Stein's Method and the Zero Bias Transformation with Application to Simple Random Sampling: Technical Report

Larry Goldstein and Gesine Reinert
April 11, 2001

Abstract

Let W be a random variable with mean zero and variance σ^2. The distribution of a variate W^*, satisfying $EWf(W) = \sigma^2 Ef'(W^*)$ for smooth functions f, exists uniquely and defines the zero bias transformation on the distribution of W. The zero bias transformation shares many interesting properties with the well known size bias transformation for non-negative variables, but is applied to variables taking on both positive and negative values. The transformation can also be defined on more general random objects. The relation between the transformation and the expression $wf'(w) - \sigma^2 f''(w)$ which appears in the Stein equation characterizing the mean zero, variance σ^2 normal σZ can be used to obtain bounds on the difference $E\{h(W/\sigma) - h(Z)\}$ for smooth functions h by constructing the pair (W, W^*) jointly on the same space. When W is a sum of n not necessarily independent variates, under certain conditions which include a vanishing third moment, bounds on this difference of the order $1/n$ for classes of smooth functions h may be obtained. The technique is illustrated by an application to simple random sampling.

1 Introduction

Since 1972, Stein's method [13] has been extended and refined by many authors and has become a valuable tool for deriving bounds for distributional approximations, in particular, for normal and Poisson approximations for sums of random variables. (In the normal case, see, for example, Ho and Chen [10], Stein [14], [15], Barbour [2], Götze [9], Bolthausen and Götze [3], Rinott [11], and Goldstein and Rinott [8]). Through the use of differential or difference equations which characterize the target distribution, Stein's method allows many different types of dependence structures to be treated, and yields computable bounds on the approximation error.

The Stein equation for the normal is motivated by the fact that $W \sim \mathcal{N}(\mu, \sigma^2)$ if and only if

$$E\left\{(W-\mu)f'(W) - \sigma^2 f''(W)\right\} = 0 \quad \text{for all smooth } f.$$

Given a test function h, let $\Phi h = Eh(Z)$ where $Z \sim \mathcal{N}(0,1)$. If W is close to $\mathcal{N}(\mu, \sigma^2)$, $Eh((W-\mu)/\sigma) - \Phi h$ will be close to zero for a large class of functions h, and $E\left\{(W-\mu)f'(W) - \sigma^2 f''(W)\right\}$ will be close to zero for a large class of functions f. It is natural then, given h, to relate the functions h and f through the differential equation

$$(x-\mu)f'(x) - \sigma^2 f''(x) = h((x-\mu)/\sigma) - \Phi h, \tag{1}$$

and upon solving for f, compute $Eh((W-\mu)/\sigma) - \Phi h$ by $E\left\{(W-\mu)f'(W) - \sigma^2 f''(W)\right\}$ for this f. A bound on $Eh((W-\mu)/\sigma) - \Phi h$ can then be obtained by bounding the difference between $E(W-\mu)f'(W)$ and $\sigma^2 Ef''(W)$.

[0]Partially supported by NSF grant DMS-9505075

Stein [15], Baldi, Rinott, and Stein [1], and Goldstein and Rinott [8], among others, were able to exploit a connection between the Stein equation (1), and the size biasing of nonnegative random variables. If $W \geq 0$ has mean $0 < EW = \mu < \infty$, we say W^s has the W-size biased distribution if for all f such that $EWf(W)$ exists,

$$EWf(W) = \mu Ef(W^s). \tag{2}$$

The connection between the Stein equation and size biasing is described in Goldstein and Rinott [8]. In brief, one can obtain a bound on $Eh((W-\mu)/\sigma) - \Phi h$ in terms of a pair (W, W^s), coupled on a joint space, where W^s has the W-size biased distribution. The terms in this bound will be small if W and W^s are close. The variates W and W^s will be close, for example, when $W = X_1 + \cdots + X_n$ is the sum of i.i.d. random variables, as then W^s can be constructed by replacing a single summand X_i by an independent variate X_i^s that has the X_i-size biased distribution. Similar constructions exist for non-identically distributed and possibly dependent variates, and are studied in [8].

As noted in [8], the size biasing method works well for combinatorial problems such as counting the number of vertices in a random graph having prespecified degrees. When the distributions approximated are counts, size biasing is natural; in particular, the counts W are necessarily nonnegative. To size bias a W which may take on both positive and negative values, it may be that for some ρ, $W + \rho$ or $-W + \rho$ is a nonnegative random variable whose mean exists. Yet if W has support on both the infinite positive and negative half lines then some truncation must be involved in order to obtain a nonnegative random variable on which the size bias transformation can be performed. This is especially unnatural if W is symmetric, as one would expect that W itself would be closer to normal than any version of itself involving translation and truncation.

The transformation and associated coupling which we study here has many similarities to the size biasing approach, yet it may be applied directly to mean zero random variables and is particularly useful for symmetric random variables or those with vanishing third moment. The transformation is motivated by the size bias transformation and the equation that characterizes the mean zero normal:

$$Z \sim \mathcal{N}(0, \sigma^2) \quad \text{if and only if} \quad EWf(W) = \sigma^2 Ef'(W). \tag{3}$$

The similarity of the latter equation to equation (2) suggests, given a mean zero random variable W, considering a new distribution related to the distribution of W according to the following definition.

Definition 1.1 *Let W be a mean zero random variable with finite, nonzero variance σ^2. We say that W^* has the W-zero biased distribution if for all differentiable f for which $EWf(W)$ exists,*

$$EWf(W) = \sigma^2 Ef'(W^*). \tag{4}$$

The existence of the zero bias distribution for any such W is easily established. For a given $g \in C_c$, the collection of continuous functions with compact support, let $G = \int_0^w g$. The quantity

$$Tg = \sigma^{-2} E\{WG(W)\}$$

exists since $EW^2 < \infty$, and defines a linear operator $T : C_c \to \mathbf{R}$. To see moreover that T is positive, take $g \geq 0$. Then G is increasing, and therefore W and $G(W)$ are positively correlated. Hence $EWG(W) \geq EWEG(W) = 0$, and T is positive. Now invoking the Riesz representation theorem (see, eg. [7]), we have $Tg = \int gd\nu$

for some unique Radon measure ν, which is a probability measure by virtue of $T1 = 1$. In fact, the W-zero biased distribution is continuous for any nontrivial W; the density of W^* is calculated explicitly in Lemma 2.1, part (2).

Definition 1.1 describes a transformation, which we term the zero bias transformation, on distribution functions with mean zero and finite variance. However, for any W with finite variance we can apply the transformation to the centered variate $W - EW$.

The zero bias transformation has many interesting properties, some of which we collect below in Lemma 2.1. In particular, the mean zero normal is the unique fixed point of the zero bias transformation. From this it is intuitive that W will be close to normal in distribution if and only if W is close in distribution to W^*.

Use of the zero bias method, as with other like techniques, is through the use of coupling and a Taylor expansion of the Stein equation; in particular, we have

$$E[Wf'(W) - \sigma^2 f''(W)] = \sigma^2 E[f''(W^*) - f''(W)],$$

and the right hand side may now immediately be expanded about W. In contrast, the use of other techniques such as size biasing requires an intermediate step which generates an additional error term (e.g., see equation (19) in [8]). For this reason, using the zero bias technique one is able to show why bounds of smaller order than $1/\sqrt{n}$ for smooth functions h may be obtained when certain additional moment conditions apply.

For distributions with smooth densities, Edgeworth expansions reveal a similar phenomenon to what is studied here. For example, (see Feller [6]), if F has a density and vanishing third moment, then an i.i.d. sum of variates with distribution F has a density which can be uniformly approximated by the normal to within a factor of $1/n$. However, these results depend on the smoothness of the parent distribution F. What we show here, in the i.i.d. case say, is that for smooth test functions h, bounds of order $1/n$ hold for any F with vanishing third moment and finite fourth moment (see Corollary 3.1).

Generally, bounds for non-smooth functions are more informative than bounds for smooth functions (see for instance Götze [9], Bolthausen and Götze [3], Rinott and Rotar [12] and Dembo and Rinott [5]); bounds for non-smooth functions can be used for the construction of confidence intervals, for instance. Although the zero bias method can be used to obtain bounds for non-smooth functions, we will consider only smooth functions for the following reason. At present, constructions for use of the zero bias method are somewhat more difficult to achieve than constructions for other methods; in particular, compare the size biased construction in Lemma 2.1 of [8] to the construction in Theorem 2.1 here. Hence, for non-smooth functions, other techniques may be easier to apply. However, under added assumptions, the extra effort in applying the zero bias method will be rewarded by improved error bounds which may not hold over the class of non-smooth functions. For example, consider the i.i.d. sum of symmetric $+1, -1$ variates; the bound on non-smooth functions of order $1/\sqrt{n}$ is unimprovable and may be obtained by a variety of methods. Yet a bound of order $1/n$ holds for smooth functions, and can be shown to be achieved by the zero bias method. Hence, in order to reap the improved error bound benefit of the zero bias method when such can be achieved, we restrict attention to the class of smooth functions.

Ideas related to the zero bias transformation have been studied by Ho and Chen [10], and Cacoullos et al. [4]. Ho and Chen consider the zero bias distribution implicitly (see equation 1.3 of [10]) in their version of one of Stein's proofs of the Berry Esseen theorem. They treat a case with a W the sum of dependent variates, and obtain rates of $1/\sqrt{n}$ for the L_p norm of the difference between the distribution function of W and the normal.

The approach of Cacoullos et al. [4] is also related to what is studied here. In the zero bias transformation, the distribution of W is changed to that of W^* on the right hand side of identity (3), keeping the form of this identity, yielding (4). In [4], the distribution of W is preserved on the right hand side of (3), and the form of the identity changed to $E[Wf(W)] = \sigma^2 E[u(W)f'(W)]$, with the function u determined by the distribution of W. Note that both approaches reduce to identity (3) when W is normal; in the first case $W^* \stackrel{d}{=} W$, and in the second, $u(w) = 1$.

The paper is organized as follows. In Section 2, we present some of the properties of the zero bias transformation and give two coupling constructions that generate W and W^* on a joint space. The first construction, Lemma 2.1, part 5, is for the sum of independent variates, and its generalization, Theorem 2.1, for possibly dependent variates. In Section 3, we show how the zero bias transformation may be used to obtain bounds on the accuracy of the normal approximation in general. In Section 4, we apply the preceding results to obtain bounds of the order $1/n$ for smooth functions h when W is a sum obtained from simple random sampling without replacement (a case of global dependence), under a vanishing third moment assumption.

2 The Zero Bias Transformation

The following lemma summarizes some of the important features of the zero bias transformation; property (4) for $n = 1$ will be of special importance, as it gives that $EW^* = 0$ whenever $EW^3 = 0$.

Lemma 2.1 *Let W be a mean zero variable with finite, nonzero variance, and let W^* have the W-zero biased distribution in accordance with Definition 1.1. Then,*

1. *The mean zero normal is the unique fixed point of the zero bias transformation.*
2. *The zero bias distribution is unimodal about zero and continuous with density function $p(w) = \sigma^{-2}E[W, W > w]$. It follows that the support of W^* is the closed convex hull of the support of W and that W^* is bounded whenever W is bounded.*
3. *The zero bias transformation preserves symmetry.*
4. *$\sigma^2 E(W^*)^n = EW^{n+2}/(n+1)$ for $n \geq 1$.*
5. *Let $X_1, \ldots, X_n$ be independent mean zero random variables with $EX_i^2 = \sigma_i^2$. Set $W = X_1 + \cdots + X_n$, and $EW^2 = \sigma^2$. Let I be a random index independent of the X's such that*
$$P(I = i) = \sigma_i^2/\sigma^2.$$
Let
$$W_i = W - X_i = \sum_{j \neq i} X_j.$$
Then $W_I + X_I^*$ has the W-zero biased distribution. (This is analogous to size biasing a sum of non-negative independent variates by replacing a variate chosen proportional to its expectation by one chosen independently from its size biased distribution; see Lemma 2.1 in [8]).

6. *Let X be mean zero with variance σ_X^2 and distribution dF. Let $(\hat{X}', \hat{X}'')$ have distribution*
$$d\hat{F}(\hat{x}', \hat{x}'') = \frac{(\hat{x}' - \hat{x}'')^2}{2\sigma_X^2} dF(\hat{x}')dF(\hat{x}'').$$
Then, with U an independent uniform variate on $[0,1]$, $U\hat{X}' + (1-U)\hat{X}''$ has the X-zero biased distribution.

Proof of claims:

1. This is immediate from Definition 1.1 and the characterization (3).
2. The function $p(w)$ is increasing for $w < 0$, and decreasing for $w > 0$. Since $EW = 0$, $p(w)$ has limit 0 at both plus and minus infinity, and $p(w)$ must therefore be nonnegative and unimodal about zero. That p integrates to 1 and is the density of a variate W^* which satisfies (4) follows by uniqueness (see the remarks following Definition 1.1), and by applying Fubini's theorem separately to $E[f'(W^*); W^* \geq 0]$ and $E[f'(W^*); W^* < 0]$, using
$$E[W; W > w] = -E[W; W \leq w],$$
which follows from $EW = 0$.
3. If w is a continuity point of the distribution function of a symmetric W, then $E[W; W > w] = E[-W; -W > w] = -E[W; W < -w] = E[W; W > -w]$ using $EW = 0$. Thus, there is a version of the dw density of W^* which is the same at w and $-w$ for almost all w $[dw]$; hence W^* is symmetric.
4. Substitute $w^{n+1}/(n+1)$ for $f(w)$ in the characterizing equation (4).
5. Using independence and equation (4) with X_i replacing W,
$$\begin{aligned} \sigma^2 Ef'(W^*) &= EWf(W) \\ &= \sum_{i=1}^n EX_i f(W) \\ &= \sum_{i=1}^n EX_i^2 Ef'(W_i + X_i^*) \\ &= \sigma^2 \sum_{i=1}^n \frac{\sigma_i^2}{\sigma^2} Ef'(W_i + X_i^*) \\ &= \sigma^2 Ef'(W_I + X_I^*). \end{aligned}$$
Hence, for all smooth f, $Ef'(W^*) = Ef'(W_I + X_I^*)$, and the result follows.
6. Let X', X'' denote independent copies of the variate X. Then,
$$\begin{aligned} \sigma_X^2 Ef'(U\hat{X}' + (1-U)\hat{X}'') &= \sigma_X^2 E\left(\frac{f(\hat{X}') - f(\hat{X}'')}{\hat{X}' - \hat{X}''}\right) \\ &= \frac{1}{2}E(X' - X'')(f(X') - f(X'')) \\ &= EX'f(X') - EX''f(X') \\ &= EXf(X) \\ &= \sigma_X^2 Ef'(X^*). \end{aligned}$$
Hence, for all smooth f, $Ef'(U\hat{X}' + (1-U)\hat{X}'') = Ef'(X^*)$.

q.e.d.

By (1) of Lemma 2.1, the mean zero normal is a fixed point of the zero bias transformation. One can also gain some insight into the nature of the transformation by observing its action on the distribution of the variate X taking the values -1 and $+1$ with equal probability. Calculating the density function of the X-zero biased variate X^* according to (2) of Lemma 2.1, we find that X^* is uniformly distributed on the interval $[-1,1]$. A similar calculation for the discrete mean zero variable X taking values $x_1 < x_2 < \cdots < x_n$ yields that the X-zero biased distribution is a mixture of uniforms over the intervals $[x_i, x_{i+1}]$. These examples may help in understanding how a uniform variate U enters in (6) of Lemma 2.1.

For a construction of W and W^* which may be applied in the presence of dependence, in the remainder of this section, we will consider the following framework. Let $X_1, \ldots, X_n$ be mean zero random variables, and with $W = X_1 + \cdots + X_n$, suppose $EW^2 = \sigma^2$ exists. For each $i = 1, \ldots, n$, assume that there exists a distribution $dF_{n,i}(x_1, \ldots, x_{i-1}, x_i', x_i'', x_{i+1}, \ldots, x_n)$ on $n+1$ variates $X_1, \ldots, X_{i-1}, X_i', X_i'', X_{i+1}, \ldots, X_n$ such that

$$(X_1, \ldots, X_{i-1}, X_i', X_i'', X_{i+1}, \ldots, X_n) \stackrel{d}{=} (X_1, \ldots, X_{i-1}, X_i'', X_i', X_{i+1}, \ldots, X_n), \tag{5}$$

and

$$(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n) \stackrel{d}{=} (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n). \tag{6}$$

(The choice $X_i' = X_i$ is natural, and then (6) is satisfied.)

Further, we will suppose that there is a ρ such that for all f for which $EWf(W)$ exists,

$$\sum_{i=1}^n EX_i' f(W_i + X_i'') = \rho EWf(W), \tag{7}$$

where $W_i = W - X_i$. We set

$$v_i^2 = E(X_i' - X_i'')^2. \tag{8}$$

Under these conditions, we have the following proposition.

Proposition 2.1

$$\rho = 1 - \frac{1}{2\sigma^2} \sum_{i=1}^n v_i^2.$$

Before proving this proposition, note that if a collection of variates already satisfies (5) and (6), and that if for each i,

$$E\{X_i' | W_i + X_i''\} = \frac{\rho}{n}(W_i + X_i''), \tag{9}$$

then

$$EX_i' f(W_i + X_i'') = \frac{\rho}{n} EWf(W),$$

and so condition (7) will be satisfied.

Proof of Proposition 2.1: Substituting $f(x) = x$ in (7) yields, by (6), that

$$\begin{aligned}
\rho\sigma^2 &= \sum_{i=1}^n EX_i'(W_i + X_i'') \\
&= \sum_{i=1}^n EX_i(W - X_i) + \sum_{i=1}^n EX_i'X_i'' \\
&= \sigma^2 - \sum_{i=1}^n EX_i^2 - \frac{1}{2}\sum_{i=1}^n \{E(X_i' - X_i'')^2 - E(X_i')^2 - E(X_i'')^2\} \\
&= \sigma^2 - \frac{1}{2}\sum_{i=1}^n v_i^2,
\end{aligned}$$

so that Proposition (2.1) follows. q.e.d.

The following theorem, generalizing (5) of Lemma 2.1, gives a coupling construction for W and W^* which may be applied in the presence of dependence under the framework of Proposition 2.1.

Theorem 2.1 *Let I be a random index independent of the X's such that*

$$P(I = i) = v_i^2 / \sum_{j=1}^n v_j^2.$$

Further, for i such that $v_i > 0$, let $\hat{X}_1, \ldots, \hat{X}_{i-1}, \hat{X}_i', \hat{X}_i'', \hat{X}_{i+1}, \ldots, \hat{X}_n$ be chosen according to the distribution

$$\begin{aligned}
&d\hat{F}_{n,i}(\hat{x}_1, \ldots, \hat{x}_{i-1}, \hat{x}_i', \hat{x}_i'', \hat{x}_{i+1}, \ldots, \hat{x}_n) \\
&= \frac{(\hat{x}_i' - \hat{x}_i'')^2}{v_i^2} dF_{n,i}(\hat{x}_1, \ldots, \hat{x}_{i-1}, \hat{x}_i', \hat{x}_i'', \hat{x}_{i+1}, \ldots, \hat{x}_n). \qquad (10)
\end{aligned}$$

Put

$$\hat{W}_i = \sum_{j \neq i} \hat{X}_j. \qquad (11)$$

Then, with U a uniform $U[0,1]$ variate which is independent of the X's and the index I,

$$U\hat{X}_I' + (1 - U)\hat{X}_I'' + \hat{W}_I$$

has the W-zero biased distribution.

In particular, when $X_1, \ldots, X_n$ are exchangeable, if one constructs exchangeable variables with distribution $dF_{n,1}$ which satisfy $v_1^2 > 0$, (6), and (9) for $i = 1$, then

$$U\hat{X}_1' + (1 - U)\hat{X}_1'' + \hat{W}_1$$

has the W-zero biased distribution.

Proof of Theorem 2.1: With Proposition 2.1 we have

$$\begin{aligned}
E\int_0^1 f'(u\hat{X}_I' + (1-u)\hat{X}_I'' + \hat{W}_I)du &= E\left(\frac{f(\hat{W}_I + \hat{X}_I') - f(\hat{W}_I + \hat{X}_I'')}{\hat{X}_I' - \hat{X}_I''}\right) \\
&= \sum_{i=1}^n \frac{v_i^2}{\sum v_j^2} E\left(\frac{f(\hat{W}_i + \hat{X}_i') - f(\hat{W}_i + \hat{X}_i'')}{\hat{X}_i' - \hat{X}_i''}\right) \\
&= \frac{1}{\sum v_j^2}\sum_{i=1}^n E(X_i' - X_i'')(f(W_i + X_i') - f(W_i + X_i''))
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\sum v_j^2} \sum_{i=1}^{n} \left(EX_i' f(W_i + X_i') - EX_i' f(W_i + X_i'') \right) \\
&= \frac{2}{\sum v_j^2} \left\{ EWf(W) - \rho EWf(W) \right\} \\
&= \frac{2(1-\rho)}{\sum v_j^2} EWf(W) \\
&= \frac{1}{\sigma^2} EWf(W) \\
&= Ef'(W^*),
\end{aligned}
$$

using Proposition (2.1) for the next to last step.

To show the claim in the case where the variates are exchangeable, set $dF_{n,i} = dF_{n,1}$ for $i = 2, \ldots, n$ and observe that the $dF_{n,i}$ so defined now satisfy the conditions of the theorem, and the distributions of the resulting $U\hat{X}_i' + (1-U)\hat{X}_i'' + \hat{W}_i$ does not depend on i.

q.e.d.

Note that if the variates $X_1, \ldots, X_n$ are independent, one can generate the collection $X_1, \ldots, X_{i-1}, X_i', X_i'', X_{i+1}, \ldots, X_n$ by letting X_i', X_i'' be independent replicates of X_i. In this case, conditions (5), (6), and (7) above are satisfied, the last with $\rho = 0$, and the construction reduces to that given in (5) of Lemma 2.1, in view of (6) of that same lemma.

3 Bounds in the Central Limit Theorem

The construction of Lemma 2.1, part 5, together with the following bounds of Barbour [2] and Götze [9] on the solution f of the differential equation (1) for a test function h with k bounded derivatives,

$$||f^{(j)}|| \le (j\sigma^j)^{-1} ||h^{(j)}|| \quad j = 1, \ldots, k, \tag{12}$$

yield the following remarkably simple proof of the Central Limit Theorem, with bounds on the approximation error, for independent possibly non-identically distributed mean zero variables $X_1, \ldots, X_n$ with variance 1 and common absolute first and third moments.

By Lemma 2.1, part (5), using independence, we can achieve W^* having the W-zero biased distribution by selecting a random index I uniformly and replacing X_I by an independent variable X_I^* having the X_I-zero biased distribution. Now, since $EWf(W) = \sigma^2 Ef'(W^*)$, using the bound (12),

$$
\begin{aligned}
|E\{h(W/\sigma) - \Phi h\}| &= \left| E\left\{ Wf'(W) - \sigma^2 f''(W) \right\} \right| \\
&= \sigma^2 |E\{f''(W^*) - f''(W)\}| \\
&\le \sigma^2 ||f^{(3)}|| E|W^* - W| \\
&\le \frac{1}{3\sigma} ||h^{(3)}|| E|X_I^* - X_I|. \qquad (13)
\end{aligned}
$$

Now, using the bound $E|X_I^* - X_I| \le E|X_I^*| + E|X_I|$ and the function $x^2\mathrm{sgn}(x)$ and its derivative $2|x|$ in equation (4) , we derive $E|X_i^*| = \frac{1}{2}E|X_i|^3$, and therefore $E|X_I^*| = \frac{1}{2}E|X_1|^3$. Next, $E|X_I| = E|X_1|$, and by Hölder's inequality and $EX_i^2 = 1$, we have $E|X_i| \le 1 \le E|X_i|^3$. Hence, since $EW^2 = n = \sigma^2$,

$$|E\{h(W/\sigma) - \Phi h\}| \le \frac{||h^{(3)}|| E|X_1|^3}{2\sqrt{n}}.$$

Thus we can obtain a bound of order $n^{-1/2}$ for smooth test functions with an explicit constant using only the first term in the Taylor expansion of $f''(W^*) - f''(W)$. For arbitrary independent mean zero variates, continuing from (13), for small additional effort the above inequality generalizes to

$$|E\{h(W/\sigma) - \Phi h\}| \leq \frac{||h^{(3)}|| E|X_I|^3}{2\sigma}.$$

The following theorem shows how the distance between an arbitrary mean zero, finite variance random variable W and a mean zero normal with the same variance can be bounded by the distance between W and a variate W^* with the W-zero biased distribution defined on a joint space. It is instructive to compare the following theorem with Theorem 1.1 of [8], the corresponding result when using the size biased transformation.

Theorem 3.1 *Let W be a mean zero random variable with variance σ^2, and suppose (W, W^*) is given on a joint probability space so that W^* has the W-zero biased distribution. Then for all h with four bounded derivatives,*

$$|Eh(W/\sigma) - \Phi h| \leq \frac{1}{3\sigma}\|h^{(3)}\|\sqrt{E\{E(W^* - W|W)^2\}} + \frac{1}{8\sigma^2}\|h^{(4)}\|E(W^* - W)^2.$$

Proof. For the given h, let f be the solution to (1). Then, using the bounds in (12), it suffices to prove

$$|E[Wf'(W) - \sigma^2 f''(W)]| \leq \sigma^2\|f^{(3)}\|\sqrt{E\{E(W^* - W|W)^2\}} + \frac{\sigma^2}{2}\|f^{(4)}\|E(W^* - W)^2.$$

By Taylor expansion, we have

$$\begin{aligned} |E[Wf'(W) - \sigma^2 f''(W)]| &= |\sigma^2 E[f''(W^*) - f''(W)]| \\ &\leq \sigma^2 |Ef^{(3)}(W)(W^* - W)| + \frac{\sigma^2}{2}\|f^{(4)}\|E(W^* - W)^2. \end{aligned}$$

For the first term, condition on W and then apply the Cauchy-Schwarz inequality;

$$\begin{aligned} |E[f^{(3)}(W)(W^* - W)]| &= |E[f^{(3)}(W)E(W^* - W|W)]| \\ &\leq \|f^{(3)}\|\sqrt{E\{E(W^* - W|W)^2\}}. \text{q.e.d.} \end{aligned}$$

For illustration only, we apply Theorem 3.1 to the sum of independent identically distributed variates to show how the the zero bias transformation leads to an error bound for smooth functions of order $1/n$, under additional moment assumptions which include a vanishing third moment.

Corollary 3.1 *Let $X, X_1, X_2, \ldots, X_n$ be independent and identically distributed mean zero, variance one random variables with vanishing third moment and EX^4 finite. Set $W = \sum_{i=1}^n X_i$. Then for any function h with four bounded derivatives,*

$$|E\{h(W/\sqrt{n}) - \Phi h\}| \leq n^{-1}\left\{\frac{1}{3}||h^{(3)}|| + \frac{1}{6}||h^{(4)}||EX^4\right\}.$$

Proof: Construct W^* as in Lemma 2.1, part (5). Then

$$E(W^* - W|W) = E(X_I^* - X_I|W) = E(X_I^*) - E(X_I|W),$$

since X_I^* and W are independent. Using the moment relation $EX^* = (1/2)EX^3$ given in Lemma 2.1, part (4), $EX_i^3 = 0$ implies that $EX_i^* = 0$, and so $EX_I^* = 0$.

Using that the X's are i.i.d., and therefore exchangeable, $E(X_I|W) = W/n$. Hence we obtain $E(X_I^* - X_I|W) = -W/n$, and

$$\sqrt{E\{E(X_I^* - X_I|W)^2\}} = \frac{1}{\sqrt{n}}.$$

For the second term in Theorem 3.1,

$$E(W^* - W)^2 = E(X_I^* - X_I)^2.$$

The moment relation property (4) in Lemma 2.1 and the assumption that EX^4 exists renders $E(X_I^* - X_I)^2$ finite and equal to $EX^4/3 + EX^2 \leq (4/3)EX^4$, by $EX^2 = 1$ and Hölder's inequality. Now using $\sigma^2 = n$ and applying Theorem 3.1 yields the assertion. q.e.d.

It is interesting to note that the constant ρ of equation (7) does not appear in the bounds of Theorem 3.1. One explanation of this phenomenon is as follows. The ρ of the coupling of Theorem 2.1 is related to the $\lambda \in (0,1)$ of a coupling of Stein [14], where a mean zero exchangeable pair (W, W'), with distribution $dF(w, w')$, satisfies $E\{W'|W\} = (1-\lambda)W$. One can show that if $(\hat{W}, \hat{W}')$ has distribution

$$d\hat{F}(\hat{w}, \hat{w}') = \frac{(\hat{w} - \hat{w}')^2}{E(W - W')^2} dF(\hat{w}, \hat{w}'),$$

then with U a uniform variate on [0,1], independent of all other variables, $U\hat{W} + (1-U)\hat{W}'$ has the W-zero bias distribution. Taking simple cases, one can see that the value of λ has no relation of the closeness of W to the normal. For instance, if W is the sum of n i.i.d. mean zero, variance one variables, then W is close to normal when n is large. However, for a given value of n, we may achieve any λ of the form j/n by taking W' to be the sum of any $n-j$ variables that make up the sum W, added to j i.i.d. variables that are independent of those that form W, but which have the same distribution.

We only study here the notion of zero biasing in one dimension; it is possible to extend this concept to any finite dimension. The definition of zero biasing in $\mathbf{R}^p$ is motivated by the following multivariate characterization. A vector $\mathbf{Z} \in R^p$ is multivariate normal with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ if and only if for all smooth test functions $f : R^p \to R$,

$$E\sum_{i=1}^{p} Z_i f_i(\mathbf{Z}) = E\sum_{i,j=1}^{p} \sigma_{ij} f_{ij}(\mathbf{Z}),$$

where $f_i, f_{ij}, \ldots$ denote the partial derivatives of f with respect to the indicated coordinates. Guided by this identity, given a mean zero vector $\mathbf{X} = (X_1, \ldots, X_p)$ with covariance matrix Σ, we say the collection of vectors $\mathbf{X}^* = (\mathbf{X}_{ij}^*)$ has the $\mathbf{X}$-zero bias distribution if

$$E\sum_{i=1}^{p} X_i f_i(\mathbf{X}) = E\sum_{i,j=1}^{p} \sigma_{ij} f_{ij}(\mathbf{X}_{ij}^*), \tag{14}$$

for all smooth f. As in the univariate case, the mean zero normal is a fixed point of the zero bias transformation; that is, if $\mathbf{X}$ is a mean zero normal vector, one may satisfy (14) by setting $\mathbf{X}_{ij}^* = \mathbf{X}$ for all i, j.

Using the definition of zero biasing in finite dimension, one can define the zero bias concept for random variables over an arbitrary index set $\mathcal{H}$ as follows. Given a collection $\{\xi(\phi), \phi \in \mathcal{H}\}$ of real valued mean zero random variables with nontrivial finite second moment, we say the collection $\{\xi^*_{\phi\psi}, \phi, \psi \in \mathcal{H}\}$ has the ξ-zero biased distribution if for all $p = 1, 2, \ldots$ and $(\phi_1, \phi_2, \ldots, \phi_p) \in \mathcal{H}^p$, the collection of p-vectors $(\mathbf{X}^*_{ij})$ has the $\mathbf{X}$-zero bias distribution, where, for $1 \le i, j \le p$,

$$(\mathbf{X}^*_{ij}) = (\xi^*_{\phi_i\phi_j}(\phi_1), \ldots, \xi^*_{\phi_i\phi_j}(\phi_p)),$$

and

$$\mathbf{X} = (\xi(\phi_1), \ldots, \xi(\phi_p)).$$

Again when ξ is normal, we may set $\xi^*_{\phi\psi} = \xi$ for all ϕ, ψ. This definition reduces to the one given above for random vectors when $\mathcal{H} = \{1, 2, \ldots, n\}$, and can be applied to, say, random processes by setting $\mathcal{H} = \mathbf{R}$, or random measures by letting $\mathcal{H}$ be a specified class of functions.

4 Application: Simple random sampling

We now apply Theorem 3.1 to obtain a bound on the error incurred when using the normal to approximate the distribution of a sum obtained by simple random sampling. In order to obtain a bound of order $1/n$ for smooth functions, we impose an additional moment condition as in Corollary 3.1.

Let $\mathcal{A} = \{a_1, \ldots, a_N\}$ be a set of real numbers such that

$$\sum_{a \in \mathcal{A}} a = \sum_{a \in \mathcal{A}} a^3 = 0; \tag{15}$$

we will often use the following consequence of (15),

$$\text{for any } E \subset \{1, \ldots, N\} \quad \text{and } k \in \{1, 3\}, \quad \sum_{a \in E} a^k = -\sum_{a \notin E} a^k. \tag{16}$$

We assume until the statement of Theorem 4.1 that the elements of $\mathcal{A}$ are distinct; this condition will be dropped in the theorem. Let $0 < n < N$, and set $N_n = N(N-1)\cdots(N-n+1)$, the n^{th} falling factorial of N. Consider the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ obtained by a simple random sample of size n from $\mathcal{A}$, that is, $\mathbf{X}$ is a realization of one of the equally likely N_n vectors of distinct elements of $\mathcal{A}$. Put

$$W = X_1 + \cdots + X_n. \tag{17}$$

Then, simply we have $EX_i = EX_i^3 = EW = EW^3 = 0$, and

$$EX_i^2 = \frac{1}{N} \sum_{a \in \mathcal{A}} a^2 = \sigma_X^2, \quad EW^2 = \frac{n(N-n)}{N(N-1)} \sum_{a \in \mathcal{A}} a^2 = \sigma^2, \quad \text{say.}$$

As we will consider the normalized variate W/σ, without loss of generality we may assume

$$\sum_{a \in \mathcal{A}} a^2 = 1; \tag{18}$$

note that (18) can always be enforced by rescaling $\mathcal{A}$, leaving (15) unchanged.

The next proposition shows how to apply Theorem 2.1 to construct W^* in the context of simple random sampling.

Proposition 4.1 *Let*

$$dF_{n,1}(x_1', x_1'', x_2, \ldots, x_n) = N_{n+1}^{-1}\mathbf{1}(\{x_1', x_1'', x_2, \ldots, x_n\} \subset \mathcal{A},\ \textit{distinct}), \tag{19}$$

the simple random sampling distribution on $n+1$ *variates from* $\mathcal{A}$*, and* $\hat{\mathbf{X}} = (\hat{X}_1', \hat{X}_1'', \hat{X}_2, \ldots, \hat{X}_n)$ *be a random vector with distribution*

$$d\hat{F}_{n,1}(\hat{\mathbf{x}}) = \frac{(\hat{x}_1' - \hat{x}_1'')^2}{2N}(N-2)_{n-1}^{-1}\mathbf{1}(\{\hat{x}_1', \hat{x}_1'', \hat{x}_2, \ldots, \hat{x}_n\} \subset \mathcal{A},\ \textit{distinct}). \tag{20}$$

Then, with U *a uniform* $[0,1]$ *random variable independent of* $\hat{\mathbf{X}}$*, and* $\hat{W}_1$ *given by (11),*

$$W^* = U\hat{X}_1' + (1-U)\hat{X}_1'' + \hat{W}_1 \tag{21}$$

has the W*-zero biased distribution.*

Proof. We apply Theorem 2.1 for exchangeable variates. With $X_1, \ldots, X_n$ a simple random sample of size n, the distributional relation (6) is immediate. Next, using the scaling (18), we see that v_1^2 given in (8) equals $2/(N-1)$, which is positive, and that furthermore, the distribution (20) is constructed from the distribution (19) according to the prescription (10). Lastly, using (16) with $k=1$, we have

$$E\{X_1' | X_1'', X_2, \ldots, , X_n\} = -\left(\frac{W_1 + X_1''}{N-n}\right),$$

and hence condition (9) is satisfied with $\rho = -n/(N-n)$. q.e.d.

We now begin to apply Theorem 3.1 by constructing W and W^* on a joint space. We achieve this goal by constructing the simple random sample $\mathbf{X} = (X_1, \ldots, X_n)$ together with the variates $\hat{\mathbf{X}} = (\hat{X}_1', \hat{X}_1'', \hat{X}_2, \ldots, \hat{X}_n)$ with distribution as in (20) of Proposition 4.1; W and W^* are then formed from these variates according to (17) and (21) respectively.

Construction of W and W^*. Start the construction with the simple random sample $\mathbf{X} = (X_1, \ldots, X_n)$. To begin the construction of $\hat{\mathbf{X}}$ with distribution (20), set

$$q(u,v) = \frac{(u-v)^2}{2N}\mathbf{1}(\{u,v\} \subset \mathcal{A}).$$

Note that variates U, V with distribution $q(u,v)$ will be unequal, and therefore we have that the distribution (20) factors as

$$d\hat{F}_{n,1}(\hat{\mathbf{x}}) = q(\hat{x}_1', \hat{x}_1'')(N-2)_{n-1}^{-1}\mathbf{1}(\{\hat{x}_2, \ldots, \hat{x}_n\} \subset \mathcal{A} \setminus \{\hat{x}_1', \hat{x}_1''\},\ \text{distinct}). \tag{22}$$

Hence, given $(\hat{X}_1', \hat{X}_1'')$, the vector $(\hat{X}_2, \ldots, \hat{X}_n)$ is a simple random sample of size $n-1$ from the $N-2$ elements of $\mathcal{A} \setminus \{\hat{X}_1', \hat{X}_1''\}$.

Now, independently of the chosen sample $\mathbf{X}$, pick $(\hat{X}_1', \hat{X}_1'')$ from the distribution $q(u,v)$. The variates $(\hat{X}_1', \hat{X}_1'')$ are then placed as the first two components in the vector $\hat{\mathbf{X}}$. How the remaining $n-1$ variates in $\hat{\mathbf{X}}$ are chosen depends on the amount of intersection between the sets $\{X_2, \ldots, X_n\}$ and $\{\hat{X}_1', \hat{X}_1''\}$. If these two sets do not intersect, fill in the remaining $n-1$ components of $\hat{\mathbf{X}}$ with $(X_2, \ldots, X_n)$. If the sets have an intersection, remove from the vector $(X_2, \ldots, X_n)$ the two variates (or single variate) that intersect and replace them (or it) with values obtained by a simple random sample of size two (one) from $\mathcal{A} \setminus \{\hat{X}_1', \hat{X}_1'', X_2, \ldots, X_n\}$. This new vector now fills in the remaining $n-1$ positions in $\hat{\mathbf{X}}$.

More formally, the construction is as follows. After generating $\mathbf{X}$ and $(\hat{X}_1', \hat{X}_1'')$ independently from their respective distributions, we define

$$R = |\{X_2, \ldots, X_n\} \cap \{\hat{X}_1', \hat{X}_1''\}|.$$

There are three cases.

Case 0: $R = 0$. In this case, set $(\hat{X}_1', \hat{X}_1'', \hat{X}_2, \ldots, \hat{X}_n) = (\hat{X}_1', \hat{X}_1'', X_2, \ldots, X_n)$.

Case 1: $R = 1$. If say, $\hat{X}_1'$ equals X_J, then set $\hat{X}_i = X_i$ for $2 \leq i \leq n, i \neq J$ and let $\hat{X}_J$ be drawn uniformly from $\mathcal{A} \setminus \{\hat{X}_1', \hat{X}_1'', X_2, \ldots, X_n\}$.

Case 2: $R = 2$. If $\hat{X}_1' = X_J$ and $\hat{X}_1'' = X_K$, say, then set $\hat{X}_i = X_i$ for $2 \leq i \leq n, i \notin \{J, K\}$, and let $\{\hat{X}_J, \hat{X}_K\}$ be a simple random sample of size 2 drawn from $\mathcal{A} \setminus \{\hat{X}_1', \hat{X}_1'', X_2, \ldots, X_n\}$.

The following proposition follows from Proposition 4.1, the representation of the distribution (20) as the product (22), and that fact that conditional on $\{\hat{X}_1', \hat{X}_1''\}$, the above construction leads to sampling uniformly by rejection from $\mathcal{A} \setminus \{\hat{X}_1', \hat{X}_1''\}$.

Proposition 4.2 *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a simple random sample of size* n *from* $\mathcal{A}$ *and let* $(\hat{X}_1', \hat{X}_2'') \sim q(u, v)$ *be independent of* $\mathbf{X}$. *If* $\hat{X}_2, \ldots, \hat{X}_n$, *given* $\hat{X}_1', \hat{X}_1'', X_2, \ldots, X_n$, *are constructed as above, then* $(\hat{X}_1', \hat{X}_1'', \hat{X}_2, \ldots, \hat{X}_n)$ *has distribution (20), and with* U *an independent uniform variate on* $[0, 1]$,

$$W^* = U\hat{X}_1' + (1 - U)\hat{X}_1'' + \hat{X}_2 + \cdots + \hat{X}_n \tag{23}$$
$$W = X_1 + \cdots + X_n \tag{24}$$

is a realization of (W, W^*) *on a joint space where* W^* *has the* $W-$*zero biased distribution.*

Under the moment conditions in (15), we have now the ingredients to show that a bound of order $1/n$ holds, for smooth functions, for the normal approximation of $W = \sum_{i=1}^n X_i$. First, define

$$\langle k \rangle = \sum_{a \in \mathcal{A}} a^k,$$
$$C_1(N, n, \mathcal{A}) = \sqrt{8}\Big(\frac{\sigma^2}{4n^2} + \langle 6 \rangle \alpha^2 + \beta^2 + \gamma^2(n-1)^2 + \eta^2\Big)^{1/2}$$

and

$$C_2(N, \mathcal{A}) = 11\langle 4 \rangle + \frac{45}{N},$$

where α, β, γ and η are given in (45), (46), (47), and (48) respectively.

Theorem 4.1 *Let* $X_1, \ldots, X_n$ *be a simple random sample of size* n *from a set of* N *real numbers* $\mathcal{A}$ *satisfying (15). Then with* $W = \sum_{i=1}^n X_i$, *for all* h *with four bounded derivatives we have*

$$|Eh(W/\sigma) - \Phi h| \leq \frac{1}{3\sigma} C_1(N, n, \mathcal{A}) ||h^{(3)}|| + \frac{1}{8\sigma^2} C_2(N, \mathcal{A}) ||h^{(4)}||. \tag{25}$$

Further, if $n \to \infty$ *so that* $n/N \to f \in (0, 1)$, *then it follows that*

$$|Eh(W/\sigma) - \Phi h| \leq n^{-1}\{B_1 ||h^{(3)}|| + B_2 ||h^{(4)}||\}(1 + o(1)),$$

where

$$B_1 = \frac{\sqrt{8}}{3}\Big(\frac{f(1-f)}{4} + n^2\langle 6 \rangle + 2\left(\frac{f}{1-f}\right)^2\Big)^{1/2} (f(1-f))^{-1/2} \text{ and } B_2 = \frac{1}{8}(11n\langle 4 \rangle + 45f)(f(1-f))^{-1}.$$

We see as follows that this bound yields a rate n^{-1} quite generally when values in $\mathcal{A}$ are "comparable." For example, suppose that $Y_1, Y_2, \ldots$ are independent copies of a nontrivial random variable Y with $EY^6 < \infty$ and $EY^2 = 1$. If N is say, even, let the elements of $\mathcal{A}$ be equal to the $N/2$ values $Y_1/(2\sum_1^{N/2} Y_j^2)^{1/2}, \ldots, Y_{N/2}/(2\sum_1^{N/2} Y_j^2)^{1/2}$ and their negatives. Then, this collection satisfies (15) and (18), and by the law of large numbers, a.s. as $N \to \infty$, the terms $n\langle 4\rangle$ and $n^2\langle 6\rangle$ converge to constants. Specifically,

$$n\langle 4\rangle \to fEY^4 \quad \text{and} \quad n^2\langle 6\rangle \to f^2EY^6,$$

and so B_1 and B_2 are asymptotically constant, and the rate $1/n$ is achieved over the class of functions h with bounded derivatives up to fourth order.

Proof. Both $Eh(W)$ and the upper bound in (25) are continuous functions of $\{a_1, \ldots, a_N\}$. Hence, since any collection of N numbers $\mathcal{A}$ is arbitrarily close to a collection of N distinct numbers, it suffices to prove the theorem under the assumption that the elements of $\mathcal{A}$ are distinct.

Constructing $\mathbf{X} = (X_1, \ldots, X_n)$, and $(\hat{X}_1', \hat{X}_1'', \hat{X}_2, \ldots, \hat{X}_n)$ according to Proposition 4.2, and (W, W^*) by (23) and (24), we realize, on a joint space, a W^* having the $W-$zero biased distribution. Recall that $R = |\{X_2, \ldots, X_n\} \cap \{\hat{X}_1', \hat{X}_1''\}|$. If $R = 1$, let J be the index in $\{2, \ldots, n\}$ for which X_J takes on the one value in $\{\hat{X}_1', \hat{X}_1''\}$; if $R = 2$, let J and K be the (distinct) indices such that $\{X_J, X_K\} = \{\hat{X}_1', \hat{X}_1''\}$. Then,

$$\begin{aligned} W^* - W &= [U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1] && (26)\\ &\quad +[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]\mathbf{1}(R=2) && (27)\\ &\quad +(\hat{X}_J - X_J)\mathbf{1}(R=1). && (28)\end{aligned}$$

We apply Theorem 3.1, and so need only show that $E\{E(W^* - W|W)\}^2$ is bounded by $C_1^2(N, n, \mathcal{A})$, and $E(W^* - W)^2$ by $C_2(N, \mathcal{A})$. Note that $EW^3 = 0$, $EW^* = 0$ by the moment relation given in part (4) of Lemma 2.1; hence the expectation of $E(W^* - W|W)$ equals zero, and the second moment of this variable may be replaced by its variance. We have,

$$\begin{aligned} &\mathrm{Var}\{E(W^* - W|W)\}\\ &= \mathrm{Var}\Big\{E(U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1|W)\\ &\quad +E([(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]\mathbf{1}(R=2) + (\hat{X}_J - X_J)\mathbf{1}(R=1)|W)\Big\}\\ &\le 2\Big\{\mathrm{Var}(E(U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1|W)) \qquad (29)\\ &\quad +\mathrm{Var}(E([(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]\mathbf{1}(R=2) + (\hat{X}_J - X_J)\mathbf{1}(R=1)|\mathbf{X}))\Big\}.\end{aligned}$$

Considering the contribution from (26) to the bound (29) first, we show that

$$E[U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1|W] = -\frac{1}{n}W. \qquad (30)$$

Because $\{\hat{X}_1', \hat{X}_1''\}$ is independent of W, the conditional expectation of term (26) given W equals $E[U\hat{X}_1' + (1-U)\hat{X}_1''] - E[X_1|W]$. From Lemma 2.1 part (6), $U\hat{X}_1' + (1-U)\hat{X}_1''$ has the X_1-zero biased distribution. Now, using $EX_1^3 = 0$ and Lemma 2.1 part (4) we obtain $E[U\hat{X}_1' + (1-U)\hat{X}_1''] = 0$. This leaves $-E(X_1|W)$, which by exchangeability, equals $-W/n$.

Consider now the contribution to (29) originating from the $R = 2$ term (27). On the event $\{R = 2\}, \{X_J, X_K\} = \{\hat{X}_1', \hat{X}_1''\}$, and so

$$[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]\mathbf{1}(R=2) = [(\hat{X}_J + \hat{X}_K) - (\hat{X}_1' + \hat{X}_1'')]\mathbf{1}(R=2). \qquad (31)$$

We now take the conditional expectation of (31) with respect to $\mathbf{X}$, beginning with

$$E\{(\hat{X}_1' + \hat{X}_1'')\mathbf{1}(R=2)|\mathbf{X}\} = E\{(\hat{X}_1' + \hat{X}_1'')\mathbf{1}(\{\hat{X}_1', \hat{X}_1''\} \subset \mathbf{Y})|\mathbf{X}\}, \qquad (32)$$

where

$$\mathbf{Y} = \{X_2, \ldots, X_n\}.$$

Hence (32) becomes

$$\begin{aligned}
\sum_{\{\hat{x}_1', \hat{x}_1''\} \subset \mathbf{Y}} (\hat{x}_1' + \hat{x}_1'')q(\hat{x}_1', \hat{x}_1'') &= \frac{1}{2N} \sum_{\{x,y\} \subset \mathbf{Y}} (x+y)(x-y)^2 \\
&= \frac{1}{N} \sum_{\{x,y\} \subset \mathbf{Y}} x(x-y)^2 \\
&= \frac{1}{N} \sum_{\{x,y\} \subset \mathbf{Y}} (x^3 - 2x^2y + xy^2) \\
&= \frac{1}{N}\{(n-1) \sum_{x \in \mathbf{Y}} x^3 - \sum_{x \in \mathbf{Y}} x \sum_{x \in \mathbf{Y}} x^2\}. \qquad (33)
\end{aligned}$$

To complete the conditional expectation of (31) with respect to $\mathbf{X}$, we need to subtract the above term from

$$\begin{aligned}
E\{(\hat{X}_J + \hat{X}_K)\mathbf{1}(R=2)|\mathbf{X}\} &= 2E\{\hat{X}_J \mathbf{1}(R=2)|\mathbf{X}\} \\
&= 2E\{\hat{X}_J|\mathbf{X}, R=2\}P(R=2|\mathbf{X}). \qquad (34)
\end{aligned}$$

Given $\mathbf{X}$ and that $\{R=2\}$, $\hat{X}_J$ is a simple random sample of size 1 from $\mathcal{A} \setminus \mathbf{Y}$. Hence, by (16),

$$2E\{\hat{X}_J|\mathbf{X}, R=2\} = 2\sum_{x \notin \mathbf{Y}} \frac{x}{(N-n+1)} = \frac{-2}{N-n+1} \sum_{x \in \mathbf{Y}} x.$$

Multiplying the above expression by $P(R=2|\mathbf{X})$ to obtain (34), we have

$$\frac{-2}{N-n+1} \sum_{x \in \mathbf{Y}} x \sum_{\{x,y\} \subset \mathbf{Y}} q(x,y) = \frac{-2}{N-n+1} \sum_{x \in \mathbf{Y}} x \sum_{\{x,y\} \subset \mathbf{Y}} \frac{(x-y)^2}{2N}.$$

Simplifying, we see that (34) equals

$$\frac{-2}{N(N-n+1)} \sum_{x \in \mathbf{Y}} x \left\{ (n-1) \sum_{x \in \mathbf{Y}} x^2 - (\sum_{x \in \mathbf{Y}} x)^2 \right\}. \qquad (35)$$

Subtracting (33) from (35) we obtain the conditional expectation $E\{[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]\mathbf{1}(R=2)|\mathbf{X}\}$,

$$\frac{1}{2N}\Big\{-2(n-1)\sum_{x \in \mathbf{Y}} x^3 + \left(\frac{2N-6n+6}{N-n+1}\right) \sum_{x \in \mathbf{Y}} x \sum_{x \in \mathbf{Y}} x^2 + \frac{4}{N-n+1}(\sum_{x \in \mathbf{Y}} x)^3\Big\}. \quad (36)$$

We now consider the term $E\{(\hat{X}_J - X_J)\mathbf{1}(R=1)|\mathbf{X}\}$ to analyze the contribution from (28) to (29). Using the symmetry between $\hat{X}_1'$ and $\hat{X}_1''$ we obtain

$$\begin{aligned}
&E\{(\hat{X}_J - X_J)\mathbf{1}(R=1)|\mathbf{X}\} \\
&= 2E\{(\hat{X}_J - X_J)\mathbf{1}(\hat{X}_1' \in \mathbf{Y}, \hat{X}_1'' \notin \mathbf{Y})|\mathbf{X}\} \\
&= 2E\{(\hat{X}_J - \hat{X}_1')\mathbf{1}(\hat{X}_1' \in \mathbf{Y}, \hat{X}_1'' \notin \mathbf{Y})|\mathbf{X}\}. \qquad (37)
\end{aligned}$$

Using (16),

$$
\begin{aligned}
& E\{\hat{X}_J \mathbf{1}(\hat{X}_1' \in \mathbf{Y}, \hat{X}_1'' \notin \mathbf{Y})|\mathbf{X}\} && (38)\\
&= \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} q(x,y)\Big(\frac{1}{N-n} \sum_{z\notin\mathbf{Y}\cup\{y\}} z\Big) \\
&= -\frac{1}{N-n} \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} q(x,y)\Big(y + \sum_{z\in\mathbf{Y}} z\Big), && (39)
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} q(x,y) &= \frac{1}{2N} \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} (x^2 - 2xy + y^2) \\
&= \frac{1}{2N}\left\{(N-n+1)\sum_{x\in\mathbf{Y}} x^2 + 2\left(\sum_{x\in\mathbf{Y}} x\right)^2 + (n-1)\sum_{y\notin\mathbf{Y}} y^2\right\} \\
&= \frac{1}{2N}\left\{(N-2n+2)\sum_{x\in\mathbf{Y}} x^2 + 2\left(\sum_{x\in\mathbf{Y}} x\right)^2 + (n-1)\right\}, \qquad (40)
\end{aligned}
$$

where the identity $\sum_{y\notin\mathbf{Y}} y^2 = 1 - \sum_{x\in\mathbf{Y}} x^2$ was applied. Moreover, using (16) again,

$$
\begin{aligned}
\sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} yq(x,y) &= \frac{1}{2N} \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} (yx^2 - 2xy^2 + y^3) \\
&= \frac{1}{2N}\left\{-\left(\sum_{x\in\mathbf{Y}} x^2\right)\left(\sum_{x\in\mathbf{Y}} x\right) - 2\left(\sum_{x\in\mathbf{Y}} x\right)\left(1 - \sum_{x\in\mathbf{Y}} x^2\right) - (n-1)\sum_{x\in\mathbf{Y}} x^3\right\} \\
&= \frac{1}{2N}\left\{\left(\sum_{x\in\mathbf{Y}} x^2\right)\left(\sum_{x\in\mathbf{Y}} x\right) - 2\sum_{x\in\mathbf{Y}} x - (n-1)\sum_{x\in\mathbf{Y}} x^3\right\}. \qquad (41)
\end{aligned}
$$

Following (39), combining (40) and (41) gives that (38) equals $1/(2N(N-n))$ times

$$
-(N-2n+3)(\sum_{x\in\mathbf{Y}} x)(\sum_{x\in\mathbf{Y}} x^2) - 2(\sum_{x\in\mathbf{Y}} x)^3 - (n-3)\sum_{x\in\mathbf{Y}} x + (n-1)\sum_{x\in\mathbf{Y}} x^3. \quad (42)
$$

Returning to (37), for $E\{\hat{X}_1' \mathbf{1}(\hat{X}_1' \in \mathbf{Y}, \hat{X}_1'' \notin \mathbf{Y})|\mathbf{X}\}$ we obtain

$$
\begin{aligned}
& \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} xq(x,y) \\
&= \frac{1}{2N} \sum_{x\in\mathbf{Y}} \sum_{y\notin\mathbf{Y}} x(x^2 - 2xy + y^2) \\
&= \frac{1}{2N}\{(N-n+1)\sum_{x\in\mathbf{Y}} x^3 + 2\sum_{x\in\mathbf{Y}} x\left(\sum_{x\in\mathbf{Y}} x^2\right) + \sum_{x\in\mathbf{Y}} x \sum_{y\notin\mathbf{Y}} y^2\} \\
&= \frac{1}{2N}\{(N-n+1)\sum_{x\in\mathbf{Y}} x^3 + \sum_{x\in\mathbf{Y}} x\left(\sum_{x\in\mathbf{Y}} x^2\right) + \sum_{x\in\mathbf{Y}} x\}. \qquad (43)
\end{aligned}
$$

Equations (42) and (43) yield that $E\{(\hat{X}_J - X_J)\mathbf{1}(R=1)|\mathbf{X}\}$ equals $(2N)^{-1}$ times

$$
\begin{aligned}
& 2\left(\frac{-(N-n)(N-n+1)+n-1}{N-n}\right)\sum_{x\in\mathbf{Y}} x^3 + \left(\frac{-4N+6n-6}{N-n}\right)\sum_{x\in\mathbf{Y}} x \sum_{x\in\mathbf{Y}} x^2 \\
&\qquad +2\left(\frac{-N+3}{N-n}\right)\sum_{x\in\mathbf{Y}} x - \frac{4}{N-n}(\sum_{x\in\mathbf{Y}} x)^3. \quad (44)
\end{aligned}
$$

Hence, from (29), using (30) and combining (44) with (36),

$$\mathrm{Var}\{E(W^* - W|W)\} \leq 2\mathrm{Var}\Big(\frac{1}{n}W\Big) + 2\mathrm{Var}(A)$$

where

$$A = \alpha \sum_{x\in\mathbf{Y}} x^3 + \beta \sum_{x\in\mathbf{Y}} x \sum_{x\in\mathbf{Y}} x^2 + \gamma(\sum_{x\in\mathbf{Y}} x)^3 + \eta \sum_{x\in\mathbf{Y}} x,$$

and

$$\alpha = \frac{n-1}{N(N-n)} - 1 \tag{45}$$

$$\beta = \frac{-2(n-1)}{N(N-n+1)} + \frac{n-3}{N(N-n)} - \frac{1}{N} \tag{46}$$

$$\gamma = \frac{-2}{N(N-n)(N-n+1)} \tag{47}$$

$$\eta = \frac{-N+3}{N(N-n)}. \tag{48}$$

We have immediately that

$$2\mathrm{Var}\Big(\frac{1}{n}W\Big) = \frac{2\sigma^2}{n^2}.$$

Moreover, twice the variance of A can be bounded by

$$8(\alpha^2 \mathrm{Var}(\sum_{x\in\mathbf{Y}} x^3) + \beta^2 \mathrm{Var}(\sum_{x\in\mathbf{Y}} x \sum_{x\in\mathbf{Y}} x^2) + \gamma^2 \mathrm{Var}((\sum_{x\in\mathbf{Y}} x)^3) + \eta^2 \mathrm{Var}(\sum_{x\in\mathbf{Y}} x)).$$

We now note that all four summands of A have mean zero; this is clear for the first and the fourth summand. For the second,

$$\begin{aligned} E(\sum_{x\in\mathbf{Y}} x \sum_{x\in\mathbf{Y}} x^2) &= E\left(\sum_{i\neq j} X_i^2 X_j + \sum_i X_i^3\right) \\ &= \frac{(n-1)_2}{N_2} \sum_{i\neq j} a_i^2 a_j \\ &= -\frac{(n-1)_2}{N_2} \sum_i a_i^3 \\ &= 0, \end{aligned}$$

where we used $EX_i^3 = 0$ and (16). The third summand can be treated similarly; with T_j the collection of all distinct indices $i_1, \ldots, i_j$ in $\{2, \ldots, n\}$,

$$E(\sum_{x\in\mathbf{Y}} x)^3 = E\left(\sum_{T_1} X_{i_1}^3 + 3\sum_{T_2} X_{i_1}^2 X_{i_2} + \sum_{T_3} X_{i_1} X_{i_2} X_{i_3}\right) = 0.$$

We can therefore compute and bound the variances by taking second moments. For the first term, we obtain

$$\mathrm{Var}(\sum_{x\in\mathbf{Y}} x^3) = E\left\{\left(\sum_{i=2}^n X_i^3\right)^2\right\}$$

$$
\begin{aligned}
&= E\left(\sum_{i=2}^{n} X_i^6\right) + E\left(\sum_{i\neq j}^{n} X_i^3 X_j^3\right) \\
&= \frac{n-1}{N}\langle 6\rangle + \frac{(n-1)_2}{N_2}\sum_{i\neq j} a_i^3 a_j^3 \\
&= \frac{n-1}{N}\langle 6\rangle - \frac{(n-1)_2}{N_2}\sum_{i} a_i^6 \\
&\leq \langle 6\rangle, \qquad (49)
\end{aligned}
$$

where we used (16). For the second summand,

$$
\begin{aligned}
\mathrm{Var}(\sum_{x\in\mathbf{Y}} x \sum_{x\in\mathbf{Y}} x^2) &= E\left\{\left(\sum_{i=2}^{n} X_i\right)^2 \left(\sum_{i=2}^{n} X_i^2\right)^2\right\} \\
&\leq E\left\{\left(\sum_{i=2}^{n} X_i\right)^2\right) \\
&\leq 1. \qquad (50)
\end{aligned}
$$

To bound the third summand we observe that $(\sum_{i=2}^{n} X_i)^2 \leq (n-1)\sum_i X_i^2 \leq (n-1)$. This applied twice gives

$$
\begin{aligned}
\mathrm{Var}((\sum_{x\in\mathbf{Y}} x)^3) &= E\left\{\left(\sum_{i=2}^{n} X_i\right)^6\right\} \\
&\leq (n-1)^2 E\left\{\left(\sum_{i=2}^{n} X_i\right)^2\right\} \\
&\leq (n-1)^2. \qquad (51)
\end{aligned}
$$

Lastly,

$$
\mathrm{Var}(\sum_{x\in\mathbf{Y}} x) \leq 1. \qquad (52)
$$

Therefore, we have achieved

$$
\mathrm{Var}(E\{W^* - W|W\}) \leq C_1^2(N, n, \mathcal{A}),
$$

where $C_1^2(N, n, \mathcal{A})$ is given in (25).

Continuing to apply Theorem 3.1, it remains to consider the term $E(W^* - W)^2$. From the representation of $W^* - W$ as given in (26), (27) and (28), we obtain

$$
\begin{aligned}
E(W^* - W)^2 \leq\ & 3\{E(U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1)^2 \qquad (53) \\
&+E[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]^2 \mathbf{1}(R=2) \qquad (54) \\
&+E(\hat{X}_J - X_J)^2 \mathbf{1}(R=1)\}. \qquad (55)
\end{aligned}
$$

We first consider (53). Recalling that $(\hat{X}_1', \hat{X}_1'')$ and X_1 are independent, and $EX_1 = 0$, we have

$$
E(U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1)^2 = E(U\hat{X}_1' + (1-U)\hat{X}_1'')^2 + EX_1^2.
$$

By Lemma 2.1, part (6), $X_1^* = U\hat{X}_1' + (1-U)\hat{X}_1''$ has the X_1 zero bias distribution, and so, by part (4) of that same lemma, $E(X_1^*)^2 = (1/3)\sum_{a\in\mathcal{A}} a^4$. Hence

$$E(U\hat{X}_1' + (1-U)\hat{X}_1'' - X_1)^2 \quad = \quad (1/3)\sum_{a\in\mathcal{A}} a^4 + 1/N. \tag{56}$$

Now we turn to term (54). As in (31), we have

$$\begin{aligned}
&E[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]^2\mathbf{1}(R=2) \\
&= \; E[(\hat{X}_J + \hat{X}_K) - (\hat{X}_1' + \hat{X}_1'')]^2\mathbf{1}(R=2) \\
&\leq \; 2\{E(\hat{X}_J + \hat{X}_K)^2\mathbf{1}(R=2) + E(\hat{X}_1' + \hat{X}_1'')^2\mathbf{1}(R=2)\}.
\end{aligned} \tag{57}$$

For the first term in (57), recall that given $\mathbf{X}$ and $\{R=2\}$, $(\hat{X}_J, \hat{X}_K)$ is a simple random sample of size 2 from $\mathcal{A}\setminus\{X_2,\ldots,X_n\}$. Writing x and y for realizations of $\hat{X}_1'$ and $\hat{X}_1''$, and s and t for those of $\hat{X}_J$ and $\hat{X}_K$, we obtain

$$\begin{aligned}
&E(\hat{X}_J + \hat{X}_K)^2\mathbf{1}(R=2) \\
&= \; E\left\{E\left((\hat{X}_J + \hat{X}_K)^2\mathbf{1}(R=2)|\mathbf{X}\right)\right\} \\
&= \; E\left\{\sum_{x,y\in\mathbf{Y}} q(x,y)\frac{1}{(N-n+1)_2}\sum_{\{s,t\}\subset\mathbf{Y}^c; s\neq t}(s+t)^2\right\} \\
&\leq \; \frac{2}{(N-n+1)_2}E\left\{\sum_{\{s,t\}\subset\mathbf{Y}^c; s\neq t}(s^2+t^2)\right\} \\
&= \; \frac{4(N-n)}{(N-n+1)_2}E\left\{\sum_{s\notin\mathbf{Y}} s^2\right\} \\
&= \; \frac{4}{(N-n+1)}(1 - E(\sum_{i=2}^n X_i^2)) \\
&= \; \frac{4}{(N-n+1)}(1-(n-1)/N) \\
&= \; 4/N.
\end{aligned}$$

For the second term in (57),

$$\begin{aligned}
E\{(\hat{X}_1' + \hat{X}_1'')^2\mathbf{1}(R=2)\} \; &\leq \; E\{[U\hat{X}_1' + (1-U)\hat{X}_1''] + [(1-U)\hat{X}_1' + U\hat{X}_1'']\}^2 \\
&\leq \; 2E(X_1^*)^2 \\
&= \; (2/3)\sum_{a\in\mathcal{A}} a^4.
\end{aligned}$$

Thus we obtain an upper bound on (54)

$$E[(\hat{X}_J - X_J) + (\hat{X}_K - X_K)]^2\mathbf{1}(R=2) \leq 8/N + (4/3)\sum_{a\in\mathcal{A}} a^4. \tag{58}$$

For (55), the situation is somewhat more involved.

$$\begin{aligned}
&E(\hat{X}_J - X_J)^2\mathbf{1}(R=1) \\
&= \; E\left\{E\left((\hat{X}_J - X_J)^2\mathbf{1}(R=1)|\mathbf{X}\right)\right\} \\
&\leq \; 2E\left\{E\left((\hat{X}_J^2 + X_J^2)\mathbf{1}(R=1)|\mathbf{X}\right)\right\}.
\end{aligned} \tag{59}$$

Again we let s be the possible outcomes for $\hat{X}_J$, and we let x and y denote the the possible outcomes for $\hat{X}_1'$ and $\hat{X}_1''$; $\mathbf{x}$ and $\mathbf{y}$ are as above. For the first term, we have that

$$\begin{aligned}
& E\left(\hat{X}_J^2 \mathbf{1}(R=1)|\mathbf{X}\right) \\
&= E\left(\hat{X}_J^2 \mathbf{1}(|\{\hat{X}_1', \hat{X}_1''\} \cap \mathbf{Y}| = 1, \hat{X}_J \notin \mathbf{Y} \cup \{\hat{X}_1', \hat{X}_1''\})|\mathbf{X}\right) \\
&= \frac{1}{N-n} \sum_{x \in \mathbf{Y}, y \notin \mathbf{Y}} q(x,y) \sum_{s \notin \mathbf{Y} \cup \{x,y\}} s^2 \\
&\leq \frac{1}{N-n} \sum_{s \notin \mathbf{Y}} s^2 \\
&= \frac{1}{N-n}(1 - \sum_{s \in \mathbf{Y}} s^2),
\end{aligned}$$

and taking expectation gives

$$E\left\{E\left(\hat{X}_J^2 \mathbf{1}(R=1)|\mathbf{X}\right)\right\} \leq \frac{1}{N-n}(1 - \frac{n-1}{N}) = \frac{N-n+1}{N(N-n)} \leq 2/N. \tag{60}$$

For the second term in (59) we have, using the symmetry of $(\hat{X}_1', \hat{X}_1'')$ and its independence of $\mathbf{X}$,

$$\begin{aligned}
E\left(X_J^2 \mathbf{1}(R=1)|\mathbf{X}\right) &= 2E\left((\hat{X}_1')^2 \mathbf{1}(\hat{X}_1' \in \mathbf{Y}, \hat{X}_1'' \notin \mathbf{Y})|\mathbf{X}\right) \\
&\leq 2E((\hat{X}_1')^2) \\
&= 2 \sum_{u,v \in \mathcal{A}} u^2 q(u,v) \\
&= \frac{1}{N} \sum_{u,v \in \mathcal{A}} (u^4 - 2u^3 v + u^2 v^2) \\
&= \sum_{a \in \mathcal{A}} a^4 + \frac{1}{N}.
\end{aligned}$$

Adding this result to (60), and multiplying by 2 as indicated in (59), we obtain

$$E\{(\hat{X}_J - X_J)^2 \mathbf{1}(R=1)\} \leq \frac{6}{N} + 2\sum_{a \in \mathcal{A}} a^4. \tag{61}$$

Adding (56), (58) and (61), and multiplying by 3 as indicated in (53), we obtain that

$$E(W^* - W)^2 \leq 11 \sum_{a \in \mathcal{A}} a^4 + \frac{45}{N}.$$

Bibliography

1. Baldi, P. Rinott, Y. and Stein, C. (1989). A normal approximations for the number of local maxima of a random function on a graph, In *Probability, Statistics and Mathematics, Papers in Honor of Samuel Karlin.* T. W. Anderson, K.B. Athreya and D. L. Iglehart eds., Academic Press, 59–81.

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

3. Bolthausen, E. and Götze, F. (1993). The rate of convergence for multivariate sampling statistics. *Ann. Statist.* **21**, 1692–1710.

4. Cacoullos, T., Papathanasiou, V. and Utev, S. (1994). Variational inequalities with examples and an application to the central limit theorem. *Ann. Probab.* **22**, 1607–1618.

5. Dembo, A. and Rinott, Y. (1996). Some examples of Normal approximations by Stein's method. In *Random Discrete Structures*, IMA volume 76, 25–44. Aldous, D. and Pemantle, R. Eds., Springer-Verlag.

6. Feller, W. (1971). An Introduction to Probability and Its Applications. Volume 2. 2^{nd} edition. John Wiley and Sons.

7. Folland, G. (1984) Real Analysis: Modern Techniques and Their Applications. John Wiley and Sons.

8. Goldstein, L. and Rinott, Y. (1996). On multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Prob.* **33**, 1–17.

9. Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19**, 724–739.

10. Ho, S.-T. and Chen. L.H.Y. (1978). An L_p bound for the remainder in a combinatorial central limit theorem. *Ann. Statist.* **6**, 231–249.

11. Rinott, Y. (1994). On normal approximation rates for certain sums of dependent random variables. *J. Computational and Applied Math.* **55**, 135–143.

12. Rinott, Y. and Rotar, V. (1996): A multivariate CLT for local dependence with $n^{-1/2}\log n$ rate and applications to multivariate graph related statistics. *J. Multivariate Analysis.* **56**, 333-350.

13. Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2**, 583–602. Univ. California Press, Berkeley.

14. Stein, C. (1986). *Approximate Computation of Expectations.* IMS, Hayward, California.

15. Stein, C. (1992). A way of using auxiliary randomization. In *Probability Theory,* 159–180. Walter de Gruyter, Berlin - New York.