The Chen-Stein Method for Convergence of Distributions

Christina Goldschmidt

New Hall, University of Cambridge

Contents

1 Introduction

In 1972, Stein [15] published a new method for bounding the distance of a random variable, X, from a standard normal random variable, Z, in terms of a test function, h. He observed that, under certain conditions, there exists a function f such that

$$\mathbb{E}[h(X) - h(Z)] = \mathbb{E}[f'(X) - Xf(X)] \tag{1}$$

For example, if $h(x) = \mathbb{I}_{\{x \in A\}}$ for some set $A \subset \mathbb{R}$, where $\mathbb{I}_{\{\cdot\}}$ denotes the indicator of a set, then we are considering

$$\mathbb{E}[h(X) - h(Z)] = \mathbb{P}(X \in A) - \frac{1}{\sqrt{2\pi}} \int_A e^{-z^2/2} dz = \mathbb{E}[f'(X) - Xf(X)]$$

Often, the right-hand side of (1) is easier to bound than the left and, in this situation, Stein's observation leads us to a new way of bounding rates of convergence in distribution.

It was quickly shown that this approach is not restricted to a normal limit. Chen [6] showed that the analogue of (1) for the Poisson distribution (i.e. $Z \sim \mathrm{Po}(\lambda)$) is

$$\mathbb{E}[h(X) - h(Z)] = \mathbb{E}[\lambda f(X+1) - Xf(X)]$$

where, this time, f and h are functions defined on $\mathbb{Z}^+$. For this reason, Stein's method applied to the Poisson limit is usually referred to as the Chen-Stein method. Since the 1970's, Stein's method has been used to bound rates of convergence to many distributions. It has, comparatively recently, been applied to stochastic processes, notably the Poisson process and diffusions.

Once we have an equation like (1), it is often reasonably straightforward to bound the right-hand side using estimates on f and its derivatives (often obtained by Taylor expansion) or, in the discrete case, backward differences. Couplings are the probabilistic tool most commonly used to implement this stage of Stein's method.

The advantage of the method over alternatives, e.g. Fourier techniques, is that we have immediate access to a bound. In certain cases, we may not be able to evaluate this bound numerically but often we can, to great effect.

Stein's method is now a well-developed field of research, with a large literature. I am not going to attempt to provide a comprehensive survey of that literature. I would, however, like to convey some feeling for how and why the method works. I will do this via a series of examples which I think demonstrate the beauty and power of the method, as well as some of its drawbacks. I also aim to give some insight into the use of couplings.

Section 2 covers Stein's method for the normal distribution. As this is the first time we use the method, we cover the major steps in detail. Initially, we use a standard approach to bound the right-hand side of (1) for a general function h. Then we specialise to the case $h(x) = \mathbb{I}_{\{x \in A\}}$, which gives us a bound on the rate of convergence in the Central Limit Theorem. This bound does not turn out to be the best possible (the correct rate is given by the Berry-Esséen Theorem, Theorem 2.4). Extra complication is required to prove the Berry-Esséen Theorem using Stein's method and we sketch this proof at the end of the section. Finally, we make a brief comparison with the usual proof of the theorem.

Section 3 covers the details of the Chen-Stein method for the Poisson limit. We consider the approximation of a sum of Bernoulli random variables by the Poisson distribution, under various conditions. Our primary aim in this section is to demonstrate the ease with which Stein's method may be adapted to deal with the distribution of sums of dependent quantities. We consider two types of dependence structure in the Bernoulli random variables. The section concludes with two simple examples of the use of couplings in the case of dependence.

Section 4 is an elegant application of Stein's method to bounding the rate of convergence of a simple random walk on the discrete circle to its stationary distribution. This section includes an example of the exchangeable pair coupling.

Section 5 is a discussion of the theoretical basis of Stein's method. We look at a procedure for applying Stein's method in general which has been proposed by Barbour, using ideas from semigroup theory. We also consider a connection between Stein's method and the Martingale Problem.

Finally, in Section 6, we look at the current state of the field.

1.1 Technicalities

We shall make much use of the supremum norm and the total variation distance.

Definition 1.1. *The supremum norm of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as follows:*

$$\|f\| = \sup_{x \in \mathbb{R}} |f(x)|$$

Definition 1.2. *The total variation distance between two probability measures μ and ν on a measurable space $(E, \mathcal{E})$ is defined to be*

$$d_{TV}(\mu, \nu) = \sup\{|\nu(A) - \mu(A)| : A \in \mathcal{E}\}$$

It should be noted that a sequence of random variables $(X_n)_{n \geq 1}$, defined on a discrete space, with associated measures $(\mu_n)_{n \geq 1}$, converges in distribution to μ if and only if $d_{TV}(\mu_n, \mu) \to 0$ as $n \to \infty$. If the random variables are defined on a continuous space, e.g. $\mathbb{R}$, then $d_{TV}(\mu_n, \mu) \to 0$ as $n \to \infty$ is, in fact, a stronger condition than convergence in distribution. Other metrics can be used but they are less natural and beyond the scope of this essay. Our use of the total variation distance means that we will mostly be interested in $h(x) = \mathbb{I}_{\{x \in A\}}$ for some set $A \in \mathcal{E}$, i.e. we want to consider the distance between the distribution of X and that of Z as

$$\mathbb{P}(X \in A) - \mathbb{P}(Z \in A)$$

As I have mentioned, couplings are very important in Stein's method. We give here a basic definition and refer the reader to Torgny Lindvall's accessible introduction to the subject [11].

Definition 1.3. *Suppose we have two random variables X and X' defined on the probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega', \mathcal{F}', \mathbb{P}')$ respectively. Then a coupling of X and X' in a measurable space $(E, \mathcal{E})$ is a random element (Y, Y'), defined on a single probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$, taking values in $(E^2, \mathcal{E}^2)$, such that $Y \sim X$ and $Y' \sim X'$.*

This definition is not very easy to comprehend at first sight and gives no idea of why couplings are useful but it will become much clearer when they are used.

2 Stein's Method for the Normal Distribution

The Central Limit Theorem is of fundamental importance in probability and statistics. However, it does not give any indication of the *rate* of convergence to a normal limit. In this section, I will discuss the use of Stein's method for finding this rate.

We will find the following characterisation of a normal random variable useful:

Proposition 2.1. *A random variable, X, is standard normal if and only if it satisfies*

$$\mathbb{E}[f'(X) - Xf(X)] = 0 \tag{2}$$

for all continuous and piecewise continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[|f'(Z)|] < \infty$ for $Z \sim$ N(0,1).

Proof. Suppose $X \sim$ N(0,1). Then

$$\begin{aligned}
\mathbb{E}[f'(X) - Xf(X)] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(x)e^{-x^2/2}dx - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xf(x)e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(x)e^{-x^2/2}dx + \left[\frac{1}{\sqrt{2\pi}}f(x)e^{-x^2/2}\right]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(x)e^{-x^2/2}dx \\
&\quad \text{by integration by parts} \\
&= 0
\end{aligned}$$

Conversely, suppose $\mathbb{E}[f'(X) - Xf(X)] = 0$ for all continuous and piecewise continuously differentiable, real-valued f such that $\mathbb{E}[|f'(Z)|] < \infty$ for $Z \sim$ N(0,1). Then this equation holds in particular for

$$f_t(x) = e^{x^2/2}\int_{-\infty}^{x}\left(\mathbb{I}_{\{y \le t\}} - \Phi(t)\right)e^{-y^2/2}dy$$

where $\Phi(t) = \int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$ and t is fixed. Then

$$f_t'(X) = Xf_t(X) + \mathbb{I}_{\{X \le t\}} - \Phi(t)$$

and so, for all $t \in \mathbb{R}$,

$$\mathbb{E}[f_t'(X) - Xf_t(X)] = \mathbb{P}(X \le t) - \Phi(t) = 0$$

which implies that $X \sim$ N(0,1). □

Proposition 2.2. *For any piecewise continuous, real-valued function h, there is a function f solving the* Stein equation

$$h(x) - \Phi h = f'(x) - xf(x) \tag{3}$$

where $\Phi h = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h(y)e^{-y^2/2}dy$ is the expectation of h with respect to the standard normal distribution.

Proof. Let $f(x) = e^{x^2/2} \int_{-\infty}^{x} (h(y) - \Phi h) e^{-y^2/2} dy$. Then

$$\begin{aligned} f'(x) &= xf(x) + e^{x^2/2}(h(x) - \Phi h) e^{-x^2/2} \\ &= xf(x) + h(x) - \Phi h \end{aligned}$$

and the result follows. □

Taking expectations in (3) gives

$$\mathbb{E}[h(X) - \Phi h] = \mathbb{E}[f'(X) - Xf(X)] \tag{4}$$

The quantity $\mathbb{E}[h(X) - \Phi h]$ gives a measure of the distance of X from the normal in terms of the test function, h. It is often easier to bound the right-hand side of the equation than the left and, when this is the case, we may proceed with Stein's method.

2.1 A First Attempt at Bounding the Rate in the Central Limit Theorem

Following Reinert [13] and Stein [16], we introduce a powerful tool: the exchangeable pair coupling. A pair of random variables defined on a single probability space is said to be exchangeable if, for all measurable sets A and A',

$$\mathbb{P}(X \in A, X' \in A') = \mathbb{P}(X \in A', X' \in A) \tag{5}$$

Suppose we have independent random variables $Y_1, Y_2, \ldots, Y_n$ with $\mathbb{E}[Y_i] = 0$, $\mathbb{E}\left[Y_i^2\right] = \sigma_i^2$, $\sum_{i=1}^n \sigma_i^2 = 1$, $\mathbb{E}\left[|Y_i|^3\right] < \infty$ and $\mathbb{E}\left[Y_i^4\right] < \infty$. Let $W = \sum_{i=0}^n Y_i$ and $W_i = W - Y_i$. Let I be a random variable, independent of the Y_i's, which is uniformly distributed on $\{1, 2, \ldots, n\}$. Now define $W' = W - Y_I + Y_I^*$ where Y_i^* is an independent copy of Y_i. Then (W, W') is an exchangeable pair. (W, W') are two copies of the same random variable which are "coupled" by the relationship (5). (W, W') satisfies

$$\mathbb{E}[W'|W] = \mathbb{E}[W - Y_I + Y_I^*|W] = \left(1 - \frac{1}{n}\right) W$$

This expression and the tower law give us

$$\mathbb{E}[Wf(W)] = \frac{n}{2}\mathbb{E}[(W - W')(f(W) - f(W'))]$$

Hence,

$$\mathbb{E}[h(W) - \Phi h] = \mathbb{E}[f'(W) - Wf(W)] = \mathbb{E}\left[f'(W) - \frac{n}{2}(W - W')(f(W) - f(W'))\right]$$

Taylor expansion gives

$$f(W) = f(W') - (W' - W)f'(W) - R$$

where $|R| \leq \frac{1}{2}\|f''\|(W-W')^2$ and $\|\cdot\|$ denotes the supremum norm. So

$$\begin{aligned}
|\mathbb{E}[h(W) - \Phi h]| &\leq \left|\mathbb{E}\left[f'(W) - \frac{n}{2}(W-W')^2 f'(W)\right]\right| + \frac{n}{4}\|f''\|\,\mathbb{E}\left[|W-W'|^3\right] \\
&\leq \mathbb{E}\left[\left|\left(1 - \frac{n}{2}\mathbb{E}\left[(W-W')^2|W\right]\right) f'(W)\right|\right] + \frac{n}{4}\|f''\|\,\mathbb{E}\left[|W-W'|^3\right] \\
&\leq \|f'\|\sqrt{\mathbb{E}\left[\left(1 - \frac{n}{2}\mathbb{E}[(W-W')^2|W]\right)^2\right]} + \frac{n}{4}\|f''\|\,\mathbb{E}\left[|W-W'|^3\right] \\
&\qquad \text{by the Cauchy-Schwarz inequality}
\end{aligned}$$

We can bound the expectations:

Proposition 2.3.

$$\begin{aligned}
\mathbb{E}\left[\left(1 - \frac{n}{2}\mathbb{E}\left[(W-W')^2|W\right]\right)^2\right] &\leq \frac{1}{4}\sum_{i=1}^{n}\left(\mathbb{E}\left[Y_i^4\right] - \sigma_i^4\right) \\
\mathbb{E}\left[|W-W'|^3\right] &\leq \frac{2}{n}\left(\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right)\right)
\end{aligned}$$

□

This proposition is proved in the Appendix. Using the bounds

$$\|f'\| \leq 2\,\|h - \Phi h\| \quad \text{and} \quad \|f''\| \leq 2\,\|h'\| \tag{6}$$

which are proved on pp.25–28 of Stein [16], we obtain

$$|\mathbb{E}[h(W) - \Phi h]| \leq \|h - \Phi h\|\sqrt{\sum_{i=1}^{n}\left(\mathbb{E}[Y_i^4] - \sigma_i^4\right)} + \|h'\|\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right)$$

Now, as set out in the introduction, we are primarily interested in $h(w) = \mathbb{I}_{\{w \leq t\}}$ for some fixed $t \in \mathbb{R}$ but we need to get around the fact that h is not differentiable at t. Following Stein [16], we introduce the function

$$h_{t,\Delta}(w) = \begin{cases} 1 & \text{if } w \leq t \\ 1 - \frac{w-t}{\Delta} & \text{if } t \leq w \leq t + \Delta \\ 0 & \text{if } w \geq t + \Delta \end{cases}$$

Then $\mathbb{E}[h_{t-\Delta,\Delta}(W)] \leq \mathbb{P}(W \leq t) \leq \mathbb{E}[h_{t,\Delta}(W)]$. It is clear that

$$\|h_{t,\Delta} - \Phi h_{t,\Delta}\| \leq 1 \quad \text{and} \quad \|h'_{t,\Delta}\| \leq \frac{1}{\Delta}$$

Then

$$\begin{aligned}
\mathbb{E}[h_{t,\Delta}(W)] &\leq \Phi h_{t,\Delta} + \|h_{t,\Delta} - \Phi h_{t,\Delta}\|\sqrt{\sum_{i=1}^{n}\left(\mathbb{E}[Y_i^4] - \sigma_i^4\right)} + \|h'_{t,\Delta}\|\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right) \\
&\leq \Phi(t) + \frac{\Delta}{\sqrt{2\pi}} + \sqrt{\sum_{i=1}^{n}\left(\mathbb{E}[Y_i^4] - \sigma_i^4\right)} + \frac{1}{\Delta}\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right)
\end{aligned}$$

We minimise this expression by differentiation:

$$\frac{d}{d\Delta}\,(\text{RHS}) = \frac{1}{\sqrt{2\pi}} - \frac{1}{\Delta^2}\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right) = 0 \quad \text{at the minimum}$$

which implies that

$$\Delta = (2\pi)^{1/4}\sum_{i=1}^{n}\left(\mathbb{E}\left[|Y_i|^3\right] + 3\sigma_i^3\right)$$

and so

$$\mathbb{E}[h_{t,\Delta}(W)] \leq \Phi(t) + \sqrt{\sum_{i=1}^{n}(\mathbb{E}[Y_i^4] - \sigma_i^4)} + \frac{2}{(2\pi)^{1/4}}\sqrt{\sum_{i=1}^{n}(\mathbb{E}[|Y_i|^3] + 3\sigma_i^3)}$$

Similarly, $\mathbb{E}[h_{t-\Delta,\Delta}(W)]$ is bounded by the same quantity and so

$$|\mathbb{P}(W \leq t) - \Phi(t)| \leq \sqrt{\sum_{i=1}^{n}(\mathbb{E}[Y_i^4] - \sigma_i^4)} + \frac{2}{(2\pi)^{1/4}}\sqrt{\sum_{i=1}^{n}(\mathbb{E}[|Y_i|^3] + 3\sigma_i^3)}$$

For the Central Limit Theorem, we want independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}\left[X_i^2\right] = 1$. Take $Y_i = \frac{1}{\sqrt{n}}X_i$ for $1 \leq i \leq n$, which means $W = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$. Hence,

$$\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \leq t\right) - \Phi(t)\right| \leq \sqrt{\frac{\mathbb{E}[X_i^4] - 1}{n}} + \frac{2}{(2\pi n)^{1/4}}\sqrt{\mathbb{E}[|X_i|^3] + 3} \tag{7}$$

This expression is $\mathcal{O}(n^{-1/4})$, which demonstrates convergence, but we know that by other methods we can obtain a bound which is $\mathcal{O}(n^{-1/2})$ so, in this case, Stein's method does not provide a sharp rate. A more complicated approach is needed to obtain the better rate.

2.2 The Berry-Esséen Theorem

We conclude this section with a sketch of the proof of the Berry-Esséen Theorem (the Central Limit Theorem with a rate) in the case of independent and identically distributed random variables. It was Bolthausen [5] who first obtained the correct rate in this theorem using Stein's Method, in 1984. We give a somewhat different version, the whole of which may be found in Stein [16].

Theorem 2.4 (Berry-Esséen). *Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with $\mathbb{E}[X_i] = 0$, $\mathbb{E}\left[X_i^2\right] = 1$ and $\mathbb{E}\left[|X_i|^3\right] < \infty$ for $1 \leq i \leq n$. Then for $n \in \mathbb{N}$ and all $x \in \mathbb{R}$,*

$$\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \leq t\right) - \Phi(t)\right| \leq \frac{9\mathbb{E}\left[|X_i|^3\right]}{\sqrt{n}}$$

Note that the bound in this theorem has a twofold advantage over our previous result, (7): the bound is $\mathcal{O}(n^{-1/2})$ and we do not require finite fourth moments.

We will need two lemmas.

Lemma 2.5. *Suppose that*

$$f_t(x) = e^{x^2/2} \int_{-\infty}^{x} \left(\mathbb{I}_{\{y \le t\}} - \Phi(t)\right) e^{-y^2/2} dy$$

Then $\|f_t\| < 1$ *and* $\|f_t'\| < 1$.

The need for this sort of lemma is typical in applications of Stein's method.

Proof.

$$\begin{aligned} f_t(x) &= e^{x^2/2} \left(\int_{-\infty}^{x \wedge t} e^{-y^2/2} dy - \sqrt{2\pi}\Phi(t)\Phi(x) \right) && (8) \\ &= \begin{cases} \sqrt{2\pi} e^{x^2/2} \Phi(x) \left(1 - \Phi(t)\right) & \text{if } x \le t \\ \sqrt{2\pi} e^{x^2/2} \Phi(t) \left(1 - \Phi(x)\right) & \text{if } x > t \end{cases} && (9) \end{aligned}$$

Now, $f_t \ge 0$ and we want to find its maximum. First note that $f_t(x)$ is increasing in x for $x \le t$ and decreasing in x for $x > t$. Hence, $f_t(x) \le f_t(t)$ for all x. Taking ϕ to be the standard normal density function,

$$\begin{aligned} \frac{d}{dt} f_t(t) &= \sqrt{2\pi} \left[t e^{t^2/2} \Phi(t)(1 - \Phi(t)) + e^{t^2/2} \phi(t)(1 - \Phi(t)) - e^{t^2/2} \phi(t)\Phi(t) \right] \\ &= \sqrt{2\pi} e^{t^2/2} \left[t\Phi(t)(1 - \Phi(t)) + \phi(t) - 2\phi(t)\Phi(t) \right] \\ &= 0 \text{ at } t = 0 \end{aligned}$$

and so we see that $f_t(t)$ is maximised at 0, where $f_0(0) = \frac{\sqrt{2\pi}}{4} < 1$. This gives the bound on $\|f_t\|$.

To obtain the bound on $\|f_t'\|$, we observe that:

For $x > 0$: $1 - \Phi(x) < \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{y}{x} e^{-y^2/2} dy = \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$

For $x < 0$: $\Phi(x) < \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \frac{y}{x} e^{-y^2/2} dy = \frac{e^{-x^2/2}}{|x|\sqrt{2\pi}}$

From (9) we see that

$$f_t'(x) = \begin{cases} (1 - \Phi(t)) \left(1 + x\sqrt{2\pi} e^{x^2/2} \Phi(x)\right) & \text{if } x \le t \\ \Phi(t) \left(x\sqrt{2\pi} e^{x^2/2}(1 - \Phi(x)) - 1\right) & \text{if } x > t \end{cases}$$

We need only consider $t \ge 0$ as $f_t(x) = f_{-t}(-x)$. Then

$$\begin{aligned} &\text{Case } x < 0: && 1 \ge (1 - \Phi(t)) \ge f_t'(x) \ge && (1 - \Phi(t)) \left[1 - |x|\sqrt{2\pi} e^{x^2/2} \frac{e^{-x^2/2}}{|x|\sqrt{2\pi}}\right] = 0 \\ &\text{Case } 0 \le x \le t: && 0 \le f_t'(x) \le && 1 - \Phi(x) + (1 - \Phi(x))\Phi(x) x\sqrt{2\pi} e^{-x^2/2} \\ &&& \le && 1 - \Phi(x) + \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \Phi(x) x\sqrt{2\pi} e^{x^2/2} = 1 \\ &\text{Case } x > t: && -1 \le -\Phi(t) \le f_t'(x) \le && \Phi(t) \left[x\sqrt{2\pi} e^{x^2/2} \frac{e^{-x^2/2}}{x\sqrt{2\pi}} - 1\right] = 0 \end{aligned}$$

Hence, $\|f_t'\| \le 1$. □

The next lemma is used at the end of the proof of the Berry-Esséen Theorem. I shall just quote it from Stein [16, p.99] as it adds nothing to a discussion of Stein's method.

Lemma 2.6 (Stein's Unsmoothing Lemma). *Suppose that F, G and H are cumulative distribution functions on the real line and that A and μ are real numbers such that $H' \leq A$, $|(F-H)*G| \leq \mu$ and $\gamma = \int |t| dG(t) < \infty$ then*

$$|F - H| \leq 3\mu + 12\gamma A$$

Now we are ready to sketch the proof of the Berry-Esséen Theorem.

Proof of Theorem 2.4. Suppose that $Y_1, Y_2, \ldots, Y_{n+1}$ are independent and identically distributed random variables with distribution function μ such that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}\left[Y_i^2\right] = \frac{1}{n}$. It can be checked that

$$g(t) = n \int_t^\infty x d\mu(x) = -n \int_{-\infty}^t x d\mu(x)$$

is a probability density function. Suppose Z_n has density g independent of $Y_1, Y_2, \ldots, Y_n$ and let $W = \sum_{i=1}^n Y_i$.

Suppose that we are given a function h and that f solves the Stein equation (3) involving h. Then some algebraic manipulation and use of this Stein equation gives

$$\mathbb{E}[h(W + Z_n)] = \Phi h + \mathbb{E}[W(f(W + Y_{n+1}) - f(W + Z_n))] - \mathbb{E}[Z_n f(W + Z_n)]$$

The idea is that, using this equation, we can bound $\mathbb{E}[h(W + Z_n)] - \Phi h$ and then use Lemma 2.6 to remove the Z_n, leaving a bound on $\mathbb{E}[h(W) - \Phi h]$.

Define $Y_i = \frac{1}{\sqrt{n}} X_i$ and take μ to be the distribution function of the Y_i. Then it can be shown that $\mathbb{E}[|Z_n|] = \frac{\mathbb{E}[|X_i|^3]}{2\sqrt{n}}$. Now let the function h be $h_t(x) = \mathbb{I}_{\{x \leq t\}}$, which corresponds to $f_t(x)$ as defined above.

By Taylor expansion, $|f_t(W + Y_{n+1}) - f_t(W + Z_n)| \leq (Y_{n+1} - Z_n)\|f_t'\|$ and so

$$\begin{aligned}
|\mathbb{P}(W + Z_n \leq t) - \Phi t| &= |\mathbb{E}[h_t(W + Z_n) - \Phi h_t]| \\
&= |\mathbb{E}[W(f_t(W + Y_{n+1}) - f_t(W + Z_n)) - Z_n f_t(W + Z_n)]| \\
&\leq \|f_t'\| \, \mathbb{E}[|W|] \, \mathbb{E}[|Y_{n+1} - Z_n|] + \|f_t\| \, \mathbb{E}[|Z_n|] \\
&\leq \mathbb{E}[|W|] \, \mathbb{E}[|Y_{n+1}| + |Z_n|] + \mathbb{E}[|Z_n|] \quad \text{by Lemma 2.5} \\
&\leq \sqrt{\mathbb{E}[W^2]} \left(\sqrt{\mathbb{E}\left[Y_{n+1}^2\right]} + \frac{\mathbb{E}\left[|X_i|^3\right]}{2\sqrt{n}} \right) + \frac{\mathbb{E}\left[|X_i|^3\right]}{2\sqrt{n}} \\
&= 1 \left(\frac{1}{\sqrt{n}} + \frac{\mathbb{E}\left[|X_i|^3\right]}{2\sqrt{n}} \right) + \frac{\mathbb{E}\left[|X_i|^3\right]}{2\sqrt{n}} \\
&= \frac{1 + \mathbb{E}\left[|X_i|^3\right]}{\sqrt{n}}
\end{aligned}$$

Applying Lemma 2.6 with $H = \Phi$, F the distribution of W, G the distribution of Z_n, $A = 1/\sqrt{2\pi}$,

$\gamma = \frac{\mathbb{E}[|X_i|^3]}{2\sqrt{n}}$ and $\mu = \frac{1+\mathbb{E}[|X_i|^3]}{\sqrt{n}}$, we obtain

$$\begin{aligned}|\mathbb{P}(W \leq t) - \Phi(t)| &\leq 3\frac{1+\mathbb{E}\left[|X_i|^3\right]}{\sqrt{n}} + 12\frac{\mathbb{E}\left[|X_i|^3\right]}{2\sqrt{n}}\frac{1}{\sqrt{2\pi}} \\ &< \frac{9\mathbb{E}\left[|X_i|^3\right]}{\sqrt{n}} \quad \text{as } 1 = \mathbb{E}\left[X_i^2\right] \leq \mathbb{E}\left[|X_i|^3\right]\end{aligned}$$

□

For the purposes of comparison, we consider briefly the original proof of this theorem by Berry (1941); see [9, pp.504–517] for the details. Berry approached the problem via Fourier techniques and an innovative smoothing method which allowed him to obtain the inequality

$$|\mathbb{P}(W \leq t) - \Phi(t)| \leq \frac{1}{\pi}\int_{-T}^{T}\left|\varphi^n\left(\frac{x}{\sqrt{n}}\right) - e^{-x^2/2}\right|\frac{1}{|x|}dx + \frac{24}{T\pi\sqrt{2\pi}}$$

for $T > 0$, where φ is the characteristic function of the X_i's. The quantities on the right-hand side may be bounded to show that

$$|\mathbb{P}(W \leq t) - \Phi(t)| \leq \frac{33}{4}\frac{\mathbb{E}\left[|X_i|^3\right]}{\sqrt{n}}$$

which gives the theorem. This method relies fundamentally on the assumption of independence which, as we will see in the next section, Stein's method does not. Moreover, it involves theory which is considerably more complicated.

3 Stein's Method for the Poisson Distribution

In this section, we outline Stein's method for Poisson approximation. The general procedure is similar to that for the normal distribution but here we emphasise the suitability of Stein's method for approximating the distributions of sums of *dependent* quantities. The ease with which the assumption of independence may be removed is one of the key advantages of the method over other rate estimation procedures.

The well-known "law of small numbers" says that the Binomial distribution $\mathrm{Bin}(n, p)$ converges to the Poisson distribution $\mathrm{Po}(\lambda)$, where $p = \lambda/n$ for some $\lambda > 0$, as $n \to \infty$. A $\mathrm{Bin}(n, p)$ random variable may be thought of as the sum of n independent and identically distributed Bernoulli random variables and so there are two constraints that we may be interested in removing: identical distribution and independence. As a result, a more general formulation is as follows: let $I_1, I_2, \ldots, I_n$ be such that $\mathbb{E}[I_i] = p_i,\ 0 \leq p_i \leq 1\ \forall i$. Let $\lambda = \sum_{i=0}^n p_i$. Then $\sum_{i=1}^n I_i$ tends to $\mathrm{Po}(\lambda)$ in distribution as $n \to \infty$. This is the result that we are interested in proving (and providing a rate for) for various types of interdependence between the I_i's. This section follows Barbour, Holst and Janson [4], the comprehensive reference on Stein's method for the Poisson distribution.

Proposition 3.1. *A random variable $Z \sim Po(\lambda)$ if and only if, for all bounded functions $g : \mathbb{Z}^+ \to \mathbb{R}$,*

$$\mathbb{E}[\lambda g(Z+1) - Zg(Z)] = 0$$

□

As in the normal case, this proposition leads us to a Stein equation:

$$h(x) - \mathbb{E}_P[h] = \lambda f(x+1) - x f(x)$$

where $\mathbb{E}_P[h] = \sum_{k=0}^{\infty} h(k) e^{-\lambda} \lambda^k / k!$. We will look in detail at the case $h(x) = \mathbb{I}_{\{x \in A\}}$ for some $A \subset Z^+$. Hence, we are looking for a function f which satisfies

$$\mathbb{I}_{\{x \in A\}} - \mathcal{P}_\lambda(A) = \lambda f(x+1) - x f(x)$$

where $\mathcal{P}_\lambda(A) = \sum_{k \in A} e^{-\lambda} \lambda^k / k!$. We can construct such a solution, f_A, recursively. Without loss of generality, we can take $f_A(0) = 0$. Let $U_k = \{0, 1, \ldots, k\}$ and let S^c denote the complement of a set S.

$$\begin{aligned}
f_A(k+1) &= \lambda^{-1}\left(\mathbb{I}_{\{k \in A\}} + k f_A(k) - \mathcal{P}_\lambda(A)\right) \\
&\vdots \\
&= \left(\lambda^{-1}\mathbb{I}_{\{k \in A\}} + k\lambda^{-2}\mathbb{I}_{\{k-1 \in A\}} + \cdots + k!\,\lambda^{-(k+1)}\mathbb{I}_{\{0 \in A\}}\right) \\
&\quad + \left(\lambda^{-1} + k\lambda^{-2} + \cdots + k!\,\lambda^{-(k+1)}\right)\mathcal{P}_\lambda(A) \\
&= k!\,\lambda^{-(k+1)} e^{\lambda}\left[\mathcal{P}_\lambda(A \cap U_k) - \mathcal{P}_\lambda(A)\,\mathcal{P}_\lambda(U_k)\right] && (10) \\
&= k!\,\lambda^{-(k+1)} e^{\lambda}[\mathcal{P}_\lambda(A \cap U_k)\,\mathcal{P}_\lambda(U_k) + \mathcal{P}_\lambda(A \cap U_k)\,\mathcal{P}_\lambda(U_k^c) \\
&\qquad - \mathcal{P}_\lambda(A \cap U_k)\,\mathcal{P}_\lambda(U_k) - \mathcal{P}_\lambda(A \cap U_k^c)\,\mathcal{P}_\lambda(U_k)] \\
&= k!\,\lambda^{-(k+1)} e^{\lambda}\left[\mathcal{P}_\lambda(A \cap U_k)\,\mathcal{P}_\lambda(U_k^c) - \mathcal{P}_\lambda(A \cap U_k^c)\,\mathcal{P}_\lambda(U_k)\right] && (11)
\end{aligned}$$

3.1 The Independent Case

Initially, suppose that the indicator random variables $I_1, I_2, \ldots, I_n$ are independent. Put $W = \sum_{i=1}^n I_i$ and $W_i = W - I_i$. Then

$$\mathbb{E}[I_i f_A(W)] = \mathbb{E}[I_i f_A(W_i+1)] = p_i \mathbb{E}[f_A(W_i+1)] \quad \text{by independence} \tag{12}$$

and so

$$\begin{aligned}
\mathbb{E}[\lambda f_A(W+1) - W f_A(W)] &= \sum_{i=1}^n \mathbb{E}[p_i f_A(W+1) - I_i f_A(W)] \\
&= \sum_{i=1}^n p_i\,\mathbb{E}[f_A(W+1) - f_A(W_i+1)] \\
&= \sum_{i=1}^n p_i^2\,\mathbb{E}[f_A(W+1) - f_A(W_i+1) | I_i = 1] \\
&= \sum_{i=1}^n p_i^2\,\mathbb{E}[f_A(W+1) - f_A(W) | I_i = 1]
\end{aligned}$$

which gives

$$|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| \;\leq\; 2 \sup_{k \geq 1} |f_A(k)| \sum_{i=1}^n p_i^2$$

and

$$|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| \le \sup_{k\ge 1} |f_A(k+1) - f_A(k)| \sum_{i=1}^n p_i^2$$

So we need to find bounds on $\|f_A\| = \sup_{k\ge1} |f_A(k)|$ and $\Delta f_A = \sup_{k\ge1} |f_A(k+1) - f_A(k)|$.

Lemma 3.2. *f_A satisfies the following bounds:*

$$\|f_A\| \le 1 \wedge \lambda^{-1/2} \quad \textit{and} \quad \Delta f_A \le \lambda^{-1}(1-e^{-\lambda}) \le 1 \wedge \lambda^{-1}$$

We will prove the bound for Δf_A; the bound on $\|f_A\|$ is proved in Section 5 by a more instructive probabilistic method.

Proof. $f_A(k) = \sum_{j\in A} f_{\{j\}}(k)$ and so we need to look at $f_{\{j\}}$ for some fixed j. Let $k \ge 1$.

$$f_{\{j\}}(k+1) = k!\lambda^{-(k+1)}e^{\lambda}\mathcal{P}_\lambda(k)\left[\mathbb{I}_{\{j\le k\}} - \mathcal{P}_\lambda(U_k)\right]$$

So $f_A(k+1)$ is negative and decreasing for $k < j$ and positive and decreasing for $k \ge j$. Therefore, the only point at which $f_{\{j\}}(k+1) - f_{\{j\}}(k) \ge 0$ in $k \ge 1$ is when $k = j$, i.e.

$$\begin{aligned}
f_{\{j\}}(j+1) - f_{\{j\}}(j) &= j!\lambda^{-(j+1)}e^{\lambda}[\mathcal{P}_\lambda(j)\,\mathcal{P}_\lambda(U_j^c)] + (j-1)!\lambda^{-j}e^{\lambda}[\mathcal{P}_\lambda(j)\,\mathcal{P}_\lambda(U_{j-1})] \\
&= \frac{1}{\lambda}\sum_{i=j+1}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!} + \frac{1}{j}\sum_{i=0}^{j-1}\frac{e^{-\lambda}\lambda^i}{i!} \\
&= \frac{e^{-\lambda}}{\lambda}\left[\sum_{i=j+1}^{\infty}\frac{\lambda^i}{i!} + \sum_{i=1}^{j}\frac{\lambda^i}{i!}\frac{i}{j}\right] \\
&\le \frac{e^{-\lambda}}{\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{i!} \\
&= \lambda^{-1}(1-e^{-\lambda})
\end{aligned}$$

So, for $j \ge 1$ we can say that $f_{\{j\}}(j+1) - f_{\{j\}}(j) \le \lambda^{-1}(1-e^{-\lambda})$. For $j = 0, k \ge 1$, $f_{\{j\}}(k+1)$ is positive and decreasing and so $f_{\{0\}}(k+1) - f_{\{0\}}(k) \le 0$. Hence, for $k \ge 1$,

$$\begin{aligned}
f_A(k+1) - f_A(k) &= \sum_{j\in A}\left(f_{\{j\}}(k+1) - f_{\{j\}}(k)\right) \\
&\le f_{\{k\}}(k+1) - f_{\{k\}}(k)
\end{aligned}$$

Now, $f_A(k) = -f_{A^c}(k)$ and so $-f_A(k) \le f_{\{k\}}(k+1) - f_{\{k\}}(k)$ which gives

$$\sup_{k\ge1}|f_A(k+1) - f_A(k)| \le \lambda^{-1}(1-e^{-\lambda}) \le 1 \wedge \lambda^{-1}$$

□

So we have proved

Theorem 3.3. *If $I_1, I_2, \dots I_n$ are independent then*

$$\begin{aligned}
|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| &\leq \lambda^{-1}(1-e^{-\lambda}) \sum_{i=1}^n p_i^2 \\
&\leq (1 \wedge \lambda^{-1}) \sum_{i=1}^n p_i^2 \\
&\to 0 \quad \textit{as } n \to \infty \textit{ as long as } \max_i p_i = o(n^{-1/2})
\end{aligned}$$

for any set $A \subset \mathbb{Z}^+$. □

3.2 The Dependent Case

Let us now relax the independence assumption. There are two ways of doing this, corresponding to two types of dependence structure in the set of random variables. The first is when each indicator I_i has a set Γ_i^s such that I_i is strongly dependent on I_j for $j \in \Gamma_i^s$ and another set Γ_i^w such that I_i is weakly dependent on I_j for $j \in \Gamma_i^w$. This is *local dependence* and the corresponding implementation of Stein's Method is referred to as the *local approach.* The alternative is that the dependence structure is global and, here, it turns out that coupling is more effective. We will begin by discussing the case of local dependence, which was first investigated by Chen [6].

3.2.1 Local Dependence

The only place in the above calculation where independence has been used is at (12). We may modify this to give

$$\mathbb{E}[I_i f(W_i + 1)] = \mathbb{E}[I_i f(Y_i + 1)] + \mathbb{E}[I_i(f(Y_i + Z_i + 1) - f(Y_i + 1))]$$

where $Z_i = \sum_{j \in \Gamma_i^s} I_j$ and $Y_i = W - I_i - Z_i = \sum_{j \in \Gamma_i^w} I_j$. The following lemma is proved in the Appendix:

Lemma 3.4.

$$\begin{aligned}
|\mathbb{E}[I_i(f(Y_i + Z_i + 1) - f(Y_i + 1))]| &\leq \Delta f \mathbb{E}[I_i Z_i] \\
|\mathbb{E}[f(Y_i + 1) - f(W + 1)]| &\leq \Delta f (p_i + \mathbb{E}[Z_i]) \\
|\mathbb{E}[I_i f(Y_i + 1) - p_i f(Y_i + 1)]| &\leq \|f\| \, \mathbb{E}\big[\big| \mathbb{E}[I_i | (I_j : j \in \Gamma_i^w)] - p_i \big| \big]
\end{aligned}$$

□

We may then obtain

Theorem 3.5.

$$\begin{aligned}
|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| \leq\ & (1 \wedge \lambda^{-1}) \sum_{i=1}^n (p_i^2 + p_i \mathbb{E}[Z_i] + \mathbb{E}[I_i Z_i]) \\
& + (1 \wedge \lambda^{-1/2}) \sum_{i=1}^n \mathbb{E}\big[\big| \mathbb{E}[I_i | (I_j : j \in \Gamma_i^w)] - p_i \big| \big] \qquad (13)
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\lambda f(W+1) - Wf(W)] &= \sum_{i=1}^{n} \mathbb{E}[p_i f(W+1) - I_i f(W)] \\
&= \sum_{i=1}^{n} \mathbb{E}[p_i f(W+1) - I_i f(W_i+1)] \\
&= \sum_{i=1}^{n} \Big(p_i \mathbb{E}[f(W+1) - f(Y_i+1)] - \mathbb{E}[I_i(f(Y_i+Z_i+1) - f(Y_i+1))] \\
&\qquad + \mathbb{E}[p_i f(Y_i+1) - I_i f(Y_i+1)] \Big)
\end{aligned}$$

which implies that

$$\begin{aligned}
\Big| \mathbb{E}[\lambda f(W+1) - Wf(W)] \Big| &\leq \sum_{i=1}^{n} \Big(\Delta f \; p_i(p_i + \mathbb{E}[Z_i]) + \Delta f \; \mathbb{E}[I_i Z_i] \\
&\qquad + \|f\| \, \mathbb{E}\big[\big| \mathbb{E}[I_i | (I_j : j \in \Gamma_i^w)] - p_i \big| \big] \Big)
\end{aligned}$$

by Lemma 3.4. This, with the bounds from Lemma 3.2, gives the result. □

Of course, the expression (13) in Theorem 3.5 simplifies considerably if I_i is independent of $\{I_j : j \in \Gamma_I^w\}$. In this case, the bound is $\mathcal{O}(\lambda^{-1})$ which is much better than the $\mathcal{O}(\lambda^{-1/2})$ bound above. One instance when this occurs is when $I_1, I_2, \ldots, I_n$ form a stationary m-dependent sequence, i.e. when the distribution of $(I_i, I_{i+1}, \ldots, I_{i+j})$ does not depend on i for any $j \geq 0$ and $(I_1, \ldots, I_k)$ and $(I_{k+m+1}, \ldots, I_{k+m+j})$ are independent for all $j, k \geq 0$. Here we take $\Gamma_i^s = \{I_j : j \neq i, |j-i| \leq m\}$ and p to be the common value of the p_i's. Then

$$\mathbb{E}[Z_i] = 2mp, \quad \mathbb{E}[I_i Z_i] = p\mathbb{E}[Z_i | I_i = 1], \quad \lambda = np$$

and so

Corollary 3.6. *For a stationary m-dependent sequence, $I_1, I_2, \ldots, I_n$,*

$$\begin{aligned}
|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| &\leq (1 \wedge (np)^{-1}) \, n \, (p^2 + 2mp^2 + p\mathbb{E}[Z_i | I_i = 1]) \\
&\leq (2m+1)p + \mathbb{E}[Z_i | I_i = 1]
\end{aligned}$$

3.2.2 Global Dependence

We now turn to the case of global dependence. We need to adapt (12) again and, here, we put

$$\mathbb{E}[I_i f(W_i + 1)] = p_i \mathbb{E}[f(W) | I_i = 1]$$

which gives

$$\begin{aligned}
\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A) &= \mathbb{E}[\lambda f(W+1) - Wf(W)] \\
&= \sum_{i=1}^{n} p_i \mathbb{E}[f(W+1) - \mathbb{E}[f(W) | I_i = 1]]
\end{aligned}$$

Now, suppose we can construct random variables U_i and V_i such that $U_i \sim W$ and $V_i + 1 \sim W | I_i = 1$. Then (U_i, V_i) is a coupling for each i.

Theorem 3.7. *Let (U_i, V_i) be any choice of couplings satisfying the conditions of the previous paragraph. Then*

$$\begin{aligned}
|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| &= \left|\sum_{i=1}^n p_i \mathbb{E}[f(U_i+1) - f(V_i+1)]\right| \\
&\leq \Delta f \sum_{i=1}^n p_i \mathbb{E}[|U_i - V_i|] \\
&\leq (1 \wedge \lambda^{-1}) \sum_{i=1}^n p_i \mathbb{E}[|U_i - V_i|]
\end{aligned}$$

Note that this reduces to the bound in Theorem 3.3 in the independent case.

If it is possible to construct pairs (U_i, V_i) such that $U_i \geq V_i$ a.s. then

$$\begin{aligned}
\sum_{i=1}^n p_i \mathbb{E}[|U_i - V_i|] &= \sum_{i=1}^n p_i \mathbb{E}[(U_i+1) - (V_i+1)] \\
&= \lambda \mathbb{E}[W+1] - \sum_{i=1}^n p_i \mathbb{E}[W | I_i = 1] \\
&= \lambda \mathbb{E}[W+1] - \sum_{i=1}^n \mathbb{E}[I_i W] \\
&= \lambda \mathbb{E}[W+1] - \mathbb{E}\left[W^2\right] \\
&= \lambda^2 + \lambda - \mathbb{E}\left[W^2\right] \\
&= \lambda - \mathrm{var}(W)
\end{aligned}$$

This leads us to the following corollary:

Corollary 3.8. *If $U_i \geq V_i$ a.s. then*

$$\begin{aligned}
|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| &\leq \lambda^{-1}(1 - e^{-\lambda})(\lambda - \mathrm{var}(W)) \\
&= (1 - e^{-\lambda}) \left(1 - \frac{1}{\lambda} \mathrm{var}(W)\right)
\end{aligned}$$

□

We are now in a position to give some examples to show that such a coupling can be constructed. We consider a two simple probabilistic models. Suppose first that we have N boxes arranged in a row and b balls, where $b < N$. Take the b balls and distribute them uniformly in the boxes. If there is a ball in the ith box then take I_i to be 1; otherwise take $I_i = 0$. Let $W = \sum_{i=1}^n I_i$ be the number of balls in the first n boxes. It is a well-known result that W has the hypergeometric distribution:

$$\mathbb{P}(W = w) = \binom{b}{w}\binom{N-b}{n-w} \Big/ \binom{N}{n} \quad \text{for } 0 \leq w \leq b$$

If $\frac{b}{N}$ and $\frac{n}{N}$ are small then it seems reasonable to approximate this distribution by $\mathrm{Po}(\lambda)$ where $\lambda = \mathbb{E}[W] = \frac{nb}{N}$. The variance is $\mathrm{var}(W) = \frac{N-n}{N-1}\frac{nb}{N}\left(1 - \frac{b}{N}\right)$ and the probability that $I_i = 1$ is b/N.

We need to construct an explicit coupling such that $U_i \geq V_i$ a.s.. We use one given by Lindvall [11, pp.65-66]. Fix a k such that $0 \leq k \leq n$ and construct a vector $(J_1^{(k)}, J_2^{(k)}, \ldots, J_n^{(k)})$ as follows. Assume that we start with all the boxes empty. Take one of the balls and put it in box k. Then distribute the remaining $b-1$ balls uniformly in the other boxes. Take $J_i^{(k)}$ to be 1 if there is a ball in box i and 0 otherwise. Then $\sum_{i=1}^n J_i^{(k)} \sim W|I_k = 1$. Now we construct a second vector $(I_1^{(k)}, I_2^{(k)}, \ldots, I_n^{(k)})$. With probability, b/N, take $(I_1^{(k)}, I_2^{(k)}, \ldots, I_n^{(k)}) = (J_1^{(k)}, J_2^{(k)}, \ldots, J_n^{(k)})$. With probability $(1-b/N)$, remove the ball from the kth box and put in one of the empty boxes (chosen uniformly from all the empty boxes). Let $I_i^{(k)}$ be 1 if box i contains a ball under the new configuration and 0 otherwise. Then $\sum_{i=1}^n I_i^{(k)} \sim W$. Take $U_i = \sum_{i=1}^n I_i^{(k)}$ and $V_i = \sum_{i \neq k} J_i^{(k)}$. By construction, $U_i \sim W$, $V_i + 1 \sim W|I_i = 1$ and $U_i \geq V_i$ a.s. for all i. Hence, we may apply Corollary 3.8 to obtain

$$
\begin{aligned}
\left|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)\right| &\leq (1-e^{-\lambda})\left(1 - \frac{N}{nb}\frac{N-n}{N-1}\frac{nb}{N}\left(1-\frac{b}{N}\right)\right) \\
&= \frac{N}{N-1}(1-e^{-\lambda})\left(\frac{n}{N} + \frac{b}{N} - \frac{1}{N} - \frac{nb}{N^2}\right) \\
&\leq \frac{n+b}{N-1}
\end{aligned}
$$

which is small, provided that $\frac{b}{N}$ and $\frac{n}{N}$ are small, as proposed above.

Our second example is Pólya's Urn. The urn contains N balls of n different colours in proportions $\pi_1, \pi_2, \ldots, \pi_n$. Balls are drawn at random. When a ball is sampled, it is put back into the urn with another ball of the same colour. Let $I_i = 1$ if no balls of colour i are drawn in the first r samplings and $I_i = 0$ otherwise. So $W = \sum_{i=1}^n I_i$ is the number of colours which have not appeared in the first r samplings. Then

$$
\begin{aligned}
p_i = \mathbb{E}[I_i] &= \mathbb{P}(\text{No balls of colour } i \text{ are drawn in the first } r \text{ samplings}) \\
&= \frac{(N-N\pi_i)}{N}\frac{(N+1-N\pi_i)}{(N+1)}\cdots\frac{(N+r-1-N\pi_i)}{(N+r-1)} \\
&= \binom{-N(1-\pi_i)}{r}\Big/\binom{-N}{r}
\end{aligned}
$$

$$
\begin{aligned}
p_{ik} = \mathbb{E}[I_i I_k] &= \mathbb{P}(\text{No balls of colours } i \text{ or } k \text{ are drawn in the first } r \text{ samplings}) \\
&= \frac{(N-N\pi_i-N\pi_k)}{N}\frac{(N+1-N\pi_i-N\pi_k)}{(N+1)}\cdots\frac{(N+r-1-N\pi_i-N\pi_k)}{(N+r-1)} \\
&= \binom{-N(1-\pi_i-\pi_k)}{r}\Big/\binom{-N}{r}
\end{aligned}
$$

We have $\lambda = \mathbb{E}[W] = \sum_{i=1}^n p_i$ and

$$
\begin{aligned}
\mathrm{var}(W) = \mathbb{E}\left[W^2\right] - \lambda^2 &= \mathbb{E}\left[\sum_{i=1}^n I_i\right] + \sum_{i\neq k}\mathbb{E}[I_i I_k] - \sum_{i=1}^n p_i^2 - \sum_{i \neq k} p_i p_k \\
&= \sum_{i=1}^n p_i(1-p_i) + \sum_{i\neq k}(p_{ik} - p_i p_k)
\end{aligned}
$$

So, if we can find couplings (U_i, V_i) such that $U_i \geq V_i$ a.s. for all i, we can apply Corollary 3.8 with these values of λ and $\mathrm{var}(W)$. We use a coupling from Barbour and Holst [3, pp.80–81]. Instead of just putting a sampled ball back in the urn with another of the same colour, imagine that we keep the sampled ball and put two of the same colour back in the urn. Then W is the number of colours which do not appear in our sample. Now fix k and construct a vector $(J_1^{(k)}, J_2^{(k)}, \ldots, J_n^{(k)})$ as follows. If $I_k = 1$ then take $(J_1^{(k)}, J_2^{(k)}, \ldots, J_n^{(k)}) = (I_1, I_2, \ldots, I_n)$. Otherwise, withdraw all balls of colour k from the urn and throw them away. Replace the balls of colour k in the sample by balls sampled as above from the remaining balls in the urn. Set $J_i^{(k)} = 1$ if the new sample has no ball of colour i and 0 otherwise. Let $U_i = W$ and let $V_i = \sum_{k \neq i} J_i^{(k)}$. Then (U_i, V_i) is a coupling which satisfies our requirements.

4 Stein's Method for the Discrete Uniform Distribution

In this section, we show that Stein's method is by no means restricted to the normal and Poisson limits and we include another example of the power of coupling.

Suppose we have a simple random walk on the discrete circle $\mathbb{Z}_p$ where p is an odd integer. The random walk is generated by the vector $X = (X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables such that

$$X_i = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

The position of the random walk is given by $S(X) \pmod p$ where $S(X) = X_1 + X_2 + \cdots + X_n$. It is intuitively obvious that $\mathbb{P}\left(S(X) = j \pmod p\right) \to \frac{1}{p}$ as $n \to \infty$ for $0 \leq j \leq p-1$. Using Stein's Method, we can prove that this is the case and bound the rate of convergence. This section follows an idea of Diaconis [7] but provides a different proof of Theorem 4.1.

Theorem 4.1. *Using the definitions of the previous paragraph,*

$$\left| \mathbb{P}\left(S(X) \pmod p \in A\right) - \frac{|A|}{p} \right| \leq \frac{p-1}{\sqrt{n}}$$

for any $A \subset \mathbb{Z}_p$.

As usual, we start by characterising the uniform distribution on $\{0, 1, \ldots, p-1\}$ by a particular equation.

Proposition 4.2. *Let p be an odd integer. Then a random variable Z is uniform on $\mathbb{Z}_p$ if and only if*

$$\mathbb{E}[f(Z) - f(Z-a)] = 0$$

for all functions $f : \mathbb{Z}_p \to \mathbb{R}$ and all $a \in \mathbb{Z}_p$ such that a and p are coprime.

Proof. Suppose that $Z \sim U\{0, 1, \ldots, p-1\}$. Then

$$\mathbb{E}[f(Z) - f(Z-a)] = \frac{1}{p} \sum_{i=0}^{p-1} [f(i) - f(i-a)] = 0$$

Conversely, suppose that $\mathbb{E}[f(Z) - f(Z-a)] = 0$ for all real-valued f. Then it holds for f_t such that $f_t(i) = \mathbb{I}_{\{i=t\}}$ for some fixed $t \in \mathbb{Z}_p$.

$$\begin{aligned}\mathbb{E}[f_t(Z) - f_t(Z-a)] &= \sum_{i=0}^{p-1} [\mathbb{I}_{\{i=t\}} - \mathbb{I}_{\{i-a=t\}}]\mathbb{P}(Z=i) \\ &= \mathbb{P}(Z=t) - \mathbb{P}(Z=t+a \pmod p) \\ &= 0\end{aligned}$$

But a and p are coprime, so the numbers $t + na \pmod p$, $n \in \mathbb{Z}$ range over all of $\mathbb{Z}_p$. Hence, $\mathbb{P}(Z=0) = \mathbb{P}(Z=1) = \cdots = \mathbb{P}(Z=p-1)$ and so $Z \sim U\{0, 1, \ldots, p-1\}$. □

For simplicity, we will take $a = 2$ which we know is always coprime to odd p.

Proposition 4.3. *For every function $h : \mathbb{Z}_p \to \mathbb{R}$, there exists a function $f : \mathbb{Z}_p \to \mathbb{R}$ such that*

$$h(W) - \mathbb{E}_U[h] = f(W) - f(W-2) \tag{14}$$

where $\mathbb{E}_U[h] = \frac{1}{p}\sum_{i=0}^{p-1} h(i)$, the expectation of h with respect to the uniform distribution.

(14) is the Stein equation for this problem.

Proof. Define f by

$$f(i) = \sum_{j=0}^{2^{-1}i} (h(2j) - \mathbb{E}_U[h]) \tag{15}$$

where 2^{-1} denotes the inverse of 2 modulo p. Then

$$\begin{aligned}f(i) - f(i-2) &= \sum_{j=0}^{2^{-1}i} (h(2j) - \mathbb{E}_U[h]) - \sum_{j=0}^{2^{-1}(i-2)} (h(2j) - \mathbb{E}_U[h]) \\ &= h(i) - \mathbb{E}_U[h]\end{aligned}$$

□

Now, we want to consider the distance between the distribution of $S(X) \pmod p$ and the uniform. We will use the total variation distance and so want to find a bound on

$$\sup_{A \subset \mathbb{Z}_p} \left| \mathbb{P}\Big(S(X) \pmod p \in A\Big) - \frac{|A|}{p} \right|$$

Take $h(w) = \mathbb{I}_{\{w \in A\}}$ for a fixed set $A \subset \mathbb{Z}_p$.

Proposition 4.4. $|f(i)| \leq \frac{p-1}{2}$ *where f is the function defined in (15).*

Proof.

$$\begin{aligned}\sum_{j=0}^{p-1}(h(j)-\mathbb{E}_U[h]) &= \sum_{j=0}^{p-1}\left(\mathbb{I}_{\{j\in A\}}-\frac{|A|}{p}\right)\\ &= |A|-|A|\\ &= 0\end{aligned}$$

and so for any set $S\subset\mathbb{Z}_p$,

$$\sum_{j\in S}(h(j)-\mathbb{E}_U[h]) = -\sum_{j\in S^c}(h(j)-\mathbb{E}_U[h])$$

Obviously, $\left|\sum_{j\in S}(h(j)-\mathbb{E}_U[h])\right|\le |S|$ and so $\sup_{S\subset\mathbb{Z}_p}\left|\sum_{j\in S}(h(j)-\mathbb{E}_U[h])\right|\le\frac{p-1}{2}$. Taking $S=\{j\in\mathbb{Z}_p: 0\le j\le 2^{-1}i\}$, we obtain

$$\begin{aligned}|f(i)| &= \left|\sum_{k\in S}(h(2k)-\mathbb{E}_U[h])\right|\\ &\le \frac{p-1}{2}\end{aligned}$$

□

As it causes no ambiguity, we introduce the convention that the arguments of f and h are taken modulo p, to simplify notation. Now, we have the equation

$$\mathbb{E}[h(S(X))-\mathbb{E}_U[h]] = \mathbb{E}[f(S(X))-f(S(X)-2)]$$

We also have a bound on $|f|$ which we need to use to bound the right-hand side of this equation. We clearly cannot just say that $|\mathbb{E}[f(S(X))-f(S(X)-2)]|\le 2|f|\le p-1$ as this does not give convergence. It turns out that it is useful to construct an exchangeable pair in this situation. The following calculation is typical when using Stein's Method; see, for example, Stein [16, p.85].

Proof of Theorem 4.1. Suppose that the random variable I is uniformly distributed on $\{1,2,\dots,n\}$ and independent of X. Define $Y=(Y_1,Y_2,\dots,Y_n)$ by

$$Y_i=\begin{cases}X_i & \text{for } i\neq I\\ 1-2X_i & \text{for } i=I\end{cases}$$

Then $\mathbb{P}(X=x,Y=y)=\mathbb{P}(X=y,Y=x)$ which is the definition of an exchangeable pair. It follows that

$$\mathbb{E}\left[f(S(X))\mathbb{I}_{\{S(X)=S(Y)-2\}}\right]=\mathbb{E}\left[f(S(Y))\mathbb{I}_{\{S(Y)=S(X)-2\}}\right] \tag{16}$$

where we do not take the quantities under the indicators modulo p. Now consider the expression

$$\begin{aligned}&2\mathbb{E}\left[f(S(X))\mathbb{I}_{\{S(X)=S(Y)-2\}}-f(S(Y))\mathbb{I}_{\{S(Y)=S(X)-2\}}|X\right]\\ &\quad=2f(S(X))\,\mathbb{P}\big(S(X)=S(Y)-2\big)\;-\;2f(S(X)-2)\,\mathbb{P}\big(S(Y)=S(X)-2\big)\end{aligned}$$

It is easy to see that

$$\mathbb{P}\left(S(X) = S(Y) - 2\right) \; = \; \mathbb{P}\left(X_I = -1\right) \; = \; \frac{N_-(X)}{n}$$

and

$$\mathbb{P}\left(S(Y) = S(X) - 2\right) \; = \; \mathbb{P}\left(X_I = +1\right) \; = \; \frac{N_+(X)}{n}$$

where $N_-(X)$ and $N_+(X)$ are the number of coordinates of X which are equal to -1 and $+1$ respectively. $N_-(X) + N_+(X) = n$ by definition and so $N_-(X) - \frac{n}{2} = \frac{n}{2} - N_+(X)$. Therefore,

$$\begin{aligned}
&2f(S(X))\,\mathbb{P}\big(S(X) = S(Y) - 2\big) \; - \; 2f(S(X) - 2)\,\mathbb{P}\big(S(Y) = S(X) - 2\big) \\
&\quad = \; 2f(S(X))\,\frac{N_-(X)}{n} \; - \; 2f(S(X) - 2)\,\frac{N_+(X)}{n} \\
&\quad = \; 2f(S(X))\,\frac{N_-(X) - \frac{n}{2}}{n} \; + \; 2f(S(X) - 2)\,\frac{N_-(X) - \frac{n}{2}}{n} \; + \; \Big[f(S(X)) - f(S(X) - 2)\Big] \\
&\quad = \; 2\,\frac{N_-(X) - \frac{n}{2}}{n}\Big[f(S(X)) + f(S(X) - 2)\Big] \; + \; \Big[f(S(X)) - f(S(X) - 2)\Big]
\end{aligned}$$

Now, $\mathbb{E}\left[2\mathbb{E}\left[f(S(X))\mathbb{I}_{\{S(X)=S(Y)-2\}} - f(S(Y))\mathbb{I}_{\{S(Y)=S(X)-2\}}|X\right]\right] = 0$, by (16) and the tower law, and so

$$\mathbb{E}[f(S(X)) - f(S(X) - 2)] = -\frac{2}{n}\mathbb{E}\left[\left(N_-(X) - \frac{n}{2}\right)\left(f(S(X)) + f(S(X) - 2)\right)\right]$$

which implies that

$$\begin{aligned}
|\mathbb{E}[f(S(X)) - f(S(X) - 2)]| \; &\leq \; \frac{2}{n}\,\mathbb{E}\left[\left|N_-(X) - \frac{n}{2}\right|\right]\,2\,\frac{p-1}{2} \quad \text{by Proposition 4.4} \\
&\leq \; \frac{2(p-1)}{n}\sqrt{\mathbb{E}\left[\left(N_-(X) - \frac{n}{2}\right)^2\right]}
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}\left[\left(N_-(X) - \frac{n}{2}\right)^2\right] \; &= \; \sum_{i=0}^{p-1}\left(i - \frac{n}{2}\right)^2\binom{n}{i}\left(\frac{1}{2}\right)^n \\
&= \; \sum_{i=0}^{p-1}\left(i^2 - ni + \frac{n^2}{4}\right)\binom{n}{i}\left(\frac{1}{2}\right)^n \\
&= \; \frac{n}{4} + \frac{n^2}{4} - \frac{n^2}{2} + \frac{n^2}{4} \\
&= \; \frac{n}{4}
\end{aligned}$$

and so

$$|\mathbb{E}[f(S(X)) - f(S(X) - 2)]| \leq \frac{2(p-1)}{n}\sqrt{\frac{n}{4}} = \frac{p-1}{\sqrt{n}}$$

Hence,

$$\left|\mathbb{P}\left(S(X) \pmod p \in A\right) - \frac{|A|}{p}\right| \leq \frac{p-1}{\sqrt{n}}$$

for all $A \subset \mathbb{Z}_p$. $\square$

Diaconis [7] generalises this result to the case of a general step-size distribution. The bound is good in that it shows that n must be roughly p^2 for the distance to be small. However, Diaconis shows that some improvement is possible and notes that, in the case $n = p^3$, Fourier analysis arguments give the bound to be $\mathcal{O}(e^{-c\sqrt{n}})$ for some $c > 0$, whereas our bound is, in this case, $\mathcal{O}(n^{-1/6})$. Thus, it is instructive to have both the Fourier analysis methods and Stein's method at our disposal.

5 Discussion

The preceeding examples will have raised some questions concerning Stein's method. In this section, I will give a more detailed discussion of the method in general. The theory mostly comes from Barbour [1, 2], Barbour, Holst and Janson [4] and Reinert [14].

5.1 The Generator Method

The heart of Stein's method is the identification of a suitable Stein equation and the bounding of the quantities on the right-hand side. But how should we go about this for a general distribution? So far, we have only dealt with simple distributions and have been able to find characterising equations and bounds easily but this is the case in general. Barbour [1] proposed a procedure which has become known as the Generator Method. He observed that the right-hand sides of Stein equations look like the infinitesimal generators of certain Markov processes. The normal Stein equation is

$$h(x) - \Phi h = f'(x) - xf(x)$$

whose right-hand side is very similar to the infinitesimal generator $\mathcal{A}f(x) = f''(x) - xf'(x)$ of an Ornstein-Uhlenbeck process. The Stein equation for the Poisson distribution is

$$h(x) - \mathcal{P}_\lambda h = \lambda f(x+1) - xf(x)$$

whose right-hand side is $\mathcal{A}g(x) = \lambda[g(x+1) - g(x)] - x[g(x) - g(x-1)]$ when we take $f(x) = g(x) - g(x-1)$, which is the infinitesimal generator of an immigration-death process. The stationary distribution of the relevant Ornstein-Uhlenbeck process is N(0,1); the stationary distribution of the immigration-death process is Po(λ). So, in the Stein equation we are using the fact that the target random variable can be imbedded in a Markov process in equilibrium. This observation not only provides us with the beginnings of a theory to justify Stein's method in general but also gives us a practical way of making estimates for the function f. For this purpose, we need some facts from semigroup theory, which may be found in Ethier and Kurtz [8].

Suppose that $(X(t))_{t\geq 0}$ is a Markov process with associated transition semigroup $(T(t))_{t\geq 0}$ and equilibrium distribution π. Then the infinitesimal generator, $\mathcal{A}$, is given by

$$\mathcal{A}h = \lim_{t\to 0} \frac{1}{t}\left(T(t)h - h\right)$$

Proposition 1.5 of Ethier and Kurtz [8, p.9] tells us that

$$T(t)h - h = \mathcal{A}\left(\int_0^\infty T(s)hds\right)$$

and that, under certain conditions, we may take the limit as $t \to \infty$ to obtain

$$\mathbb{E}_\pi[h] - h = \mathcal{A}\left(\int_0^\infty T(s)h ds\right)$$

But the Stein equation is

$$h - \mathbb{E}_\pi[h] = \mathcal{A}f$$

for some f and so we expect that we can take

$$\begin{aligned} f &= -\int_0^\infty T(s)h ds \\ \text{i.e. } f(x) &= -\int_0^\infty \mathbb{E}_x[h(X(s))]ds \end{aligned} \tag{17}$$

where $\mathbb{E}_x[h(X(s))] = \mathbb{E}[h(X(s))|X(0) = x]$. This expression does not necessarily provide the easiest way to determine f: for example, in the Poisson case, it was very easy to construct a solution to the Stein equation recursively when $h(x) = \mathbb{I}_{\{x \in A\}}$. However, by the use of appropriate couplings, we can at least estimate quantities associated with f. Often, it is useful to produce a coupling of the Markov process started at one point with the same process started at a different point, e.g. equilibrium.

In order to demonstrate the effectiveness of this approach, I shall look at the Poisson example in detail. First, I will show that the expression $\lambda[g(x+1) - g(x)] - x[g(x) - g(x-1)]$ is the generator equation of an immigration-death process, $(X(t))_{t\geq 0}$ with immigration rate λ and unit *per capita* death rate. The transition probabilities, $p_{i,j}(h) = \mathbb{P}(X(t+h) = j|X(t) = i)$, are

$$\begin{aligned} p_{i,i+1}(h) &= \lambda h + o(h) \\ p_{i,i-1}(h) &= ih + o(h) \\ p_{i,i}(h) &= 1 - (\lambda + i)h + o(h) \\ p_{i,j}(h) &= o(h) \text{ for } |j - i| > 1 \end{aligned}$$

Let the matrix $P(h) = (p_{i,j}(h))$. Then the generator matrix, A, is given by

$$A = \lim_{h\to 0} \frac{1}{h}(P(h) - I) = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 1 & -1-\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 2 & -2-\lambda & \lambda & 0 & \cdots \\ 0 & 0 & 3 & -3-\lambda & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and so, taking $g = (g(0), g(1), \dots)^T$,

$$\begin{aligned} Ag &= xg(x-1) - (x+\lambda)g(x) + \lambda g(x+1) \\ &= \lambda[g(x+1) - g(x)] - x[g(x) - g(x-1)] \end{aligned}$$

In order to find the stationary distribution, π, we solve the equation $\pi A = 0$. This gives

$$\pi_n = \frac{\lambda^n e^{-n}}{n!}$$

which is $\text{Po}(\lambda)$, as expected.

When we considered the Poisson example in Section 3, in Lemma 3.2 we needed to find a bound on $\|f\| = \sup_{x \geq 0} |f(x+1)|$, i.e. a bound on $\Delta g = \sup_{x \geq 0} |g(x+1) - g(x)|$. By (17), we should have

$$g(x) = -\int_0^\infty \mathbb{E}_x [h(X(t))] dt$$

We take $h(x) = \mathbb{I}_{\{x \in A\}}$ for some set $A \subset \mathbb{Z}^+$. Then

$$\begin{aligned} g(x+1) - g(x) &= \int_0^\infty \Big(\mathbb{E}_x [h(X(t))] - \mathbb{E}_{x+1} [h(X(t))] \Big) dt \\ &= \int_0^\infty \Big(\mathbb{P}_x (X(t) \in A) - \mathbb{P}_{x+1} (X(t) \in A) \Big) dt \end{aligned}$$

Adapting the method used in Barbour, Holst and Janson [4, p.209, pp.221–223] for the Poisson process, we construct processes X_1 and X_2 with distributions $\mathbb{P}_x$ and $\mathbb{P}_{x+1}$. Let X_0 be the immigration-death process with the same rates as X but started from 0 (so X_0 has distribution $\mathbb{P}_0$) and D be a pure death process with unit *per capita* death rate such that $D(0) = x$. Then define

$$\begin{aligned} X_1(t) &= X_0(t) + D(t) \\ X_2(t) &= X_0(t) + D(t) + \mathbb{I}_{\{E > t\}} \end{aligned}$$

where E is an exponential random variable with mean 1, independent of everything else, which represents the extra starting particle under $\mathbb{P}_{x+1}$. Then

$$\begin{aligned} g(x+1) - g(x) &= \int_0^\infty \Big(\mathbb{P}(X_1(t) \in A) - \mathbb{P}(X_2(t) \in A) \Big) dt \\ &= \int_0^\infty \mathbb{P}(E > t) \big[\mathbb{P}(X_1(t) \in A) - \mathbb{P}(X_1(t) + 1 \in A) \big] dt \\ &= \int_0^\infty e^{-t} \big[\mathbb{P}(X_1(t) \in A) - \mathbb{P}(X_1(t) + 1 \in A) \big] dt \end{aligned}$$

from which it is obvious that $|g(x+1) - g(x)| \leq \int_0^\infty e^{-t} dt = 1$. But we can find another bound, involving $\lambda^{-1/2}$.

$$\begin{aligned} \mathbb{P}(X_1(t) \in A) - \mathbb{P}(X_1(t) + 1 \in A) &= \sum_{n=0}^{x} \mathbb{P}(D(t) = n) \left[\mathbb{P}(X_0(t) + n \in A) - \mathbb{P}(X_0(t) + n + 1 \in A) \right] \\ &= \sum_{n=0}^{x} \mathbb{P}(D(t) = n) \sum_{k \in A} \left(\mathbb{P}(X_0(t) = k - n) - \mathbb{P}(X_0(t) = k - n - 1) \right) \\ &\leq \left(\sum_{n=0}^{x} \mathbb{P}(D(t) = n) \right) \max_{k \geq 0} \mathbb{P}(X_0(t) = k) \\ &= \max_{k \geq 0} \mathbb{P}(X_0(t) = k) \end{aligned}$$

Now, $X_0(t) \sim \mathrm{Po}(\lambda(1 - e^{-t}))$: $(X_0(t))$ may be viewed as an M/M/∞ queue with arrival rate λ and service rate 1; see Norris [12, pp.191–192] for the distributional result. By Proposition A.2.7 of Barbour, Holst and Janson [4], if $Z \sim \mathrm{Po}(\nu)$ then $\max_{k \geq 0} \mathbb{P}(Z = k) \leq \frac{1}{\sqrt{2e\nu}}$ and so

$$\max_{k \geq 0} \mathbb{P}(X_0(t) = k) \leq \frac{1}{\sqrt{2e\lambda(1 - e^{-t})}}$$

Hence,

$$\begin{aligned}
|g(x+1)-g(x)| &\leq \frac{1}{\sqrt{2e\lambda}} \int_0^\infty \frac{e^{-t}}{\sqrt{1-e^{-t}}} dt \\
&= \frac{1}{\sqrt{2e\lambda}} \left[2(1-e^{-t})^{1/2} \right]_0^\infty \\
&= \sqrt{\frac{2}{e\lambda}} \\
&\leq \lambda^{-1/2}
\end{aligned}$$

So, overall, $\|f\| = \Delta g \leq 1 \wedge \lambda^{-1/2}$, which was the bound given in Lemma 3.2.

In a comparatively simple manner, we have obtained a good bound for use in an approximation theorem by using the Generator Method. The calculation above is used in Barbour, Holst and Janson [4] to obtain bounds for Poisson process approximation and similar calculations may be used for general distributions. Of course, there may be more than one Markov Process which has the target distribution as its stationary distribution and which one we should use is still an open problem.

5.2 The Martingale Problem

Suppose we have a generator equation and we want to find the corresponding Markov process. Then we need to solve the *Martingale Problem* for the process $(X(t))_{t \geq 0}$: under certain conditions (which can be made precise — see Ethier and Kurtz [8]),

$$f(X(t)) - \int_0^t \mathcal{A} f(X(s)) ds$$

is a (possibly local) martingale for all "nice" functions f.

Suppose we have $\mathcal{A} f(x) = f''(x) - x f'(x)$, which is the right-hand side of the normal Stein equation. We restrict attention to continuous processes X and functions $f \in C^2(\mathbb{R})$. Then, Itô's formula tells us that

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) dA_s + \int_0^t f'(X_s) dM_s + \frac{1}{2} \int_0^t f''(X_s) d\langle M \rangle_s$$

where we use $X_t = X(t)$ for convenience of notation and X has Doob-Meyer decomposition $X_t = X_0 + A_t + M_t$ with A a continuous process of finite variation and M a continuous (local) martingale. Thus,

$$\begin{aligned}
f(X_t) - \int_0^t \mathcal{A} f(X(s)) ds &= f(X_0) + \int_0^t f'(X_s) dA_s + \int_0^t f'(X_s) dM_s + \frac{1}{2} \int_0^t f''(X_s) d\langle M \rangle_s \\
&\quad - \int_0^t f''(X_s) ds + \int_0^t X_s f'(X_s) ds
\end{aligned}$$

is a (local) martingale. Hence, the finite variation part must be zero:

$$\int_0^t f'(X_s) dA_s + \frac{1}{2} \int_0^t f''(X_s) d\langle M \rangle_s - \int_0^t f''(X_s) ds + \int_0^t X_s f'(X_s) ds = 0$$

This must hold for all $f \in C^2(\mathbb{R})$ and so we must have

$$dA_s = -X_s ds \quad \text{and} \quad \frac{1}{2} d\langle M \rangle_s = ds$$
$$\Rightarrow M_s = \sqrt{2} B_s$$

where B is a standard Brownian Motion. This gives

$$dX_t = -X_s ds + \sqrt{2} dB_s,$$

an Ornstein-Uhlenbeck process. If $X_0 \sim$ N(0,1) then X is a stationary zero-mean Gaussian process with $\mathrm{cov}(X_t, X_s) = e^{-|t-s|}$ ie. $X_t \sim$ N(0,1) (see Karatzas and Shreve [10, p.358]).

Thus, we can apply the theory that has been developed for dealing with the Martingale Problem to obtain a Markov Process when we are given a generator.

6 Concluding Remarks

Stein's method is applicable in many more situations than just those which have been demonstrated above. Reinert [13] gives a good summary of the different distributions which have been covered. Many applications are combinatorial in nature (see, for example, Stein [16]). Stein's method can often be used where other methods have failed, particularly, as in Section 3, when dependent random variables are considered. It does not, however, necessarily provide sharp rates without some extra work, as shown in Section 2.

Much recent research has focused on Stein's method for processes and it is here that the Generator Method has proved invaluable (see Barbour, Holst and Janson [4] for the Poisson process and Barbour [2] for diffusions). Reinert [14] has developed Stein's method for measure-valued random elements and here, too, the Generator Method comes into play. It is clear that Stein's method is an invaluable technique for finding rates of convergence in distribution which is still finding new areas of application.

7 Appendix

This section contains the proofs of some results that I have used above. The proofs consist of unilluminating calculations.

Proof of Proposition 2.3.

$$\begin{aligned}
\mathbb{E}\left[\left(1-\frac{n}{2}\mathbb{E}\left[(W-W')^2|W\right]\right)^2\right] &\leq \mathbb{E}\left[\left(1-\frac{n}{2}\mathbb{E}\left[(Y_I-Y_I^*)^2|Y_I\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(1-\frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[Y_i^2-2Y_iY_i^*+Y_i^{*2}|Y_i\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(1-\frac{1}{2}\sum_{i=1}^n (Y_i^2+\sigma_i^2)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n \sigma_i^2-\frac{1}{2}\sum_{i=1}^n (Y_i^2+\sigma_i^2)\right)^2\right] \\
&= \frac{1}{4}\sum_{i=1}^n \left(\mathbb{E}\left[Y_i^4\right]-\sigma_i^4\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[|W-W'|^3\right] &= \mathbb{E}\left[|Y_I-Y_I^*|^3\right] \\
&= \mathbb{E}\left[|Y_I^3-3Y_I^2Y_I^*+3Y_IY_I^{*2}\right] \\
&\leq 2\mathbb{E}\left[|Y_I|^2\right]+6\mathbb{E}\left[|Y_I^2Y_I^*|\right] \\
&\leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[|Y_i|^3\right]+6\mathbb{E}\left[Y_I^2\right]\sqrt{\mathbb{E}[Y_I^2]} \\
&\qquad \text{as } Y_I^* \sim Y_I \text{ and they are independent} \\
&= \frac{2}{n}\left(\sum_{i=1}^n \left(\mathbb{E}\left[|Y_i|^3\right]+3\sigma_i^3\right)\right)
\end{aligned}$$

□

Proof of Lemma 3.4.

$$\begin{aligned}
|\mathbb{E}[I_i(f(Y_i+Z_i+1)-f(Y_i+1))]| &= |\mathbb{E}\left[I_i\left(f(Y_i+I_{\gamma_1}+\cdots+I_{\gamma_k})-f(Y_i+I_{\gamma_1}+\cdots+I_{\gamma_{k-1}})\right.\right. \\
&\qquad +f(Y_i+I_{\gamma_1}+\cdots+I_{\gamma_{k-1}})-f(Y_i+I_{\gamma_1}+\cdots+I_{\gamma_{k-2}}) \\
&\qquad +\cdots-f(Y_i+1))]| \\
&\leq \Delta f\mathbb{E}\left[I_i\sum_{j\in\Gamma_i^s} I_j\right] \\
&= \Delta f\mathbb{E}[I_iZ_i]
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}[f(Y_i+1)-f(W+1)]| &= |\mathbb{E}[f(W-I_i-Z_i+1)-f(W+1)]| \\
&\leq \Delta f\mathbb{E}[Z_i+I_i] \\
&= \Delta f(p_i+\mathbb{E}[Z_i])
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}[I_if(Y_i+1)-p_if(Y_i+1)]| &= |\mathbb{E}[f(Y_i+1)\mathbb{E}[I_i-p_i|(I_j:j\in\Gamma_i^w)]]| \\
&\leq \|f\|\mathbb{E}[|\mathbb{E}[I_i|(I_j:j\in\Gamma_i^w)]-p_i|]
\end{aligned}$$

□

References

[1] Barbour, A.D. (1988). Stein's Method and Poisson Process Convergence. *Journal of Applied Probability* **25(A)**, 175–184.

[2] Barbour, A.D. (1990). Stein's Method for Diffusion Approximations. *Probability Theory and Related Fields* **84**, 297–322.

[3] Barbour, A.D., Holst, L. (1989). Some Applications of the Stein-Chen Method for Proving Poisson Convergence. *Advances in Applied Probability* **21**, 74–90.

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

[5] Bolthausen, E. (1984). An Estimate of the Remainder in a Combinatorial Central Limit Theorem. *Zeitschrift für Warscheinlichkeitstheorie und verwandte Gebiete* **66**, 379–386.

[6] Chen, L.H.Y. (1975). Poisson Approximation for Dependent Trials. *Annals of Probability* **3**, 534–545.

[7] Diaconis, P. (1991). An Example for Stein's Method. Technical Report No. 384, Department of Statistics, Stanford University, Stanford, California.

[8] Ethier, S.N., Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence.* Wiley.

[9] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. II. Wiley.

[10] Karatzas, I., Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*, 2nd Ed. Springer-Verlag.

[11] Lindvall, T. (1992). *Lectures on the Coupling Method.* Wiley.

[12] Norris, J.R. (1997). *Markov Chains.* Cambridge University Press.

[13] Reinert, G. (1998). Couplings for Normal Approximations with Stein's Method. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **41**, AMS, 193–207.

[14] Reinert, G. (1998). Stein's Method in Application to Empirical Measures. V Simposio De Probililidad Y Procesos Estocasticos.

[15] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* **2**, 583–602. University of California Press, Berkeley.

[16] Stein, C. (1986). *Approximate Computation of Expectations.* Institute of Mathematical Statistics, Hayward, California.