Stein's Method: Expository Lectures and Applications

Persi Diaconis and Susan Holmes, Editors

March 22nd, 2004

3

Stein's Method for Birth and Death Chains

Susan Holmes

Abstract
This article presents a review of Stein's method applied to the case of discrete random variables. We attempt to complete one of Stein's open problems, that of providing a discrete version for chapter 6 of his book. This is illustrated by first studying the mechanics of comparison between two distributions whose characterizing operators are known, for example the binomial and the Poisson. Then the case where one of the distributions has an unknown characterizing operator is tackled. This is done for the hypergeometric which is then compared to a binomial. Finally the general case of the comparison of two probability distributions that can be seen as the stationary distributions of two birth and death chains is treated and conditions of the validity of the method are conjectured.

3.1 Overview

Stein's method provides ways of proving weak-convergence results using test functions and approximations to expectations. It is a method that many have found quite difficult to infiltrate because it does not use any of the more classical tools such characteristic functions.

My thanks go to Charles Stein who painstakingly led me through the intricacies of his approach while I was visiting Stanford in 1993, and to Persi Diaconis who first tried to explain his picture of the method to me. I made my own picture of the procedure by trying to make a discrete version of chapter 6 of Charles' book (stein, 1986) upon his suggestion.

A little history: Stein's method of exchangeable pairs and characterizing operators, not to be confused with shrinkage, was first used by Charles in the early 70's, at the 6th Berkeley Symposium to prove central limit theorems for dependent random variables (Stein, 1992).

His approach was a complete innovation, because he does not use characteristic functions. Instead Charles based his argument on what he called a characterizing operator for the normal distribution.

Here is how this characterization is stated in his book (Stein, page 21, 1986).

Proposition 3.1.1 *A random variable has a standard normal distribution iff for all* $h : \mathbb{R} \longrightarrow \mathbb{R}$*, piecewise continuously differentiable whose absolute value of first derivative has a finite expectation* *with regards to the normal $N|h'| < \infty$ we have :*

$$E\{h'(W) - Wh(W)\} = 0$$

which we will also write

$$\forall h \in \mathcal{F}_N, E(T_N h)(W) = 0, \text{ where } T_N h(x) = h'(x) - xh(x). \tag{3.1}$$

T_N is a function from the space of piecewise continuously differentiable functions $\mathcal{F}_N$ to the space of continuously differentiable functions, we will call it the characterizing operator for the normal. Nh denotes the expectation of h with regards to the normal.

After following Stein's proof of the central limit theorem, I realized that he not only associated an operator to the normal, but also built one for the other distribution then compared the two operators, bounding the expectation of their difference on special test functions.

Following Charles' work, many authors have built characterizing operators, Chen (1974) for the Poisson, Loh (1992) for the multinomial, Diaconis (1998) for the uniform, Mann (1995) and Reinert (1997) for the χ^2. Barbour, Holst and Janson(1992) have written a book on the use of the method in the context of Poisson approximation.

The question arises of how to construct the characterizing operator for any given distribution. Once this has been done and certain properties have been proved for both operator and inverse, limit theorems become reasonably straightforward to prove. Let us start exploring with this latter part of the method. How can we compare two distributions for which the characterizing operators are well studied?

We begin with the binomial distribution as the **target**, playing the same role as the normal in Charles' first work. The Poisson will be the random variable we want to approximate.

3.2 Examples

3.2.1 Bounds on the distance between Poisson and binomial

As a first motivation we will show the procedure for proving a bound for the total variation distance between a Poisson $\mathcal{P}(\lambda)$ and a binomial $\mathcal{B}(n,p)$ distribution. Of course, to make the distributions close we will suppose that $\lambda = np$.

We will not worry about how to build the characterizing operators for the time being, and we will just use the fact that Charles Stein (1986) proved the following:

Proposition 3.2.1 *A random variable is binomial $\mathcal{B}(n,p)$ if and only if for every bounded function f, the expectation computed with respect to that random variable $\mathbb{E}(T_0 f)$ is zero, where*

$$T_0 f(w) = p(n-w)f(w+1) - w(1-p)f(w).$$

This T_0 is called the characterizing operator of the binomial $\mathcal{B}(n,p)$ distribution.

Remark on notation. In this example, our target distribution is the binomial we will denote anything related to the target with the index 0, for instance the expectation under the binomial will be $\mathbb{E}_0$, this is a convention that extends to the sequel as well, where the target distributions will not necessarily be the binomial, but will always be identified by the index 0.

The other distribution is Poisson with matching mean $\mu = np$ which also has a characterizing operator denoted $T\alpha$, which we will prove later to be:

$$(T\alpha)f(w) = npf(w+1) - wf(w) = \mu f(w+1) - wf(w). \tag{3.2}$$

We expect the fit to depend on p, in particular, the fit should be good for p small, of an order $\frac{1}{n}$. All we have to remember about these operators for the moment is that

1. $ET\alpha f = 0$, for all f, iff the expectation is computed with respect to the Poisson, and
2. $ET_0 f = 0$, for all f, iff the expectation is taken with respect to a binomial distribution.

For any function f defined on $[0, 1, \ldots, n]$ the difference between these operators is

(3.3) $$(T\alpha - T_0)f(w) = pw(f(w+1) - f(w)) = pw\Delta f(w), \text{ defined for } w \in [0, 1, \ldots, n-1]$$

Where $\Delta f(w) = f(w+1) - f(w)$ denotes the first order difference for f at w.

If we take as our function f a function whose image by T_0 is precisely $\mathbb{I}_m - P_0(m) = T_0 f$, where $\mathbb{I}_m$ denotes the indicator function for the set $\{m\}$, by computing the expected value of the difference between operators we will obtain the difference in expectations at that function f:

$$P(m) - P_0(m) = (E - E_0)(\mathbb{I}_m - P_0(m)) = ET_0 f = E(T_0 - T_\alpha)f$$

Using (2.3), the right hand side will be easy to bound if for this particular f, we can bound its increase $|\Delta f|$.

It has been proved by Stein (1986), in the case $p = 1/2$ and by Barbour, Holst and Janson, (1992), page 190, for general p, that for f such that: $\mathbb{I}_m - P_0(m) = T_0 f$, we have:

(3.4) $$|\Delta f(w)| < \frac{1}{npq}, \qquad \forall w.$$

This provides the following uniform pointwise bound:

$$\text{For every } m, \ |P(m) - P_0(m)| < \mathbb{E}pw(\frac{1}{npq}) = \frac{p}{q}, \text{ where } q = 1 - p$$

Remarks:

1. This bound would usually be used when p is small (for instance $p = \frac{1}{n}$, so q close to 1).
2. This translates to the following inequality:

$$\left| \frac{(np)^k e^{-(np)}}{k!} - \binom{n}{m} p^m q^{n-m} \right| < \frac{p}{q}$$

   which might not be so easy to prove by simple calculus. We will see in section 4, that there is a better bound available, reversing the roles of the two distributions, making the Poisson the target and using the bound on the first order difference of its pseudo-inverse.

3. Actually one is usually more interested in the total variation distance between the two distributions then in the pointwise distance. The bound on $|\Delta f|$ is available even for f a solution to $T_0 f = \mathbb{I}_A - P_0(A)$, for any $A \subset [0, n]$. Barbour, Holst and Janson (1992) proved that (2.4) still holds for these more general f. This provides the bound:

$$d_{TV}(P, P_0) < \frac{p}{q}.$$

4. Bounds such as (2.4) are crucial properties of characterizing operators that must be proved anew for each new target distribution.

3.2.2 Hypergeometric and binomial

Now, we will do a more original example. Suppose that the target is binomial $\mathcal{B}(n,p)$ again, but that the other distribution's characterizing operator is unknown.

We will bound the distance between a binomial $\mathcal{B}(n,p)$ distribution and the hypergeometric $\mathcal{H}(N,R,n)$. That is, suppose we are picking n balls without replacement from an urn of N balls of which R are red and we look at the distribution of the number of red balls denoted k. We are going to compare it to a binomial $\mathcal{B}(n,\frac{R}{N})$.

We start by finding an exchangeable pair denoted (k,k'). This is used to define a characterizing operator for the hypergeometric distribution with parameters R, N,n: $\mathcal{H}(N,R,n)$.

Suppose we lay out the balls uniformly at random in a line, the left n ones are the ones in the sample, among which k are red and $n-k$ are black. Now suppose that we exchange two different balls picked uniformly at random and then count again how many red ones among the left n, this will be our random variable k'.

(k,k') is an exchangeable pair because the procedure is obviously reversible. Repeating these switches defines a birth and death chain: k will at most change by one. Call β_k the probability, given that the variable was at k, that it will go to $k+1$ after one move, and δ_k the probability that it will go down one given that it was at k. Thus $\beta_k = P($ hit a black ball among left n and switch it with a red ball among the right $N-n)$.
This gives

$$\beta_k = P(k'=k+1|k) = \frac{n-k}{N}\frac{R-k}{N-1}\times 2,$$

$\delta_k = P($ hit a red ball among the left n ones and switch it with a black ball in the right ones$)$,

$$\delta_k = P(k'=k-1|k) = \frac{k}{N}\frac{N-R-(n-k)}{N-1}\times 2.$$

From these definitions we may compute

$$E^k(k'-k) = \beta_k - \delta_k = -\lambda(k-\mu), \qquad \text{with } \lambda = \frac{2}{N-1} \tag{3.5}$$

because

$$\begin{aligned}
E^k(k'-k) &= (+1)\beta_k + (-1)\delta_k = \beta_k - \delta_k \\
&= \frac{2}{N(N-1)}\left[nR - nk - kR + k^2 - kN - k^2 + kn + Rk\right] \\
&= \frac{2}{N(N-1)}\left[nR - kN\right] = -\frac{1}{N-1}(k - n\frac{R}{N}) \\
&= -\lambda(k-\mu) \text{ with } \mu = n\frac{R}{N} \text{ and } \lambda = \frac{2}{N-1}
\end{aligned}$$

In order to construct the characterizing operator we define a map α from functions f on $\{0,1,\ldots,n\}$ to the space of antisymmetric functions as

$$(\alpha f)(w,w') = \frac{1}{\lambda}\left(f(w')\mathbb{I}_{w'=w+1} - f(w)\mathbb{I}_{w=w'+1}\right)$$

Note: Given any exchangeable pair (w,w') of random variables, any antisymmetric real function A defined for all pairs (w,w') with finite expectation, has to satisfy $EA(w,w')=0$. This is going to provide a good way of finding functions in the kernel of E. To construct T, as is usual in Stein's method of use of exchangeable pairs, we take an exchangeable pair (w, w') and an antisymmetric F and define: $(TF)(w) = E^w F(w, w')$. In this case:

$$(T\alpha)f(k) = \frac{\beta_k}{\lambda}\Delta f(k) + \frac{\beta_k - \delta_k}{\lambda} f(k) \tag{3.6}$$

Note that by construction, because (w, w') is exchangeable and (αf) antisymmetric, we have $ET\alpha = 0$.

This operator $T\alpha$ is a characterizing operator for the hypergeometric. We will see that for any function f the difference between $T\alpha$ and T_0 will again be of particular utility. In fact, originally the characterizing operator of the binomial was constructed in a similar fashion and can be written as:

$$T_0 f(k) = p(n-k)\Delta f - (k - np) f(k)$$

Here we will have $p = \frac{R}{N}$. We compare the two operators:

$$(T\alpha - T_0)f(k) = \left[\frac{\beta_k}{\lambda} - (n-k)\frac{R}{N}\right]\frac{1}{n} \tag{3.7}$$

$$= \frac{(n-k)}{N}(-k)\Delta f(k). \tag{3.8}$$

First consider the simple case bounding pointwise probabilities, say at the point m. We would like to bound $|P(m) - P_0(m)| = |\mathbb{E}\mathbb{I}_m - E_0\mathbb{I}_m|$, this is done taking f in the equations above to be the solution to:

$$T_0 f(k) = \mathbb{I}_m(k) - P_0(m) \tag{3.9}$$

Theorem 3.1 (Distance between binomial and hypergeometric) *Let $P_{\mathcal{H}}$ denote the hypergeometric probability distribution and P_0 the binomial $\mathcal{B}(n,p)$ then:*

$$|P_{\mathcal{H}}(m) - P_0(m)| \le \frac{n-1}{N-1}$$

Proof. Through (2.4) we have, for f the solution to (2.9), a bound on Δf.
Note that if k has a $\mathcal{H}(N, R, n)$ distribution and $p = \frac{R}{N}$:

$$\begin{aligned}
\mathbb{E}(k^2) = var(k) + n^2p^2 &= npq\{1 - \frac{n-1}{N-1}\} + n^2p^2 \\
\mathbb{E}(T\alpha - T_0)f(k) &\le \frac{1}{N}\frac{1}{npq}(\mathbb{E}kn - \mathbb{E}k^2) \\
&\le \frac{1}{N}\frac{1}{npq}(n^2p - n^2p^2 - npq\{1 - \frac{n-1}{N-1}\}) \\
&\le \frac{1}{N}(n - 1 + \frac{n-1}{N-1}) \\
&\le \frac{n-1}{N-1}
\end{aligned}$$

Remark 3.2.1 Actually if we are more ambitious and want to bound the TV distance between the two distributions, as in the first example, exactly the same argument follows through, replacing (2.9) by:

$$T_0 f_A = \mathbb{I}_A - P_0(A)$$

where A is a set of $[0, n]$. As above Barbour et al. [1992] show that we still have the bound

$$\Delta f_A < \frac{1}{npq}$$

and all the other computations are the same. This proves:

Theorem 3.2 *The total variation distance between the hypergeometric $\mathcal{H}(N, R, n)$ and the relevant binomial $\mathcal{B}(n, \frac{R}{N})$ is bounded by $(n-1)/(N-1)$, uniformly in R, for $R > n$.*

This can be compared to Diaconis and Freedman (1981):

$$d_{TV(P_{\mathcal{H}}, P_0)} \leq \frac{4n}{N},$$

which they proved to be sharp, up to constants.

Let us now generalize each step of this procedure. The next section sets the scene for extensions of the method from situations where we know the characterizing operators to cases where we need to build them and the 'pseudo-inverse' for a new target and bound the increase in this 'pseudo-inverse'.

3.3 Notation and Context

Suppose we have a probability space $(\Omega, \mathcal{B}, P)$ we will call $\mathbb{E} : \mathcal{X} \to \mathbb{R}$ the expectation associated to P on $\mathcal{X}$, the space of real-valued random variables defined on Ω that have finite expectation.

We will be trying to compute EZ the expectation of some random variable or an approximation thereof. To this end we will consider the null-space of E: $ker\ E = \{y : Ey = 0\}$, we will look for a random variable close to $Z - c$ (c a constant). Thus we will be able to say $EZ \approx c$. We will call $\mathcal{X}_0$ the space of real valued functions that have finite expectation with the target distribution. Here, $\mathcal{X}_0$ will be considered a subset of $\mathcal{X}$ and β will denote a natural embedding of $\mathcal{X}_0$ into $\mathcal{X}$.

3.3.1 Exchangeable Variables

Strange as it may seem, the study of $ker\ \ E$ is done through a pair of exchangeable variables, the definition of which I recall to be:

(X, X') is a pair of exchangeable variables iff the joint distribution of the pair (X, X') is identical to the distribution of (X', X), written sometimes $(X, X') \stackrel{d}{=} (X', X)$.

In what follows (X, X') is used to denote an exchangeable pair.

3.3.2 Operators of Antisymmetric Functions

Call $\mathcal{F}$ the set of antisymmetric functions defined on Ω^2.

In what follows we will denote by T the operator $T : \mathcal{F} \longrightarrow \mathcal{X}$ which associates to every antisymmetric F in $\mathcal{F}$ the function:

$$TF \text{ such that } TF(x) = E^{X=x} F(X, X')$$

where E^X is the conditional expectation given X.

A simple computation shows $ImF \subset kerE$, the reverse is also true as long as any two elements of the state space can be connected through a sequence of exchangeable pairs. (Diaconis, personal communication)

Then $ImT = kerE$ and the following diagram is exact:

$$\mathcal{F} \xrightarrow{T} \mathcal{X} \xrightarrow{E} \mathbb{R} \longrightarrow 0$$

Thus the image of T completely defines the null space $kerE$. Now if we're trying to find the distribution of $W = \psi(X)$ we may try and give an approximation of $Eh(\psi(X))$ for functions h such as indicators.

3.3.3 A Characterizing Operator for the Target Distribution

It has to be the case that we have an idea about the relevant target. That is, we know which approximation to choose. In most cases the expectations with respect to this distribution are denoted E_0.

We will define an operator T_0 that characterizes the target distribution. Later we will explain more in detail how such a characterization is built. For the time being, we will look at cases where this operator is known.

3.3.4 A Useful Diagram

The point of view we are going to stress here starts through the comparison of the two exact sequences:

$$\mathcal{F} \xrightarrow{T\alpha} \mathcal{X} \xrightarrow{E} \mathbb{R} \longrightarrow 0$$

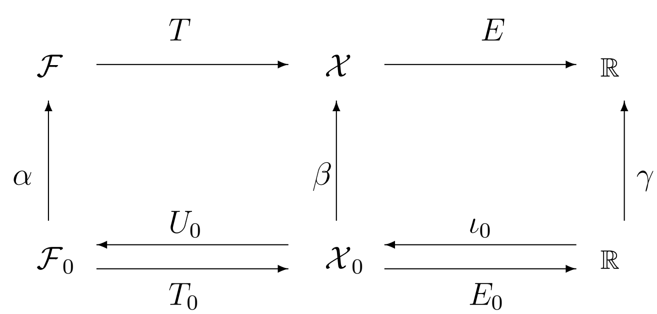
$$\mathcal{F}_0 \xrightarrow{T_0} \mathcal{X}_0 \xrightarrow{E_0} \mathbb{R} \longrightarrow 0$$

we will show that we can write

$$Eh(x) = E_0 h + E(Qh)(x, x')$$

where the last term on the right provides an indication of how good the approximation is. It is especially important to notice that this residual is an expectation with regards to the 'unknown' distribution. Bounds on $|Qh|$ will provide bounds for the approximation.

We will now detail this decomposition through what will be called the basic diagram:



The top part of the diagram contains the sets $\mathcal{F}$ and $\mathcal{X}$ and the operators T and E defined above.

$\mathcal{X}_0$ is a subspace of $\mathcal{X}$. $\mathcal{F}_0$ is a subspace of $\mathcal{F}$ and α will denote a natural embedding of $\mathcal{F}_0$ into $\mathcal{F}$.

The function ι_0 transforms a real number into the random variable always equal to that real value.

U_0 is the 'pseudo-inverse' for the function $\mathbb{I}_m$. In the examples above, we looked for a function f such that $T_0 f = \mathbb{I}_m - P_0(m) = \mathbb{I}_m - E_0\mathbb{I}_m$, this can be expressed as the condition that for any g in $\mathcal{X}_0$ we can define $U_0(g)$ such that

$$T_0 \circ U_0 g = g - \iota_0 \circ E_0(g)$$

We will call U_0 the 'pseudo-inverse' of T_0 in all that follows.

An algebraic lemma of Stein [1986] is the basis for the approximations used here.

Lemma 3.1 (Commutation of the Diagram) *When the sets of the diagram are vector spaces and the functions linear and when the following conditions are fulfilled:*

- $E \circ T = 0$
- $\iota_0 \circ E_0 + T_0 \circ U_0 = I_{\mathcal{X}_0}$

- $E \circ \beta \circ \iota_0 \circ E_0 = \gamma \circ E_0$

Then we can write:

$$E \circ \beta - \gamma \circ E_0 = E \circ (T \circ \alpha - \beta \circ T_0) \circ U_0$$

It is often possible to bound the right hand side of this equation.

3.4 Birth and Death Chains

3.4.1 Exchangeable Pairs

We will start with the case of a random variable taking its values in $\{0, \ldots, n\}$. We suppose that this random variable W has a distribution:

$$P(W = k) = p_k, \text{ for } 0 \leq k \leq n$$

and a mean denoted

$$\mu = EW = \sum_{k=0}^{n} p_k k$$

We would like to find a W' such that:

- (W, W') is exchangeable
- a general contraction property is satisfied, i. e.

  (3.10) $$\boxed{E^W W' - W = -\lambda(W - \mu), \text{ with } 0 < \lambda < 1}.$$

  This is the generalization of (2.5).
- W and W' differ by at most 1 (thus forming a birth and death chain).

These conditions specify enough equations so that we can define as follows:

$$\begin{aligned} P(W' = W + 1 | W = k) &= \beta_k \\ P(W' = W - 1 | W = k) &= \delta_k \\ \beta_n = 0 \text{ and } \delta_0 &= 0 \\ \delta_k + \beta_k &\leq 1 \end{aligned}$$

(W, W') exchangeable implies: $$\begin{aligned} \beta_0 p_0 &= \delta_1 p_1 \\ \beta_1 p_1 &= \delta_2 p_2 \\ \beta_k p_k &= \delta_{k+1} p_{k+1} \\ \vdots \quad & \quad \vdots \end{aligned}$$

We rewrite the contraction property:

$$\beta_k - \delta_k = -\lambda(k - \mu)$$

δ_k and β_k have to be of the form:

(3.11) $$\begin{aligned} \beta_0 &= \lambda\mu \\ \delta_k &= -\frac{\lambda}{p_k} \sum_{j=0}^{k-1} p_j (j - \mu), \quad 1 \leq k \leq n \\ \beta_k &= -\frac{\lambda}{p_k} \sum_{j=0}^{k} p_j (j - \mu), \quad 1 \leq k \leq n - 1 \end{aligned}$$

In order for this to be possible λ must satisfy $0 \leq \delta_k + \beta_k \leq 1$. This is equivalent to

$$0 \leq \lambda \leq \frac{-p_k}{2\sum_{j=0}^{k-1} p_j(j-\mu) + p_k(k-\mu)}$$

3.4.2 A generalization of Todhunter's formula

Mills ratio type bounds for binomial tail probabilities can be derived from the following formula due to Todhunter (see Diaconis and Zabell (1991)). We can generalize this idea to give bounds for the stationary distribution of the birth and death chains constructed above.

$$\sum_{\ell}^{m}(i-np)p_i = (1-p)\ell p_\ell - (n-m)p_m, \qquad \forall m, \forall \ell$$

The definition of the birth rate in the above birth and death chains enables us to write:

$$\sum_{\ell}^{m}(i-\mu)p_i = \frac{1}{\lambda}(\beta_{\ell+1}p_{\ell+1} - \beta_m p_m)$$

We will now look at how this can be used in examples.

Uniform Distribution on $\{0,\ldots,n\}$

In this case we have:

$$\begin{aligned}
p_k &= \frac{1}{n+1} \\
\mu &= \frac{n}{2} \\
\beta_k &= \frac{\lambda}{2}(k+1)(n-k) \\
\delta_k &= \frac{\lambda}{2}k(n-k+1) \\
\text{If } n \text{ is even we must have:} \lambda &\leq \frac{4}{n^2+2n} \\
\text{If } n \text{ is odd we must have:} \lambda &\leq \frac{4}{n^2+2n-1}
\end{aligned}$$

In the appendix some of the numerical simulations show how the value of λ influences the speed of convergence to stationarity.

Note the "standard birth and death chain with a uniform stationary distribution is the random walk on a path with holding $\frac{1}{2}$ at each end. This does not give $E(W) = (1-\lambda)W$ for any λ.

Binomial Distribution

The algebraic construction obtained through the above formula gives exactly the exchangeable pair we find for the binomial by using the construction:

Define the exchangeable pair (W, W') as follows:

- Write $W = \sum_{i=1}^{n} X_i$, sum of independent Bernoulli variables with $p = P(X_i = 1)$
- Choose a random I uniformly in $\{1\ldots n\}$

- X_I is changed into X'_I with $P(X'_I = 1) = p$
- $W' = W - X_I + X'_I$

The computations then give in this case:$\beta_k = p \times (n-k)$ and $\delta_k = q \times k$

Poisson Distribution

In our development we have not used the fact that the random variable is bounded. By induction we can generalize the definition of β and δ to $\mathbb{N}$. For example if the stationary distribution is Poisson:

$$p_k = \frac{\mu^k e^{-\mu}}{k!} \qquad \text{and } \frac{p_j}{p_{j+1}} = \frac{j+1}{\mu}$$

By induction as before:

$$\begin{aligned}
\beta_0 &= \lambda\mu \\
\delta_1 &= \frac{p_0}{p_1}\beta_0 = \lambda \\
\beta_1 &= \delta_1 - \lambda(1-\mu) = \lambda\mu \\
\delta_2 &= \frac{p_1}{p_2}\beta_1 = 2\lambda \\
\beta_2 &= \delta_2 - \lambda(2-\mu) = \lambda\mu \\
\vdots &\quad \vdots \\
\delta_k &= k\lambda \\
\beta_k &= \lambda\mu
\end{aligned}$$

Hypergeometric Distribution

This construction provides the same birth and death chain the exchangeable pair did, with $\lambda = \frac{1}{N-1}$. The general form gives in this case:

$$\begin{aligned}
\delta_k &= \lambda k \frac{(N-R-(n-k))}{N} \\
\beta_k &= \lambda(\frac{(n-k)(R-k)}{N})
\end{aligned}$$

3.4.3 Characterizing Operators

From any function f we build an antisymmetric function (αf) defined 'locally' as:

$$(\alpha f)(w,w') = \frac{1}{\lambda}\left(f(w')\mathbb{I}_{w'=w+1} - f(w)\mathbb{I}_{w=w'+1}\right)$$

For T, we take an exchangeable pair (w,w') and an antisymmetric F and define: $(TF)(w) = E^w F(w,w')$ so that in this case:

$$T(\alpha f)(w) = E^w(\alpha f)(w,w') = \frac{1}{\lambda}\left(\beta_w f(w+1) - \delta_w f(w)\right). \tag{3.12}$$

Because of the exchangeability and the antisymmetry we will have $ETf = 0$, for all f so that:

$$ImT \subset KerE.$$

Further, if the birth and death chain is connected $ImT = KerE$ and $T\alpha$ is a characterizing operator. In the four examples considered above this gives:

Uniform

$$T\alpha f(w) = (w+1)(n-w)f(w+1) - w(n-w+1)f(w)$$

Binomial(n,p)

$$T\alpha f(w) = p \times (n-w)f(w+1) - q \times wf(w)$$

Poisson(μ)

$$T\alpha f(w) = \mu f(k+1) - kf(k)$$

Hypergeometric((n, N, R))

$$T\alpha f(w) = \frac{(n-k)(R-k)}{N} f(k+1) - \frac{k((N-R)-(n-k))}{N} f(k)$$

It is sometimes a good idea, given the contraction property to rewrite (2.12) as follows:

(3.13) $$T(\alpha f)(w) = \frac{\beta_w}{\lambda}\Delta f(w) + \frac{\beta_w - \delta_w}{\lambda} f(w) = \frac{\beta_w}{\lambda}\Delta f(w) - (w-\mu)f(w).$$

Because then, in comparisons between two birth and death chains whose means are equal, the second part of the right hand side cancels. The following section presents a few specific examples. We will return to the general birth and death chains and the definition of the inverse to $T\alpha$, and its bounds in the section 5.3.

3.4.4 Examples

Comparison of the binomial and Poisson

Just to illustrate how the machinery we installed works formally, we can turn over the first example, taking the target to be Poisson. We will show how the algebraic lemma and the properties of the pseudo-inverse g of an indicator function $\mathbb{I}_A$ provide bounds for the distances between these two distributions.

Let's define the elements of the diagram. In this case the target is Poisson with mean np (largely developed in the book by Barbour, Holst and Jansen (1992)) the characterizing operator is:

(3.14) $$T_0 f(w) = npf(w+1) - wf(w).$$

We will take the binomial characterization obtained above

(3.15) $$T(\alpha f)(w) = p(n-w)f(w+1) - w(1-p)f(w).$$

$\mathcal{X}_O$ is the space of functions $\mathbb{N} \longrightarrow \mathbb{R}$ having at most exponential increase,
$\mathcal{F}_0 = \mathcal{X}_0 \cap \{f : f(0) = 0\}$. $\mathcal{X}$ the same as $\mathcal{X}_0$ but restricted to functions defined on $\{0...n\}$. β is the relevant restriction function:

$$\beta f(w) = \begin{cases} f(w) & \text{if } w \le n \\ 0 & \text{if } w > n \end{cases}$$

By taking $f = \mathbb{I}_k$ and $U_0 f = g$ defined such that:

$$T_0 \circ U_0(\mathbb{I}_k) = \mathbb{I}_k - p_\mu(k)$$
$$npg(w+1) - wg(w) = f(w) - E_0 f = f(w) - E_0 f, \forall w.$$

Lemma 3.2 (Bound on the pseudo-inverse and its increase) *For g_A the solution to the equation:*

$$\mu g_A(w+1) - w g_A(w) = \mathbb{I}_A(w) - Po(A),$$

we have the bounds:

$$\|g\| = \sup_j g(j) \le \min(1, \frac{1}{\sqrt{\mu}}) \tag{3.16}$$

$$\Delta g = \sup_j |g(j+1) - g(j)| \le \min(1, \frac{1}{\mu}). \tag{3.17}$$

For a proof one can look at Barbour, Holst and Jansen (page 7 and page 223).

Then, the algebraic lemma implies that for any set A and function g_A defined as above:

$$\begin{aligned} |P(A) - Po(A)| &= E[T\alpha g_A - T_0 g_A] \\ &= E[npg(w+1) - wg(w) - npg(w+1) \\ &\qquad + wpg(w+1) + wg(w) - wpg(w)] \\ &= E[wp(g(w+1) - g(w))] \\ &\le E(wp)\Delta \\ &\le np^2\Delta \le np^2 \min(1, \frac{1}{\mu}). \end{aligned}$$

This is sharper than the result in remark 3 of Section 3.2.1.

The number of ones in the binary expansion of an integer

This is an example treated in different ways by Diaconis (1977), Stein (1986) and Barbour and Chen (1992). This presentations follows the first two authors closely.

Let n be fixed. Choose uniformly an integer X between 0 and n. We want to study:
W = Number of 1's in the binary expansion of X.
Let's write this expansion: $X_m X_{m-1} \dots X_1$ with m the maximal number of possible digits that X could take:

$$m = [\log_2 n] + 1$$

following Diaconis (1977) we will call $Q(x)$ the number of 0's in x's binary expansion which can't be changed without making the new number bigger than n. For instance $Q(17) = 2$ if $n = 23$.

Exchangeable Pair

Choose I uniformly in $\{0 \dots m\}$. Change X_I into its contrary as long as this doesn't make the new integer larger than n.

$$\begin{cases} W' &= W - X_I + (1 - X_I) \quad \text{if } X + (1-2X_I)2^{I-1} \le n \\ W' &= W \quad \text{otherwise} \end{cases}$$

(W, W') is exchangeable, and this example is the first we define that is a birth and death chain.

$$\begin{aligned} EE^W(W' - W) &= 0. \\ E^W(W' - W) &= \frac{m - W - Q}{m} - \frac{W}{m} \end{aligned} \tag{3.18}$$

$$
\begin{aligned}
(2.18) \quad &\Longrightarrow \quad E\left(\frac{m-2W-Q}{m}\right)=0 \\
&\Longrightarrow \quad E(W)=\frac{1}{2}(m-E(Q))
\end{aligned}
$$

The function $(w,w') \longrightarrow (w'-w)(w+w')$ being antisymmetric, we have

$$
\begin{aligned}
&(3.19) & EE^W(W'^2-W^2) &= 0 \\
&(3.20) & \text{Thus var } (W) &= \frac{1}{2}EE^W(W'-W)^2 \\
&(3.21) & \text{as } E^W(W'-W)^2 &= \frac{m-Q}{m}.
\end{aligned}
$$

We will take for our operator $T\alpha$:

$$
\begin{aligned}
T(\alpha f)(w) &= \frac{\beta_w f(w+1)-\delta_w f(w)}{\frac{2}{m}} \\
&= \frac{m}{2}\left(\frac{m-w-Q}{m}f(w+1)-\frac{w}{m}f(w)\right) \\
&= \frac{m-w}{2}f(w+1)-\frac{w}{2}f(w)-\frac{Q}{2}f(w+1) \\
&= T_0 f(w)-\frac{Q}{2}f(w+1) \\
\text{Where } T_0 f(w) &= \frac{m-w}{2}f(w+1)-\frac{w}{2}f(w)
\end{aligned}
$$

is the *characterizing operator* of the binomial $\mathcal{B}(m,\frac{1}{2})$.

For g the solution to

$$
\mathbb{I}_k - P_0(k) = \frac{m-w}{2}g(w+1)-\frac{w}{2}g(w)
$$

Stein (1980) shows $|g(w)| \leq \frac{4}{m}$.

$$
\begin{aligned}
\text{And } P(Q>k) &\leq \frac{1}{2^k} = P(X \geq n-2^{m-k}) \\
\text{implies } EQ &= \sum_{k=0}^{\infty} P(Q>k) \leq \frac{1}{1-\frac{1}{2}} = 2
\end{aligned}
$$

$$
\text{Therefore} \qquad |p(k)-P_0(k)| \leq \frac{4}{m}
$$

Contingency Tables

Diaconis and Saloff-Coste(1996) take the following example to show how Nash inequalities can be used to bound rates of convergence of Markov Chains.

Call $\mathcal{M}_n^2$ the set of all $n \times n$ contingency tables whose margins are all equal to 2. We are going to consider W= Number of 2's in M, a table chosen uniformly among tables of $\mathcal{M}_n^2$. For n large these are sparse tables with 2 a rare event.

In this case we will start by creating an approximate birth and death chain through construction of an exchangeable pair, this will make clear what the mean and variance are. Seeing that they are equal points to a Poisson target. We then explore the distance to the Poisson using the bound we have on the inverse to the Poisson characterizing operator.

Exchangeable Pair

We will use the pair (M', M) constructed as a reversible Markov chain for generating uniformly such tables as our basis for the exchangeable pair (W', W).

Note: When we have a procedure for generating a reversible Markov chain, we will always have an exchangeable pair. See Chapter 1 of this book.

- Choose a pair of different rows at random
- Choose a pair of different columns at random
- As long as it doesn't make any table value negative make the following change to the 2 by 2 square thus defined: $\begin{pmatrix} + & - \\ - & + \end{pmatrix}$ or $\begin{pmatrix} - & + \\ + & - \end{pmatrix}$ choosing one of the above with probability $\frac{1}{2}$. Otherwise the chain stays at the original table.

An exchangeable pair (W', W) is thus defined naturally from the pair (M, M').

Let's compute $\beta_w = P(W' = W + 1|W)$, the probability that the number of 2's increases by 1. For that to happen a configuration of the $\begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix}$ must be chosen as the 2 by 2 square, this can be decomposed into the product:

-Probability of choosing two columns without any 2's.

$$\frac{(n-w)(n-w-1)}{(n-1)n}$$

-and the probability of choosing the two 1's among n, when there are only two of them:

$$\frac{2}{n(n-1)},$$

-and the probability that the second column is $(0, 1)$:

$$\frac{(n-2)}{(n-1)}\frac{1}{(n-1)}.$$

There are four configurations of this type (four positions of 0's), only half of which will be compatible with the choice of $+$ and $-$ patterns to enable a step, thus:

$$\beta_w = \frac{4(n-w)(n-w-1)(n-2)}{n^2(n-1)^4} = \frac{4}{n(n-1)^2}[(1-\frac{w}{n})(1-\frac{w}{n-1})(1-\frac{1}{n-1})]$$

For the probability that the number of 2's to decreases by 1, we look for the probability of a configuration of a: $\begin{vmatrix} 2 & 0 \\ 0 & 1 \end{vmatrix}$ configuration.

By a similar decomposition as above, this configuration has probability

$$\frac{w}{n} \times \frac{(n-w)}{(n-1)} \times \frac{1}{n} \times \frac{2}{n-1}$$

Of which only two out of four will produce a move, thus:

$$\delta_w = P(W' = w-1|W=w) = \frac{4w(n-w)}{n^2(n-1)^2}$$

Note that

$$P(W' = w-2|W=w) = \frac{w(w-1)}{2n^2(n-1)^2}$$

is of an order n^{-1} smaller, we are going to ignore it, as well as

$$P(W' = w+2|W=w) = \frac{4(n-w)(n-w-1)}{n^3(n-1)^3}.$$

In fact, if the original chain is modified to hold when W jumps by two, the following calculations are all valid. We can start by computing the mean simply by exchangeability:

$$\begin{aligned}
EE^wW' - W &= E(\beta_w - \delta_w) = 0 \\
\beta_w - \delta_w &= \frac{4(n-w)}{n^2(n-1)^2}\left[\frac{(n-w-1)(n-2)}{(n-1)^2} - w\right] \\
&= \frac{4(n-w)}{n^3(n-1)^3}\left[(2-w)n^2 - n(w+6) + 4(w+1)\right] \\
&= \frac{4}{n(n-1)^2}\left[-(w-1) - \frac{w(1-w)}{n} - \frac{(1+w)}{n-1}\frac{\cdots}{n^2}\right]
\end{aligned}$$

thus $E(\beta_w - \delta_w) \doteq 0$ implies $E(W) \doteq 1$.

We remark that:

$$\beta_w - \delta_w \doteq \frac{4}{n(n-1)^2}[-(w-1)]$$

providing an approximate contraction property

$$\begin{aligned}
E^wW' - W &\doteq -\frac{1}{\lambda}(W - E(W)) \\
\text{with } \lambda &= \frac{4}{n(n-1)^2}
\end{aligned}$$

For the variance we can use the fact that:

$$\begin{aligned}
E(W'^2 - W^2) = E(W'-W)(W'+W) &= 0 \text{ by antisymmetry} \\
E^w(W'-W)^2 &= E^w(W'^2 - W^2) + 2WE^w(W - W') \\
&= E^w(W'^2 - W^2) - 2WE^w(W' - W) \\
\text{thus var}(W) &\doteq \frac{(W - E(W))}{2(\beta_w - \delta_w)}EE^w(W'-W)^2 \\
\text{Then } EE^w(W'-W)^2 &\doteq 2\lambda EW(W - E(W)) \\
\text{But } E^w(W'-W)^2 &\doteq \frac{4}{n(n-1)^2}\left[(w+1) - \frac{w}{n}(w+2) + \frac{\cdots}{n^2}\right] \\
\text{var}(W) &\doteq \frac{1}{2}E(w+1) \doteq 1
\end{aligned}$$

As usual we define:

$$(\alpha f)(w,w') = \frac{1}{\lambda}\left[\mathbb{I}_{w'=w+1} f(w') - \mathbb{I}_{w=w'+1} f(w)\right]$$

$$T(\alpha f)(w) = E^w(\alpha f)(w,w') = \frac{\beta_w}{\lambda} f(w+1) - \frac{\delta_w}{\lambda} f(w)$$

As suggested by the first two moments, a Poisson(1) approximation seems appropriate, so we are going to compare the operator constructed above with the Poisson(1) operator: $T_0 f(w) = 1 \times f(w+1) - wf(w)$:

$$\begin{aligned}
T\alpha(f)(w) - T_0 f(w) &= (\frac{\beta_w}{\lambda} - 1)\Delta f(w) + (\frac{\beta_w - \delta_w}{\lambda}) f(w) + (w-1) f(w) \\
\frac{\beta_w}{\lambda} - 1 &= -\frac{w}{n} - \frac{w}{n-1} - \frac{1}{n-1} + \frac{\cdots}{n^2} \\
\frac{\beta_w - \delta_w}{\lambda} &\doteq \left[-(w-1) + \frac{w(w-2)}{n}\right] \\
T\alpha(f)(w) - T_0 f(w) &= -\frac{2w+1}{n}\Delta f(w) - \frac{w(w-2)}{n} f(w)
\end{aligned}$$

The pseudo-inverse f and its increase are both bounded by 1. To bound the total variation distance between these two measures, for any set A its measure is the expectation of the indicator $\mathbb{I}_A$, call the Poisson one $Po(A)$ and denote be $g_A = U_0 \mathbb{I}_A$ the solution to the equation:

$$T_0 g = 1 \times g_A(w+1) - w g_A(w) = \mathbb{I}_A - Po(A).$$

Lemma 5 of Barbour, Holst and Janson (1992) provides, as shown above:

$$\Delta g_A \leq 1 \text{ and } \|g\| \leq 1.$$

Thus bounding the expectation of $T\alpha(g_A)(w) - T_0 g_A(w)$ gives:

$$d_{TV}(P, Po) \leq |E(\frac{2w+1}{n}) + E(\frac{(w-2)w}{n})| \leq \frac{3}{n} + O(\frac{1}{n^2}).$$

3.5 General Discrete Target distribution

This section is the discrete version of chapter 6 of Stein (1986) which he suggests for development in the section on open problems.

For a given target distribution (2.14) provides a general form of characterizing operator. In order for the method to be useful we need to define and bound the inverse of some very specific test functions such as $\mathbb{I}_A - p_A$.

First we will define the inverse, then we will give conditions on the stationary distribution that will ensure that the increase in the solution is bounded. This section concludes with a large class of new examples, related to distance regular graphs, where our conditions are satisfied.

Pseudo-Inverse for T_0

Suppose that the target distribution is also defined from a birth and death chain (δ_k, β_k defined in (2.11)) that is we have T_0 of the form :

$$T_0 f(w) = \frac{1}{\lambda}\left(\beta_w f(w+1) - \delta_w f(w)\right) \tag{3.22}$$

Given a function g defined on $\{0 \dots n\}$, how and when can we define its inverse by T_0? This can be reduced to a set of recurrence equations, setting $f_k = f(k)$ and $g_k = g(k)$, we want:

$$\left\{ \begin{array}{ll} \beta_k f_{k+1} - \delta_k f_k = \lambda g_k & 0 < k < n \\ \beta_0 f_1 = \lambda g_0 & \\ -\delta_n f_n = \lambda g_n & \end{array} \right. \tag{3.23}$$

The last condition of the recurrence implies:

$$f_{n-k} = -\frac{1}{\delta_{n-k}} \left(\frac{\beta_{n-k}}{\delta_{n-k+1}} \frac{\beta_{n-k+1}}{\delta_{n-k+2}} \cdots \frac{\beta_{n-2}}{\delta_{n-1}} \frac{\beta_{n-1}}{\delta_n} g_n + \cdots g_{n-k} \right).$$

Exchangeability of the chain imposes :

$$\beta_k p_k = \delta_{k+1} p_{k+1}.$$

From this we can do the simplifications :

$$\frac{\beta_1}{\delta_2} \frac{\beta_2}{\delta_3} \cdots \frac{\beta_{n-1}}{\delta_n} = \frac{p_n}{p_1}$$

In particular

$$f_1 = -\frac{\lambda}{\delta_1} \left(\frac{p_n}{p_1} g_n + \frac{p_{n-1}}{p_1} g_{n-1} + \cdots + g_1 \right).$$

Coherent with the initial condition $\beta_0 f_1 = g_0$ if

$$\sum_{k=0}^{n} p_k g_k = 0.$$

When this is fulfilled the general form of the inverse is:

$$f_k = -\frac{\lambda}{\delta_k p_k} (p_k g_k + \dots p_n g_n) = \frac{(p_k g_k + \dots p_n g_n)}{\sum_{j=k}^{n} p_j (j - \mu)} \tag{3.24}$$

Because of the definition (2.11) of δ_k. For such a definition, (Barbour, Holst and Janson (1992), page 189, Lemma 9.2.1) give a general bound for Δf, under the condition that the β_k are non-increasing and the δ_k non-decreasing. This bound is valid for the inverse of the indicator of any set A : If f satisfies $T_0 f(k) = \mathbb{I}_A - p(A)$ then

$$\Delta f = \max_j |f(j+1) - f(j)| \leq \max_j \min(\frac{\lambda}{\beta_j}, \frac{\lambda}{\delta_j}). \tag{3.25}$$

Again taking into account the definitions of β_k and δ_k, we have

$$\begin{aligned} \frac{\lambda}{\delta_j} &= \frac{-p_k}{\sum_{j=0}^{k-1} p_j (j-\mu)} = \frac{p_k}{\sum_{j=k}^{n} p_j (j-\mu)} \\ \frac{\lambda}{\beta_k} &= \frac{-p_k}{\sum_{j=0}^{k} p_j (j-\mu)} = \frac{p_k}{\sum_{j=k+1}^{n} p_j (j-\mu)} \end{aligned}$$

We define the pseudo inverse of any function f by

$$g(k) = Uf(k) = \frac{\lambda}{\beta_{k-1} p_{k-1}} \sum_{i=0} p_j (f(j) - Ef)$$

It is easy to check that such a g satisfies: $Tg(k) = f(k) - Ef$ as before. Suppose the test function f of interest is the indicator function $f = \mathbb{I}_{\{k_0\}}$. In this case the expectation of f will be $Ef = p(k_0) = p_{k_0}$ and

$$U\mathbb{I}_{\{k_0\}}(k) = \begin{cases} -\frac{\lambda p_{k_0}}{\beta_{k-1}p_{k-1}} \sum_0^{k-1} p_j, & k \leq k_0 \\ \frac{\lambda p_{k_0}}{\beta_{k-1}p_{k-1}} \sum_{j=k}^{n} p_j, & k \leq k_0 \end{cases} .$$

We know that if we match up the means of the distributions, we only need to bound the first order difference of $U\mathbb{I}_{k_0}$, which we will denote by $\Delta U\mathbb{I}_{k_0} = U\mathbb{I}_{k_0}(k+1) - U\mathbb{I}_{k_0}(k)$.

There are three possible cases for the form that this can take on, depending on where k is situated with regards to k_0:

1. $$\text{If } k < k_0, \text{ then } \Delta U\mathbb{I}_{k_0}(k) = -\lambda p_{k_0}\left(\frac{S_k}{\beta_k p_k} - \frac{S_{k-1}}{\beta_{k-1}p_{k-1}}\right)$$

2. $$\Delta U\mathbb{I}_{k_0}(k_0) = -\lambda p_{k_0}\left(\frac{(1 - S_{k_0}}{\beta_{k_0} p_{k_0}} + \frac{S_{k_0-1}}{\beta_{k_0-1}p_{k_0-1}}\right)$$

3. $$\text{If } k > k_0, \text{ then } \Delta U\mathbb{I}_{k_0}(k) = \lambda p_{k_0}\left(\frac{(1 - S_k)}{\beta_k p_k} - \frac{(1 - S_{k-1})}{\beta_{k-1}p_{k-1}}\right)$$

Proposition 3.5.1 *For β_k decreasing and δ_k increasing, then the only case where $\Delta U\mathbb{I}_{k_0}(k) > 0$ is for $k = k_0$.*

Proof.

We are going to look at :

$$\left(\frac{S_k}{\beta_k p_k} - \frac{S_{k-1}}{\beta_{k-1}p_{k-1}}\right)$$

and prove that it is always positive.

$$\begin{aligned}
\left(\frac{S_k}{\beta_k p_k} - \frac{S_{k-1}}{\beta_{k-1}p_{k-1}}\right) &= \sum_{j=0}^{k} \frac{p_j}{p_k \beta_k} - \sum_{j=0}^{k-1} \frac{p_j}{\beta_{k-1}p_{k-1}} \\
\overset{=}{exch.} & \quad \frac{p_0}{p_k\beta_k} + \sum_1^k \left(\frac{p_j}{p_k\beta_k} - \frac{p_{j-1}}{p_k\delta_k}\right) \\
&= \frac{p_0}{p_k\beta_k} + \frac{1}{p_k\beta_k}\sum_1^k p_j\left(1 - \frac{p_{j-1}}{p_j}\frac{\beta_k}{\delta_k}\right) \\
&= \frac{p_0}{p_k\beta_k} + \frac{1}{p_k\beta_k}\sum_1^k p_j\left(1 - \frac{\delta_j}{\beta_{j-1}}\frac{\beta_k}{\delta_k}\right)
\end{aligned}$$

Under the monotonicity conditions above for β_k and δ_k, this last parenthesis on the right will always be positive, thus proving that $\Delta\mathbb{I}_{k_0}(k) < 0$ for all $k < k_0$.

A very similar argument gives the same result in case $k > k_0$ ■

Corollary 3.5.1 *For β_k decreasing and δ_k increasing and for any subset $A \subset \{0, 1, 2, \ldots, n\}$:*

$$|\Delta U\mathbb{I}_{k_0}(A)| \leq \Delta U\mathbb{I}_{k_0}(k_0).$$

Proof.

$$\sum_{j=0}^{n-1} \Delta U\mathbb{I}_{k_0}(j) = U\mathbb{I}_{k_0}(n) - U\mathbb{I}_{k_0}(0) = \lambda \frac{p_{k_0} p_n}{\beta_{n-1} p_{n-1}} = \lambda \frac{p_{k_0}}{\delta_n} > 0$$

So the overall sum of the sequence is positive, thus the one positive element is larger than any combination of the others. ■

A large class of examples of distributions on $\{0, 1, 2, \ldots d\}$ where the appropriate monotonicity conditions for a natural birth and death chain are satisfied is the class of distance regular graphs. These are connected graphs γ with vertex set Ω. Let $d(x, y)$ be the graph distance between vertices x and y. Let $[(i, x)]$ be the vertices at distance i from $X_i, 0 \leq i \leq d$, with d the diameter of the graph. Then γ is *distance regular* if there are numbers $c_i, a_i, b_i, 0 \leq i \leq d$ such that if $d(x, y) = i$, then the number of neightbors of y which lie at distance $i - 1, i, i + 1$ from x are c_i, a_i, b_i the nearest neighbor random walk on a distance regular graph generates a birth and death chain by looking at the distance from the starting state. This chain has stationary distribution $\pi(i)$, proportional to $\frac{b_0 b_1 \ldots b_{i-1}}{c_1 c_2 \ldots c_i}$. A theorem of Smith Smith, D. [1971] says that the birth and death rates satisfy $c_1 \leq c_2 \ldots \leq c_d$ and $b_0 \geq b_1 \geq \ldots \geq b_{d-1}$ (note that c_0 and b_d are undefined for distance regular graphs). This result shows that our bounds on the inverse are in force for all of these birth and death chains.

The classification of distance regular graphs is one of the most active topics in algebraic combinatorics. Well-known examples include the hypercube (with binomial stationary distribution) and the k-sets of an n-set (with hypergeometric stationary distribution). For a splendid introduction to the subject see Cameron (1999) chapter three. The definite work on the subject is by Brouwer, Cohen and Neumaier (1984). This contains hundreds of families of examples. Andries Brouwer (`www.win.tue.nl/˜aeb`) maintains a website dedicated to this subject.

3.6 Appendix:Some Numbers

Here is the matrix of the birth and death chain that converges to uniform, with $\lambda = 0.08$, $n = 6$.

```
 bd2(n = 7, lambda = 0.08)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 0.76 0.24    0    0    0    0    0
[2,] 0.24 0.36 0.40    0    0    0    0
[3,]    0 0.40 0.12 0.48    0    0    0
[4,]    0    0 0.48 0.04 0.48    0    0
[5,]    0    0    0 0.48 0.12 0.40    0
[6,]    0    0    0    0 0.40 0.36 0.24
[7,]    0    0    0    0    0 0.24 0.76
```

Here are a few powers showing how long it takes to converge:

```
puissance(bd2(n = 7, lambda = 0.08),2ˆ5)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.1652 0.1577 0.1503 0.1428 0.1354 0.1280 0.1206
[2,] 0.1577 0.1528 0.1478 0.1429 0.1379 0.1329 0.1280
[3,] 0.1503 0.1478 0.1454 0.1429 0.1404 0.1379 0.1354
[4,] 0.1428 0.1429 0.1429 0.1429 0.1429 0.1429 0.1428
[5,] 0.1354 0.1379 0.1404 0.1429 0.1454 0.1478 0.1503
[6,] 0.1280 0.1329 0.1379 0.1429 0.1478 0.1528 0.1577
[7,] 0.1206 0.1280 0.1354 0.1428 0.1503 0.1577 0.1652
puissance(bd2(n = 7, lambda = 0.08),2ˆ6)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.1444 0.1439 0.1434 0.1429 0.1423 0.1418 0.1413
```

```
[2,] 0.1439 0.1435 0.1432 0.1429 0.1425 0.1422 0.1418
[3,] 0.1434 0.1432 0.1430 0.1429 0.1427 0.1425 0.1423
[4,] 0.1429 0.1429 0.1429 0.1429 0.1429 0.1429 0.1429
[5,] 0.1423 0.1425 0.1427 0.1429 0.1430 0.1432 0.1434
[6,] 0.1418 0.1422 0.1425 0.1429 0.1432 0.1435 0.1439
[7,] 0.1413 0.1418 0.1423 0.1429 0.1434 0.1439 0.1444
```

For a smaller λ, it's slower :

```
bd2(n = 7, lambda = 0.04)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 0.88 0.12    0    0    0    0    0
[2,] 0.12 0.68 0.20    0    0    0    0
[3,]    0 0.20 0.56 0.24    0    0    0
[4,]    0    0 0.24 0.52 0.24    0    0
[5,]    0    0    0 0.24 0.56 0.20    0
[6,]    0    0    0    0 0.20 0.68 0.12
[7,]    0    0    0    0    0 0.12 0.88
 puissance(bd2(n = 7, lambda = 0.04),2^6)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.1665 0.1586 0.1507 0.1428 0.1349 0.1271 0.1194
[2,] 0.1586 0.1533 0.1481 0.1429 0.1376 0.1324 0.1271
[3,] 0.1507 0.1481 0.1455 0.1429 0.1403 0.1376 0.1349
[4,] 0.1428 0.1429 0.1429 0.1429 0.1429 0.1429 0.1428
[5,] 0.1349 0.1376 0.1403 0.1429 0.1455 0.1481 0.1507
[6,] 0.1271 0.1324 0.1376 0.1429 0.1481 0.1533 0.1586
[7,] 0.1194 0.1271 0.1349 0.1428 0.1507 0.1586 0.1665
puissance(bd2(n = 7, lambda = 0.04),2^7)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.1446 0.1440 0.1434 0.1429 0.1423 0.1417 0.1411
[2,] 0.1440 0.1436 0.1432 0.1429 0.1425 0.1421 0.1417
[3,] 0.1434 0.1432 0.1430 0.1429 0.1427 0.1425 0.1423
[4,] 0.1429 0.1429 0.1429 0.1429 0.1429 0.1429 0.1429
[5,] 0.1423 0.1425 0.1427 0.1429 0.1430 0.1432 0.1434
[6,] 0.1417 0.1421 0.1425 0.1429 0.1432 0.1436 0.1440
[7,] 0.1411 0.1417 0.1423 0.1429 0.1434 0.1440 0.1446
```

For n=10, 11 possible values and $\lambda = 0.03 < 1/30$

```
round(bd2(n=11,lambda=0.03),3)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
 [1,] 0.85 0.15 0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00
 [2,] 0.15 0.58 0.27 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00
 [3,] 0.00 0.27 0.37 0.36 0.00 0.00 0.00 0.00 0.00  0.00  0.00
 [4,] 0.00 0.00 0.36 0.22 0.42 0.00 0.00 0.00 0.00  0.00  0.00
 [5,] 0.00 0.00 0.00 0.42 0.13 0.45 0.00 0.00 0.00  0.00  0.00
 [6,] 0.00 0.00 0.00 0.00 0.45 0.10 0.45 0.00 0.00  0.00  0.00
 [7,] 0.00 0.00 0.00 0.00 0.00 0.45 0.13 0.42 0.00  0.00  0.00
 [8,] 0.00 0.00 0.00 0.00 0.00 0.00 0.42 0.22 0.36  0.00  0.00
 [9,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.36 0.37  0.27  0.00
[10,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.27  0.58  0.15
[11,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.15  0.85

 round(puissance(bd2(n=11,lambda=0.03),8),3)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
```

```
 [1,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [2,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [3,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [4,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [5,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [6,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [7,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [8,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [9,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
[10,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
[11,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091

> round(puissance(bd2(n=11,lambda=0.03),7),3)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
 [1,] 0.096 0.095 0.094 0.093 0.092 0.091 0.090 0.089 0.088 0.087 0.086
 [2,] 0.095 0.094 0.093 0.092 0.092 0.091 0.090 0.089 0.089 0.088 0.087
 [3,] 0.094 0.093 0.093 0.092 0.091 0.091 0.090 0.090 0.089 0.089 0.088
 [4,] 0.093 0.092 0.092 0.092 0.091 0.091 0.091 0.090 0.090 0.089 0.089
 [5,] 0.092 0.092 0.091 0.091 0.091 0.091 0.091 0.091 0.090 0.090 0.090
 [6,] 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091 0.091
 [7,] 0.090 0.090 0.090 0.091 0.091 0.091 0.091 0.091 0.091 0.092 0.092
 [8,] 0.089 0.089 0.090 0.090 0.091 0.091 0.091 0.092 0.092 0.092 0.093
 [9,] 0.088 0.089 0.089 0.090 0.090 0.091 0.091 0.092 0.093 0.093 0.094
[10,] 0.087 0.088 0.089 0.089 0.090 0.091 0.092 0.092 0.093 0.094 0.095
[11,] 0.086 0.087 0.088 0.089 0.090 0.091 0.092 0.093 0.094 0.095 0.096
```

Here is the bd chain for the hypergeometric:

```
pi.hyper <- dhyper(0:5, 5, 7, 5)
pi.hyper
 0.0265 0.221 0.442 0.265 0.0442 0.00126

puissance(bd2(n=6,p=pi.hyper,lambda=1/3),3)
       [,1]  [,2]  [,3]  [,4]   [,5]     [,6]
[1,] 0.0325 0.246 0.445 0.240 0.0355 0.000892
[2,] 0.0296 0.234 0.444 0.252 0.0396 0.001063
[3,] 0.0267 0.222 0.442 0.264 0.0438 0.001244
[4,] 0.0240 0.210 0.440 0.276 0.0481 0.001433
[5,] 0.0213 0.198 0.438 0.289 0.0526 0.001632
[6,] 0.0187 0.186 0.435 0.301 0.0571 0.001840

 puissance(bd2(n=6,p=pi.hyper,lambda=1/3),4)
       [,1]  [,2]  [,3]  [,4]   [,5]    [,6]
[1,] 0.0267 0.222 0.442 0.264 0.0438 0.00125
[2,] 0.0266 0.221 0.442 0.265 0.0440 0.00125
[3,] 0.0265 0.221 0.442 0.265 0.0442 0.00126
[4,] 0.0264 0.221 0.442 0.266 0.0443 0.00127
[5,] 0.0263 0.220 0.442 0.266 0.0445 0.00128
[6,] 0.0262 0.220 0.442 0.267 0.0447 0.00128
 puissance(bd2(n=6,p=pi.hyper,lambda=1/10),4)
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
[1,] 0.071 0.346 0.416 0.152 0.016 0.000
[2,] 0.041 0.287 0.446 0.201 0.025 0.001
[3,] 0.025 0.223 0.453 0.259 0.039 0.001
[4,] 0.015 0.168 0.432 0.321 0.062 0.002
```

```
[5,] 0.009 0.124 0.392 0.373 0.098 0.004
[6,] 0.006 0.090 0.343 0.407 0.144 0.010
puissance(bd2(n=6,p=pi.hyper,lambda=1/10),5

      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
[1,] 0.032 0.244 0.444 0.243 0.037 0.001
[2,] 0.029 0.233 0.443 0.253 0.040 0.001
[3,] 0.027 0.222 0.442 0.264 0.044 0.001
[4,] 0.024 0.211 0.440 0.275 0.048 0.001
[5,] 0.022 0.201 0.437 0.286 0.052 0.002
[6,] 0.020 0.191 0.434 0.296 0.057 0.002
```

Acknowledgements. This work would not have been possible without the patient help of Charles Stein and Persi Diaconis, who both explained their very different pictures of the method.

I would also like to thank David Freedman, Marc Coram and David Wilson who both read early versions of this article and helped me clarify some of the notation.

References

Barbour, A. D. and Chen,L. H. Y. On the binary expansion of a random integer. *Statistics & Probability Letters*, 14:235–241, 1992.

Barbour, A.D., Holst, L. and Janson,S. *Poisson Approximation*. Clarendon Press, 1992.

Bolthausen, E. An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66:379–386, 1984.

Brouwer, A. Cohen, A. and Neumaier, A. *Distance Regular Graphs* Springer, NY. 1984

Cameron, P. *Permutation Groups* Cambridge University Press, Cambridge, 1999.

Diaconis, P. The distribution of leading digits and uniform distribution mod 1. *The Annals of Probability*, 5:72–81, 1977.

Diaconis, P. An example of Stein's method. Chapter 2, in this volume *Expository lectures and applications of Stein's method*, 2004.

Diaconis, P. and Freedman, D. Finite exchangeable sequences. *The Annals of Probability*, 8:745–764, 1980.

Diaconis, P. and Saloff-Coste, L. Nash inequalities for finite Markov chains. *Journal of Theoretical Probability*, 9:459–510, 1996.

Diaconis, P. and Zabell, S. Closed form summation for classical distributions: Variations on a theme of De Moivre. *Statistical Science*, 6:284–302, 1991.

Götze, F. On the rate of convergence in the multivariate Clt. *The Annals of Probability*, 19:724–739, 1991.

Wei-Liem Loh. Stein's method and multinomial approximation. *Annals of Applied Probability*, 2: 536–554, 1992.

Mann, B. PhD thesis, Harvard, 1996.

Reinert, G. Approximation for the χ^2using stein's method. Technical report, 1998.

Smith, D. Primitive and Imprimitive Graphs. *Quat. Jour. Math.* 22 ,551-557, 1971.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symposium in Mathematical Statistics and Probability*, volume 2, pages 583–602, 1972.

Stein, C. *Approximate computation of expectations*. Institute of Mathematical Statistics, 1986.