Stein's Method: Expository Lectures and Applications
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 46 (2004) 26–41

2. Stein's method for Markov chains: first examples

Persi Diaconis[1]

Stanford University

Abstract: Charles Stein has introduced a general approach to proving approximation theorems in probability theory. The method is being actively used for normal and Poisson approximation. This paper uses the method to derive rates of convergence of some simple Markov chains to their stationary distribution. The main purpose is to present Stein's general approach in a simple setting where the many choices can be examined and compared.

2.1. Introduction

Charles Stein has introduced a general approach to proving approximation theorems in probability theory. The method has been developed and applied for normal and Poisson approximation by Louis Chen, Andrew Barbour, and others. As the applications become more complex and refined, the overall approach fades into the background.

This paper applies Stein's method to get rates of convergence to uniformity for random walk on the discrete circle. The intent is largely expository, offering an example where the many choices made can be examined and compared. The example is developed in simplest from in Section 2. Section 3 extends the results to a general step size distribution. It gives new results that seem inaccessible with other tools such as Fourier analysis and coupling. Section 4 applies the approach to the Ehrenfest Urn. Section 5 gives a connection between Stein's method and Fourier analysis.

Stein's method proves approximation theorems like the central limit theorem by means of characterizing operators. For example, a real random variable W is standard normal if and only if $E\{Wf(W) - f'(W)\} = 0$ for every smooth f with compact support. The operator $Tf(W) = Wf((W) - f'(W)$ is characterizing. One shows that a random variable W_n is approximately normal if E_nTf is close to zero. To implement this, one introduces an exchangeable pair (W_n, W_n'). These basic ingredients, a characterizing operator and an exchangeable pair are often not hard to find for problems of interest. The clearest development of this approach is Stein (1992) which is my recommendation for a place to start reading. This may be followed by Stein (1986).

Stein's method has been cleaned up and smoothed over for routine use for Poisson approximation. Building on work of Chen (1975), Barbour, Holst, and Janson (1992) have a book length treatment with a remarkable collection of examples. This is based on a coupling approach to Stein's method which has a life of its own as developed by Goldstein, Rinnott, and Reinert, among others. Reinert (1998) is an excellent recent survey.

Bolthausen (1984) has introduced new ways of using characterizing operators and couplings. These result in a three page completely self contained proof of the Berry–Esseen theorem (without Fourier analysis) and the solution of a long open

[1]Department of Mathematics and Statistics, Stanford University, Stanford, CA 94305, USA.

problem, the right Berry–Esseen rate for Hoeffding's combinatorial central limit theorem. These ideas have been developed by Götze (1991) who used them to prove the best available multivariate Berry–Esseen type theorems.

One further important theme is the development of the generator method by Andrew Barbour and others. This allows process approximation and a wide variety of special cases. See Barbour (1997) for an overview.

For a beginner, Stein (1986), (1992) followed by perusal of the surveys of Barbour (1997) and Reinert (1998) may be the best way to start. There is much further to do in applying and developing Stein's method.

2.2. Convergence to the uniform distribution for simple random walk

Consider simple random walk on the discrete circle $\mathbb{Z}_p$ – the integers mod p. The walk is generated by independent random variables $X_1, \cdots, X_n$ where $X_i = \pm 1$ with probability $\frac{1}{2}$. If $S_n = X_1 + \cdots + X_n \pmod p$, it is well known that $P\{S_n = j\} \to \frac{1}{p}$ for large n provided p is odd. Rates of convergence for this limit theorem in total variation will be derived by Stein's method.

Theorem 2.2.1. *Let $Q^{\star n}$ be the law of simple random walk on $\mathbb{Z}_p$ for p odd. Let U be the uniform distribution. For all n and p*

$$\left\|Q^{\star n} - U\right\| \leq \frac{p-1}{\sqrt{n}}. \tag{2.1}$$

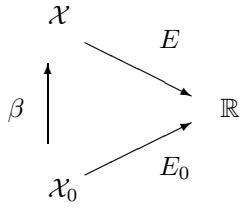
Here $\|Q^{\star n} - U\| = \max_A |Q^{\star n}(A) - U(A)|$.

The proof will be given as a series of steps with discussion. These steps appear quite generally. The characterizing operator is introduced in the first Lemma. The exchangeable pair is introduced above (9). Let $\mathbb{Z}_2^n$ be the space of sequences of $\{\pm 1\}$ of length n. Let $\mathcal{X} = L(\mathbb{Z}_2^n)$ be the space of real valued functions on $\mathbb{Z}_2^n$. Let E be the expectation operator $E : \mathcal{X} \to \mathbb{R}$ via $Ef = E(f(X_1, \ldots, X_n))$.

Let $\mathcal{X}_0 = L(\mathbb{Z}_p)$ be the real valued functions of $\mathbb{Z}_p$. Define E_0 as the expectation operator under the uniform distribution. Thus $E_0 : \mathcal{X}_0 \to \mathbb{R}$ given by

$$E_0 f = \frac{1}{p} \sum_{j=0}^{p-1} f(j). \tag{2.2}$$

Functions in $\mathcal{X}_0$ can be carried into functions in $\mathcal{X}$ by summation. Define $\beta : \mathcal{X}_0 \to \mathcal{X}$ via $\beta f(x_1, \ldots, x_n) = f(x_1 + \cdots + x_n)$, where the sum is taken mod p. These definitions can be summarized by a diagram

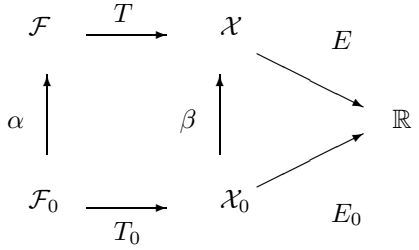


(2.3)

To say that S_n has an approximate uniform distribution is the same thing as saying that the diagram approximately commutes:

$$E\beta \doteq E_0.$$

Stein's method introduces a second stage to the diagram.



(2.4)

This will be defined precisely below. One feature of the construction is that if the left square approximately commutes, so $T\alpha \doteq \beta T_0$, then the right triangle approximately commutes.

The first step of rigorous argument involves constructing the diagram. This is done in three stages: bottom row, top row, and the map α. Following this, Stein's lemma makes the approximate commutation precise. Theorem 2.2.1 follows easily from these considerations and the weak law of large numbers.

The Bottom Row. Let $\mathcal{F}_0 = \mathcal{X}_0 = L(\mathbb{Z}_p)$. Define $T_0 : \mathcal{F} \to \mathcal{X}_0$ by

$$T_0 f(i) = f(i) - f(i-2). \tag{2.5}$$

The operator T_0 characterizes the uniform distribution in the following sense.

Lemma 2.2.1. *Let p be an odd integer. A probability Q on $\mathbb{Z}_p$ is uniform if and only if for each $f \in \mathcal{F}_0$, $QT_0 f = 0$.*

Proof. If $Q(j) = \frac{1}{p}$ for all $j \in \mathbb{Z}_p$, the $QT_0 f = 0$ for all f. Taking $f(j) = \delta_{j_0}(j)$, the Kronecker delta for fixed j_0, shows $Q(j_0) = Q(j_0+2) = Q(j_0+4) = \cdots = Q(j_0+2\ell)$ for any ℓ. Since the numbers 2ℓ run over all of $\mathbb{Z}_p$, $Q(j) = \frac{1}{p}$.

Remark. Lemma 1 shows that the bottom row of the diagram (2.3) is exact:

$$\text{Im } T_0 = \text{Ker } E_0.$$

Indeed, the lemma clearly implies Im $T_0 \subset$ Ker E_0. The linear map E_0 is onto $\mathbb{R}$ and so Ker E_0 has co-dimension 1 and $\mathcal{X}_0 =$ Ker $E_0 \oplus \{\text{constants}\}$. If Im T_0 is properly contained in Ker E_0, then $\mathcal{X}_0 =$ Im $T_0 \oplus R \oplus \{\text{constants}\}$, with R a nontrivial subspace. Define a linear map $L : \mathcal{X}_0 \to \mathbb{R}$ to preserve constants, be zero on Im T_0, and nonzero on R. Then $LT_0 = 0$. But L can be chosen arbitrarily on R and so need not be uniform. Now, using L for Q in Lemma 1 gives a contradiction. The next lemma shows Im $T_0 \supset$ Ker E_0 more explicitly by producing an inverse to T_0 on Ker E_0. This is a map $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$ defined by

$$U_0 f(i) = \sum_{0 \le j \le i/2} \big(f(2j) - E_0 f\big). \tag{2.6}$$

Here the sum is over $0, 1, 2, 3, \ldots, i/2$ and 2 has an inverse because p is odd.

Lemma 2.2.2. *For E_0, U_0, and T_0 defined at* (2.2), (2.5), *and* (2.6)

$$T_0 U_0(j) = f(j) - E_0 f. \tag{2.7}$$

Proof. If $T_0U_0f(j) = U_0f(j) - U_0f(j-2) = f(j) - E_0f$. The argument works at $j = 0$ because $U_0f(-2) = 0$.

Remarks. 1) Lemma 2 shows Im $T_0 \supset$ KerE_0: if $f \in$ KerE_0, $T_0U_0f = f$.
2) Clearly the operator $T_0f(j) = f(j) - f(j-a)$ characterizes the uniform distribution whenever a is relatively prime to p. An investigation of the various choices appears in Section 4.
3) If $\mathcal{F}_0$ is taken as $\{f : \mathbb{Z}_p \to \mathbb{R} : f(p-2) = 0\}$, the bottom row can be extended to the short exact sequence $0 \to \mathcal{F}_0 \stackrel{T_0}{\to} \mathcal{X}_0 \stackrel{E_0}{\to} 0$. Indeed, T_0 only kills constant functions and this choice of $\mathcal{F}_0$ forces such to be zero.

The following bound on U_0 is important. Again, in great generality, one needs an $L^\infty \to L^\infty$ bound on the inverse of the characterizing operator.

Lemma 2.2.3. *For U_0 defined by* (2.6)*, and $S \subset \mathbb{Z}_p$ with indicator δ_S,*

$$\big|U_0\delta_S(j)\big| \le \frac{p-1}{2}. \tag{2.8}$$

Proof. Define $f(j) = \delta_S(j) - E_0\delta_S$. Then $\sum_{j=0}^{p-1} f(j) = 0$ so for any set $A \subset \mathbb{Z}_p$

$$\sum_{j\in A} f(j) = -\sum_{j\in A^c} f(j).$$

Clearly

$$\left|\sum_{j\in A} f(j)\right| \le |A| \qquad \text{so} \qquad \sup_A \left|\sum_{j\in A} f(j)\right| \le \frac{p-1}{2}. \quad \square$$

The Top Row. Let $\mathcal{F}$ be the space of antisymmetric functions $f : \mathbb{Z}_2^n \times \mathbb{Z}_2^n \to \mathbb{R}$. Thus $f \in \mathcal{F}$ satisfies $f(x,y) = -f(y,x)$. The object is to construct a map T from $\mathcal{F}$ to $\mathcal{X}$ such that Im $T =$ Ker E. The construction uses an exchangeable pair (X,Y) with X uniformly distributed on $\mathbb{Z}_2^n$. While many choices for the joint distribution work (see Section 3), one simple choice picks $I \in \{1,2,\dots,n\}$ uniformly independent of X and forms $Y = (X_1, \dots, 1 - 2X_I, \dots, X_n)$. Thus $Y_i = X_i$ if $i \ne I$ and Y_I is the opposite of X_I mod 2. Clearly $P\{X = x, Y = y\} = P\{X = y, Y = x\}$. Define $T : \mathcal{F} \to \mathcal{X}$ by

$$Tf(x) = E\big(f(X,Y)|X = x\big). \tag{2.9}$$

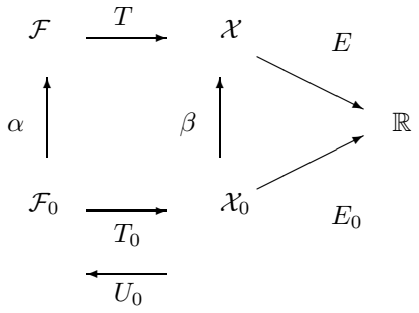
Since $ETf = E\{f(X,Y)\} = -E\{f(Y,X)\} = -E\{f(X,Y)\}$, Im $T \subset$ Ker E. This is all that is needed for Stein's lemma proved below. Stein (1990) investigates when Im $E =$ Ker T. This does not hold in the present case, but would hold if the exchangeable pair was defined by choosing X_I' as ± 1 at random.

The Map α. To complete the description of the basic diagram a map α from $\mathcal{F}_0$ to $\mathcal{F}$ must be chosen. This takes a function on $\mathbb{Z}_p$ into an anti-symmetric function on $\mathbb{Z}_2^n \times \mathbb{Z}_2^n$. After some experimentation, discussed below, the following choice emerged. Let $S(x) = x_1 + \cdots + x_n$. Define $\alpha : \mathcal{F}_0 \to \mathcal{F}$ by

$$\alpha f(x,y) = cf\big(S(x)\big)\delta_{S(x)}\big(S(y) - 2\big) - cf\big(S(y)\big)\delta_{S(y)}\big(S(x) - 2\big)\big). \tag{2.10}$$

On the right $S(x)$ is take mod p inside f but in $\mathbb{Z}$ inside δ and $c > 0$ is a constant to be chosen presently.

Stein's Lemma. The following simple computation gives an exact expression for the error term in saying if the left square commutes then the right square commutes. Consider the diagram



(2.11)

with

$$\text{Im } T \subset \text{ Ker } E, \qquad \text{Im } T_0 \subset \text{ Ker } E_0 \quad \text{ and } T_0 U_0 = \text{ id } - E_0. \tag{2.12}$$

Lemma 2.2.4 (Stein). *Let (2.11) be a diagram of linear spaces and maps satisfying (2.12). Then*

$$\beta E = E_0 + E\{\beta T_0 - T\alpha\}U_0. \tag{2.13}$$

Proof. $0 = ET\alpha = ET\alpha - E\beta T_0 + E\beta T_0$. Multiplying on the right by U_0 and use $T_0 U_0 = \text{ id } - E_0$ gives

$$0 = E(T\alpha - E\beta T_0)U_0 + E\beta - E_0.$$

Rearranging terms gives the result. □

Remark. If the left square commutes, the term $\{\quad\}$ in (2.13) is zero, so the right triangle commutes. The lemma gives an explicit expression for the error which can be usefully bounded.

Proof of Theorem 2.2.1. To begin, a careful expression for the map $\beta T_0 - T\alpha$ must be derived. Let f be an arbitrary function. Clearly

$$\beta T_0 f(x) = f\big(S(x)\big) - f\big(S(x) - 2\big).$$

For $T\alpha$,

$$\begin{aligned} T\alpha f(x) &= cE\big(f(S(X)\delta_{S(X)}\big(S(Y) - 2\big) - f\big(S(Y)\big)\delta_{S(Y)}\big(S(X) - 2\big)|X = x\big) \\ &= cf\big(S(x)\big)P\big\{S(Y) = S(x) + 2|X = x\big\} \\ &\quad - cf\big(S(x) - 2\big)P\big\{S(Y) = S(x) - 2|X = x\big\} \\ &= cf\big(S(x)\big)\frac{N_-(x)}{n} - cf\big(S(x) - 2\big)\frac{N_+(x)}{n} \end{aligned}$$

where $N_+(x)$ is the number of plus signs in x and $N_+ + N_- = n$. This last expression can be centered to give

$$\begin{aligned} T\alpha f(x) &= \frac{c}{2}\big\{f\big(S(x)\big) - f\big(S(x) - 2\big)\big\} + cf\big(S(x)\big)\left(\frac{N_-(x) - \frac{n}{2}}{n}\right) \\ &\quad - cf\big(S(x) - 2\big)\left(\frac{N_+(x) - \frac{n}{2}}{n}\right) \\ &= \frac{c}{2}\big\{f\big(S(x)\big) - f\big(S(x) - 2\big)\big\} \\ &\quad + \frac{c(N_-(x) - \frac{n}{2})}{n}\big\{f\big(S(x)\big) + f\big(S(x) - 2\big)\big\}. \end{aligned}$$

The lead term here is $\frac{c}{2} \cdot \beta T_0$. To make things cancel, take $c = 2$. Then

$$(\beta T_0 - T\alpha)f(x) = -\frac{2(N_- - \frac{n}{2})}{n}\big\{f\big(S(x)\big) + f\big(S(x) - 2\big)\big\}.$$

From lemma 3, for $f = U_0\delta_S$, $|f(j)| \le \frac{p-1}{2}$ for all j. From this and Stein's lemma, for $S \subset \mathbb{Z}_p$,

$$\left|P\{S_n \in S\} - \frac{|S|}{p}\right| \le 2(p-1)E\left|\frac{N_-(X) - \frac{n}{2}}{n}\right| \le \frac{p-1}{\sqrt{n}}.$$

The last inequality used

$$E\left|N_-(X) - \frac{n}{2}\right| \le \sqrt{E\left(N_-(X) - \frac{n}{2}\right)^2} = \frac{1}{2}\sqrt{n}. \quad \square$$

Remarks. 1) The result gives a clean bound which gives the right rate in the sense that it shows that n must be of order p^2 to make the variation distance small. Note, however, that if, e.g., $n = p^3$, the bound gives $O(\frac{1}{\sqrt{n}})$ as an error while the Fourier analysis arguments of Diaconis (1988) give $O(e^{-c\sqrt{n}})$ for $c > 0$. A refinement would have to use a less crude bound on $f(S(x)) + f(S(x) - 2)$ in the critical range. Note that an exponential rate follows from Theorem 2.2.1 and the elementary fact that total variation is submultiplicative.
2) The probability content amounts to the law of large numbers. A related way to prove this theorem uses the central limit theorem coupled with the result that a normal (μ, σ) variable mod 1 tends to uniform as σ tends to ∞. The only known way to get rates requires the Berry–Essen theorem. Thus here the formalism seems to be useful.
3) Instead of using the Cauchy-Schwarz inequality at the end of the proof, De Moivre's formula gives the exact expression. Fix n and let ν be the unique integer with $\frac{n}{2} < \nu \le \frac{n}{2} + 1$. Then

$$E\left|N_-(X) - \frac{n}{2}\right| = \frac{\nu}{4}\,\frac{\binom{n}{\nu}}{2^n}.$$

Using this, the right side of (2.1) can be slightly improved to a quantity asymptotic to $\frac{p-1}{4\sqrt{2\pi n}}$ which gives a slight improvement. Diaconis and Zabell (1991) discuss De Moivre's formula and a different connection to Stein's method.

2.3. An extension to general measures

This section develops bounds when the basic step distribution is a general probability Q on $\mathbb{Z}_p$. The following examples may help motivate the general result.

Example 1. Fix an integer k and let

$$Q(j) = \begin{cases} \dfrac{1}{2k+1} & \text{for } -k \le j \le k \\ 0 & \text{else} \end{cases}$$

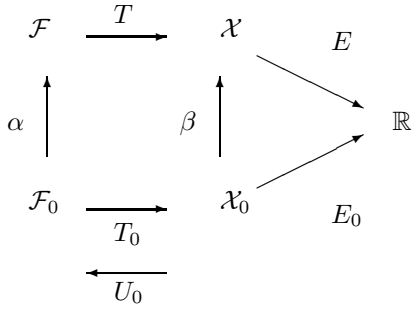
Then, for any $k \ge 1$, all integers $p > 2k$, and all $n \ge 1$ on $\mathbb{Z}_p$

$$\big\|Q^{\star n} - U\big\| \le \frac{p-1}{\sqrt{2kn}}. \tag{2.14}$$

Example 2. Fix $\theta \in (0,1)$ and let $Q(1) = \theta,\ Q(-1) = 1-\theta$ on $\mathbb{Z}_p$. Then, for all odd integers p and all $n \geq 1$, on $\mathbb{Z}_p$

$$\left\|Q^{\star n} - U\right\| \leq \frac{p-1}{2\sqrt{n\theta(1-\theta)}}. \tag{2.15}$$

These results follow from making the appropriate choices in the basic diagram



(2.16)

For the righthand triangle, choose $\mathcal{X} = L(\mathbb{Z}_p^n)$ and E as expectation under the product measure Q^n. Choose $\mathcal{X}_0 = L(\mathbb{Z}_p)$ and E as expectation under the uniform distribution. The map β is defined as before $\beta f(x) = f(x_1 + \cdots + x_n)$.

For the bottom row, choose $\mathcal{F}_0 = L(\mathbb{Z}_p)$ and for an integer r chosen later

$$T_0 f(i) = f(i) - f(i-r) \text{ for } r \text{ relatively prime to } p. \tag{2.17}$$

For the inverse of T_0, take

$$U_0 f(i) = \sum_{0 \leq j \leq i/r} \big(f(rj) - E_0 f\big). \tag{2.18}$$

As in Lemma 2 of Section 2,

$$T_0 U_0 = \text{ id } - E_0 \tag{2.19}$$

and for $S \subset \mathbb{Z}_p$

$$|U_0 \delta_S(i)| \leq \frac{p-1}{2}. \tag{2.20}$$

This completes the specification of the bottom row of the diagram.

For the top row, $\mathcal{F}$ is taken as the anti-symmetric functions on $\mathbb{Z}_p^n \times \mathbb{Z}_p^n$. To define T, an exchangeable pair (X, Y) must be created. Choose $X \in \mathbb{Z}_p^n$ from the product measure Q^n. Choose I uniform and independent on $\{1, 2, \ldots, n\}$, set $Y_i = X_i$ for $i \neq I$ and choose Y_I from the measure $P(X_I, \cdot)$ with $P(i,j)$ a transition kernel. Exchangeability of (X, Y) requires

$$QP = Q \qquad \text{and} \qquad Q(i)P(i,j) = Q(j)P(j,i). \tag{2.21}$$

Thus P generates a reversible Markov chain with stationary distribution Q. Note that the rows of P can always be taken as copies of Q. This corresponds to taking Y_I as an independent pick from Q. This will be called the *independence coupling*. With this notation, a general bound can be stated.

Theorem 2.3.1. *Let p be an integer and Q a probability on $\mathbb{Z}_p$. Then, for any n*

$$\big\|Q^{\star n} - U\big\| \le \frac{p-1}{2\sqrt{n}}\, K$$

with

$$K = \left\{ \frac{\Sigma Q(j) P^2(j, j+r)}{(\Sigma Q(j) P(j, j+r))^2} - 1 \right\}^{\frac{1}{2}} + \left\{ \frac{\Sigma Q(j) P^2(j, j-r)}{(\Sigma Q(j) P(j, j-r))^2} - 1 \right\}^{\frac{1}{2}}. \tag{2.22}$$

and P defined at (2.21) *and any fixed $r \in \mathbb{Z}_p$ relatively prime to p. If the term in the denominator in* (2.22) *is zero, K is defined as ∞ and the bound is vacuous.*

Examples 1 and 2. Before proving the result, let us see how it yields the bounds announced in examples 1 and 2 above. For Example 1, take the independence coupling $P(i,j) = Q(j)$ and $r = 1$. Each displayed term in (2.22) is $\frac{1}{\sqrt{2k}}$ which gives (2.14). For Example 2, take the independence coupling for the measure Q and $r = 2$. Then

$$K = \left\{ \frac{\theta}{1-\theta} \right\}^{\frac{1}{2}} + \left\{ \frac{1-\theta}{\theta} \right\}^{\frac{1}{2}} = \frac{1}{\sqrt{\theta(1-\theta)}}.$$

This gives (2.15).

Proof of Theorem 2.3.1. The map α in the diagram (2.16) is taken as

$$\alpha f(x,y) = c f\big(S(x)\big)\delta_{S(x)}\big(S(y) - r\big) - c f\big(S(y)\big)\delta_{S(y)}\big(S(x) - r\big) \tag{2.23}$$

with c to be chosen and $S(x) = x_1 + \cdots + x_n$. From Stein's lemma

$$\beta E = E_0 + E\{\beta T_0 - T\alpha\} U_0. \tag{2.24}$$

Here

$$\begin{aligned}
\beta T_0 f(x) &= f\big(S(x)\big) - f\big(S(x) - r\big), \\
T\alpha f(x) &= cE\big\{ f\big(S(X)\big)\delta_{S(X)}\big(S(Y) - r\big) - f\big(S(Y)\big)\delta_{S(Y)}\big(S(X) - r\big) | X = x \big\} \\
&= cf\big(S(x)\big) P^x\big(S(Y) = S(x) + r\big) \\
&\quad - cf\big(S(x) - r\big) P^x\big(S(Y) = S(x) - r\big),
\end{aligned}$$

where $P^x(\cdot) = P\{\cdot | X = x\}$. From the construction of (X, Y)

$$\begin{aligned}
P^x\big\{S(Y) = S(x) + r\big\} &= \frac{1}{n} \sum_j N_j P(j, j+r) \\
&= \sum_j \frac{(N_j - nQ(j))}{n}\, P(j, j+r) + \sum_j Q(j) P(j, j+r)
\end{aligned}$$

while $N_j = N_j(x)$ equals the number of coordinates in x equal to j. Similarly

$$P^x\big\{S(Y) = S(x) - r\big\} = \sum_j \frac{(N_j - nQ(j))}{n} P(j, j-r) + \sum_j Q(j) P(j, r-r).$$

Further, reversibility implies

$$\sum_j Q(j) P(j, j-r) = \sum_j Q(j-r) P(j-r, j) = \sum_k Q(k) P(k, k+r) = C.$$

Here C is the term appearing in the denominators of K at (2.22). It is the chance a Markov chain started with distribution Q and having transitions given by P goes up by r in two steps. If C is zero, the bound of Theorem 2.3.1 is satisfied. Thus assume that P and r have been chosen so C is positive. Then take the constant $c = C^{-1}$ in (2.23). This allows the lead terms to cancel and yields

$$\begin{aligned}(\beta T_0 - T\alpha)f(x) \\ &= \frac{c}{n}\Big\{ f(S(x)\sum_j (N_j - nQ(j))P(j, j+r) \\ &\qquad - f(S(x)-r)\sum_j (N_j - nQ(j))P(j, j-r)\Big\}. \end{aligned} \tag{2.25}$$

For $f = U_0\delta_S$, (2.20) yields $|f(j)| \le \frac{p-1}{2}$. Thus (2.24) gives

$$\begin{aligned}&\big|(\beta E - E_0)f\big| \\ &\quad \le \frac{c}{n}\frac{p-1}{2} E\big\{\big|\big(N_j - nQ(j)\big)\big|P(j, j+r) + \big|N_j - nQ(j)\big|P(j, j-r)\big\}. \end{aligned} \tag{2.26}$$

The random variables N_j have the law of n balls dropped into p boxes with chance $Q(j)$ of being dropped into box j. This has $p \times p$ covariance matrix $n(\Delta - QQ^t)$ with Δ a diagonal matrix having $Q(j)$ as diagonal entries and Q treated as a column vector. It follows that $\sum_j (N_j - nQ(j))P(j, j+r)$ has mean zero and variance $n\{\sum_j Q(j)P^2(j, j+r) - C^2\}$. Similarly $\sum_j (N_j - nQ(j))P(j, j-r)$ has mean zero and variance $n\{\sum_j Q(j)P^2(j, j-r) - C^2\}$. Using these results and $E|X| \le \{E|X|^2\}^{\frac{1}{2}}$ in (2.26) completes the proof of Theorem 2.3.1. □

Remarks. 1) The bound for Example 2 is sharp. A matching lower bound follows from the arguments of Chapter 3 of Diaconis (1986). The bound for Example 1 is not sharp when k is permitted to grow with p. Fourier analysis shows that if $\frac{p}{k\sqrt{n}}$ is small, the variation distance is small. Since Stein's method remains an identity up to the final bound it is instructive to see where the precision is lost.

Begin with (2.25). Using the independence coupling for example 2, $C^{-1} = \frac{2k}{(2k+1)^2}$ and (2.25) simplifies to

$$\begin{aligned}&(\beta T_0 - T\alpha)f(x) \\ &\quad = \frac{2k+1}{2kn}\left\{ f(S(x))\left\{\frac{n}{2k+1} - N_k\right\} - f(S(x)-1)\left\{\frac{n}{2k+1} - N_{-k}\right\}\right. . \end{aligned} \tag{2.27}$$

Here $f = U_0\delta_S$ satisfies $\|f\|_\infty \le \frac{p-1}{2}$ and $f(S(x)) - f(S) - 1) = \delta_S(S(x)) - E_0\delta_S$. If f is bounded by $\frac{p-1}{2}$, no substantial improvement over (2.14) seems possible. Of course, roughly, $f(S(x))$ and $N_k - \dfrac{n}{2k+1}$, are independent and this last term has mean 0, so the expected value of the left side should be small. To make this precise requires rather precise knowledge about the random walk.

As an indication of the information available, consider example 1 with k only slightly smaller than $p/2$. Then, elementary considerations show that $n = 1$ is sufficient to make the variation distance small. To see this from Stein's method, write (2.26) as equal to

$$\frac{-(2k+1)}{2k}\big\{f\big(S(x)\big)N_k - f\big(S(x)-1\big)N_{-k}\big\} + \frac{1}{2k}\big\{f\big(S(x)\big) - f\big(S(x)-1\big)\big\}.$$

The second term in brackets is bounded by $1/2k$. Taking expectations of the first,

$$E\big\{f\big(S(x)\big)N_k - f\big(S(x)-1\big)N_{-k}\big\} = \frac{f(k)-f(-k-1)}{2k+1}.$$

This last term is bounded above in absolute value by $2(\frac{p}{2}-k)/(2k+1)$. Combining bounds gives

$$\big|Q(S)-U(S)\big| \le \left(2\left(\frac{p}{2}-k\right)-1\right)/2k.$$

It follows that one step is enough if $(p/2-k)/k$ is small. Extending this argument to a wider range of k values seems tricky.
2) The two denominators in (2.22) are equal because of (2.25).
3) In example 2, all possible transition matrices P lead to the bound (2.15). Similarly, in example 1, the exchangeable pair from $P(i,i+1) = P(i,i-1) = \frac{1}{2}$ for $-k < i < k$ with $P(k,-k) = P(-k,k) = P(k,k-1) = P(-k,-k+1) = \frac{1}{2}$ and $P(i,j) = 0$ else leads again to the bound (2.14). This indicates that the independence coupling may be a reasonable first choice.

Example 3. To conclude this section, consider a set $A \subset \mathbb{Z}_p$ of size $|A| = \alpha$, and let

$$Q(j) = \begin{cases} \dfrac{1}{\alpha} & \text{if } j \in A \\ 0 & \text{otherwise .} \end{cases} \tag{2.28}$$

Let $\beta(r)$ be the number of $a \in A$ such that $a + r \in A$, let $\beta^\star = \max_r \beta(r)$ and assume there is an $r^\star$ relatively prime to p such that the maximum is achieved. Then, Stein's method, with the independence coupling and $r = r^\star$ gives

$$\big\|Q^{\star n} - U\big\| \le \frac{p-1}{\sqrt{n}}\left(1-\frac{\beta^\star}{\alpha}\right)^{1/2}.$$

Remarks. This bound reduces to (2.14) for A an interval $[-k,k]$. It is not even sharp for certain small sets A. For large primes p, most sets A with $|A| = 3$ have $\|Q^{\star n} - U\| \le \theta\frac{p}{n}$ for θ universal.

Examples 1 and 3 show that there is room for improvement in the bounds suggested above. Hopefully this will deepen our understanding of Stein's method. The bounds aimed for have the delicacy of Berry–Essen bounds so it is not surprising they are elusive.

2.4. The Ehrenfest Urn

The Ehrenfest Urn is a well studied Markov chain on the state space $\mathcal{S} = \{0,1,\dots,d\}$ with transition probabilities

$$\begin{aligned} P(i,i) &= \frac{1}{d+1} \\ P(i,i+1) &= 1 - \frac{i}{d+1} \\ P(i,i-1) &= \frac{i}{d+1}. \end{aligned} \tag{2.29}$$

and stationary distribution

$$\pi(i) = \frac{\binom{d}{i}}{2^d}\frac{1}{n} \tag{2.30}$$

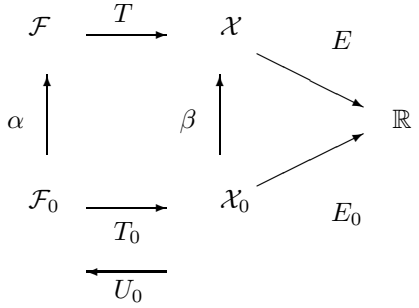
Here a fixed holding probability has been introduced to get rid of parity problems. Stein's method will be used to derive the following result:

Theorem 2.4.1. *For the Markov chain (2.29) all d and n, there is a universal constant b such that*

$$\left\|P^n(0,\cdot)-\pi(\cdot)\right\| \le \sqrt{\frac{d+1}{n}} + b\sqrt{\frac{d(d+1)}{n}}. \tag{2.31}$$

This shows that $n \gg d^2$ steps suffice to achieve uniformity. As shown in Diaconis-Graham- Morrison (1990) $n \gg d\log d$ suffice.

Proof. As is well known, the Ehrenfest chain is the distance from zero process for nearest neighbor random walk on a d-dimensional hypercube $\mathbb{Z}_2^d$ with $\mathbb{Z}_2 = \{0,1\}$. This motivates the following choices in the basic diagram



(2.32)

Take $\mathcal{F}_0 = \mathcal{X}_0 = L(\mathcal{S})$ where $\mathcal{S} = \{0,1,\dots,d\}$. Take $E_0(f) = \sum_{i=0}^d f(i)\pi(i)$. The characterizing operator T_0 can be chosen as

$$T_0 f(j) = (d-j)f(j) - jf(j-1).$$

Stein (1986) shows $E_0 T_0 = 0$. For the inverse, define, for $0 \le j \le d-1$,

$$U_0 f(j) = \frac{1}{\binom{d}{j}(d-j)} \sum_{k=0}^{j} \binom{d}{k} \big(f(k) - E_0(f)\big).$$

This satisfies

$$T_0 U_0 f(j) = f(j) - E_0(f) \qquad \text{for } 0 \le j \le d.$$

Standard bounds on binomial probabilities show that

$$\text{for } \|f\|_\infty \le 1, \qquad \|U_0 f\|_\infty \le \frac{b}{\sqrt{d}} \qquad \text{for universal } b.$$

This completes the specification of the bottom row. For the top row, let $\mathcal{X} = L((\mathbb{Z}_2^d)^n)$, let Q be defined on $\mathbb{Z}_2^d$ by $Q(0) = Q(e_1) = \cdots = Q(e_d) = \frac{1}{d+1}$, where e_i are the standard Euclidean basic vectors. Define E as expectation under Q^n : $Ef = E(f(X_1 \cdots X_n))$ with X_i being independent and identically distributed with respect to Q. Define $\beta : \mathcal{X}_0 \to \mathcal{X}$ via $\beta f(x) = f(S(x))$ where $S(x) = |x_1 + \cdots + x_n|$ and $|v|$ is the number of ones in the binary vector v.

Let $\mathcal{F}$ be the antisymmetric functions on $(\mathbb{Z}_2^d)^n \times (\mathbb{Z}_2^d)^n$. To construct T, construct an exchangeable pair of random variables (X,Y) by choosing X from Q^n

on $(\mathbb{Z}_2^d)^n$, choosing a coordinate I uniformly in $\{1,2,\ldots,n\}$, and setting $Y_i = X_i$, $i \neq I$ with Y_I and independent pick from Q. Define

$$Tf(x) = E\big(f(X,Y)|X=x\big).$$

Finally, define $\alpha : \mathcal{F}_0 \to F$ by

$$\alpha f(x,y) = cf\big(S(x)-1\big)\delta_{S(x)}\big(S(y)+1\big) - cf\big(S(y)-1\big)\delta_{S(y)}\big(S(x)+1\big)$$

with c to be chosen later. Note $f(-1)$ may be defined arbitrarily as zero. This choice will not enter the final computations.

For use in Stein's lemma, $\beta T_0 - T\alpha$ must be computed.

$$\begin{aligned}
\beta T_0 f(x) &= \big(d - S(x)\big)f\big(S(x)\big) - S(x)f\big(S(x)-1\big).\\
T\alpha f(x) &= cf\big(S(x)-1\big)P^x\big(S(Y)=S(x)-1\big)\\
&\quad - cf\big(S(x)\big)P^x\big(S(Y)=S(x)+1\big).
\end{aligned}$$

To calculate the conditional probabilities, let $N_i(x)$ be the number of coordinates in x equal to e_i and let $N_0(x)$ be the number of coordinates in x equal to 0. Let $\delta_1(N_i)$ be one or zero as N_i is odd or even. Then

$$\begin{aligned}
P^x\big(S(Y)=S(x)-1\big) &= \frac{N_0}{n}\frac{S(x)}{d+1} + \sum_{i=1}^d \frac{N_i}{n}\frac{\delta_1(N_i)}{d+1}\\
P^x\big(S(Y)=S(x)+1\big) &= \frac{N_0}{n}\frac{d-S(x)}{d+1} + \sum_{i=1}^d \frac{N_i}{n}\frac{1-\delta_1(N_i)}{d+1}.
\end{aligned}$$

For example, the first expression follows from the following considerations. The chance that $S(Y) = S(x) - 1$ given x is the chance that the random coordinate I hits a zero and is replaced by one of the coordinates where the final sum is 1 plus the chance of hitting an odd coordinate and replacing it by zero. These expressions centered are:

$$\begin{aligned}
P^x\big(S(Y)=S(x)-1\big) &= \frac{2S(x)}{(d+1)^2} + \frac{S(x)(N_0 - \frac{n}{d+1})}{n(d+1)} + \sum_{i=1}^d \frac{(N_i - \frac{n}{d+1})\delta_1(N_i)}{n(d+1)}\\
P^x\big(S(Y)=S(x)+1\big) &= \frac{2(d-S(x))}{(d+1)^2} + \frac{(d-S(x))(N_0 - \frac{n}{d+1})}{n(d+1)}\\
&\quad + \sum_{i=1}^d \frac{(N_i - \frac{n}{d+1})(1-d_1(N_i))}{n(d+1)}.
\end{aligned}$$

To cancel lead terms, c must be chosen as $(d+1)^2/2$. Then

$$\begin{aligned}
&(\beta T_0 - T\alpha)f(x)\\
&\quad = \frac{d+1}{2n}\Big\{f\big(S(x)-1\big)\Big[S(x)\Big(N_0 - \frac{n}{d+1}\Big) + \sum_{i=1}^d \Big(N_i - \frac{n}{d+1}\Big)\delta_i(N_i)\Big]\\
&\qquad - f\big(S(x)\big)\Big[\big(d-S(x)\big)\Big(N_0 - \frac{n}{d+1}\Big) + \sum_{i=1}^d \Big(N_i - \frac{n}{d+1}\Big)\big(1-\delta_i(N_i)\big)\Big]\Big\}.
\end{aligned}$$

To bound the expected absolute value of this error term take $f = U_0\delta_A$ for $A \subset \mathbb{Z}_2^d$. The terms involving N_0 are

$$\frac{d+1}{2n}\left\{\big[S(x)f\big(S(x)-1\big) - \big(d-S(x)\big)f\big(S(x)\big)\big]\left(N_0 - \frac{n}{d+1}\right)\right\}$$
$$= \frac{d+1}{2n}\left\{T_0U_0\delta_A(x)\left(N_0 - \frac{n}{d+1}\right)\right\}.$$

Using $\|T_0U_0\delta_A\|_\infty \le 2$ and $E|N_0 - \frac{n}{d+1}| \le \sqrt{n/(d+1)}$ shows that this part of the error is bounded by $\sqrt{\frac{d+1}{n}}$.

For the sums, use $\|f\|_\infty \le \frac{b}{\sqrt{d}}$ and $E|N_i - \frac{n}{d+1} \le \sqrt{n/(d+1)}$ to give

$$\frac{d+1}{2n}\frac{b}{\sqrt{d}}d\sqrt{\frac{n}{d+1}} = \frac{b}{2}\sqrt{\frac{d(d+1}{n}}.$$

Combining bounds completes the proof of Theorem 2.4.1. □

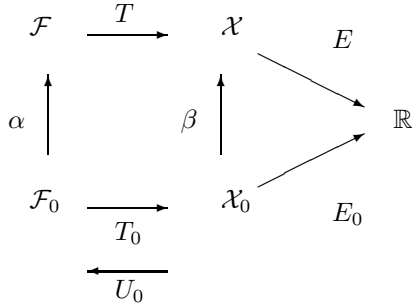
Remark. The commutativity of $\mathbb{Z}_2^d$ underlies the argument above. A similar argument can be given for any walks involving Gelfand pairs where the convolution operation is commutative. See Diaconis (1987, Chapter 3-G).

2.5. A Fourier connection

This section shows how a natural choice of the map α can force consideration of the eigenvalues of the underlying walk. Let G be a finite Abelian group. Let $S = S^{-1}$ be a set containing the identity and generating G. Let

$$Q(t) = \delta_S(t)/|S| \qquad \text{for } t \in G \tag{2.33}$$

be the corresponding measure on G. To study convolution powers of Q by Stein's method, choices must be made in the basic diagram



(2.34)

Here $\mathcal{X} = L(G^n)$, the operator E is expectation with respect to the n-fold produce measure $Q \times Q \cdots \times Q, \mathcal{X}_0 = L(G)$ and E_0 is expectation with respect to the uniform distribution. The map $\beta f(x_1 \cdots x_n) = f(x_1 + \cdots + x_n)$ and T is based on an exchangeable pair formed by choosing a random coordinate and replacing it with a random choice in S. All of these choices are just as seen in Section 1. Further $\mathcal{F}_0 = \mathcal{X}_0$ and $\mathcal{F}$ is taken as the anti-symmetric functions from $G^n \times G^n$.

The map α seems crucial. One natural choice is

$$\alpha f(\underline{x}, \underline{y}) = f\big(S(\underline{x})\big) - f\big(S(\underline{y})\big), \qquad \text{with } S(\underline{x}) = x_1 + \cdots + x_n. \tag{2.35}$$

This gives

$$\alpha f(\underline{x}) = f\big(S(\underline{x})\big) - \frac{1}{n|S|}\sum_{s,s'} N_s f\big(S(\underline{x}) + s' - s\big). \tag{2.36}$$

In (2.36) the sum is over all pairs $s, s' \in S \times S$ and N_s is the number of times s appears in x. This suggests choosing T_0, the characterizing map, as

$$T_0 f(t) = \big(\delta_{id} - Q \star \widetilde{Q} \star f(t)\big) \qquad \text{for } t \in G. \tag{2.37}$$

In (2.37), the convolution of functions is $f_1 \star f_2(t) = \sum_{s\in G} f_1(t-s)f_2(s)$, $\widetilde{Q}(t) = Q(-t)$, and δ_{id} is point mass at the identity of G. It is clear that $E_0 T_0 = 0$. The lemma below will show that $\operatorname{Ker} E_0 = \operatorname{Im} T_0$ and provide an explicit inverse U_0 as required by Stein's formalism, provided that the Fourier transform of Q does not vanish.

To set things up, recall that a character of G is a map $\chi : G \to \mathbb{C}$ such that $\chi(s+t) = \chi(s)\chi(t)$ for all $s, t \in G$. The set of distinct characters is denoted $\widehat{G}$. The Fourier transform is defined by

$$\widehat{Q}(\chi) = \sum_{s\in G} Q(s)\chi(s).$$

As usual $Q_1 \star Q_2(\chi) = Q_1(\chi)Q_2(\chi)$ and the uniform distribution has zero transform except at the character $\chi(s) \equiv 1$ (the trivial character). Convergence to uniformity can often be studied by bounding the transform and showing how fast its powers tend to zero. As shown, e.g., in Diaconis (1988, Chapter 3E), the eigenvalues of the Markov chain associated to the random walk are precisely the number $\widehat{Q}(\chi)$ and χ varies in $\widehat{G}$. In particular, the random walk converges to the uniform distribution if and only if $|\widehat{Q}(\chi)| \neq 1$ for χ non-trivial.

To invert the map T_0 of (2.37) involves solving for g in the equation $T_0 g = f$, or $(\delta_0 - Q \star \widetilde{Q}) \star g = f$. Taking transforms,

$$\big(1 - \big|\widehat{Q}(\chi)\big|^2\big)\widehat{g}(\chi) = \widehat{f}(\chi).$$

This defines g uniquely (up to constants), since $|\widehat{Q}(\chi)| \neq 1$. To summarize:

Lemma 2.5.1. *Under the assumptions above, assume $\hat{Q}(\chi) \neq 1$ for $\chi \neq 1$. Let T_0 be defined by (2.37). Define*

$$U_0 f(t) = \frac{1}{|G|}\sum_{\chi\in\widehat{G}-1} \frac{\widehat{f}(\chi)}{(1-|\widehat{Q}(\chi)|^2}\chi(t^{-1}).$$

Then

$$T_0 U_0 f(t) = f(t) - E_0 f. \tag{2.38}$$

Proof. The maps U_0 and T_0 are linear. Each vanishes on constant functions. Consider $T_0\widehat{U_0}f(\chi) = \widehat{f}(\chi)$ for χ non-trivial. Thus both sides of (2.38) have the same transform at all representations. □

Corollary 2.5.1. *If $|\widehat{Q}(\chi)| \neq 1$ for χ non-trivial Ker $E_0 =$ Im T_0.*

Let us plug these computations into Stein's lemma (Lemma 4). From (2.36) and (2.37)

$$\begin{aligned}(\beta T_0 - T\alpha)f(\underline{x}) \;&=\; \big(\delta_{id} - Q\star\widetilde{Q}\big)\star f\big(S(\underline{x})\big) \\ &\quad - \Big\{ f\big(S(\underline{x})\big) - \frac{1}{n|S|}\sum_{s,s'} N_s f\big(S(\underline{x}) + s' - s\big)\Big\}\end{aligned}$$

with

$$f = U_0 \delta_A,$$

where U_0 is defined in Lemma 5.

The procedure above seems convoluted. To turn it into a bound of some sort, note that for any A and non-trivial χ, $|\widehat{\delta}_A(\chi)| \le \frac{|G|}{2}$ so that

$$\big|f(s)\big| \le \frac{1}{2} \sum_{\chi \in \widehat{G}-1} \frac{1}{1-|\widehat{Q}(\chi)|^2} \le \frac{|G|}{2} \frac{1}{1-\pi_\star^2} \tag{2.39}$$

with $\pi_\star = \max_\chi |\widehat{Q}(\chi)|$, the largest non-trivial eigenvalue. As above, for any s, $|N_s - \frac{n}{|S|}| \le \sqrt{\frac{n}{|S|}}$. Combining bounds,

$$\big\|Q^{\star n} - U\big\| \le \sqrt{\frac{|S|}{n}} \frac{|G|}{2}, \frac{1}{1-\pi_\star^2}. \tag{2.40}$$

Example 1. Take $G = \mathbb{Z}_p$ for p odd and $S = \{0, 1, -1\}$, the bound becomes

$$\big\|Q^{\star n} - U\big\| \le \frac{Cp^3}{\sqrt{n}}$$

with universal C. Here, the easily proved result $\pi^\star = 1 - \frac{C'}{p2}$ for C' constant, was used. This shows that $n \gg p^6$ steps are sufficient to have small variation distance small. As is well known, $n \gg p^2$ steps are necessary and suffice.

The main reason for including this section is to indicate a connection between Stein's method and the eigenvalues. The argument does underscore the value of choosing a characterizing operator with a simple inverse.

References

[1] Barbour, A. D. (1997). Stein's method. *Encyclopedia of Statistical Science, Update Vol. 1*, Wiley, NY, 513–521. MR1469744

[2] Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson Approximation.* Oxford University Press, Oxford. MR1163825

[3] Bolthausen, E. (1984). An Estimate of the remainder in a combinatorial central limit theorem. *Z. Wahr. Verw. Geb.* **66**, 379–386. MR751577

[4] Chen, L. (1975). Poisson Approximation for dependent trials. *Ann. Probab.* **3**, 534–545. MR428387

[5] Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA. MR964069

[6] Diaconis, P., Graham, R. L., and Morrison, J. (1990). Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures and Algorithms* **1**, 51–72. MR1068491

[7] 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

[8] Goetze, F.(1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19**, 724–739. MR1106283

[9] Reinert, G.(1998). Coupling for normal approximations with Stein's method. In D. Aldous and J. Propp (eds), *Microsurveys in Discrete Probability*, DIMACS, vol. 41, 193–208. MR1630415

[10] Stein, C. (1986). *Approximate Computation of Expectations*, Institute of Mathematical Statistics, Hayward, CA. MR882007

[11] Stein, C. (1992). A way of using auxiliary randomization. In *Probability Theory.* L. Chen, K. Choi, K. Hu, and J. Lou, eds. de Gruyter, Berlin, 159–180. MR1188718