1 & 0 < p \le 2, \\ 2^{1/2}(\Gamma((p+1)/2)/\sqrt{\pi})^{1/p} & 2 < p < \infty. \end{cases}
$$

In particular $B_3^3 \le 1.6$, and using the fact that X_i is independent of the event $\{X_i \ne X_i'\}$, we obtain

$$
E\big|\Delta_i\psi(\mathbf{X})\big|^3 = 4E\left| \sum_{j=1}^n a_{ij}X_j \right|^3 \le 7\left(\sum_{j=1}^n a_{ij}^2 \right)^{3/2}. \qquad \square
$$

To consider some further examples, we make the following definition. With $\mathcal{X}$ the space in which our random variables take values, given $n \in \mathbb{N}$ suppose there is a map $\mathcal{G}$, or 'graphical rule', which to every $\mathbf{x} \in \mathcal{X}^n$ assigns an undirected graph, that is, a collection of edges $\mathcal{G}(\mathbf{x})$ on the vertices $\{1,\dots,n\}$. We will say the map $\mathcal{G}$ is symmetric if it respects the action of permutations, that is, if for every permutation π of $\{1,\dots,n\}$ and any $(x_1,\dots,x_n) \in \mathcal{X}^n$,

$$
\begin{aligned}
&\big\{\{i,j\}\colon \{i,j\} \in \mathcal{G}(x_{\pi(1)},\dots,x_{\pi(n)})\big\} \\
&\quad = \big\{\big\{\pi(i),\pi(j)\big\}\colon \{i,j\} \in \mathcal{G}(x_1,\dots,x_n)\big\}.
\end{aligned}
$$

Now fixing $m > n$, we say the vector $\mathbf{x} \in \mathcal{X}^n$ is embedded in the vector $\mathbf{y} \in \mathcal{X}^m$ if there exist distinct indices $i_1,\dots,i_n$ in $\{1,\dots,m\}$ with $x_k = y_{i_k}$ for $1 \le k \le n$. A graphical rule $\mathcal{G}'$ on $\mathcal{X}^m$ will be called an extension of the rule $\mathcal{G}$ if whenever the vector $\mathbf{x} \in \mathcal{X}^n$ is embedded in $\mathbf{y} \in \mathcal{X}^m$ the graph $\mathcal{G}(\mathbf{x})$ on $\{1,\dots,n\}$ is the naturally induced subgraph of $\mathcal{G}(\mathbf{y})$ on $\{1,\dots,m\}$.

Now let $\mathbf{x}$ and $\mathbf{x}'$ be any two elements of $\mathcal{X}^n$. For every $i \in \{1,\dots,n\}$, let $\mathbf{x}^i$ be the vector obtained by replacing x_i by x_i' in $\mathbf{x}$, and, for i and j distinct elements of

$\{1, \ldots, n\}$, let $\mathbf{x}^{ij}$ be similarly obtained be replacing x_i and x_j in $\mathbf{x}$ by x_i' and x_j', respectively. With $\psi : \chi^n \to \mathbb{R}$, we say the coordinates i and j are non-interacting with respect to the triple $(\psi, \mathbf{x}, \mathbf{x}')$ if

$$\psi(\mathbf{x}) - \psi\big(\mathbf{x}^j\big) = \psi\big(\mathbf{x}^i\big) - \psi\big(\mathbf{x}^{ij}\big).$$

We will say that $\mathcal{G}$ is an interaction rule for a function ψ if for any choice of $\mathbf{x}, \mathbf{x}'$ and i, j, the event that $\{i, j\}$ is not an edge in the graphs $\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x}^i), \mathcal{G}(\mathbf{x}^j), \mathcal{G}(\mathbf{x}^{ij})$ implies that i and j are non-interacting vertices with respect to $(\psi, \mathbf{x}, \mathbf{x}')$. With these definitions in hand, we can now state the following theorem; we present the proof, from Chatterjee (2008), at the end of this section.

Theorem 4.12 *Let the symmetric map $\mathcal{G}$ be an interaction rule for $\psi : \mathcal{X}^n \to \mathbb{R}$, and $\mathbf{X} = (X_1, \ldots, X_n)$ a vector of i.i.d. $\mathcal{X}$ valued variates such that $W = \psi(\mathbf{X})$ has mean zero and variance* 1. *For each $i \in \{1, \ldots, n\}$ define*

$$\Delta_i \psi(\mathbf{X}) = \psi(\mathbf{X}) - \psi\big(\mathbf{X}^i\big)$$

where $\mathbf{X}'$ is an independent copy of $\mathbf{X}$, and let

$$M = \max_{i=1,\ldots,n} \big|\Delta_i \psi(\mathbf{X})\big|. \tag{4.161}$$

Let $\mathcal{G}'$ be any extension of $\mathcal{G}$ on $\mathcal{X}^{n+4}$, and set

$$\delta = 1 + \textit{degree of vertex } 1 \textit{ in } \mathcal{G}'(X_1, \ldots, X_{n+4}). \tag{4.162}$$

Then for some universal constant C,

$$\big\|\mathcal{L}(W) - \mathcal{L}(Z)\big\|_1 \le Cn^{1/2} E\big(M^8\big)^{1/4} E\big(\delta^4\big)^{1/4} + \frac{1}{2}\sum_{i=1}^{n} E\big|\Delta_i \psi(\mathbf{X})\big|^3.$$

Following Chatterjee (2008), we apply Theorem 4.12 to prove an L^1 bound to the normal for two problems which stem from the theory of coverage processes; the volume of the region covered by the union of n balls with random centers and some radius, and the number of such centers that are isolated at some radius; see Hall (1988) and Penrose (2003) for more background. Generally, we may work in a separable metric space $(\mathcal{X}, \rho)$, and for the first case, we take as given one endowed with measure λ. Let the components of $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. with values in $\mathcal{X}$. For some fixed radius $r > 0$, let $\mathcal{R}$ be given by

$$\mathcal{R} = \mathcal{R}(\mathbf{X}) \quad \text{where } \mathcal{R}(\mathbf{x}) = \bigcup_{i=1}^{n} B(x_i, r), \tag{4.163}$$

with $B(x, r)$ the closed ball of radius r centered at x. Proposition 4.8 gives an L^1 bound to the normal for the 'covered volume' $\lambda(\mathcal{R}(\mathbf{X}))$ in terms of

$$K_V = \sup_{u \in \mathcal{X}} \lambda\big(B(u, r)\big), \tag{4.164}$$

an upper bound to the volume of any ball of radius r.

By very similar reasoning, we also derive an L^1 bound to the normal for the number $\mathcal{S}$ of isolated points, or singletons, given by

$$\mathcal{S} = \mathcal{S}(\mathbf{X}) \quad \text{where } \mathcal{S}(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{1}\big(\{x_1, \ldots, x_n\} \cap B(x_i, 2r) = \{x_i\}\big), \quad (4.165)$$

that is, the number of points of $\mathbf{X}$ such that the ball $B(X_i, r)$ had empty intersection with $B(X_j, r)$ for all $j \neq i$. Proposition 4.8 gives an L^1 bound to the normal for $\mathcal{S}$ in terms of

$$K_S = \sup\big\{k\colon\ \exists \mathbf{x} \in \mathcal{X}^{k+1} \text{ such that } B(x_{k+1}, r) \cap B(x_i, r) \neq \emptyset, \text{ and } B(x_i, r) \cap B(x_j, r) = \emptyset \text{ for all distinct } 1 \leq i, j \leq k\big\}, \quad (4.166)$$

which is an upper bound to the number of points in any collection from $\mathcal{X}$ which may become isolated upon the removal of a single point. In Euclidean space, the number K_S is a lower bound to the so called kissing number, the maximum number of spheres of radius 1 that can simultaneously touch the unit sphere at the origin; see Zong (1999), Conway and Sloane (1999), and Leech and Sloane (1971) for estimates on the kissing number. For example, in two dimensions $K_S = 5$, since at most five unit circles can intersect another unit circle without intersecting each other, while the kissing number in two dimensions is 6.

Proposition 4.8 *With $p = P(\rho(X_1, X_2) \leq 2r)$ we have*

$$\big\|\mathcal{L}(W_V) - \mathcal{L}(Z)\big\|_1 \leq \frac{Cn^{1/2}K_V^2(1+np)}{\sigma_V^2} + \frac{nK_V^3}{2\sigma_V^3}$$

for some universal constant C, with $\mu_V = EY_V$, $\sigma_V^2 = \mathrm{Var}(Y_V)$ and $W_V = (Y_V - \mu_V)/\sigma_V$, when $Y_V = \lambda(\mathcal{R})$ with $\mathcal{R}$ as given in (4.163) and K_V as in (4.164). The same bound holds for $Y_S = \mathcal{S}$ in (4.165) and $W_S = (Y_V - \mu_S)/\sigma_S$ where $\mu_S = EY_S$ and $\sigma_S^2 = \mathrm{Var}(Y_S)$, with the same constant C, upon replacing σ_V and K_V by σ_S and K_S, respectively.

Proof It suffices to prove the theorem when the variables standardized to have mean zero and variance one; we apply Theorem 4.12. First we consider $\mathcal{R}$, and let $\psi(x) = \lambda(\mathcal{R}(\mathbf{x}))$ for $\mathbf{x} \in \mathcal{X}^n$. Let $\mathcal{G}(\mathbf{x})$ be the graph on $\{1, \ldots, n\}$ with edges between points i and j if and only if $\rho(x_i, x_j) \leq 2r$. Clearly the graphical rule $\mathcal{G}$ is symmetric, as distances are unchanged by relabeling.

We verify that $\mathcal{G}$ is an interaction rule as follows. With $\mathbf{x}$ and $\mathbf{x}'$ any points in $\mathcal{X}^n$, let $\mathbf{x}^i$ and $\mathbf{x}^{ij}$ be obtained by replacing the ith, or both the ith and jth, coordinate respectively of $\mathbf{x}$ by those of $\mathbf{x}'$. Writing B_j and B'_j for $B(x_j, r)$ and $B(x'_j, r)$ respectively, we let

$$R_i = \bigcup_{j \neq i} B_j \quad \text{so that} \quad \psi(\mathbf{x}) = R_i \cup B_i \quad \text{and} \quad \psi(\mathbf{x}') = R_i \cup B'_i.$$

Hence,

$$\begin{aligned}\psi(\mathbf{x}) - \psi\big(\mathbf{x}^i\big) &= \lambda(R_i \cup B_i) - \lambda\big(R_i \cup B_i'\big) \\ &= \lambda\big((R_i \cup B_i) \cap \big(R_i \cup B_i'\big)^c\big) - \lambda\big(\big(R_i \cup B_i'\big) \cap (R_i \cup B_i)^c\big) \\ &= \lambda\big(B_i \cap \big(B_i'\big)^c \cap R_i^c\big) - \lambda\big(B_i' \cap B_i^c \cap R_i^c\big) \\ &= \lambda\big(B_i \cap R_i^c\big) - \lambda\big(B_i' \cap R_i^c\big),\end{aligned}$$

where we obtain the last inequality by adding and subtracting $\lambda(B_i \cap B_i' \cap R_i^c)$. Hence, with $N_i(\mathbf{x})$ be the set of indices $j \neq i$ of the neighbors of x_i in the graph $\mathcal{G}(\mathbf{x})$,

$$\psi(\mathbf{x}) - \psi\big(\mathbf{x}^i\big) = \lambda\Bigg(B_i \cap \Bigg(\bigcup_{j \in N_i(\mathbf{x})} B_j\Bigg)^c\Bigg) - \lambda\Bigg(B_i' \cap \Bigg(\bigcup_{j \in N_i(\mathbf{x}^j)} B_j\Bigg)^c\Bigg). \quad (4.167)$$

The pair $\{i, j\}$ fails to be an edge in the graphs $\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x}^i), \mathcal{G}(\mathbf{x}^j), \mathcal{G}(\mathbf{x}^{ij})$ if and only no member of $\{x_i, x_i'\}$ is a neighbor of $\{x_j, x_j'\}$, in which case $N_i(\mathbf{x}) = N_i(\mathbf{x}^j)$ and $N_i(\mathbf{x}^i) = N_i(\mathbf{x}^{ij})$, and $\psi(\mathbf{x}) = \psi(\mathbf{x}^i)$ and $\psi(\mathbf{x}^i) = \psi(\mathbf{x}^{ij})$. Thus $\mathcal{G}$ is an interaction rule. In addition, (4.167) shows that for all $\mathbf{x} \in \mathcal{X}^n$ and all $i = 1, \ldots, n$

$$\big|\Delta_i \psi(\mathbf{x})\big| = \big|\psi(\mathbf{x}) - \psi\big(\mathbf{x}^i\big)\big| \leq \lambda\big(B(x_i, r)\big) \leq K_V. \quad (4.168)$$

Hence we may take $M = K_V$ in the first term in the bound of Theorem 4.12, and also apply this same estimate to the second term.

Defining the graph $\mathcal{G}'$ on $(x_1, \ldots, x_{n+4}) \in \mathcal{X}^{n+4}$ by placing edges between any two points using the same rule as for $\mathcal{G}$, the rule $\mathcal{G}'$ clearly extends $\mathcal{G}$. As each of the $n+3$ points $x_2, \ldots, x_{n+4}$ is independently a neighbor of x_1 with probability p_r, we have that $\delta - 1 \sim \mathrm{Bin}(n+3, p_r)$. As $E(\delta - 1)^4 = n^4 p^4 + O(n^3)$, we may bound $(E\delta^4)^{1/4}$ by some constant times $1 + np$, completing the argument for $\mathcal{R}$.

The calculation for $\mathcal{S}$ is similar. Let $\psi(\mathbf{x}) = \mathcal{S}(\mathbf{x})$ and take $\mathcal{G}$ to be the same graphical rule as the one used for $\mathcal{R}$. As the removal of a point from $\mathbf{x} \in \mathcal{X}^n$ can cause at most K_S points to become isolated,

$$\big|\Delta_i \psi(\mathbf{x})\big| = \big|\psi(\mathbf{x}) - \psi\big(\mathbf{x}^i\big)\big| \leq K_S.$$

As the graph for $\mathcal{S}$ is the same as for $\mathcal{R}$, the distribution and bounds for the degree δ are the same as for $\mathcal{R}$. □

To test the quality of the bounds, we specialize to Euclidean space, and in the case of V, let λ be the Lebesgue measure. Specializing a bit further, we take the points $X_1, \ldots, X_n$ uniformly and independently in the cube $C_n = [0, n^{1/d})^d$ in $\mathbb{R}^d$, with periodic boundary conditions. Then letting $v_\rho = \rho^d \pi^{d/2} / \Gamma(1 + d/2)$, the volume of the radius ρ ball in dimension d, we have $K_V = v_r$. Now assuming $r \leq n^{1/d}/2$ we have $p = v_{2r}/n$. By Goldstein and Penrose (2010),

$$\lim_{n \to \infty} n^{-1} \sigma_V^2 = g_V \quad (4.169)$$

with an explicit $g_V > 0$, showing the bound of Proposition 4.8 to be of order $n^{-1/2}$.

Similar remarks apply to S. In particular, K_S, as a lower bound on the kissing number, is bounded in any dimension as $n \to \infty$, and (4.169) holds for some

$g_S > 0$ when σ_V^2 is replaced by σ_S^2. At the cost of considerable more effort, Goldstein and Penrose (2010) apply Theorem 5.6 to obtain bounds of order $n^{-1/2}$ for the Kolmogorov distance for both the standardized V and S, with explicit constants.

Though Chatterjee's approach might at first glance seem to bear little connection to the methods already presented, and (4.158) indeed appears a bit mysterious, Chen and Röllin (2010) have an interpretation which fits it into a general framework that contains a number of previous techniques mentioned, the exchangeable pair and size bias methods in particular. Chen and Röllin (2010) consider an identity of the form

$$E\big[Gf(W') - Gf(W)\big] = E\big[Wf(W)\big], \tag{4.170}$$

for some triple (W, W', G) of square integrable random variables. If W', W is a λ-Stein pair then by (2.35) identity (4.170) is satisfied with

$$G = \frac{1}{2\lambda}(W' - W).$$

If Y^s is on the same space as Y and has the Y-size biased distribution, and if $EY = \mu$ and $\mathrm{Var}(Y) = \sigma^2$, then by (2.64) the variables $W = (Y - \mu)/\sigma$ and $W' = (Y^s - \mu)/\sigma$ satisfy (4.170) with $G = \mu/\sigma$.

Chatterjee's approach is also included in the framework of Chen and Röllin (2010), by the method of 'interpolation to independence', as follows. Suppose W is a mean zero, variance 1 random variable, and for each $i \in \{1, \ldots, n\}$ we have a random variable W_i' which is close in some sense to W. Suppose there exists a sequence of random variables $V_0, \ldots, V_n$ such that $V_0 = W$, that V_0 and V_n are independent, and that

$$\big((W, V_{i-1}), (W_i', V_i)\big) =_d \big((W_i', V_i), (W, V_{i-1})\big) \quad \text{for all } i = 1, \ldots, n.$$

Note in particular we must therefore have $W =_d W_i'$ and $V_i =_d V_{i-1}$, so all elements of the sequence $V_0, \ldots, V_n$ are equal in distribution, and have mean $E[V_0] = EW = 0$. Given such variables, letting I be uniform over $\{1, \ldots, n\}$ and independent of the remaining variables and

$$G = \frac{n}{2}(V_I - V_{I-1}),$$

we have, by telescoping the sum, using the independence of V_n and W on the first term and taking conditional expectation with respect to W on the second, that

$$\begin{aligned} E\big[Gf(W)\big] &= \frac{1}{2}\sum_{i=1}^{n}(V_i - V_{i-1})f(W) \\ &= \frac{1}{2}(V_n - V_0)f(W) \\ &= -\frac{1}{2}E\big[Wf(W)\big], \end{aligned}$$

while

$$\begin{aligned} E\big[Gf(W')\big] &= \frac{1}{2}\sum_{i=1}^{n}(V_i - V_{i-1})f(W') \\ &= -\frac{1}{2}\sum_{i=1}^{n}(V_i - V_{i-1})f(W) \\ &= -\frac{1}{2}E\big[Gf(W)\big] \\ &= \frac{1}{2}E\big[Wf(W)\big]. \end{aligned}$$

Hence (4.170) is satisfied with $W' = W'_I$.

Now when $W = \psi(\mathbf{X})$, a mean zero, variance one function of i.i.d. variables $X_1, \ldots, X_n$, one can construct the required sequence $V_0, \ldots, V_n$ by setting V_i to be the function ψ evaluated on $X'_1, \ldots, X'_i, X_{i+1}, \ldots, X_n$ where X'_i is an independent copy of X_i. Let also $W'_i = \psi(\mathbf{X}^i)$, where $\mathbf{X}^i$ is the vector $\mathbf{X}$ with X'_i replacing X_i. It is clear that $V_0 = W$, and is independent of V_n. In the notation of (4.156) we have

$$W'_i = \psi\big(\mathbf{X}^i\big) \quad \text{and} \quad V_i = \psi\big(\mathbf{X}^{\{1,\ldots,i\}}\big).$$

Now consider the variation where π is a random permutation independent of the remaining variables, and we interpolate to independence in the order determined by π, that is,

$$W'_i = \psi\big(\mathbf{X}^{\pi(i)}\big) \quad \text{and} \quad V_i = \psi\big(\mathbf{X}^{\{\pi(1),\ldots,\pi(i)\}}\big).$$

Then (4.170) is satisfied with

$$G = \frac{1}{2n}\big(W'_{\pi(I)} - W'_{\pi(I-1)}\big),$$

where I is an independent index chosen uniformly from $\{1, \ldots, n\}$. Moreover, bounds to the normal in this framework involve conditional expectations such as (4.144), and in particular $E(G(W' - W)|\mathbf{X}, \mathbf{X}')$ is the expression (4.158), see Chen and Röllin (2010) for details.

We now present the proof of Theorems 4.11 and 4.12, starting with some preliminary lemmas.

Lemma 4.8 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with independent χ valued components. Then, for any functions $\phi, \psi : \chi^n \to \mathbb{R}$ such that $E\phi(\mathbf{X})^2$ and $E\psi(\mathbf{X})^2$ are both finite,*

$$\mathrm{Cov}\big(\phi(\mathbf{X}), \psi(\mathbf{X})\big) = \frac{1}{2}\sum_{\substack{A\subset\{1,\ldots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A} E\big[\Delta_j\phi(\mathbf{X})\Delta_j\psi\big(\mathbf{X}^A\big)\big].$$

Proof First, we claim that

$$\begin{aligned}
&\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)} \sum_{j\notin A} \Delta_j \psi\big(\mathbf{X}^A\big) \\
&= \sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)} \sum_{j\notin A} \big(\psi\big(\mathbf{X}^A\big) - \psi\big(\mathbf{X}^{A\cup j}\big)\big) \\
&= \psi(\mathbf{X}) - \psi(\mathbf{X}'). \qquad (4.171)
\end{aligned}$$

In particular, note that for any set $A \subset \{1,\dots,n\}$, except $A = \{1,\dots,n\}$, as there are $n - |A|$ elements $j \notin A$, these set appear in (4.171) with a positive sign a total of

$$\frac{1}{\binom{n}{|A|}(n-|A|)} \times \big(n-|A|\big) = \frac{1}{\binom{n}{|A|}}$$

times. Similarly, any set $B \subset \{1,\dots,n\}$, except $B = \emptyset$, can be represented as $B = A \cup j$ for $|B|$ different sets A, so these sets appear with a negative sign a total of

$$\frac{1}{\binom{n}{|B|-1}(n-|B|+1)} \times |B| = \frac{1}{\binom{n}{|B|}}$$

times. Hence only the terms $A = \emptyset$ and $A \cup j = \{1,\dots,n\}$ do not cancel out, the first one appearing with a coefficient of $1/\binom{n}{0} = 1$, and the latter with coefficient $-1/\binom{n}{n} = -1$.

Now, for a fixed A and $j \notin A$ let $U = \phi(\mathbf{X})\Delta_j\psi(\mathbf{X}^A)$, a function of the random vectors $\mathbf{X}$ and $\mathbf{X}'$. Note that upon interchanging X_j and X'_j the joint distribution of $(\mathbf{X}, \mathbf{X}')$ is unchanged, while U becomes $U' = -\phi(\mathbf{X}^j)\Delta_j\psi(\mathbf{X}^A)$. Thus,

$$EU = EU' = \frac{1}{2}E(U+U') = \frac{1}{2}\big[\Delta_j\phi(\mathbf{X})\Delta_j\psi\big(\mathbf{X}^A\big)\big].$$

Combining these observations yields

$$\begin{aligned}
\operatorname{Cov}\big(\phi(\mathbf{X}),\psi(\mathbf{X})\big) &= E\big[\phi(\mathbf{X})\psi(\mathbf{X})\big] - E\big[\phi(\mathbf{X})\big]E\big[\psi(\mathbf{X})\big] \\
&= E\big[\phi(\mathbf{X})\big(\psi(\mathbf{X}) - \psi(\mathbf{X}')\big)\big] \\
&= \sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)} \sum_{j\notin A} E\big[\phi(\mathbf{X})\Delta_j\psi\big(\mathbf{X}^A\big)\big] \\
&= \frac{1}{2} \sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)} \sum_{j\notin A} E\big[\Delta_j\phi(\mathbf{X})\Delta_j\psi\big(\mathbf{X}^A\big)\big],
\end{aligned}$$

as desired. □

Lemma 4.9 *Let $W = \psi(\mathbf{X})$ with $EW = 0$ and $\operatorname{Var}(W) = 1$ where $\mathbf{X} = (X_1,\dots,X_n)$ is a vector of χ valued, independent components, and let T be given by* (4.158).

Then, for any twice continuously differentiable function f with bounded second derivative, we have

$$\left|E\big(f(W)W\big) - E\big(f'(W)T\big)\right| \le \frac{\|f''\|}{4}\sum_{j=1}^{n} E\left|\Delta_j\psi(\mathbf{X})\right|^3,$$

where T is given by (4.158).

Proof For each $A \subset \{1,\dots,n\}$ and $j \notin A$, let

$$R_{A,j} = \Delta_j(f\circ\psi)(\mathbf{X})\Delta_j\big(\psi\big(\mathbf{X}^A\big)\big)$$

and

$$\widetilde{R}_{A,j} = f'\big(\psi(\mathbf{X})\big)\Delta_j\psi(\mathbf{X})\Delta_j\big(\psi\big(\mathbf{X}^A\big)\big).$$

By Lemma 4.8 with $g = f\circ\psi$, we have

$$E\big[f(W)W\big] = \frac{1}{2}\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A} E R_{A,j}. \tag{4.172}$$

By the mean value theorem, and Hölder's inequality, we have

$$\begin{aligned} E|R_{A,j} - \widetilde{R}_{A,j}| &\le \frac{\|f''\|}{2}E\left|\big(\Delta_j\psi(\mathbf{X})\big)^2\Delta_j\big(\psi\big(\mathbf{X}^A\big)\big)\right| \\ &\le \frac{\|f''\|}{2}E\left|\Delta_j\big(\psi\big(\mathbf{X}^A\big)\big)\right|^3. \end{aligned} \tag{4.173}$$

From the definition of T,

$$f'(W)T = \frac{1}{2}\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A} \widetilde{R}_{A,j}. \tag{4.174}$$

Combining (4.172), (4.174) and (4.173), we obtain

$$\begin{aligned} &E\left|f(W)W - Ef'(W)T\right| \\ &= \left|\frac{1}{2}\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A} E(R_{A,j} - \widetilde{R}_{j,A})\right| \\ &\le \frac{\|f''\|}{4}\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\neq n}} \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A} E\left|\Delta_j\psi(\mathbf{X})\right|^3 \\ &= \frac{\|f''\|}{4}\sum_{j=1}^{n} E\left|\Delta_j\psi(\mathbf{X})\right|^3, \end{aligned}$$

as claimed. □

Proof of Theorem 4.11 Let h be any absolutely continuous function with $\|h'\| \le 1$, and let f be the solution to the Stein equation for h,

$$Eh(W) - Nh = E\big[f'(W) - Wf(W)\big].$$

By (2.13) of Lemma 2.4, we have that $\|f'\| \le \sqrt{2/\pi}$ and $\|f''\| \le 2$. Setting $\phi = \psi$ in Lemma 4.8, we obtain $ET = EW^2 = 1$. Therefore

$$\begin{aligned}
\big|Eh(W) - Nh\big| &\le E\big|f'(W) - Wf(W)\big| \\
&\le E\big|f'(W) - f'(W)T\big| + E\big|f'(W)T - Wf(W)\big| \\
&\le \sqrt{2/\pi}\, E\big|E(T|W) - 1\big| + E\big|f'(W)T - Wf(W)\big| \\
&\le \sqrt{2/\pi}\big[\mathrm{Var}\big(E(T|W)\big)\big]^{1/2} + \frac{1}{2}\sum_{j=1}^{n} E\big|\Delta_j \psi(\mathbf{X})\big|^3,
\end{aligned}$$

by the Cauchy–Schwarz inequality, and Lemma 4.9. The proof is completed by taking supremum over h, noting (4.8). □

We now proceed to the proof of Theorem 4.12. By Theorem 4.11, it suffices to bound $\mathrm{Var}(E(T|\mathbf{X}))$. For this reason, the proof of Theorem 4.12 follows quickly from the following upper bound.

Lemma 4.10 *Let* $\mathbf{X}$ *be a vector of i.i.d. variates,* $A \subset \{1, \dots, n\}$ *with* $|A| \ne n$, *and* T_A, M *and* δ *given by* (4.158), (4.161) *and* (4.162), *respectively. Then there exists a constant* C *such that*

$$\mathrm{Var}\big(E(T_A|\mathbf{X})\big) \le C\big(EM^8\big)^{1/2}\big(E\delta^4\big)^{1/2}\sqrt{n(n-|A|)}.$$

For the remainder of this section, we make the convention that constants C need not be the same at each occurrence. Deferring the proof of Lemma 4.10, we present the proof of Theorem 4.12.

Proof By the definition of T and Minkowski's inequality, we obtain

$$\big[\mathrm{Var}\big(E(T|\mathbf{X})\big)\big]^{1/2} \le \frac{1}{2}\sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\ne n}} \frac{[\mathrm{Var}(E(T_A|\mathbf{X}))]^{1/2}}{\binom{n}{|A|}(n-|A|)}.$$

Substituting the bound from Lemma 4.10 yields

$$\begin{aligned}
\big[\mathrm{Var}\big(E(T|\mathbf{X})\big)\big]^{1/2} &\le C\big(EM^8\big)^{1/4}\big(E\delta^4\big)^{1/2} \sum_{\substack{A\subset\{1,\dots,n\}\\ |A|\ne n}} \frac{n^{1/4}(n-|A|)^{1/4}}{\binom{n}{|A|}(n-|A|)} \\
&= C\big(EM^8\big)^{1/4}\big(E\delta^4\big)^{1/4}\sum_{k=1}^{n} n^{1/4}k^{-3/4} \\
&= C\big(EM^8\big)^{1/4}\big(E\delta^4\big)^{1/4} n^{1/2}.
\end{aligned}$$

Now invoking Theorem 4.11 completes the proof. □

It remains to prove Lemma 4.10. We proceed by way of the following preliminary result.

Lemma 4.11 *Suppose that $\mathcal{G}$ is a symmetric graphical rule on χ^n and $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of i.i.d. χ-valued random variables. Let d_1 be the degree of vertex 1 in $\mathcal{G}(\mathbf{X})$, and, for any $k \leq n-1$, let $i, i_1, \ldots, i_k$ be any collection of $k+1$ distinct elements of $\{1, \ldots, n\}$. Then*

$$P\big(\{i, i_l\} \in \mathcal{G}(\mathbf{X}) \text{ for all } 1 \leq l \leq k\big) = \frac{E(d_1)_k}{(n-1)_k}, \tag{4.175}$$

where $(r)_k$ stands for the falling factorial $r(r-1)\cdots(r-k+1)$.

Proof Since $\mathcal{G}$ is a symmetric rule and $X_1, \ldots, X_n$ are i.i.d., the probability

$$P\big(\{i, i_l\} \in \mathcal{G}(\mathbf{X}) \text{ for all } 1 \leq l \leq k\big)$$

does not depend on $i, i_1, \ldots, i_k$. Hence

$$\begin{aligned} &P\big(\{i, i_l\} \in \mathcal{G}(\mathbf{X}) \text{ for all } 1 \leq l \leq k\big) \\ &\quad = \frac{1}{(n-1)_k} \sum_{\substack{\{j_1, \ldots, j_k\} \subset \{1, \ldots, n\} \setminus \{i\} \\ |\{j_1, \ldots, j_k\}| = k}} P\big(\{i, j_l\} \in \mathcal{G}(\mathbf{X}) \text{ for all } 1 \leq l \leq k\big). \end{aligned}$$

Lastly, note that

$$\sum_{\substack{\{j_1, \ldots, j_k\} \subset \{1, \ldots, n\} \setminus \{i\} \\ |\{j_1, \ldots, j_k\}| = k}} \mathbf{1}\big(\{i, j_l\} \in \mathcal{G}(\mathbf{X}) \text{ for all } 1 \leq l \leq k\big) = (d_i)_k,$$

where d_i is the degree of vertex i. As d_i and d_1 have the same distribution, the argument is complete. □

To prove Lemma 4.10 we require the following result, the Efron–Stein inequality, see Efron and Stein (1981), and Steele (1986).

Lemma 4.12 *Let $U = g(Y_1, \ldots, Y_m)$ be a function of independent random objects $Y_1, \ldots, Y_m$, and let Y_i' be an independent copy of Y_i for $i = 1, \ldots, m$. Then*

$$\operatorname{Var}(U) \leq \frac{1}{2} \sum_{i=1}^{m} E\big(g\big(Y_1, \ldots, Y_{i-1}, Y_i', Y_{i+1}, \ldots, Y_m\big) - g(Y_1, \ldots, Y_m)\big)^2.$$

Proof of Lemma 4.10 Fix $A \subset \{1, \ldots, n\}$ with $|A| \neq n$. For each $j \notin A$, let

$$\begin{aligned} R_j &= \Delta_j \psi(\mathbf{X}) \Delta_j \psi\big(\mathbf{X}^A\big) \\ &= \big(\psi(\mathbf{X}) - \psi\big(\mathbf{X}^j\big)\big)\big(\psi\big(\mathbf{X}^A\big) - \psi\big(\mathbf{X}^{A \cup j}\big)\big). \end{aligned}$$

Let $Y = (Y_1, \ldots, Y_n)$ be a copy of $\mathbf{X}$, which is independent of both $\mathbf{X}$ and $\mathbf{X}'$. For a fixed $i \in \{1, \ldots, n\}$ let

$$\widetilde{\mathbf{X}} = (X_1, \ldots, X_{i-1}, Y_i, X_{i+1}, \ldots, X_n).$$

Similarly, for each $B \subset \{1, \ldots, n\}$, let

$$\widetilde{\mathbf{X}}^B = \begin{cases} (X_1^B, \ldots, X_{i-1}^B, Y_i, X_{i+1}^B, \ldots, X_n^B) & \text{if } i \notin B, \\ \mathbf{X}^B & \text{if } i \in B. \end{cases}$$

Now let

$$R_{ji} = \big(\psi(\widetilde{\mathbf{X}}) - \psi\big(\widetilde{\mathbf{X}}^j\big)\big)\big(\psi\big(\widetilde{\mathbf{X}}^A\big) - \psi\big(\widetilde{\mathbf{X}}^{A\cup j}\big)\big),$$

and put

$$h_i = E\bigg(\sum_{j\notin A}(R_j - R_{ji})\bigg)^2.$$

It follows from inequality (4.143) and Lemma 4.12 that

$$\mathrm{Var}\big(E(T_A|\mathbf{X})\big) \le \mathrm{Var}(T_A) \le \frac{1}{2}\sum_{i=1}^n h_i. \tag{4.176}$$

Hence, we turn our attention to bounding h_i, and note that we need only consider $j \notin A$. When $j \neq i$ let

$$\begin{aligned} d_{ji}^1 &= \mathbf{1}\big(\{i,j\} \in \mathcal{G}(\mathbf{X})\big), \\ d_{ji}^2 &= \mathbf{1}\big(\{i,j\} \in \mathcal{G}\big(\mathbf{X}^j\big)\big), \\ d_{ji}^3 &= \mathbf{1}\big(\{i,j\} \in \mathcal{G}(\widetilde{\mathbf{X}})\big) \quad \text{and} \\ d_{ji}^4 &= \mathbf{1}\big(\{i,j\} \in \mathcal{G}\big(\widetilde{\mathbf{X}}^j\big)\big). \end{aligned}$$

Suppose in a particular realization we have $d_{ji}^1 = d_{ji}^2 = d_{ji}^3 = d_{ji}^4 = 0$. Since $\mathcal{G}$ is an interaction rule for ψ, on this event we have

$$\psi(\mathbf{X}) - \psi\big(\mathbf{X}^j\big) = \psi(\widetilde{\mathbf{X}}) - \psi\big(\widetilde{\mathbf{X}}^j\big).$$

If we now take $\mathbf{X}^A$ and $\widetilde{\mathbf{X}}^A$ in place of $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, and define $e_{ji}^1, e_{ji}^2, e_{ji}^3$ and e_{ji}^4 analogously, then when $e_{ji}^1 = e_{ji}^2 = e_{ji}^3 = e_{ji}^4 = 0$ we have

$$\psi\big(\mathbf{X}^A\big) - \psi\big(\mathbf{X}^{A\cup j}\big) = \psi\big(\widetilde{\mathbf{X}}^A\big) - \psi\big(\widetilde{\mathbf{X}}^{A\cup j}\big),$$

whether $i \in A$ or not. Now, let

$$L_i = \max_{j\notin A}\big|\Delta_j\psi(\mathbf{X})\Delta_j\psi\big(\mathbf{X}^A\big) - \Delta_j\psi(\widetilde{\mathbf{X}})\Delta_j\psi\big(\widetilde{\mathbf{X}}^A\big)\big|.$$

From the preceding considerations, when $j \neq i$

$$|R_j - R_{ji}| \le L_i\sum_{k=1}^4\big(d_{ji}^k + e_{ji}^k\big).$$

When $j = i$ then $i \notin A$ and we have $|R_j - R_{ji}| \le L_i$. The Cauchy–Schwarz inequality now yields

$$h_i \le \bigg[EL_i^4E\bigg(\mathbf{1}(i\notin A) + \sum_{j\notin A\cup i}\sum_{k=1}^4\big(d_{ji}^k + e_{ji}^k\big)\bigg)^4\bigg]^{1/2}. \tag{4.177}$$

Applying the inequality $(\sum_{i=1}^r a_i)^4 \le r^3 \sum_{i=1}^r a_i^4$, we obtain

$$E\left(\mathbf{1}(i \notin A) + \sum_{j \notin A \cup i} \sum_{k=1}^{4} \left(d_{ji}^k + e_{ji}^k\right)\right)^4$$
$$\le 9^3 \mathbf{1}(i \in A) + 9^3 \sum_{k=1}^{4} E\left(\sum_{j \notin A \cup i} d_{ji}^k\right)^4 + 9^3 \sum_{k=1}^{4} E\left(\sum_{j \notin A \cup i} e_{ji}^k\right)^4.$$

To handle the first term in the first sum, from Lemma 4.11, for any j, k, l and m,

$$E\left(d_{ji}^1 d_{ki}^1 d_{li}^1 d_{mi}^1\right) \le C \frac{E\delta_1^r}{n^r},$$

where r is the number of distinct indices among j, k, l, m, and δ_1 is the degree of vertex 1 in $\mathcal{G}(\mathbf{X})$. Recall the definition of δ from (4.162), and observe that $\delta \ge \delta_1 + 1$. It follows easily that

$$E\left(\sum_{j \notin A \cup i} d_{ji}^1\right)^4 \le C E\left(\delta^4\right)\left(\frac{n - |A|}{n}\right).$$

Now we consider bounding $E(d_{ji}^2 d_{ki}^2 d_{li}^2 d_{mi}^2)$. First suppose that j, k, l, m are distinct. Now let $\widetilde{\mathbf{X}}$ be the random vector in χ^{n+4} given by

$$\widetilde{\mathbf{X}} = \left(X_1, \ldots, X_n, X_j', X_k', X_l', X_m'\right).$$

Note that if $d_{ji}^2 = d_{ki}^2 = d_{li}^2 = d_{mi}^2 = 1$ then $\{i, n+1\}$, $\{i, n+2\}$, $\{i, n+3\}$ and $\{i, n+4\}$ are all edges in the extended graph $\mathcal{G}'(\widetilde{\mathbf{X}})$. Since $\mathcal{G}'$ is a symmetric rule and the components of $\widetilde{\mathbf{X}}$ are i.i.d., it follows from Lemma 4.10 that

$$E\left(d_{ji}^2 d_{ki}^2 d_{li}^2 d_{mi}^2\right) \le C \frac{E\delta^4}{n^4}.$$

Now, suppose j, k, l are distinct, and that $m = l$. Let $s \in \{1, \ldots, n\}$ be distinct from j, k and l, and define

$$\widetilde{\mathbf{X}} = \left(X_1, \ldots, X_n, X_j', X_k', X_l', X_s'\right)$$

and argue as before to conclude that in this case

$$E\left(d_{ji}^2 d_{ki}^2 d_{li}^2 d_{mi}^2\right) = E\left(d_{ji}^2 d_{ki}^2 d_{li}^2\right) \le C \frac{E\delta^3}{n^3}.$$

In general, if r is the number of distinct elements among j, k, l, m, then

$$E\left(d_{ji}^2 d_{ki}^2 d_{li}^2 d_{mi}^2\right) \le C \frac{E\delta^r}{n^r}.$$

From this inequality we obtain as before that

$$E\left(\sum_{j \notin A \cup i} d_{ji}^2\right)^4 \le C E\left(\delta^4\right)\left(\frac{n - |A|}{n}\right).$$

The d^3, e^1 and e^3 terms can be bounded as the d^1 term, while the d^4, e^2 and e^4 terms like the d^2 term. Combining, we conclude

$$E\left(\mathbf{1}(i \notin A) + \sum_{j \notin A \cup i} \sum_{k=1}^{4} \left(d_{ji}^k + e_{ji}^k\right)\right)^4 \leq CE\left(\delta^4\right)\left(\mathbf{1}(i \notin A) + \frac{n - |A|}{n}\right).$$

As $M = \max_j |\Delta_j \psi(\mathbf{X})|$ have $EL_i^4 \leq CEM^8$, and applying these bounds in (4.177), along with the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for nonnegative x and y, we obtain

$$h_i \leq C\left(EM^8\right)^{1/2}\left(E\delta^4\right)^{1/2}\left(\mathbf{1}(i \notin A) + \sqrt{\frac{n - |A|}{n}}\right).$$

Substituting this bound in (4.176), we obtain

$$\begin{aligned}\operatorname{Var}\left(E(T_A|\mathbf{X})\right) &\leq C\left(EM^8\right)^{1/2}\left(E\delta^4\right)^{1/2}\left(n - |A| + \sqrt{n\left(n - |A|\right)}\right) \\ &\leq C\left(EM^8\right)^{1/2}\left(E\delta^4\right)^{1/2}\sqrt{n(n - |A|)},\end{aligned}$$

completing the proof. □

4.7 Locally Dependent Random Variables

In this section we consider L^1 bounds for sums of locally dependent random variables. We being by recalling that an m-dependent sequence of random variables ξ_i, $i \in \mathbb{N}$, is one with the property that, for each i, the sets of random variables $\{\xi_j,\ j \leq i\}$ and $\{\xi_j,\ j > i + m\}$ are independent. Independent random variables are the special case of m-dependence when $m = 0$. Local dependence generalizes the notion of m-dependence to collections of random variables indexed more generally. The concept of local dependence is applicable, for example, to random variables indexed by the vertices of a graph such that the collections $\{\xi_i,\ i \in I\}$ and $\{\xi_j,\ j \in J\}$ are independent whenever $I \cap J = \emptyset$ and the graph contains no edges $\{i, j\}$ with $i \in I$ and $j \in J$.

Let $\mathcal{J}$ be a finite index set of cardinality n, and let $\{\xi_i, i \in \mathcal{J}\}$ be a random field, that is, an indexed collection of random variables, with zero means and finite variances. Define $W = \sum_{i \in \mathcal{J}} \xi_i$, and assume that $\operatorname{Var}(W) = 1$. For any $A \subset \mathcal{J}$ let

$$A^c = \{j \in \mathcal{J}\colon\ j \notin A\} \quad \text{and} \quad \xi_A = \{\xi_i\colon\ i \in A\}.$$

We introduce the following two conditions, corresponding to different degrees of local dependence.

(LD1) For each $i \in \mathcal{J}$ there exists $A_i \subset \mathcal{J}$ such that ξ_i and $\xi_{A_i^c}$ are independent.
(LD2) For each $i \in \mathcal{J}$ there exist $A_i \subset B_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{A_i^c}$ and ξ_{A_i} is independent of $\xi_{B_i^c}$.

Clearly (LD2) implies (LD1). Whenever (LD1) or (LD2) hold we set

$$\eta_i = \sum_{j\in A_i} \xi_j \quad \text{and} \quad \tau_i = \sum_{j\in B_i} \xi_j \tag{4.178}$$

respectively. Note that when $\{\xi_i,\ i \in \mathcal{J}\}$ are independent (LD2) holds with $A_i = B_i = \{i\}$, in which case $\eta_i = \tau_i = \xi_i$.

Theorem 4.13 *Let $\{\xi_i,\ i \in \mathcal{J}\}$ be a random field with mean zero and* $\mathrm{Var}(W) = 1$ *where $W = \sum_{i\in\mathcal{J}} \xi_i$. If* (LD1) *holds then, then with η_i as in* (4.178),

$$\big\|\mathcal{L}(W) - \mathcal{L}(Z)\big\|_1 \le \sqrt{\frac{2}{\pi}}\, E\left|\sum_{i\in\mathcal{J}} \big\{\xi_i\eta_i - E(\xi_i\eta_i)\big\}\right| + \sum_{i\in\mathcal{J}} E\big|\xi_i\eta_i^2\big|, \tag{4.179}$$

and if (LD2) *holds, then with η_i and τ_i as in* (4.178),

$$\big\|\mathcal{L}(W) - \mathcal{L}(Z)\big\|_1 \le 2\sum_{i\in\mathcal{J}} \big(E|\xi_i\eta_i\tau_i| + \big|E(\xi_i\eta_i)\big|E|\tau_i|\big) + \sum_{i\in\mathcal{J}} E\big|\xi_i\eta_i^2\big|. \tag{4.180}$$

We remark that for independent random variables, applying Hölder's inequality to the bound in (4.180) yields $5\sum_{i\in\mathcal{J}} E|\xi_i|^3$, somewhat larger than the constant of 1 given by Corollary 4.2.

Proof Assume (LD1) holds and let $f = f_h$ be the solution of the Stein equation (2.4) for an absolutely continuous function h satisfying $\|h'\| \le 1$. By the independence of ξ_i and $W - \eta_i$, and that $E\xi_i = 0$, we have

$$E\big\{Wf(W)\big\} = \sum_{i\in\mathcal{J}} E\xi_i f(W) = \sum_{i\in\mathcal{J}} E\xi_i\big[f(W) - f(W-\eta_i)\big].$$

Now adding and subtracting yields

$$\begin{aligned} E\big\{Wf(W)\big\} &= \sum_{i\in\mathcal{J}} E\big\{\xi_i\big[f(W) - f(W-\eta_i) - \eta_i f'(W)\big]\big\} \\ &\quad + E\left\{\left(\sum_{i\in\mathcal{J}} \xi_i\eta_i\right) f'(W)\right\}. \end{aligned} \tag{4.181}$$

Now, using again that $E\xi_i = 0$ for all i, from (LD1) it follows that

$$1 = EW^2 = \sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}} E\{\xi_i\xi_j\} = \sum_{i\in\mathcal{J}} E\{\xi_i\eta_i\},$$

and so

$$\begin{aligned} E\big\{f'(W) - Wf(W)\big\} &= -E\left(\sum_{i\in\mathcal{J}} \big\{\xi_i\eta_i - E(\xi_i\eta_i)\big\} f'(W)\right) \\ &\quad - \sum_{i\in\mathcal{J}} E\big\{\xi_i\big[f(W) - f(W-\eta_i) - \eta_i f'(W)\big]\big\}. \end{aligned} \tag{4.182}$$

By (2.13), $\|f'\| \le \sqrt{2/\pi}$ and $\|f''\| \le 2$. Therefore it follows from (4.182) and a Taylor expansion that

$$\big|Eh(W) - Eh(Z)\big| \le \sqrt{\frac{2}{\pi}} E\left|\sum_{i\in\mathcal{J}} \big\{\xi_i\eta_i - E(\xi_i\eta_i)\big\}\right| + \sum_{i\in\mathcal{J}} E\big|\xi_i\eta_i^2\big|.$$

Now (4.179) follows from (4.8).

When (LD2) is satisfied, $f'(W-\tau_i)$ and $\xi_i\eta_i$ are independent for each $i \in \mathcal{J}$. Hence, using (4.182), we can write

$$\begin{aligned}
&\big|Eh(W) - Eh(Z)\big| \\
&\quad\le \left|E\sum_{i\in\mathcal{J}} \big\{\xi_i\eta_i - E(\xi_i\eta_i)\big\}\big(f'(W) - f'(W-\tau_i)\big)\right| + \sum_{i\in\mathcal{J}} E\big|\xi_i\eta_i^2\big| \\
&\quad\le 2\sum_{i\in\mathcal{J}} \big(E|\xi_i\eta_i\tau_i| + \big|E(\xi_i\eta_i)\big|E|\tau_i|\big) + \sum_{i\in\mathcal{J}} E\big|\xi_i\eta_i^2\big|,
\end{aligned}$$

as desired. □

We provide two examples of locally dependent random variables. We refer to Baldi and Rinott (1989), Rinott (1994), Baldi et al. (1989), Dembo and Rinott (1996), and Chen and Shao (2004) for more details.

Example 4.1 (Graphical dependence) Consider a set of random variables $\{\xi_i, i \in \mathcal{V}\}$ indexed by the vertices of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The graph $\mathcal{G}$ is said to be a dependency graph if, for any pair of disjoint sets Γ_1 and Γ_2 in $\mathcal{V}$ such that no edge in $\mathcal{E}$ has one endpoint in Γ_1 and the other in Γ_2, the sets of random variables $\{\xi_i, i \in \Gamma_1\}$ and $\{\xi_i, i \in \Gamma_2\}$ are independent. Let

$$A_i = \{i\} \cup \big\{j \in \mathcal{V}\colon \{i, j\} \in \mathcal{E}\big\}$$

and $B_i = \bigcup_{j\in A_i} A_j$. Then $\{\xi_i, i \in \mathcal{V}\}$ satisfies (LD2). Hence (4.180) holds.

Example 4.2 (The number of local maxima on a graph) Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (which is not necessary a dependency graph) and independent and identically distributed continuous random variables $\{Y_i, i \in \mathcal{V}\}$. For $i \in \mathcal{V}$ define the indicator variable

$$\xi_i = \begin{cases} 1 & \text{if } Y_i > Y_j \text{ for all } j \in \mathcal{N}_i, \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{N}_i = \{j \in \mathcal{V}\colon \{i, j\} \in \mathcal{E}\}$. Hence $\xi_i = 1$ indicates that Y_i is a local maximum and $W = \sum_{i\in\mathcal{V}} \xi_i$ is the total number of local maxima. Letting

$$A_i = \{i\} \cup \mathcal{N}_i \cup \bigcup_{j\in\mathcal{N}_i} \mathcal{N}_j \quad \text{and} \quad B_i = \bigcup_{j\in A_i} A_j$$

we find that $\{\xi_i, i \in \mathcal{V}\}$ satisfies (LD2), and therefore (4.180) holds. Bounds in L^∞ for this problem are considered in Example 6.4.

4.8 Smooth Function Bounds

In defining a distance $\|\mathcal{L}(X) - \mathcal{L}(Y)\|_{\mathcal{H}}$ through (4.1) one typically chooses $\mathcal{H}$ to be a convergence determining class of functions, that is, a collection of functions such that if $\{X_n\}_{n\geq 0}$ is any sequence of random variables then

$$Eh(X_n) \to Eh(X_0) \quad \text{for all } h \in \mathcal{H} \text{ implies } X_n \to_d X_0.$$

A convergence determining class can consist of functions all of which are very smooth, such as the collection of all infinity differentiable functions with compact support.

To describe the collection of functions we consider in this section, following E.M. Stein (1970), let $L_m^\infty(\mathbb{R})$ be all functions $h : \mathbb{R} \to \mathbb{R}$ satisfying $\|h\|_{L_m^\infty(\mathbb{R})} < \infty$ where

$$\|h\|_{L_m^\infty(\mathbb{R})} = \max_{0\leq k\leq m} \left\|h^{(k)}\right\|.$$

That is $L_m^\infty(\mathbb{R})$ consists of all functions possessing m bounded derivatives. Now let $\|\mathcal{L}(W) - \mathcal{L}(Z)\|_{\mathcal{H}_{m,\infty}}$ be the distance which is obtained through (4.1) by setting

$$\mathcal{H}_{m,\infty} = \left\{h \in L_m^\infty(\mathbb{R}):\ \|h\|_{L_m^\infty(\mathbb{R})} \leq 1\right\}. \tag{4.183}$$

In the following section we show how fast rates of convergence can be obtained under a vanishing third moment assumption when inducing our distance by $\mathcal{H}_{4,\infty}$. In Chap. 12 we prove a smooth function theorem in $\mathbb{R}^p$ using a multidimensional generalization of the distances defined here, and produce bounds in that distance for the problem of counting the number of vertices in a random graph that have specified degree counts.

4.8.1 Fast Rates for Smooth Functions

In this section we first prove Theorem 4.14, a smooth function theorem parallel to Theorem 4.9, for the zero bias coupling as discussed in Sect. 2.3.3. Comparing Theorems 4.9 and 4.14, we see that the latter requires the computation of a conditional expectation of a difference, rather than of a difference squared, and that the second, or remainder term is of a square, rather than a cube. Lastly, Theorem 4.9 requires the linearity condition (4.108) to be satisfied, whereas Theorem 4.14 does not. After the proof we apply Theorem 4.14 in an independent case to show that fast rates of convergence for smooth functions are obtained when fourth moments exist and third moment vanishes.

Theorem 4.14 *Let W be a mean zero, variance 1 random variable and suppose that the pair (W, W^*) is given on a joint probability space so that W^* has the W-zero biased distribution. Then*

$$\left\|\mathcal{L}(W) - \mathcal{L}(Z)\right\|_{\mathcal{H}_{4,\infty}} \leq \frac{1}{3}E\left|E\left(W^* - W|W\right)\right| + \frac{1}{8}E(W^* - W)^2.$$

Proof Let g be the solution to (2.19) for a given $h \in \mathcal{H}_{4,\infty}$. By the bounds in Lemma 2.6

$$\|g^{(3)}\| \le \frac{1}{3} \quad \text{and} \quad \|g^{(4)}\| \le \frac{1}{4}. \tag{4.184}$$

By (2.19), (2.51), and Taylor expansion,

$$\begin{aligned} \big|Eh(W) - Nh\big| &= \big|E\big(g''(W) - Wg'(W)\big)\big| \\ &= \big|E\big(g''(W^*) - g''(W)\big)\big| \\ &\le \big|Eg^{(3)}(W)(W^* - W)\big| + \left|E\int_W^{W^*} g^{(4)}(t)(W^* - t)dt\right|. \end{aligned}$$

Conditioning on W we may bound the first term as

$$\big|E\big[g^{(3)}(W)E\big(W^* - W|W\big)\big]\big| \le \|g^{(3)}\|\big|E|E\big(W^* - W|W\big)\big|.$$

For the second term

$$\left|E\int_W^{W^*} g^{(4)}(t)(W^* - t)dt\right| \le \frac{1}{2}\|g^{(4)}\|E(W^* - W)^2.$$

Applying (4.184) completes the proof. □

We now apply Theorem 4.14 to the sum of independent identically distributed variables and show how the zero bias transformation leads to an error bound for smooth functions of order n^{-1}, under additional moment assumptions which include a vanishing third moment.

Corollary 4.4 *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed mean zero, variance one random variables with vanishing third moment and $EX^4 < \infty$. Then, for $W = n^{-1/2}\sum_{i=1}^n X_i$,*

$$\big\|\mathcal{L}(W) - \mathcal{L}(Z)\big\|_{\mathcal{H}_{4,\infty}} \le \frac{1}{24n}\big(11 + EX^4\big).$$

Proof For $i = 1, \ldots, n$ let X_i^* have the X_i-zero biased distribution and be independent of X_j, $j = 1, \ldots, n$, and I a random index independent of X_i, X_i^*, $i = 1, \ldots, n$ with distribution

$$P(I = i) = 1/n.$$

Then, by Lemma 2.8 and the scaling property (2.59),

$$W^* = W - X_I/\sqrt{n} + X_I^*/\sqrt{n}$$

has the W-zero biased distribution.

From substituting $f(x) = x^2/2$ into (2.51), for every $i = 1, \ldots, n$ we have

$$EX_i^* = (1/2)EX_i^3 = 0 \quad \text{and therefore} \quad EX_I^* = 0. \tag{4.185}$$

Next, using that $X_1, \ldots, X_n$'s are i.i.d., and therefore exchangeable, $E(X_I|W) = W/\sqrt{n}$.

Now, by the independence of X_I^* and W, and (4.185), we obtain

$$\begin{aligned} E(W^* - W|W) &= n^{-1/2} E\big(X_I^* - X_I|W\big) \\ &= n^{-1/2}\big(E\big(X_I^*\big) - E(X_I|W)\big) \\ &= -n^{-1/2} E(X_I|W) \\ &= -n^{-1} W. \end{aligned}$$

Therefore

$$E\big|E(W^* - W|W)\big| = n^{-1} E|W| \le \frac{1}{n}.$$

For the second term in Theorem 4.14, application of (2.51) with $f(x) = x^3/3$ yields

$$E\big(X_I^*\big)^2 = E\big(X_i^*\big)^2 = \frac{1}{3} E X_i^4.$$

Since X_I and X_I^* are independent, and the latter variable has mean zero,

$$E(W^* - W)^2 = \frac{1}{n} E\big(X_I^* - X_I\big)^2 = \frac{1}{n}\big(E\big(X_I^*\big)^2 + E X_I^2\big) = \frac{1}{n}\left(\frac{E X^4}{3} + 1\right).$$

Applying Theorem 4.14 now yields the claim. □

Under more special assumptions a fast rates may be obtained for distances induced by classes of non-smooth functions. In particular, Klartag (2009) demonstrates a bound of order $1/n$ for cases which include the sum of independent symmetric random variables whose density is log concave.

Appendix

Proof of Lemma 4.7 Let π be uniform on $\mathcal{S}_n$ and $I^\dagger, J^\dagger, K^\dagger, L^\dagger$ be independent of π with distribution (4.134). Constructing Y from π and $Y^\dagger$ and $Y^\ddagger$ from $\pi^\dagger$ and $\pi^\ddagger$ respectively, as in Lemma 4.6, we have

$$\begin{aligned} Y^* - Y &= U Y^\dagger + (1-U) Y^\ddagger - Y \\ &= U \sum_{i=1}^n a_{i,\pi^\dagger(i)} + (1-U) \sum_{i=1}^n a_{i,\pi^\ddagger(i)} - \sum_{i=1}^n a_{i,\pi(i)}. \end{aligned}$$

With

$$\mathcal{I} = \big\{I^\dagger, J^\dagger, \pi^{-1}\big(K^\dagger\big), \pi^{-1}\big(L^\dagger\big)\big\}, \tag{4.186}$$

we see from (4.126) in Lemma 4.5, and from $\pi^\ddagger = \pi^\dagger \tau_{I^\dagger, J^\dagger}$, that if $m \notin \mathcal{I}$, then $\pi(m) = \pi^\dagger(m) = \pi^\ddagger(m)$. Hence, setting $V = Y^* - Y$, we have

$$V = \sum_{i \in \mathcal{I}} \big(U a_{i,\pi^\dagger(i)} + (1-U) a_{i,\pi^\ddagger(i)} - a_{i,\pi(i)}\big). \tag{4.187}$$

Further, letting

$$R = \left|\left\{\pi(I^\dagger), \pi(J^\dagger)\right\} \cap \{K^\dagger, L^\dagger\}\right|$$

and $\mathbf{1}_k = \mathbf{1}(R = k)$, since $P(R \leq 2) = 1$, we have

$$\begin{aligned} V &= V\mathbf{1}_2 + V\mathbf{1}_1 + V\mathbf{1}_0, \\ &\text{and therefore} \quad E|V| \leq E|V|\mathbf{1}_2 + E|V|\mathbf{1}_1 + E|V|\mathbf{1}_0. \end{aligned} \tag{4.188}$$

The three terms on the right hand side of (4.188) give rise to the three components of the bound in the theorem.

For notational simplicity, the following summations in this section are performed over all indices which appear, whether in the summands or in a (possibly empty) collection of restrictions. In what follows, we will apply equalities and bounds such as

$$\begin{aligned} \sum |a_{il}|\big[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})\big]^2 &= \sum |a_{il}|\big(a_{ik}^2 + a_{jl}^2 + a_{il}^2 + a_{jk}^2\big) \\ &\leq 4n^2\gamma. \end{aligned} \tag{4.189}$$

Due to the form of the terms being squared on the left-hand side, if the factors in a cross term agree in their first index, they will have differing second indices, and likewise if their second indices agree. This gives cross terms which are zero by virtue of (4.120), since there will be at least one unpaired index outside the absolute value over which to sum, for instance, the index k in the term $\sum |a_{il}|a_{ik}a_{il}$. Hence the equality. To obtain the inequality, on each of the four terms are argue as for the first,

$$\sum_{i,j,k,l} |a_{i,l}|a_{i,k}^2 \leq \left(\sum_{j,k} |a_{il}|^3\right)^{1/3} \left(\sum_{j,l} |a_{ik}|^3\right)^{2/3} = n^2\gamma. \tag{4.190}$$

Generally, the power of n in such an inequality, in this case 2, will be 2 less than the number of indices of summation, in this case 4.

Calculation on $R = 2$ On $\mathbf{1}_2$ we have $\{\pi(I^\dagger), \pi(J^\dagger)\} = \{K^\dagger, L^\dagger\}$ and therefore $\mathcal{I} = \{I^\dagger, J^\dagger, \pi^{-1}(K^\dagger), \pi^{-1}(L^\dagger)\} = \{I^\dagger, J^\dagger\}$. As the intersection which gives $R = 2$ can occur in two different ways, we make the further decomposition

$$V\mathbf{1}_2 = V\mathbf{1}_{2,1} + V\mathbf{1}_{2,2},$$

where

$$\begin{aligned} \mathbf{1}_{2,1} &= \mathbf{1}\big(\pi(I^\dagger) = K^\dagger,\ \pi(J^\dagger) = L^\dagger\big) \\ \text{and} \quad \mathbf{1}_{2,2} &= \mathbf{1}\big(\pi(I^\dagger) = L^\dagger,\ \pi(J^\dagger) = K^\dagger\big). \end{aligned}$$

Since $\pi^\dagger = \pi$ on $\mathbf{1}_{2,1}$ by (4.125), following (4.187) we have

$$
\begin{aligned}
V\mathbf{1}_{2,1} &= \sum_{i\in\{I^\dagger,J^\dagger\}} \big(Ua_{i,\pi^\dagger(i)} + (1-U)a_{i,\pi^\ddagger(i)} - a_{i,\pi(i)}\big)\mathbf{1}_{2,1} \\
&= \big[U(a_{I^\dagger,\pi^\dagger(I^\dagger)} + a_{J^\dagger,\pi^\dagger(J^\dagger)}) + (1-U)(a_{I^\dagger,\pi^\ddagger(I^\dagger)} + a_{J^\dagger,\pi^\ddagger(J^\dagger)}) \\
&\quad - (a_{I^\dagger,\pi(I^\dagger)} + a_{J^\dagger,\pi(J^\dagger)})\big]\mathbf{1}_{2,1} \\
&= \big[U(a_{I^\dagger,\pi(I^\dagger)} + a_{J^\dagger,\pi(J^\dagger)}) + (1-U)(a_{I^\dagger,\pi(J^\dagger)} + a_{J^\dagger,\pi(I^\dagger)}) \\
&\quad - (a_{I^\dagger,\pi(I^\dagger)} + a_{J^\dagger,\pi(J^\dagger)})\big]\mathbf{1}_{2,1} \\
&= (1-U)(a_{I^\dagger,\pi(J^\dagger)} + a_{J^\dagger,\pi(I^\dagger)} - a_{I^\dagger,\pi(I^\dagger)} - a_{J^\dagger,\pi(J^\dagger)})\mathbf{1}_{2,1} \\
&= (1-U)(a_{I^\dagger,L^\dagger} + a_{J^\dagger,K^\dagger} - a_{I^\dagger,K^\dagger} - a_{J^\dagger,L^\dagger})\mathbf{1}_{2,1}. \qquad (4.191)
\end{aligned}
$$

Due to the presence of the indicator $\mathbf{1}_{2,1}$, taking the expectation of (4.191) requires a joint distribution which includes the values taken on by π at $I^\dagger$ and $J^\dagger$, say s and t, respectively. Since these images can be any two distinct values, and are independent of $I^\dagger$, $J^\dagger$, $K^\dagger$ and $L^\dagger$, we have, with p_1 and p_2 given in (4.122) and (4.134), respectively,

$$
\begin{aligned}
p_3(i,j,k,l,s,t) &= P\big((I^\dagger, J^\dagger, K^\dagger, L^\dagger, \pi(I^\dagger), \pi(J^\dagger)\big) = (i,j,k,l,s,t)\big) \\
&= p_2(i,j,k,l)\,p_1(s,t) \\
&= \frac{[(a_{ik}+a_{jl})-(a_{il}+a_{jk})]^2}{4n^3(n-1)^2\sigma^2}\mathbf{1}(s\neq t). \qquad (4.192)
\end{aligned}
$$

Now bounding the absolute value of the first term in (4.191) using (4.189), we obtain

$$
\begin{aligned}
E\big|(1-U)a_{I^\dagger,L^\dagger}\big|\mathbf{1}_{2,1} &= \frac{1}{2}\sum |a_{il}|\mathbf{1}(s=k,t=l)\,p_3(i,j,k,l,s,t) \\
&= \frac{1}{2}\sum |a_{il}|\,p_3(i,j,k,l,k,l) \\
&= \frac{1}{8n^3(n-1)^2\sigma^2}\sum |a_{il}|\big[(a_{ik}+a_{jl})-(a_{il}+a_{jk})\big]^2 \\
&\le \frac{\gamma}{2n(n-1)^2\sigma^2}.
\end{aligned}
$$

Using the triangle inequality in (4.191) and applying the same reasoning to the remaining three terms shows that $E|V|\mathbf{1}_{2,1} \le 2\gamma/(n(n-1)^2\sigma^2)$. Since by symmetry the term $V\mathbf{1}_{2,2}$ can be handled the same way, we obtain

$$
E|V|\mathbf{1}_2 \le \frac{4\gamma}{n(n-1)^2\sigma^2} \le \frac{4\gamma}{(n-1)^3\sigma^2}. \qquad (4.193)
$$

Calculation on $R=1$ As the event $R=1$ can occur in four different ways, depending on which element of $\{\pi(I^\dagger),\pi(J^\dagger)\}$ equals an element of $\{K^\dagger, L^\dagger\}$, we decompose $\mathbf{1}_1$ to yield

$$
V\mathbf{1}_1 = V\mathbf{1}_{1,1} + V\mathbf{1}_{1,2} + V\mathbf{1}_{1,3} + V\mathbf{1}_{1,4}, \qquad (4.194)
$$

where $\mathbf{1}_{1,1} = \mathbf{1}(\pi(I^\dagger) = K^\dagger \text{ and } \pi(J^\dagger) \neq L^\dagger)$, specifying the remaining three indicators in (4.194) similarly.

On $\mathbf{1}_{1,1}$ we have, from (4.186), that $\mathcal{I} = \{I^\dagger, J^\dagger, \pi^{-1}(L^\dagger)\}$, and from (4.125) that $\pi^\dagger = \pi\tau_{\pi^{-1}(L^\dagger),J^\dagger}$ and so $\pi^\ddagger = \pi\tau_{\pi^{-1}(L^\dagger),J^\dagger}\tau_{J^\dagger,I^\dagger}$, yielding $\pi^\ddagger(\pi^{-1}(L)) = \pi^\dagger(\pi^{-1}(L)) = \pi(J)$. Now, using (4.187),

$$\begin{aligned}
V\mathbf{1}_{1,1} &= \sum_{i\in\{I^\dagger,J^\dagger,\pi^{-1}(L^\dagger)\}} \big(Ua_{i,\pi^\dagger(i)} + (1-U)a_{i,\pi^\ddagger(i)} - a_{i,\pi(i)}\big)\mathbf{1}_{1,1} \\
&= \big[U(a_{I^\dagger,\pi^\dagger(I^\dagger)} + a_{J^\dagger,\pi^\dagger(J^\dagger)} + a_{\pi^{-1}(L^\dagger),\pi^\dagger(\pi^{-1}(L^\dagger))}) \\
&\quad + (1-U)(a_{I^\dagger,\pi^\ddagger(I^\dagger)} + a_{J^\dagger,\pi^\ddagger(J^\dagger)} + a_{\pi^{-1}(L^\dagger),\pi^\ddagger(\pi^{-1}(L^\dagger))}) \\
&\quad - (a_{I^\dagger,\pi(I^\dagger)} + a_{J^\dagger,\pi(J^\dagger)} + a_{\pi^{-1}(L^\dagger),\pi(\pi^{-1}(L^\dagger))})\big]\mathbf{1}_{1,1} \\
&= \big[U(a_{I^\dagger,K^\dagger} + a_{J^\dagger,L^\dagger} + a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)}) \\
&\quad + (1-U)(a_{I^\dagger,L^\dagger} + a_{J^\dagger,K^\dagger} + a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)}) \\
&\quad - (a_{I^\dagger,K^\dagger} + a_{J^\dagger,\pi(J^\dagger)} + a_{\pi^{-1}(L^\dagger),L^\dagger})\big]\mathbf{1}_{1,1} \\
&= \big[Ua_{J^\dagger,L^\dagger} + (1-U)(a_{I^\dagger,L^\dagger} + a_{J^\dagger,K^\dagger} - a_{I^\dagger,K^\dagger}) \\
&\quad - a_{J^\dagger,\pi(J^\dagger)} - a_{\pi^{-1}(L^\dagger),L^\dagger} + a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)}\big]\mathbf{1}_{1,1}. \qquad (4.195)
\end{aligned}$$

For the first term in (4.195), dropping the restriction $t \neq l$ and summing over t to obtain the first inequality, and then applying (4.189) with $|a_{il}|$ replaced by $|a_{jl}|$, we obtain

$$\begin{aligned}
EU|a_{J^\dagger,L^\dagger}|\mathbf{1}_{1,1} &= \frac{1}{2}\sum |a_{jl}|\mathbf{1}(s=k, t\neq l)\, p_3(i,j,k,l,s,t) \\
&\leq \frac{1}{8n^2(n-1)^2\sigma^2}\sum |a_{jl}|\big[(a_{ik}+a_{jl}) - (a_{il}+a_{jk})\big]^2 \\
&\leq \frac{\gamma}{2(n-1)^2\sigma^2}. \qquad (4.196)
\end{aligned}$$

The second, third and fourth terms in (4.195) also may be bounded by (4.196) upon replacing $|a_{jl}|$ by $|a_{il}|$, $|a_{jk}|$ and $|a_{ik}|$, respectively, yielding

$$E\big|Ua_{J^\dagger,L^\dagger} + (1-U)(a_{I^\dagger,L^\dagger} + a_{J^\dagger,K^\dagger} - a_{I^\dagger,K^\dagger})\big|\mathbf{1}_{1,1} \leq \frac{2\gamma}{(n-1)^2\sigma^2}. \qquad (4.197)$$

For the fifth term in (4.195), that is, for $-a_{J^\dagger,\pi(J^\dagger)}$, reasoning similarly,

$$\begin{aligned}
E|a_{J^\dagger,\pi(J^\dagger)}|\mathbf{1}_{1,1} &= \sum |a_{jt}|\mathbf{1}(s=k, t\neq l)\, p_3(i,j,k,l,s,t) \\
&\leq \frac{1}{4n^3(n-1)^2\sigma^2}\sum |a_{jt}|\big[(a_{ik}+a_{jl}) - (a_{il}+a_{jk})\big]^2 \\
&\leq \frac{\gamma}{(n-1)^2\sigma^2}. \qquad (4.198)
\end{aligned}$$

Note that for the final inequality, though the sum being bounded is not of the form (4.189), having the index t, the same reasoning applies and that, moreover, the five indices of summation require that n^2 in (4.190) be replaced by n^3.

To handle the sixth term in (4.195), $-a_{\pi^{-1}(L^\dagger),L^\dagger}$, we need the joint distribution

$$\begin{aligned} &p_4(i,j,k,l,s,t,u) \\ &\quad = P\big(\big(I^\dagger, J^\dagger, K^\dagger, L^\dagger, \pi(I^\dagger), \pi(J^\dagger), \pi^{-1}(L^\dagger)\big) = (i,j,k,l,s,t,u)\big), \end{aligned}$$

accounting for the value u taken on by $\pi^{-1}(L^\dagger)$. If l equals s or t, then u is already fixed at i or j, respectively; otherwise, $\pi^{-1}(L^\dagger)$ is free to take any of the remaining available $n-2$ values, with equal probability. Hence, with p_3 given by (4.192), we deduce that

$$p_4(i,j,k,l,s,t,u) = \begin{cases} p_3(i,j,k,l,s,t), & \text{if } (l,u) \in \{(s,i),(t,j)\}, \\ p_3(i,j,k,l,s,t)\frac{1}{n-2}, & \text{if } l \notin \{s,t\} \text{ and } u \notin \{i,j\}, \\ 0, & \text{otherwise.} \end{cases}$$

Note, for example, that on $\mathbf{1}_{1,1}$, where $\pi(I^\dagger) = K^\dagger$ and $\pi(J^\dagger) \neq L^\dagger$, the value u of $\pi^{-1}(L^\dagger)$ is neither $I^\dagger$ nor $J^\dagger$, so the second case above is the relevant one and the vanishing of the first sum on the third line of the following display is to be expected.

Now, applying the density p_4 we may bound the sixth term in (4.195) as follows,

$$\begin{aligned} &E|a_{\pi^{-1}(L^\dagger),L^\dagger}|\mathbf{1}_{1,1} \\ &\quad = \sum |a_{ul}|\mathbf{1}(s=k, t\neq l)\, p_4(i,j,k,l,s,t,u) \\ &\quad = \sum_{t\neq l} |a_{ul}|\, p_4(i,j,k,l,k,t,u) \\ &\quad = \sum |a_{ik}|\, p_3(i,j,k,k,k,t) + \frac{1}{n-2} \sum_{l\notin\{k,t\}, u\notin\{i,j\}} |a_{ul}|\, p_3(i,j,k,l,k,t) \\ &\quad = \frac{1}{n-2} \sum_{l\neq t, u\notin\{i,j\}} |a_{ul}|\, p_2(i,j,k,l)\, p_1(k,t) \\ &\quad = \frac{1}{(n)_3} \sum_{t\notin\{l,k\},\, u\notin\{i,j\}} |a_{ul}|\, p_2(i,j,k,l) \qquad (4.199) \\ &\quad = \frac{1}{(n)_2} \sum_{u\notin\{i,j\}} |a_{ul}|\, p_2(i,j,k,l) \\ &\quad \le \frac{1}{4n^3(n-1)^2\sigma^2} \sum |a_{ul}|\big[(a_{ik}+a_{jl}) - (a_{il}+a_{jk})\big]^2 \\ &\quad \le \frac{\gamma}{(n-1)^2\sigma^2}, \qquad (4.200) \end{aligned}$$

where the final inequality is achieved using (4.189) in the same way as for (4.198).

The computation for the seventh term in (4.195) begins as that for the sixth, yielding (4.199) with a_{ut} replacing a_{ul}, so that

$$\begin{aligned} E|a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)}|\mathbf{1}_{1,1} &= \frac{1}{(n)_3} \sum_{t\notin\{l,k\},u\notin\{i,j\}} |a_{ut}|\, p_2(i,j,k,l) \\ &\le \frac{1}{4(n)_3 n^2(n-1)\sigma^2} \sum |a_{ut}|\big[(a_{ik}+a_{jl}) - (a_{il}+a_{jk})\big]^2 \end{aligned}$$

$$\leq \frac{n^2\gamma}{(n)_3(n-1)\sigma^2}$$
$$\leq \frac{3\gamma}{(n-1)^2\sigma^2}, \tag{4.201}$$

where we have applied reasoning as in (4.189), replaced n^2 by n^4 in (4.190) due to the sum over six indices, and recalled our assumption that $n \geq 3$.

Returning to (4.195) and adding the contribution (4.197) from the first four terms together with (4.198), (4.200) and (4.201) from the fifth, sixth and seventh, respectively, we obtain $E|V|\mathbf{1}_{1,1} \leq 7\gamma/((n-1)^2\sigma^2)$. Since, by symmetry, all four terms on the right-hand side of (4.194) can be handled in the same way as the first, we obtain the following bound on the event $R = 1$:

$$E|V|\mathbf{1}_1 \leq \frac{28\gamma}{(n-1)^2\sigma^2}. \tag{4.202}$$

Calculation on $R = 0$ We may write the indicator of the event that $R = 0$ as

$$\mathbf{1}_0 = \mathbf{1}\big(\pi(I^\dagger) \notin \{K^\dagger, L^\dagger\}, \pi(J^\dagger) \notin \{K^\dagger, L^\dagger\}\big),$$

and we see from (4.186) that $\mathcal{I} = \{I^\dagger, J^\dagger, \pi^{-1}(K^\dagger), \pi^{-1}(L^\dagger)\}$, a set of size 4, on $R = 0$. Hence, from (4.187),

$$\begin{aligned} V\mathbf{1}_0 &= \sum_{i \in \{I^\dagger, J^\dagger, \pi^{-1}(K^\dagger), \pi^{-1}(L^\dagger)\}} \big(U a_{i,\pi^\dagger(i)} + (1-U)a_{i,\pi^\ddagger(i)} - a_{i,\pi(i)}\big)\mathbf{1}_0 \\ &= \big[U(a_{I^\dagger,K^\dagger} + a_{J^\dagger,L^\dagger}) + (1-U)(a_{I^\dagger,L^\dagger} + a_{J^\dagger,K^\dagger}) \\ &\quad + a_{\pi^{-1}(K^\dagger),\pi(I^\dagger)} + a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)} \\ &\quad - (a_{I^\dagger,\pi(I^\dagger)} + a_{J^\dagger,\pi(J^\dagger)} + a_{\pi^{-1}(K^\dagger),K^\dagger} + a_{\pi^{-1}(L^\dagger),L^\dagger})\big]\mathbf{1}_0. \end{aligned} \tag{4.203}$$

Since the first four terms in (4.203) have the same distribution, we bound their contribution to $E|V|\mathbf{1}_0$, using (4.189), by

$$\begin{aligned} 4EU|a_{I^\dagger,K^\dagger}|\mathbf{1}_0 \leq 4EU|a_{I^\dagger,K^\dagger}| &= 2\sum |a_{ik}|p_2(i,j,k,l) \\ &= \frac{1}{2n^2(n-1)\sigma^2}\sum |a_{ik}|\big[(a_{ik} + a_{jl}) - (a_{il} + a_{jk})\big]^2 \\ &\leq \frac{2\gamma}{(n-1)\sigma^2}. \end{aligned} \tag{4.204}$$

The sum of the contributions from the fifth and sixth terms of (4.203) can be bounded as

$$\begin{aligned} &2E|a_{\pi^{-1}(L^\dagger),\pi(J^\dagger)}|\mathbf{1}_0 \\ &\quad = 2\sum_{s \notin \{k,l\}, t \notin \{k,l\}} |a_{ut}|p_4(i,j,k,l,s,t,u) \end{aligned}$$

$$= \frac{2}{n-2} \sum_{s\notin\{k,l\},t\notin\{k,l\},u\notin\{i,j\},s\neq t} |a_{ut}|\, p_3(i,j,k,l,s,t)$$

$$\leq \frac{n-3}{2(n-2)n^3(n-1)^2\sigma^2} \sum |a_{ut}| \big[(a_{ik}+a_{jl})-(a_{il}+a_{jk})\big]^2 \tag{4.205}$$

$$\leq \frac{2n(n-3)\gamma}{(n-2)(n-1)^2\sigma^2}$$

$$\leq \frac{2\gamma}{(n-1)\sigma^2}, \tag{4.206}$$

where the second equality follows from the form of p_4 and that $l \notin \{s,t\}$ implies $(l,u) \notin \{(s,i),(t,j)\}$, inequality (4.205) is obtained by summing over the $n-3$ choices of s and dropping the remaining restrictions, and the next inequality by following the reasoning of (4.189).

Similarly, for the sum of the contributions from the seventh and eighth terms of (4.203), summing over the $n-3$ choices of t and then dropping the remaining restrictions to obtain the first inequality, we have

$$2E|a_{I^\dagger,\pi(I^\dagger)}|\mathbf{1}_0 = 2 \sum_{s\notin\{k,l\},t\notin\{k,l\}} |a_{is}|\, p_3(i,j,k,l,s,t)$$

$$= \frac{1}{2n^3(n-1)^2\sigma^2} \sum_{s\notin\{k,l\},t\notin\{k,l\},s\neq t} |a_{is}| \big[(a_{ik}+a_{jl})-(a_{il}+a_{jk})\big]^2$$

$$\leq \frac{n-3}{2n^3(n-1)^2\sigma^2} \sum |a_{is}| \big[(a_{ik}+a_{jl})-(a_{il}+a_{jk})\big]^2$$

$$\leq \frac{2(n-3)\gamma}{(n-1)^2\sigma^2}$$

$$\leq \frac{2\gamma}{(n-1)\sigma^2}. \tag{4.207}$$

The total contribution of the ninth and tenth terms together can be bounded like the sum of the fifth and sixth, yielding (4.205) with $|a_{ul}|$ replacing $|a_{ut}|$, and then summing over the n choices of t to give

$$2E|a_{\pi^{-1}(L^\dagger),L^\dagger}|\mathbf{1}_0 \leq \frac{n-3}{2(n-2)n^2(n-1)^2\sigma^2} \sum |a_{ul}| \big[(a_{ik}+a_{jl})-(a_{il}+a_{jk})\big]^2$$

$$\leq \frac{2n(n-3)\gamma}{(n-2)(n-1)^2\sigma^2}$$

$$\leq \frac{2\gamma}{(n-1)\sigma^2}. \tag{4.208}$$

Adding up the bounds for the first four terms (4.204), the fifth and sixth terms (4.206), the seventh and eighth terms (4.207) and the ninth through tenth terms (4.208) yields

$$E|V|\mathbf{1}_0 \leq \frac{8\gamma}{(n-1)\sigma^2}. \tag{4.209}$$

Now, from (4.188), adding up the contributions from (4.193), (4.202) and (4.209) from $R = 2$, $R = 1$, and $R = 0$, respectively, for this coupling of Y^* and Y we find that

$$E|Y^* - Y| \leq \frac{\gamma}{(n-1)\sigma^2}\left(8 + \frac{28}{(n-1)} + \frac{4}{(n-1)^2}\right).$$

The proof of the lemma may now be completed by noting that $E|Y^* - Y|$ is an upper bound on the L^1 norm $\|\mathcal{L}(Y^*) - \mathcal{L}(Y)\|_1$, by the dual form of the L^1 norm 4.6.

□