

1. Use of exchangeable pairs in the analysis of simulations

Charles Stein¹, Persi Diaconis², Susan Holmes³, and Gesine Reinert⁴

Stanford University, INRA-Montpellier, and University of Oxford

Abstract: The method of exchangeable pairs has emerged as an important tool in proving limit theorems for Poisson, normal and other classical approximations. Here the method is used in a simulation context. We estimate transition probabilities from the simulations and use these to reduce variances. Exchangeable pairs are used as control variates.

Finally, a general approximation theorem is developed that can be complemented by simulations to provide actual estimates of approximation errors.

1.1. Introduction

A basic computational problem of the theory of probability may be formulated in the following way. Let \mathcal{X} and \mathcal{W} be two finite sets and let ω be a function on \mathcal{X} to \mathcal{W} . We know (except possibly for the normalizing factor) the distribution of a random variable X taking values in \mathcal{X} , and want to study the distribution of the random variable $W = \omega(X)$, perhaps to evaluate or approximate the expectation $\mathbf{E}f(W)$ with f a given real-valued function on \mathcal{W} . Often \mathcal{X} is a space of functions (in particular sequences or graphs) and \mathcal{W} is a subset of \mathbf{R}^P . In typical situations, \mathcal{X} is so large and complicated that direct computation of $\mathbf{E}f(W)$ is intractable. An example to keep in mind is the classical Ising model on an $N \times N \times N$ size grid. Here \mathcal{X} is the space of 2^{N^3} labelings of the grid by $\{\pm 1\}$. If $W = \omega(X)$ is the sum of all the grid labels (the so-called *magnetization*), direct or theoretical evaluation of $\mathbf{E}W$ is impossible e.g. when $N = 10$.

These problems can be studied by simulation methods such as Markov chain Monte Carlo. This paper discusses three techniques which can be used in conjunction with standard simulation procedures to get increased accuracy. The techniques are all based on creating exchangeable pairs (X, X') . These pairs give rise to classes of identities which suggest new estimators.

In Section 1.2, exchangeable pairs are introduced. The relation with reversible Markov chains is recalled. A basic identity for an exchangeable pair (W, W') , as given in Proposition 2 is :

$$\frac{p(w')}{p(w)} = \frac{p(w'|w)}{p(w|w')}.$$

¹Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA. e-mail: stein@stat.stanford.edu

²Department of Mathematics and Statistics, Stanford University, Stanford, CA 94305, USA.

³Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, USA, and Unité de Biométrie, INRA-Montpellier, France. e-mail: susan@stat.stanford.edu

⁴Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK. e-mail: reinert@stats.ox.ac.uk

This suggests that the ratios $\frac{p(w')}{p(w)}$ can be estimated by counting $w \rightarrow w'$ transitions in a sequence of pairs. In the Markov chain context this is the transition matrix Monte Carlo technique of Wang *et al.* [29]. The technique is illustrated on two examples in Section 1.3: the distribution of the number of ones in Poisson-binomial trials and the Ising model. It works well in the first example and modestly in the second example.

Section 1.4 uses exchangeable pairs (X, X') to make control variates $\mathbf{E}^X(W')$ for W . This is used to improve the naive estimate $\frac{1}{N} \sum_{i=1}^N W_i$ of $\mathbf{E}W$, obtained by N simulations of W . New estimates of $\text{Var}(W)$ are also suggested.

Section 1.5 uses exchangeable pairs to derive a closed form expression for the error of a classical approximation (e.g., normal or Poisson) for the distribution of W . The error is an explicit function of (W, W') . It can thus be estimated from a sequence of such pairs and used to correct the classical approximation. A normal example is worked through in detail. A general approximation theorem for an essentially arbitrary limit is also derived and used to suggest non-parametric alternate estimators.

Exchangeable pairs have been used to derive a class of limiting approximations via versions of “Stein’s method”. The basic ratio identities of Section 1.4 were used to derive approximations to the number of Latin rectangles (Stein [23]) and to derive combinatorial formulae for balls and boxes and cycle lengths in random permutations (Stein [27], Chapter 5). The idea is that the ratios $\frac{p(w'|w)}{p(w)}$ may be much easier to work with than $\frac{p(w')}{p(w)}$. In Section 1.4 we find versions of these ratios which are easily computible. The explicit remainder terms of Section 1.5 appear in the earliest versions of Stein’s method. In previous work, calculus and probability estimates were used to bound the remainders, giving Berry–Esséen like errors. Here the emphasis will be on applications to the output of a simulation.

1.2. Exchangeable pairs

We first define exchangeable pairs and give examples and a basic ratio identity. Then the connection with reversible Markov chains is given.

1.2.1. Definitions

An ordered pair (X, X') of random variables taking values in the finite set \mathcal{X} is defined to be *exchangeable* if, for all x_1 and x_2 in \mathcal{X} ,

$$\mathbf{P}\{X = x_1 \text{ and } X' = x_2\} = \mathbf{P}\{X = x_2 \text{ and } X' = x_1\}. \quad (1.1)$$

The *graph* of an exchangeable pair

$$(\mathcal{X}, \mathcal{G}) \quad (1.2)$$

associated with (X, X') has vertex set \mathcal{X} and edge set \mathcal{G} the set of all two element subsets $\{x_1, x_2\}$ of \mathcal{X} such that $\mathbf{P}\{X = x_1 \text{ and } X' = x_2\} > 0$. It is convenient to use the abbreviations

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ p(x_2|x_1) = p_{X'|X}(x_2|x_1) &= \mathbf{P}\{X' = x_2|X = x_1\}. \end{aligned}$$

The following two propositions will be used without comment throughout. Their proofs are immediate from the definitions.

Proposition 1.1. *Let (X, X') be an exchangeable pair taking values in a finite \mathcal{X} . Let ω be a function on \mathcal{X} to another set \mathcal{W} . Define random variables W and W' by $W = \omega(X)$ and $W' = \omega(X')$. Then (W, W') is an exchangeable pair.*

Proposition 1.2. *Let (X, X') be an exchangeable pair taking values in a finite set \mathcal{X} . Let $(\mathcal{X}, \mathcal{G})$ be the associated graph. Then, for all x_1 and x_2 with $\{x_1, x_2\}$ in \mathcal{G} ,*

$$\frac{p_X(x_2)}{p_X(x_1)} = \frac{p_{X'|X}(x_2|x_1)}{p_{X'|X}(x_1|x_2)}. \quad (1.3)$$

As a partial converse, if the associated graph $(\mathcal{X}, \mathcal{G})$ is connected and (1.3) holds for all x_1 and x_2 , then (X, X') is exchangeable.

Example 1.1 (Poisson–Binomial trials). Let X be a random function on a finite set \mathcal{S} with the collection $(X(s), s \in \mathcal{S})$ independent Bernoulli($p(s)$, $s \in \mathcal{S}$). Let S be a random element of \mathcal{S} , independent of X (not necessarily uniformly distributed) and define X' by setting $X'(s) = X(s)$ for s not equal to S but letting $X'(S)$ be distributed according to the conditional distribution of $X(S)$ given S . Then (X, X') is an exchangeable pair. The associated graph is connected if for all s , $p(s) > 0$. For this example $W = \sum_{s \in \mathcal{S}} X(s)$ is studied in Section 1.3; see also Stein [26].

Example 1.2 (Random permutations). Let X be a random permutation of $\{1, 2, \dots, n\}$, uniformly distributed. Let $X' = (I, J)X$ where the transposition (I, J) is uniformly chosen, then (X, X') is an exchangeable pair and the associated graph is connected. This exchangeable pair was used in the very first application of “Stein’s method” to prove the limiting normality in Hoeffding’s Combinatorial Limit Theorem (Stein [25], Stein [27], Chapter 3). Instead of multiplying by a random transposition, X' can be built from x by multiplying by any random permutation chosen from a symmetric probability distribution. The construction of an appropriate exchangeable pair may depend on the function w of interest; the computations are simpler if W' is close to W . See Fulman [11] for an instructive example. The idea can be used for any group. Stein [24] employed it for studying the trace of a random orthogonal matrix.

Many further examples are given in Section 1.2.2. There is a large literature on exchangeability. Informative treatments are in Kingman [16], Aldous [1], Diaconis [6]. Most of this literature deals with potentially infinite exchangeable sequences and is not relevant for present purposes.

1.2.2. Reversible Markov chains

Let \mathcal{X} be a finite set and $\pi(x)$ a probability defined on \mathcal{X} . A stochastic matrix $K(x_1, x_2)$ is *reversible* with respect to π if

$$\pi(x_1)K(x_1, x_2) = \pi(x_2)K(x_2, x_1) \quad \text{for all } x_1, x_2 \in \mathcal{X}. \quad (1.4)$$

In the physics literature Condition (1.4) is called *detailed balance*. Comparing (1.3) and (1.4) we see the following result.

Proposition 1.3. *Let π, K be respectively a probability and stochastic matrix on a finite set \mathcal{X} . Define a pair of random variables X and X' by*

$$\mathbf{P}(X = x_1, X' = x_2) = \pi(x_1)K(x_1, x_2).$$

Then (X, X') is an exchangeable pair if and only if K is reversible with respect to π .

Proposition 1.3 allows the rich variety of techniques for constructing reversible Markov chains to be adapted for constructing exchangeable pairs.

Example 1.3 (Metropolis algorithm). Let \mathcal{X} be a finite set. Suppose we are given a probability distribution $p_X(x)$ known to within a constant factor. We are also given a stochastic matrix $\alpha(x, y)$ with $\alpha(x, y) > 0$ if and only if $\alpha(y, x) > 0$. As given, the matrix α has no relation to $p_X(x)$. We can change the stationary distribution of α to $p_X(x)$ by accepting transitions from x_1 to x_2 with probability $\beta(x_1, x_2)$ and thus staying at x_1 with probability $1 - \beta(x_1, x_2)$. If (X, X') denote successive states of the new chain with X distributed as $p_X(x)$, the exchangeability condition (1.3) becomes

$$\begin{aligned} \frac{p_X(x_2)}{p_X(x_1)} &= \frac{p_{X'|X}(x_2|x_1)}{p_{X'|X}(x_1|x_2)} \\ &= \frac{\alpha(x_1, x_2)\beta(x_1, x_2)}{\alpha(x_2, x_1)\beta(x_2, x_1)}. \end{aligned} \quad (1.5)$$

This condition can be satisfied in many ways, but most conveniently by

$$\beta(x_1, x_2) = \min\left(\frac{p_X(x_2)\alpha(x_2, x_1)}{p_X(x_1)\alpha(x_1, x_2)}, 1\right). \quad (1.6)$$

The Metropolis algorithm originated as a device for sampling from a stationary distribution p_X known to within a constant factor. The exchangeable pair constructed above gives a p_X -reversible Markov chain

$$K(x_1, x_2) = p_{X'|X}(x_2|x_1). \quad (1.7)$$

For history and a literature review on the Metropolis [18] algorithm see Billera and Diaconis [4]. A large collection of algebraic techniques for constructing reversible Markov chains for problems such as contingency tables with fixed row and column sums appears in Diaconis and Sturmfels [7].

Rinott and Rotar [21] have used the connection between exchangeable pairs and reversible Markov chains in their work on normal approximation. Of course, techniques like the Gibbs sampler (also known as the heat-bath algorithm) can be similarly used. Fishman [9] and Liu [17] give current accounts of a host of other methods for constructing reversible Markov chains.

In the following sections we will suggest running the associated Markov chains as a way of estimating probabilities $p_X(x)$ via the ratio identity (1.3) in Proposition 1.2. Then, convergence issues become important. We will not try to summarize the developing literature. See Aldous and Fill [2], Fishman [9], Liu [17] or Newman and Barkema [19].

To conclude this section, we call attention to two widely used techniques of computational statistical mechanics which seem seldom employed by statisticians. The first is a method for dealing with large holding probabilities for algorithms such as the Metropolis algorithm. For some problems the holding probability can be explicitly computed. The current state can be weighted by the inverse of the holding probability and a different state can be chosen. This is explained as “continuous time Monte Carlo” in Newman and Barkema [19], Section 2.4. An example is in Section 3.2 below. Here is a brief description.

Instead of spending a large proportion of time holding at some state, we can change the Markov chain to another one, that never holds by redistributing the diagonal probability among the other states.

In more detail, we define a new Markov chain

$$\tilde{K}(x, x') = \begin{cases} \frac{K(x, x')}{1 - K(x, x')} & \text{if } x' \neq x \\ 0 & \text{otherwise} \end{cases}$$

This new chain is reversible with regards to the unnormalized weight $\pi(x)(1 - K(x, x))$:

$$\begin{aligned} \pi(x)(1 - K(x, x)) \frac{K(x, x')}{1 - K(x, x')} &= \pi(x)K(x, x') = \pi(x')K(x', x) \\ &= \pi(x')(1 - K(x', x')) \frac{K(x', x)}{1 - K(x', x)}. \end{aligned}$$

If we run the original chain X_1, X_2, \dots, X_R and estimate $\int f d\pi$ by $\frac{1}{R} \sum_{i=1}^R f(X_i)$ we count each X_i that holds J times with weight $1 + J$. If $X_i = x$, then let J_x be the holding time at x . If the probability of holding at x is denoted by $h(x)$, then $P(J_x = J) = h(x)^J(1 - h(x))$ and

$$E(J_x) = \frac{h(x)}{1 - h(x)} \quad \text{and} \quad E(1 + J_x) = \frac{1}{1 - h(x)}.$$

Thus, if $\bar{x}_0, \bar{x}_1, \dots, \bar{x}_R$ is the realization of the \tilde{K} chain, the appropriate estimator is

$$\frac{1}{R} \sum_{i=1}^R \frac{f(\bar{x}_i)}{1 - h(\bar{x}_i)}. \quad (1.8)$$

The second idea is a method of estimating expected values under a range of parameter values from simulation at one (or a few) parameter values. The rough idea is to use exponential tilting to reweight the samples. For this to work, the original samples must be chosen from a broad distribution to avoid uncovered parts of the space. These ideas are explained as entropic sampling methods (Section 6.3) and flat histogram methods (Sections 8.1, 8.2) in Newman and Barkema [19]. Wang *et al.* [29] is a recent extension. An example is in Section 3.2 below.

For both techniques, the computational effort can be considerably diminished by maintaining an additional book-keeping array along with the current state X . For example, the book-keeping array for the 2-dimensional Ising model is the number of $+$ vertices with a given neighborhood pattern, and the number of $-$ vertices with a given neighborhood pattern.

1.3. First examples

This section sets out the basic machinery of transition matrix Monte Carlo. Two examples are considered in 1.3.2: the number of ones in Poisson–Binomial trials is studied, while the most straightforward application of exchangeable pairs offer little improvement, eliminating, holding and tilting give large gains over naive Monte Carlo. In 1.3.3, transition rate Monte Carlo for a variety of Ising model simulations are summarized.

1.3.1. Transition matrix Monte Carlo

Consider the simulation problem described in the Introduction. Consider $X_1, X_2, X_3, \dots, X_N$ with X_i distributed as $p_X(x)$. The joint distribution of the X_i may be arbitrary, for example independent and identically distributed or the realization of a Markov chain. The naive estimate of $\mathbf{E}f(W)$ is

$$\frac{1}{N} \sum_{i=1}^N f(\omega(X_i)). \quad (1.9)$$

Suppose we construct an exchangeable pair (X, X') as described in Section 1.2 above and can calculate $\mathbf{P}^X(W' = w)$ with $W' = \omega(X')$. Then as an estimate of $p_{W'|W}(w_2|w_1)$, abbreviated by $p(w_2|w_1)$, we can use

$$\widehat{p(w_2|w_1)} = \frac{\sum_{i=1}^N \delta_{W_i=w_1} \mathbf{P}^{X_i}(W'_i = w_2)}{\sum_{i=1}^N \delta_{W_i=w_1}}. \quad (1.10)$$

Then, for all w_1 and w_2 for which both $\hat{p}(w_2|w_1)$ and $\hat{p}(w_1|w_2)$ are positive we estimate the ratio $\frac{\mathbf{P}(W=w_2)}{\mathbf{P}(W=w_1)}$ by

$$\frac{\hat{p}(w_2|w_1)}{\hat{p}(w_1|w_2)}.$$

From these ratio estimates all ratios of all probabilities, and so all probabilities, can be estimated, provided the sample is large enough for the connectedness of the graph (1.2) to be reflected in the sample. We assume throughout that the graph of the exchangeable pair is connected

To go from ratios to probabilities, form a matrix with rows and columns indexed by \mathcal{W} having (w, w') entry

$$\frac{\hat{p}(w'|w)}{\hat{p}(w|w')}.$$

In applications, this is often a sparse matrix. For example, for W a birth and death chain, the matrix is tridiagonal. For (w, w') with zero entry in the matrix there may be many paths in the graph giving estimates of $\frac{p(w')}{p(w)}$.

Fitzgerald *et al.* [10] have suggested reconciling these various estimates by least squares. Treat $p(w)$ as parameters in

$$\frac{p(w)}{p(w')} = \frac{\hat{p}(w|w')}{\hat{p}(w'|w)}.$$

Take logarithms on both sides

$$\ell(w) - \ell(w') = \ell(w|w') - \ell(w'|w)$$

and solve for $\ell(w)$ by minimizing

$$\sum (\ell(w) - \ell(w') - \ell(w|w') + \ell(w'|w))^2$$

with the sum over pairs (w, w') with $\hat{p}(w|w')\hat{p}(w'|w) \neq 0$.

A more careful reconciliation of different estimators is complicated by correlation and inhomogeneity of variances.

| W | Ordinary MC | Ratio | Truth |
|------------|-------------|-----------|-------------|
| 0 | 0.08350 | 0.0829528 | 0.0909091 |
| 1 | 0.2607 | 0.2605 | 0.266270 |
| 2 | 0.3176 | 0.3180 | 0.319504 |
| 3 | 0.2064 | 0.2066 | 0.210676 |
| 4 | 0.0956 | 0.0957 | 0.0856013 |
| 5 | 0.0304 | 0.0304 | 0.0225984 |
| 6 | 0.0052 | 0.0052 | 0.00395255 |
| 7 | 0.0006 | 0.0006 | 0.000454696 |
| Total Var. | 0.0888 | 0.0868 | 0 |

Table 1: Table for $d = 10$, $N=10,000$

| W | Ordinary MC | Ratio | Truth |
|------------|-------------|----------|-----------|
| 0 | 0.04720 | 0.046644 | 0.06250 |
| 1 | 0.2128 | 0.2129 | 0.2074 |
| 2 | 0.3220 | 0.3222 | 0.2947 |
| 3 | 0.2289 | 0.2290 | 0.2417 |
| 4 | 0.1245 | 0.1246 | 0.129372 |
| 5 | 0.04670 | 0.04672 | 0.04826 |
| 6 | 0.01530 | 0.01531 | 0.01304 |
| 7 | 0.00250 | 0.00250 | 0.00261 |
| 8 | 0.00010 | 0.00010 | 0.0003923 |
| Total Var. | 0.0700 | 0.0706 | 0 |

Table 2: Table for $d = 15$, $N = 10,000$

A version of this idea was applied by Wang *et al.* [29] who implemented it for the Ising model with substantial success. They chose X_1, \dots, X_N from the Metropolis algorithm and used the proportion of (w_1, w_2) transitions to estimate $p(w_2|w_1)$. A clear exposition with variations close to (1.10) is given by Fitzgerald *et al.* [10]. Some of their numerical results are described in Section 1.3.3 below.

1.3.2. A Poisson–Binomial example

Let \mathcal{X} be the space of binary d -tuples $x = (x_{(1)}, \dots, x_{(d)})$. Fix $\theta_i, 1 \leq i \leq d$ with $0 < \theta_i < 1$. In our numerical illustrations below $\theta_i = \frac{1}{i+1}$. Let $\mathcal{W} = \{0, 1, \dots, d\}$ and $\omega(X) = W = \sum_{i=1}^d X_{(i)}$ with $X_{(i)} \sim Be(\theta_i), i = 1, \dots, d$. We form X_0, X_1, \dots by running a reversible Markov chain on \mathcal{X} . This proceeds by choosing a coordinate I uniformly in $1 \leq i \leq d$ and replacing the I^{th} coordinate of the current vector by an independent binary random variable with chance of success θ_I . The chain is started in stationarity. Tables 1, 2, 3 show results of a small trial for $d = 10, 15, 18$.

Remarks. We do not see any difference between the transition matrix approach and naive Monte Carlo. Neither approach reached points in the extreme tails of the distribution and for the bulk of the distribution they seem equivalent. Since this ratio estimator is computationally costly, there is not much to recommend it here.

| W | Ordinary MC | Ratio | Truth |
|------------|-------------|----------|-----------|
| 0 | 0.06590 | 0.06489 | 0.05263 |
| 1 | 0.18780 | 0.18698 | 0.18395 |
| 2 | 0.29730 | 0.29713 | 0.27960 |
| 3 | 0.24110 | 0.24168 | 0.24926 |
| 4 | 0.12810 | 0.12898 | 0.14757 |
| 5 | 0.055700 | 0.056082 | 0.06208 |
| 6 | 0.018600 | 0.018728 | 0.01934 |
| 7 | 0.004300 | 0.00433 | 0.00459 |
| 8 | 0.001100 | 0.001107 | 0.0008419 |
| 9 | 0.000100 | 0.000101 | 0.000121 |
| Total Var. | 0.0701 | 0.0661 | 0 |

Table 3: Table for $d = 18$, $N = 10,000$

We next compare the transition Monte Carlo approach with Naive Monte Carlo for the chain run without holding. Call this chain Y_0, Y_1, Y_2, \dots , following (1.8) above we have

$$P(Y' = y' | Y = y) = \frac{P(X' = y' | X = y)}{1 - h(y)}. \quad (1.11)$$

In our example:

$$\begin{aligned} h(y) &= \sum_{\ell: y_\ell=1} \frac{\theta_\ell}{d} + \sum_{j: y_j=0} \frac{1 - \theta_j}{d} \\ &= \frac{1}{d} (\Theta - 2\gamma(y) + n - \omega(y)) \\ &\text{where } \Theta = \sum_{i=1}^d \theta_i \quad \text{and} \quad \gamma(y) = \sum_{j: y_j=0} \theta_j. \end{aligned}$$

To describe the complete procedure, choose a binary vector Y_0 by flipping coins with probability of success θ_i , $1 \leq i \leq d$. The process updates each time according to the following rules giving Y_1, Y_2, \dots . Let $\omega(Y_i) = W_i$ be the sum of elements in Y_i .

- With probability $P_{up}(y) = \frac{\gamma(y)}{d(1-h(y))}$ the chain goes up and an index j at which y_j is zero is turned into a 1, j is chosen with probabilities $\frac{\theta_j}{\gamma(y)}$.
- With probability $P_{down}(y) = 1 - \frac{\gamma(y)}{d(1-h(y))}$ the chain goes down and an index ℓ at which y_ℓ is one is turned into a zero, ℓ is chosen with probabilities

$$\frac{1 - \theta(\ell)}{d(1 - h(y))} \frac{1}{P_{down}(y)} = \frac{1 - \theta(\ell)}{\omega(y) - \Theta + \gamma(y)}.$$

This construction satisfies (1.11).

Remark. Instead of going up or down, we can also directly choose the index of Y to change by choosing index i with probability

$$\frac{\theta_i^{(1-y_i)} (1 - \theta_i)^{(y_i)}}{d(1 - h(y))}.$$

At each time τ record the probability $P_{up}(Y(\tau), \tau)$ of going up if $Y_\tau = y(\tau)$ is observed, and the holding times $\beta(y(\tau)) = 1/(1 - h(y(\tau)))$. To simplify notation, we write $P_{up}(\tau)$ for $P_{up}(Y(\tau), \tau)$, $\beta(\tau)$ for $\beta(y(\tau))$, $h(\tau)$ for $h(y(\tau))$ and $\omega(\tau)$ for $\omega(y(\tau))$.

We observe Y_1, \dots, Y_N . At the end of the run the naive estimate (incorporating a speedup without holding) is

$$\hat{p}(w) = \frac{\sum_{\tau \in \{1, \dots, N\}: \omega(\tau)=w} \beta(\tau)}{\sum_{\tau=1}^N \beta(\tau)}. \quad (1.12)$$

The ratio estimators are

$$\begin{aligned} \hat{p}(w-1|w) &= \frac{\sum_{\tau: \omega(\tau)=w} P_{down}(\tau)}{\sum_{\tau: \omega(\tau)=w} \beta(\tau)} \\ \hat{p}(w+1|w) &= \frac{\sum_{\tau: \omega(\tau)=w} P_{up}(\tau)}{\sum_{\tau: \omega(\tau)=w} \beta(\tau)}. \end{aligned}$$

Then our estimator is built from the ratios:

$$\hat{\rho}(w) = \frac{\widehat{p(w)}}{p(w-1)} = \frac{\hat{p}(w|w-1)}{\hat{p}(w-1|w)}$$

together with $\sum_{w=0}^d \hat{p}(w) = 1$ to obtain $\hat{p}(w)$. Specifically, write $\hat{p}_1(0) = c$, $\hat{p}_1(j) = \hat{\rho}(j) \times \hat{p}_1(j-1)$ and then

$$\hat{p}(j) = \frac{\hat{p}_1(j)}{\sum_{i=0}^l \hat{p}_1(i)}. \quad (1.13)$$

Simulation results are given in Tables 4 and 5. We see a marked improvement:

- First, eliminating holding gives an improvement of about a factor of 3 (compare the first columns of Tables 1 and 4).
- Second, the transition matrix approach gives improvements of an order of 10 (compare the first two columns of Table 4 or the first two columns of Table 5).

As a third variation, we employ the flat histogram method outlined at the end of Section 2. In Table 1 above $p(10) = \mathbf{P}(W = 10) \doteq 2.50521 \times 10^{-8}$. It is not surprising that there were no Monte Carlo trials with ten successes. One way of investigating the tails is to sample from X^* where

$$\mathbf{P}(X^* = x) = Z^{-1} \eta(\omega(x)) \mathbf{P}(X = x)$$

with a known weight function $\eta(\omega)$, chosen to tilt the distribution to large values of ω . A natural choice is $\eta(\omega)$ proportional to the reciprocals of conjectured values of $\mathbf{P}(W = \omega)$. In the example to follow, $\eta(\omega)$ was taken as the inverse of $Pois_\lambda(w)$ with λ the mean of W . The Metropolis algorithm was used to sample from the distribution of X^* . The probability that $W^* = w$ was estimated by the ratio method. Then these numbers were multiplied by $\eta(\omega)$ and renormalized to sum to one.

As an example, for $d = 10$ with $\lambda = 2.5$, a Markov chain of length $N = 10^4$ produced the values given in Table 6.

Comparing with the true values, there is a big improvement in the estimates of the upper tail values. The sum of absolute errors is .00346312. This shows some deterioration. Perhaps a compromise can be used to reduce this effect. Very similar improvements were observed in trials with $d = 20$ (e.g. $\mathbf{P}(S_{20} = 20) = 1.95729 \times 10^{-20}$, $\hat{p}(20) = 6.32623 \times 10^{-21}$, $\hat{\rho}(20) = 1.13414 \times 10^{-20}$ based on 10^4 trials).

| W | No-hold MC | Ratio | Truth |
|------------|-------------|-------------|------------------|
| 0 | 0.089593 | 0.090624 | 0.090909 |
| 1 | 0.26896 | 0.26621 | 0.26627 |
| 2 | 0.31977 | 0.32032 | 0.31950 |
| 3 | 0.20793 | 0.21047 | 0.21068 |
| 4 | 0.086466 | 0.085350 | 0.085601 |
| 5 | 0.023173 | 0.022639 | 0.022598 |
| 6 | 0.0037734 | 0.0039319 | 0.0039525 |
| 7 | 0.00034131 | 0.00045885 | 0.00045470 |
| 8 | . | . | 0.00003306878307 |
| 9 | . | . | 0.00000137786596 |
| 10 | . | . | 0.00000002505211 |
| Total Var. | 0.013190217 | 0.001309314 | 0 |

Table 4: Table for $d = 10$, $N = 10,000$

| W | No-hold MC | Hold-Ratio | Truth |
|------------|--------------|--------------|------------|
| 0 | 0.055758 | 0.053261 | 0.052632 |
| 1.0 | 0.17837 | 0.18409 | 0.18395 |
| 2.0 | 0.27270 | 0.27890 | 0.27960 |
| 3.0 | 0.24988 | 0.24883 | 0.24926 |
| 4.0 | 0.15004 | 0.14743 | 0.14757 |
| 5.0 | 0.066245 | 0.062305 | 0.062078 |
| 6.0 | 0.021428 | 0.019493 | 0.019344 |
| 7.0 | 0.0047119 | 0.0046758 | 0.0045865 |
| 8.0 | 0.00082380 | 0.00089075 | 0.00084194 |
| 9.0 | 0.000044097 | 0.00012792 | 0.00012093 |
| Total Var. | 0.0294353652 | 0.0018954746 | 0 |

Table 5: Table for $d = 18$, $N=10,000$

| j | 0 | 1 | 2 | 3 | 4 | 5 |
|--------------------|-------|-------|-------|-------|-------|-------|
| $p(j)$ | .0909 | .2663 | .3195 | .2107 | .0856 | .0226 |
| $\hat{p}(j)$ | .0867 | .2637 | .3217 | .2110 | .0893 | .0238 |
| $\hat{\hat{p}}(j)$ | .0913 | .2673 | .3298 | .2196 | .0847 | .0220 |

| j | 6 | 7 | 8 | 9 | 10 |
|--------------------|-------|-----------|-----------|--------------------------|--------------------------|
| $p(j)$ | .0040 | .00045470 | .00003307 | 1.37787×10^{-6} | 2.50521×10^{-8} |
| $\hat{p}(j)$ | .0043 | .00053721 | .00004043 | 1.5738×10^{-6} | 3.14043×10^{-8} |
| $\hat{\hat{p}}(j)$ | .0038 | .00043782 | .00003184 | 1.27737×10^{-6} | 2.37551×10^{-8} |

Table 6: Comparison of estimates in Poisson-Binomial case, $d = 10$, $\lambda = 2.5$

1.3.3. Another example: The Ising model

The Ising model may well be the most thoroughly studied object of theoretical physics. A huge number of techniques have been invented for simulation and analysis. Because of this, it makes a good testing ground for new ideas. Here we set out the basic approach of exchangeable pairs. Closely related ideas have been previously developed (Wang *et al.* [29], Fitzgerald *et al.* [10]) and we give a brief report of these simulation results.

Let $(\mathcal{V}, \mathcal{G})$ be a regular graph of degree $d > 0$. Let m be the number of elements in the vertex set \mathcal{V} . In the examples below, the graph is an n by n square lattice on a torus with $d = 4, m = n^2$. Let X be a random function on \mathcal{V} to the two-point set $\{-1, 1\}$, uniformly distributed. Let $H_1 = \sum_v X_v, W = \sum_{v_1, v_2} X_{v_1} X_{v_2}$, where the first sum is over all elements of \mathcal{V} and the second sum is over all edges $\{v_1, v_2\}$ of \mathcal{G} . We are interested primarily in the case where m is large. We want to study the joint distribution of H_1 and W or, equivalently, their moment generating function

$$Z(\lambda, \nu) = \mathbf{E}e^{\lambda W + \mu H_1}.$$

Physicists call Z the *partition function* and study its various logarithmic derivatives and other related functions. For simplicity we study the special case $Z(\lambda, 0)$ which gives the distribution of W alone. We focus on estimating the logarithmic derivative of $Z(\lambda, 0)$ at a particular value of λ . This is called the *energy* in the physics literature.

Let (X, X') be an exchangeable pair obtained from X by setting X' equal to the result of changing the sign of X_V where V is uniformly distributed in \mathcal{V} independent of X . Let W' be related to X' as W is to X . Our aim is to study the transition probabilities

$$\mathbf{P}\{W' = w_2 | W = w_1\} \tag{1.14}$$

from which the pointwise distribution of W can be reconstructed. The analysis will be based on the exchangeable pair described above. Note that the Markov chain used to simulate realizations may be very different from the single site dynamics which underly our exchangeable pair. Thus the Markov chain may be generated by the Swendsen-Wang algorithm or, in the case of a bipartite graph $(\mathcal{V}, \mathcal{G})$ by an alternating (checkerboard) algorithm. To compute an estimate of (1.14) consider the random variables

$$\begin{aligned} Y_v &= \sum_{v': (v, v') \text{ neighbors}} X_{v'} \delta_{v, v'}(\mathcal{G}) \\ W &= \frac{1}{2} \sum_v X_v Y_v. \end{aligned}$$

Then

$$W' - W = \omega(X') - \omega(X) = -X_v Y_v.$$

Thus the conditional distribution of $W' - W$ given X is given by $\mathbf{P}^X\{W' - W = d\} = \frac{s(d, x)}{m}$, where $s(d, x) = |\{v : X_v Y_v = -d\}|$. This gives the needed ingredients to take the output of a Markov chain X_0^*, X_1^*, \dots , where

$$\mathbf{P}\{X_i^* = x\} = Z^{-1}(\lambda, 0) e^{\lambda \omega(x)} \mathbf{P}(X = x).$$

Then, the procedure outlined in Sections 1.2.1, 1.2.2 can be used. This first derives estimates of ratios in (1.14) and then of $\mathbf{P}(W = w)$. These may be used to estimate $\frac{Z'}{Z}$ by

$$\frac{\sum_w w e^{\lambda w} \hat{p}_W(w)}{\sum_w e^{\lambda w} \hat{p}_W(w)}.$$

(Here Z' denotes the derivative.)

A version of this approach has been implemented by Fitzgerald *et al.* [10]. They carried out a large simulation to assess the improvement in mean-square error due to their version of the transition density method. They studied the expected value of H_1^2 (*magnetic susceptibility*) when $\lambda = .42$ and $\mu = 0$. This is just slightly above the critical temperature. Their Markov chain was the result of a single sweep through the 900 sites. In this case the true expectation is known. They chose $N = 5 \times 10^6$ sweeps and repeated the entire run 500 times. They calculated the average error for $t = 1, \dots, 5 \times 10^6$. They found relatively smooth decrease of the mean-squared error in t . The transition density method improved mean-squared error over the naive estimator by about 25%.

They carried out a similar experiment for another functional (specific heat) and found an improvement of about 7%.

Fitzgerald *et al.* [10] report a more naive method of estimating $p(w'|w)$ based on counting the proportion of w to w' transitions in a chain generated by single site updates showed *no* improvement over the naive estimator. We hope to try adjusting for holding times in later work.

1.4. Exchangeable pairs as auxiliary variates

This section develops the use of the exchangeable pairs (X, X') and (W, W') constructed in Section 1.2 for estimating the mean $\xi = \mathbf{E}W$ and variance $\sigma^2 = \mathbf{E}(W - \xi)^2$. The idea is to use $\mathbf{E}^W(W' - W)$ as an auxiliary variate combining it with observed values of W by linear regression, making use of negative correlation. Because these estimates (especially that of the variance) are motivated by pretending that the joint distribution of (W, W') is normal, they cannot be expected to work well in all situations, but they are not strongly dependent on the assumption of normality. Estimates of mean and variance are needed to apply the more refined developments of later sections.

Techniques for combining estimates to reduce the variance are known variously as *control variates*, *antithetic variates*, or *regression methods*. They are discussed and illustrated in the books of Hammersley and Hanscomb [13] or Fishman [9]. We have not found the exact suggestions below in previous literature.

Section 1.4.1 sets out the needed formulae.

1.4.1. Basic formulae

As usual, we have an exchangeable pair (X, X') of random variables taking values in a finite set \mathcal{X} . We want to estimate the mean ξ and variance σ^2 of $W = \omega(X)$ where ω is a real-valued function on \mathcal{X} . We have available the results of a simulation X_1, X_2, \dots, X_S which is marginally distributed as X . To implement the techniques of this section we must be able to compute or approximate

$$D_{1,i} = \mathbf{E}^{X_i}(W'_i - W_i) \quad \text{and} \quad D_{2,i} = \mathbf{E}^{X_i}(W'_i - W_i)^2. \quad (1.15)$$

As will be seen below, $D_{1,i}$ is negatively correlated with W_i . It is natural to seek a linear combination which has smaller variance than the naive estimator

$$\bar{W} = \frac{1}{S} \sum_{i=1}^S W_i. \quad (1.16)$$

This will be done using classical regression to estimate the best linear combination from the data. Using identities for exchangeable pairs we can also give a natural estimate for the variance. We first describe our estimators and then give their derivation.

Let

$$\bar{D}_1 = \frac{1}{S} \sum_{i=1}^S D_{1,i} \quad \text{and} \quad \bar{D}_2 = \frac{1}{S} \sum_{i=1}^S D_{2,i}. \quad (1.17)$$

An estimate $\hat{\xi}$ for $\xi = \mathbf{E}W$ is

$$\hat{\xi} = \bar{W} + \hat{a}\bar{D}_1, \quad \text{with } \hat{a} = -\frac{\sum_{i=1}^S (W_i - \bar{W})(D_{1,i} - \bar{D}_1)}{\sum_{i=1}^S (D_{1,i} - \bar{D}_1)^2}. \quad (1.18)$$

An estimate $\hat{\sigma}^2$ for $\sigma^2 = \text{Var}W$ is

$$\hat{\sigma}^2 = -\frac{1}{2S} \frac{(\sum_{i=1}^S D_{2,i})(\sum_{i=1}^S (W_i - \bar{W})^2)}{\sum_{i=1}^S (W_i - \bar{W})(D_{1,i} - \bar{D}_1)}. \quad (1.19)$$

To begin, let us show that W and $\mathbf{E}^X(W' - W)$ are negatively correlated. For this assume without loss of generality that the mean $\xi = 0$. First, $(W' + W)(W' - W)$ is an antisymmetric function of (W, W') , so that $\mathbf{E}(W' + W)(W' - W) = 0 = \mathbf{E}W'^2 - \mathbf{E}W^2$. Thus $\mathbf{E}W'^2 = \mathbf{E}W^2$. Then

$$\begin{aligned} \mathbf{E}(W' - W)^2 &= \mathbf{E}(W')^2 + \mathbf{E}W^2 - 2\mathbf{E}WW' \\ &= 2\mathbf{E}W^2 - 2\mathbf{E}WW' = -2\mathbf{E}(W(W' - W)) \\ &= -2\mathbf{E}(W\mathbf{E}^X(W' - W)). \end{aligned}$$

It follows that $\mathbf{E}(W\mathbf{E}^X(W' - W)) \leq 0$, with strict inequality unless $W = W'$.

To motivate the estimate $\hat{\xi}$ of (1.18) observe that both \bar{W} and $\bar{W} + \bar{D}_1$ are unbiased estimates of $\xi = \mathbf{E}W$. It is reasonable to estimate ξ by a linear combination of these with coefficients adding to 1 determined from the data in the same way as a regression coefficient. This leads to

$$\hat{\xi}_a = \hat{a}(\bar{W} + \bar{D}_1) + (1 - \hat{a})\bar{W} = \bar{W} + \hat{a}\bar{D}_1,$$

with \hat{a} given in (1.18).

This is related to the problem of finding the best linear predictor of W using $\mathbf{E}^X(W' - W)$. Indeed, writing

$$W = \xi + a\mathbf{E}^X(W' - W) + R \quad (1.20)$$

with $\mathbf{E}R = 0$, $\mathbf{E}RW = 0$, the coefficient yielding the smallest variance between observed and predicted is

$$a = \frac{\text{Cov}(W, \mathbf{E}^X(W' - W))}{\text{Var}(\mathbf{E}^X(W' - W))}.$$

Estimating a leads to (1.18). Note that estimating

$$\tilde{\xi} = W - a\mathbf{E}^{\mathbf{X}}(W' - W), \quad (1.21)$$

we obtain

$$\text{Var}(\tilde{\xi}) = \text{Var}W(1 - \text{Corr}^2(W, \mathbf{E}^{\mathbf{X}}(W' - W))).$$

Note that this quantity is smaller than $\text{Var}W$, and thus improves on the standard estimate of estimating ξ by W .

To understand this approach better, we now focus on the *perfect case*. Suppose we have an exchangeable pair (W, W') and a constant λ , $0 < \lambda < 1$, such that

$$\mathbf{E}^W(W' - W) = -\lambda(W - \xi). \quad (1.22)$$

There are many examples when (1.22) is satisfied, see [27]. Because $w' - w$ is an antisymmetric function in (w, w') we have

$$\mathbf{E}\mathbf{E}^W(W' - W) = 0 = -\lambda\mathbf{E}(W - \xi),$$

yielding $\xi = \mathbf{E}W$. Note that ξ can also be written as

$$\xi = W + \frac{1}{\lambda}\mathbf{E}^W(W' - W). \quad (1.23)$$

We see this as the sum of two antithetic random variables because

$$\mathbf{E}(W' - W)^2 = -2\mathbf{E}W\mathbf{E}^W(W' - W),$$

thus $\mathbf{E}W\mathbf{E}^W(W' - W) < 0$, so W and $\mathbf{E}^W(W' - W)$ are negatively correlated. Under (1.22), we have

$$\mathbf{E}(W' - W)^2 = -2\lambda\mathbf{E}W(W - \xi) = -2\lambda(\mathbf{E}W^2 - \xi^2) = -2\lambda\text{Var}W,$$

so that the two components have covariance

$$\text{Cov}\left(W, \frac{1}{\lambda}\mathbf{E}^W(W' - W)\right) = \frac{1}{\lambda}\mathbf{E}W\mathbf{E}^W(W' - W) = -\text{Var}W.$$

We also remark that given (1.22) we know that

$$\lambda = \frac{1}{2} \frac{\mathbf{E}(W' - W)^2}{\text{Var}W}.$$

We estimate λ using the regression approach :

$$\hat{\lambda} = \frac{\sum_i (D_{1,i} - \bar{D}_1)(W_i - \bar{W})}{\sum_i (W_i - \bar{W})^2}$$

and

$$\begin{aligned} \hat{\sigma}^2 &= \frac{E(\widehat{W' - W})^2}{2\hat{\lambda}} \\ 2\hat{\lambda} &= -\frac{\sum_i D_{2,i} \sum_i' (W_i' - \bar{W})^2}{2S \sum_i (D_{1,i} - \bar{D}_1)(W_i - \bar{W})}. \end{aligned}$$

This leads to (1.19).

Approximate case

Suppose now that

$$\mathbf{E}^W(W' - W) = -\lambda(W - \xi) + R. \quad (1.24)$$

Here, (1.24) and exchangeability imply that if $\mathbf{E}W = \xi$ then $\mathbf{E}R = 0$ and conversely if $\mathbf{E}R = 0$ then $\mathbf{E}W = \xi$.

If we want to estimate ξ we can write

$$\xi = W + \frac{1}{\lambda}\mathbf{E}^W(W' - W) - \frac{1}{\lambda}R.$$

The right hand side leads to the antithetic variables $W - \frac{1}{\lambda}R$ and $\frac{1}{\lambda}\mathbf{E}^W(W' - W)$:

$$\begin{aligned} \text{Cov}\left(W - \frac{1}{\lambda}R, \frac{1}{\lambda}\mathbf{E}^W(W' - W)\right) &= \mathbf{E}\left(W - \frac{1}{\lambda}R - \xi, \frac{1}{\lambda}\mathbf{E}^W(W' - W)\right) \\ &= -\mathbf{E}\left[\left(W - \frac{1}{\lambda}R - \xi\right)\left(W - \frac{1}{\lambda}R - \xi\right)\right] \\ &= -\text{Var}\left(W - \frac{1}{\lambda}R\right) < 0. \end{aligned}$$

As to the estimate of variance; if R is small, it can be effectively neglected and calculations for the perfect case above are in force; yet a further justification for $\hat{\sigma}$ is given next.

As a regression problem

Write $\hat{\xi} = W - \beta(\mathbf{E}^W W' - W)$, this is an unbiased estimate of ξ . For all β to minimize its variance:

$$\begin{aligned} \text{Var}(\hat{\xi}) &= \text{Var}W - 2\beta\text{Cov}(W, \mathbf{E}^W W' - W) + \beta^2\text{Var}(\mathbf{E}^W W' - W) \\ \text{Choose } \beta &= \frac{\text{Cov}(W, \mathbf{E}^W W' - W)}{\text{Var}(\mathbf{E}^W W' - W)} \end{aligned}$$

In fact, with our perfect case notation

$$\lambda = -\frac{1}{\beta} = -\frac{\text{Var}\mathbf{E}^W W' - W}{\text{Cov}(W, \mathbf{E}^W W' - W)}.$$

This can be estimated by:

$$\hat{\lambda} = -\frac{\sum_i (D_{i,1} - D_1)^2}{\sum (W_i - \bar{W})(D_{i,1} - D_1)}.$$

Another extension is the following. To simplify we have been conditioning on the values of $W_i = \omega(X_i)$. It is also possible to rewrite all the above conditioning on the larger state X_i ; this is what is suggested in practice.

1.5. Distributional approximations

The basic theorem of this section is an identity which provides an explicit expression for the error of an approximation to the distribution of a real random variable by a continuous distribution coming from a rather large class, which contains the normal distribution as well as the uniform distribution, for example. A corollary provides simple bounds for the error of the normal approximation to the expectation of a smooth function, as can be found in Stein [27]. This same idea has been applied by many people to obtain bounds of Berry–Esséen type for the error of the normal approximation. In this section the aim is to explore the possible application of this idea to the analysis of simulations. As in Stein [27], Chapter 6, we first derive a characterization for a continuous distribution. This is obtained essentially by integration by parts. Let $I = [a, b]$ be a real interval, where $-\infty \leq a < b \leq \infty$. For abbreviation, we call a real function f on I *regular* if f is finite on I and, at any interior point of I , f possesses a right-hand limit and a left-hand limit. Further, f possesses a right-hand limit $f(a+)$ at the point a and a left-hand limit $f(b-)$ at the point b . Thus the set of discontinuity points of f is countable.

Proposition 1.4. *Let p be a regular, strictly positive density on an interval $I = [a, b]$, where $-\infty \leq a < b \leq \infty$. Suppose p has a derivative p' that is regular on I , having only countably many sign changes and being continuous at the sign changes. Suppose*

$$\int_I p(x) |\ln(p(x))| dx < \infty. \quad (1.25)$$

Let

$$\psi(x) = \frac{p'(x)}{p(x)}, \quad (1.26)$$

and suppose that ψ is regular. Let \mathcal{F} be the class of all regular functions on I possessing (piecewise) a regular derivative on I such that

$$\int_I |f'(x)| p(x) dx < \infty \quad (1.27)$$

$$\int_I |f(x)\psi(x)| p(x) dx < \infty. \quad (1.28)$$

Then, in order that a random variable Z be distributed according to the density p it is necessary and sufficient that, for all functions $f \in \mathcal{F}$ we have

$$\mathbf{E}(f'(Z) + \psi(Z)f(Z)) = f(b-)p(b-) - f(a+)p(a+). \quad (1.29)$$

Note that from (1.27) we have that $\mathbf{E}f'(Z)$ exists, and (1.28) ensures that $\mathbf{E}\psi(Z)f(Z)$ exists.

Example 1.4. For the standard normal density ϕ we have $\phi'(x) = -x\phi(x)$, and ϕ, ϕ' are regular on $(-\infty, \infty)$; $\psi(x) = -x$ is regular on $(-\infty, \infty)$, and

$$\int \phi(x) |\ln \phi(x)| dx = \frac{1}{2\sqrt{2\pi}} \int x^2 \phi(x) dx = \frac{1}{2\sqrt{2\pi}}.$$

We obtain that Z is standard normal if and only if, for all functions $f \in \mathcal{F}$ we have

$$\mathbf{E}(f'(Z) - Zf(Z)) = 0.$$

This can be found in Stein [22] and has been explored by many authors.

Example 1.5. For the uniform $U[a, b]$, $-\infty < a < b < \infty$, we have $\phi'(x) = 0$ on $[a, b]$, and ϕ, ϕ' are regular on $[a, b]$; $\psi(x) = 0$ is regular, and

$$\int p(x) |\ln p(x)| dx = \ln(b-a) < \infty.$$

We obtain that Z is $U[a, b]$ if and only if, for all functions $f \in \mathcal{F}$ we have

$$\mathbf{E}(f'(Z)) = f(b-) - f(a+).$$

Example 1.6. For exponential $\exp(\lambda)$, $I = [0, \infty)$, we have $\phi'(x) = -\lambda\phi(x)$ on $[0, 1]$, and ϕ, ϕ' are regular on $[0, 1]$; $\psi(x) = -\lambda$ is regular, and

$$\int p(x) |\ln p(x)| dx = \int_0^\infty \lambda e^{-\lambda x} (\lambda x + |\ln \lambda|) dx < \infty.$$

We obtain that Z is $\exp(\lambda)$ if and only if, for all functions $f \in \mathcal{F}$ we have

$$\mathbf{E}(f'(Z) - \lambda f(Z)) = -\lambda f(0+).$$

Example 1.7. For the arcsine law $p(x) \propto (x(1-x))^{-\frac{1}{2}}$, $I = [0, 1]$, the density p is not finite at the endpoints of I , so p is not regular, and Proposition 1.4 does not apply.

See Diaconis and Zabell [8] and Hudson [15] for more characterizations.

Proof of Proposition 1.4. *Proof of necessity*

From (1.27) we know that $\int_I f'(x)p(x) dx$ exists, and from (1.28) we know that $\int_I f(x)p'(x) dx$ exists, so we may apply integration by parts. We have

$$\begin{aligned} \mathbf{E}f'(Z) &= \int_I f'(z)p(z) dz \\ &= f(b-)p(b-) - f(a+)p(a+) - \int_I f(z)p'(z) dz \\ &= f(b-)p(b-) - f(a+)p(a+) - \int_I f(z)\psi(z)p(z) dz \\ &= f(b-)p(b-) - f(a+)p(a+) - \mathbf{E}f(Z)\psi(Z). \end{aligned}$$

Proof of sufficiency

Let Z be a real random variable such that, for all functions $f \in \mathcal{F}$, (1.29) holds, and let h be an arbitrary measurable function for which

$$\int_I |h(z)|p(z) dz < \infty. \quad (1.30)$$

Let f be the particular solution of the differential equation

$$f'(z) + \psi(z)f(z) = h(z) - Ph \quad (1.31)$$

given by

$$f(z) = \frac{\int_a^z (h(x) - Ph)p(x) dx}{p(z)}, \quad (1.32)$$

where

$$Ph = \int_I h(z)p(z) dz.$$

We want to show that $f \in \mathcal{F}$, for then, (1.29) holds, yielding

$$\begin{aligned} 0 &= \mathbf{E}(f'(Z) + \psi(Z)f(Z)) - f(b-)p(b-) + f(a+)p(a+) \\ &= \mathbf{E}h(Z) - Ph. \end{aligned}$$

As the class of all measurable regular functions h satisfying (1.30) contains the indicator functions of Borel sets and hence is measure-determining for p , this would prove that Z has density p .

From (1.32) we have that f is regular and $f(b-)p(b-) = f(a+)p(a+) = 0$ and

$$\int_I |f'(z)|p(z) dz \leq \int_I |h(z)|p(z) dz + Ph + \int_I |f(z)\psi(z)|p(z) dz,$$

so that it suffices to prove that (1.28) holds. We have

$$\begin{aligned} \int_I |f(z)\psi(z)|p(z) dz &= \int_I |f(z)p'(z)| dz \\ &\leq \int_I \frac{|p'(z)|}{p(z)} \int_z^b |h(x) - Ph|p(x) dx dz. \end{aligned}$$

Denote by $c_1 < c_2 < \dots$ the sign change points of p' and hence of ψ , and note that due to the continuity assumption $\psi(c_i) = 0, i = 1, 2, \dots$. Let $A_i = (a_{i_1}, a_{i_2}), i = 1, 2, \dots$ be the intervals where $\psi > 0$ and let $B_j = (b_{j_1}, b_{j_2}), j = 1, 2, \dots$ be the intervals where $\psi \leq 0$. Then

$$\begin{aligned} \int_I \frac{|p'(z)|}{p(z)} \int_z^b |h(x) - Ph|p(x) dx dz &= \sum_{i=1}^{\infty} \int_{A_i} \psi(z) \int_z^b |h(x) - Ph|p(x) dx dz \\ &\quad - \sum_{j=1}^{\infty} \int_{B_j} \psi(z) \int_z^b |h(x) - Ph|p(x) dx dz. \end{aligned}$$

Note that $\psi(z) = (\ln p(z))'$ and $\ln p(z)$ is regular, so we can apply integration by parts again to obtain that the above equals

$$\begin{aligned} &\sum_{i=1}^{\infty} \left\{ \int_{A_i} |h(x) - Ph|p(x) \ln(p(x)) dx - [|h(x) - Ph|p'(x)]_{a_{i_1}}^{a_{i_2}} \right\} dx dz \\ &\quad - \sum_{j=1}^{\infty} \int_{B_j} \{ |h(x) - Ph|p(x) \ln(p(x)) dx - [|h(x) - Ph|p'(x)]_{b_{j_1}}^{b_{j_2}} \} dx dz \\ &\leq \int_I |h(x) - Ph|p(x) \ln(p(x)) dx \\ &\quad + |h(b-) - Ph|p'(b-) + |h(a+) - Ph|p'(a+) \\ &< \infty, \end{aligned}$$

due to (1.25). □

Proposition 1.4 will be used to obtain a general approximation theorem. Under the assumption of Proposition 1.4, let for convenience

$$\phi(x) = -\psi(x). \quad (1.33)$$

Note that, from (1.29),

$$\mathbf{E}\psi(Z) = p(b-) - p(a+)$$

and

$$\mathbf{E}\psi(Z)Z = bp(b-) - ap(a+) - 1.$$

We will often have the case that

$$\mathbf{E}\phi(Z) \approx 0, \quad \mathbf{E}\phi(Z)Z \approx 1.$$

Theorem 1.1. *Assume that Z is a random variable having distribution with probability density function p satisfying the assumptions of Proposition 1.4. Let (W, W') be an exchangeable pair of real random variables such that $\mathbf{E}(\phi(W))^2 = \sigma^2 < \infty$, with ϕ defined at (1.33) and let*

$$\lambda = \frac{\mathbf{E}(\phi(W') - \phi(W))(W' - W)}{2\sigma^2}. \quad (1.34)$$

Then, for all piecewise continuous functions h on \mathbf{R} to \mathbf{R} for which $\mathbf{E}|h(Z)| < \infty$,

$$\begin{aligned} & \mathbf{E}h(W) - \mathbf{E}h(Z) \\ &= \mathbf{E}f'(W) - \frac{1}{2\lambda\sigma^2}\mathbf{E}(\phi(W') - \phi(W))(f(W') - f(W)) \\ & \quad - \mathbf{E}\mathbf{E}^W\left(\frac{\phi(W') - (1 - \lambda\sigma^2)\phi(W)}{\lambda\sigma^2}\right)f(W), \end{aligned} \quad (1.35)$$

where f is defined by

$$f(w) = \frac{\int_a^z (h(x) - Ph)p(x) dx}{p(z)} = (Uh)(w) \quad (1.36)$$

and

$$f'(w) = (Vh)(w) = (Uh)'(w). \quad (1.37)$$

Remark. In the normal case, the second summand in (1.37) can be viewed as $\mathbf{E}(Vh)(Y)$, where Y is distributed according to the probability density function π defined by

$$\pi(y) = \mathbf{E}\frac{\phi(W') - \phi(W)}{\lambda\sigma^2}\delta_{\{W < y < W'\}}.$$

for all y . This distribution has been called the *zero bias* distribution by Goldstein and Reinert [12], but has appeared many times before in the literature in disguise; see Goldstein and Reinert [12] for references.

Remark. It is useful to think about how (1.35) could be small. One instance when it is small is if

$$\begin{aligned}
& \mathbf{E}f'(W) - \frac{1}{2\lambda\sigma^2}\mathbf{E}(\phi(W') - \phi(W))(f(W') - f(W)) \\
&= \mathbf{E}f'(W) - \frac{1}{2\lambda\sigma^2}(\phi(W') - \phi(W)) \int_W^{W'} f'(w) dw \\
&\approx \mathbf{E}f'(W) \left(1 - \frac{1}{2\lambda\sigma^2}(\phi(W') - \phi(W))(W' - W)\right).
\end{aligned}$$

From (1.34) we have that

$$\frac{1}{2\lambda\sigma^2}\mathbf{E}(\phi(W') - \phi(W))(W' - W) = 1,$$

so that

$$\mathbf{E}f'(W) - \frac{1}{2\lambda\sigma^2}\mathbf{E}(\phi(W') - \phi(W))(f(W') - f(W)) \approx 0.$$

Moreover, if

$$\mathbf{E}^W \phi(W') = (1 - \lambda\sigma^2)\phi(W)$$

then

$$\mathbf{E}\mathbf{E}^W \left(\frac{\phi(W') - (1 - \lambda\sigma^2)\phi(W)}{\lambda\sigma^2} \right) f(W) = 0$$

relating to Condition (1.22).

Proof of Theorem 1.1. Let $f \in \mathcal{F}$ be a function on I to \mathbf{R} , where \mathcal{F} is as in Proposition 1.4. For any antisymmetric function F on \mathbf{R}^2 to \mathbf{R} ,

$$\mathbf{E}F(W, W') = 0. \tag{1.38}$$

Applying this to the function F defined by

$$\begin{aligned}
F(w_1, w_2) &= \frac{(\phi(w_2) - \phi(w_1))(f(w_1) + f(w_2))}{2\lambda\sigma^2} \\
&= \frac{\phi(w_2) - \phi(w_1)}{\lambda\sigma^2} f(w_1) \\
&\quad + \frac{\phi(w_2) - \phi(w_1)}{2\lambda\sigma^2} (f(w_2) - f(w_1)),
\end{aligned}$$

we obtain

$$\mathbf{E} \left[\frac{\phi(W') - \phi(W)}{\lambda\sigma^2} f(W) - \frac{\phi(W') - \phi(W)}{2\lambda\sigma^2} (f(W') - f(W)) \right] = 0.$$

This can be rewritten in the form

$$\begin{aligned}
& \mathbf{E} \left[-\phi(W)f(W) + \frac{\phi(W') - (1 - \lambda\sigma^2)\phi(W)}{\lambda\sigma^2} f(W) \right. \\
& \quad \left. + \frac{\phi(W') - \phi(W)}{2\lambda\sigma^2} \int_W^{W'} f'(w) dw \right] = 0.
\end{aligned}$$

By Proposition 1.4, the distribution of Z is characterized by the property that, for all functions $f \in \mathcal{F}$,

$$\mathbf{E}(f'(Z) + \psi(Z)f(Z)) = f(b-)p(b-) - f(a+)p(a+).$$

This suggests that, in order to prove that $\mathbf{E}h(W)$ is approximately equal to $\mathbf{E}h(Z)$, it is appropriate to substitute for f a solution

$$f(w) = (Uh)(w)$$

of the differential equation

$$f'(w) - \phi(w)f(w) - f(b-)p(b-) + f(a+)p(a+) = h(w) - \mathbf{E}h(Z). \quad (1.39)$$

We use the solution given in (1.36), so that $f(b-)p(b-) = f(a+)p(a+) = 0$. We substitute $f'(W) - (h(W) - \mathbf{E}h(Z))$ for $\phi(W)f(W)$ in (1.29) and rearrange terms, obtaining

$$\begin{aligned} & \mathbf{E}h(W) - \mathbf{E}h(Z) \\ &= \mathbf{E} \left[f'(W) - \frac{\phi(W') - \phi(W)}{2\lambda\sigma^2} \int_W^{W'} f'(w) dw - \frac{\phi(W') - (1 - \lambda\sigma^2)\phi(W)}{\lambda\sigma^2} f(W) \right]. \end{aligned}$$

Using the definition of V in (1.37), we obtain (1.35). This finishes the proof. \square

In connection with simulations, we suggest using Theorem 1.1 for simulating the error in the distributional transformation by simulating the quantities on the right-hand side of (1.35). Let us concentrate on the standard normal case. Many more examples will be necessary to fully understand this method. Suppose we want to estimate $\mathbf{E}h_0(W)$ where h_0 is a reasonable piecewise continuous function and W is a random variable which we suspect has an approximately normal distribution. In principle, we apply Theorem 1.1 to the function h defined by

$$h\left(\frac{w - \xi}{\sigma}\right) = h_0(w). \quad (1.40)$$

We estimate σ^2 and ξ and λ as in Section 1.4 before. In the following, we write $\alpha \rightarrow \beta$ for “ α is replaced by β ”.

$$\begin{aligned} \mathbf{E}f'\left(\frac{W - \xi}{\sigma}\right) &\rightarrow \frac{1}{r} \sum_t f'\left(\frac{W_t - \hat{\xi}}{\hat{\sigma}}\right) \\ \mathbf{E}f'\left(\frac{W' - \xi}{\sigma}\right) &\rightarrow \frac{1}{r} \sum_t \mathbf{E}^{X_t} f'\left(\frac{W'_t - \hat{\xi}}{\hat{\sigma}}\right) \end{aligned}$$

and

$$\begin{aligned} & \mathbf{E} \frac{W' - W}{2\lambda\sigma} \left(f\left(\frac{W' - \xi}{\sigma}\right) - f\left(\frac{W - \xi}{\sigma}\right) \right) \\ & \rightarrow \frac{1}{r} \sum_t \mathbf{E}^{X_t} \frac{W'_t - W_t}{2\hat{\lambda}\hat{\sigma}} \left(\mathbf{E}^{X_t} f\left(\frac{W'_t - \hat{\xi}}{\hat{\sigma}}\right) - f\left(\frac{W_t - \hat{\xi}}{\hat{\sigma}}\right) \right) \end{aligned}$$

and

$$\mathbf{E} \left(\frac{W' - W}{\lambda\sigma} + \frac{W - \xi}{\sigma} \right) f\left(\frac{W - \xi}{\sigma}\right) \rightarrow \frac{1}{r} \sum_t \left(\frac{D_{1,t}}{\hat{\lambda}\hat{\sigma}} + \frac{W_t - \hat{\xi}}{\hat{\sigma}} \right) f\left(\frac{W_t - \hat{\xi}}{\hat{\sigma}}\right),$$

pretending that $\hat{\xi}$, $\hat{\lambda}$, and $\hat{\sigma}$ are constants.

In an elementary case with $\mathbf{E}W = 0$, $\mathbf{E}W^2 = 1$, $\mathbf{E}^W(W') = (1 - \lambda)W$, and $W' \in \{W - c, W, W + c\}$ (where c could be small, of the order $n^{-\frac{1}{2}}$), for a given h with $\mathbf{E}h(Z) = 0$ we would need to numerically approximate the function

$$f(w) = e^{\frac{w^2}{2}} \int_{-\infty}^w h(x) e^{-\frac{x^2}{2}} dx.$$

Then we put

$$f'(w) = h(w) + wf(w).$$

Given W_t we generate Y_t uniformly from the interval $(W_t, W_t + c)$, and we can estimate the error in the standard normal approximation by

$$\frac{1}{R} \sum_{t=1}^R (f'(W_t) - f'(Y_t)).$$

If c is small then this sum will be small.

Often there might not be an obvious candidate for a distributional approximation. Let (W, W') be an exchangeable pair. We want to approximate the distribution of W . Put

$$\begin{aligned} \alpha_1(w) &= \mathbf{E}^{W=w}(W' - W) \\ \alpha_2(w) &= \frac{1}{2} \mathbf{E}^{W=w}(W' - W)^2 \end{aligned}$$

and define the density

$$p(w) = \frac{c}{\alpha_2(w)} e^{\int_0^w \frac{\alpha_1(z)}{\alpha_2(z)} dz}, \quad -\infty < w < \infty,$$

where c is determined by the condition that $\int p(w) dw = 1$. Note that

$$\psi(w) = \frac{p'(w)}{p(w)} = \frac{\alpha_1(w) - \alpha_2'(w)}{\alpha_2(w)}$$

is of Pearson type. If p satisfies the assumptions of Proposition 1.4 with $p(-\infty) = p(\infty) = 0$, then any random variable Z has density p if and only if, for all $f \in \mathcal{F}$,

$$\mathbf{E}f'(Z) + \phi(Z)f(Z) = 0.$$

Theorem 1.2. *In the above situation, let Z have density p defined by (1.41). Then, for all regular functions h such that $\int |h(x)|p(x)dx < \infty$ we have*

$$\mathbf{E}h(W) = \mathbf{E}h(Z) = -\mathbf{E}\left\{R_1\left(\frac{g}{\alpha_2}\right)(W, W')\right\},$$

where

$$R_1(f)(w, w') = \frac{1}{2}(w' - w)(f(w') - f(w)) - \frac{1}{4}(w' - w)^2(f'(w') - f'(w)) \quad (1.41)$$

and

$$g(z) = \frac{1}{p(z)} \int_{-\infty}^z (h(x) - Ph)p(x) dx. \quad (1.42)$$

Proof of Theorem 1.2. We use the antisymmetric function

$$\begin{aligned}
F(w, w') &= \frac{1}{2}(w' - w)(f(w') + f(w)) \\
&= (w' - w)f(w) + \frac{1}{2}(w' - w)(f(w') - f(w)) \\
&= (w' - w)f(w) + \frac{1}{2}(w' - w)^2 \frac{f(w') + f(w)}{2} \\
&\quad + \frac{1}{2}(w' - w)(f(w') - f(w)) - \frac{1}{4}(w' - w)^2(f(w') + f(w)) \\
&= (w' - w)f(w) + \frac{1}{2}(w' - w)^2 \frac{f(w') + f(w)}{2} + R_1(f)(w, w'),
\end{aligned}$$

where $R_1(f)(w, w')$ is given in (1.41). Thus, from (1.38),

$$0 = \mathbf{E}F(W, W')$$

giving

$$\begin{aligned}
0 &= \mathbf{E}\mathbf{E}^W(W' - W)f(W) + \frac{1}{2}\mathbf{E}\mathbf{E}^W(W' - W)^2 \frac{f(W') + f(W)}{2} \\
&\quad + \mathbf{E}R_1(f)(W, W').
\end{aligned}$$

Put $g(w) = \alpha_2(w)f(w)$, so that

$$f'(w) = \frac{g'(w)}{\alpha_2(w)} - g(w) \frac{\alpha_2'(w)}{(\alpha_2(w))^2}.$$

We obtain

$$\begin{aligned}
\mathbf{E}R_1(f)(W, W') &= \mathbf{E}\left\{R_1\left(\frac{g}{\alpha_2}\right)(W, W')\right\} \\
&= \mathbf{E}\left\{\frac{\alpha_1(W)}{\alpha_2(W)}g(W) + \frac{\alpha_2(W)g'(W) - g(W)\alpha_2'(W)}{\alpha_2(W)}\right\} \\
&= \mathbf{E}\left\{\frac{\alpha_1(W) - \alpha_2'(W)}{\alpha_2(W)}g(W) + g'(W)\right\} \\
&= \mathbf{E}\psi(W)g(W) + g'(W) \\
&= \mathbf{E}h(W) - \mathbf{E}h(Z).
\end{aligned}$$

Here, h and g are related through g given in 1.42. □

In particular,

$$\begin{aligned}
\mathbf{E}h(W) &= \mathbf{E}h(Z) + \mathbf{E}\mathbf{E}^W R_1\left(\frac{g}{\alpha_2}\right)(W, W') \\
&= \mathbf{E}h(Z) + \mathbf{E}R_2(g)(W),
\end{aligned}$$

where

$$R_2(g)(w) = \mathbf{E}^{W=w} R_1\left(\frac{g}{\alpha_2}\right)(W, W').$$

Thus, from R observations we can estimate

$$\hat{\mathbf{E}}h(W) = \mathbf{E}h(Z) + \frac{1}{R} \sum_{t=1}^R \mathbf{E}^{X_t} R_1\left(\frac{g}{\alpha_2}\right)(W_t, W'_t).$$

References

- [1] Aldous, D. J. (1981). Representations for Partially Exchangeable Arrays of Random Variables. *J. Multivariate Anal.* **11**, 581–598. MR637937
- [2] Aldous, D. and Fill, J. A. (2003). Markov Chains, book available on the web at: <http://www.stat.berkeley.edu/~aldous/>
- [3] Arnold, H. J., Bucher, B. D., Trotter, H. F. and Tukey, J. W. (1956). Monte Carlo techniques in a complex problem about normal samples. *Symposium on Monte Carlo methods*. H. A. Meyer, ed., Wiley, New York; 80–88. MR79826
- [4] Billera, L. J. and Diaconis, P. (2001). A geometric interpretation of the Metropolis-Hastings algorithm. *Statistical Science* **16**(4), 335–339. MR1888448
- [5] Diaconis, P. (1989). An example for Stein’s method. in this volume, Chapter 2.
- [6] Diaconis, P. (1988). Recent progress in de Finetti’s notions of exchangeability. In *Bayesian Statistics* vol. 3, J. Bernardo et al. (eds), Oxford Press, Oxford, 111–125. MR1008047
- [7] Diaconis, P. and Sturmfels. (1998). Algebraic Algorithms for Sampling from Conditional Distributions. *Ann. Statist.* **26**, 363–397. MR1608156
- [8] Diaconis, P. and Zabell, S. (1991). Closed form summation for classical distributions: Variations on a theme of de Moivre *Statistical Science* **6**, 284–302. MR1144242
- [9] Fishman, G. S. (1996). *Monte Carlo : Concepts, Algorithms, and Applications*. Springer, New York etc. MR1392474
- [10] Fitzgerald, M., Picard, R. R. and Silver, R. N. (2000). Monte Carlo transition dynamics and variance reduction. *J. Statist. Phys.* **98**, 321–345.
- [11] Fulman, J. (2001). A Stein’s method proof of the asymptotic normality of descents and inversions in the symmetric group. In this volume, Chapter 4.
- [12] Goldstein, L. and Reinert, G. (1997). Stein’s method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* **7**, 935–952. MR1484792
- [13] Hammersley, J. M. and Hanscomb, D. C. (1965). *Monte Carlo Methods*. Methuen, London.
MR223065
- [14] Huber, M. and Reinert, G. (2000). The stationary distribution in the antivoter model: exact sampling and approximations. In this volume, Chapter 5.
- [15] Hudson, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6**, 473–484.
MR467991
- [16] Kingman, J. C. F. (1978). Uses of exchangeability. *Ann. Probab.* **6**, 183–197. MR494344
- [17] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

- [18] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.
- [19] Newman, M. E. J. and Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics*. Clarendon Press, Oxford. MR1691513
- [20] de Oliveira, P. M. C., Penna, T. J. P. and Herrmann, H. J. (1998). Broad histogram Monte Carlo. *Eur. Phys. J. B* **1**, 205–208. MR1644520
- [21] Rinott, Y. and Rotar, V. (1997). On coupling constructions and rates in the clt for dependent summands with applications to the anti-voter model and weighted U-statistics. *Ann. Applied Probability* **7**, 1080–1105. MR1484798
- [22] 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. MR402873
- [23] Stein, C. (1978). Asymptotic evaluation of the number of Latin rectangles. *J. Combinat. Theory A* **25**, 38–49. MR499035
- [24] Stein, C. (1996). Trace of random matrix, Stanford Statistics Dept, Technical Report, 1996.
- [25] Stein, C. (1967). Class notes from course taught at Stanford, notes taken by Lincoln Moses.
- [26] Stein, C. (1988). Application of Newton’s identities to a generalized birthday problem and to the Poisson binomial distribution. Stanford, Technical Report.
- [27] Stein, C. (1986). *Approximate Computation of Expectations*. IMS, Hayward, California. MR882007
- [28] Trotter, H. F. and Tukey, J. W. (1956). Conditional Monte Carlo for normal samples. *Symposium on Monte Carlo methods*. H. A. Meyer, ed., Wiley, New York, 64–79. MR79825
- [29] Wang, J. S., Tay, T. K. and Swendsen, R. H. (1999). Transition matrix Monte Carlo reweighting and dynamics. *Phys. Rev. Lett.* **82**, 476–479.