The Stein–Chen Method

Coupling Techniques for Probability Approximations

Fraser Daly[*]
Heriot–Watt University

April 2020

Contents

1 Introduction **2**

2 Gaussian approximation and related topics **3**

2.1 Introduction . . . 3
2.2 Sums of independent random variables . . . 7
2.3 Coupling and other approaches to Stein's method for Gaussian approximation 9
2.3.1 Local dependence . . . 9
2.3.2 Exchangeable pairs . . . 11
2.3.3 The zero-biased coupling . . . 12
2.3.4 The size-biased coupling . . . 16
2.3.5 The Malliavin–Stein method . . . 16
2.4 Variance bounds using zero biasing . . . 17

3 Exponential approximation **19**

3.1 Introduction . . . 19
3.2 Approximation of geometric sums . . . 21

4 Poisson approximation and related topics **23**

4.1 Poisson approximation by maximal coupling . . . 23
4.2 The Stein–Chen method for Poisson approximation . . . 24
4.2.1 Independent summands . . . 26
4.2.2 Dependent summands: the local approach . . . 27
4.3 The size-biased coupling and monotonicity . . . 28
4.3.1 Negative dependence . . . 29
4.3.2 Positive dependence . . . 32

[*] f.daly@hw.ac.uk

4.4 Further applications of monotone size-biased couplings 33
4.4.1 Bounds on the Poincaré constant 33
4.4.2 Entropy bounds 35
4.5 Concentration via size-biased couplings 36
4.6 Extensions and generalisations of Poisson approximation results 38

5 Approximation by geometric sums 40
5.1 The Stein equation and its solution 41
5.2 Markov chain hitting times 42
5.3 Geometric approximation for IFR random variables 46

6 Infinitely divisible distributions 48
6.1 Introduction 48
6.2 Application: Gumbel approximation for the maximum of exponential data . . 50

1 Introduction

The study of limit theorems in probability theory has a long and rich history: results related to the central limit theorem date back to de Moivre in the 1730s, who used a normal distribution to approximate probabilities associated with binomial random variables, and the convergence of the binomial distribution $\mathrm{Bin}(n, \lambda/n)$ to the Poisson distribution $\mathrm{Po}(\lambda)$ as $n \to \infty$ was first established by Poisson in 1837. Generalizations and extensions of these prototypical examples continue to find applications in diverse fields.

There are numerous techniques which may be used to establish limit theorems in probability. In these lectures we will focus on one, the Stein–Chen method (also referred to as Stein's method). Compared to many other techniques, Stein's method has three principal advantages:

(i) It is applicable in a wide variety of settings, including univariate, multivariate and stochastic process limits;

(ii) It can handle dependence between the underlying random variables; and

(iii) Alongside limit theorems, we can (usually) establish explicit bounds on the error in the approximation.

In these lectures we will mostly look at applications of Stein's method in univariate settings. Most of our attention will focus on coupling techniques that may be applied in conjunction with Stein's method, though we will also mention several other approaches to Stein's method.

Charles Stein's seminal 1972 paper [59] first introduced his technique in the setting of Gaussian approximation for sums of weakly dependent random variables. This was followed by work of Louis Chen [17], a student of Stein, who applied the same ideas to prove Poisson approximation results. Since then, these same techniques have been applied in a wide range of settings.

Good surveys of Stein's method can be found in the book edited by Barbour and Chen [8] and the paper by Ross [55]. The book [19] by Chen, Goldstein and Shao gives an extensive treatment of Stein's method for Gaussian approximation. The book [10] by Barbour, Holst and Janson is dedicated to Poisson approximations, with an emphasis on coupling-based approaches. These books all also present material beyond the case of univariate approximations.

We will begin with the case of Gaussian approximation (and some related topics) in Section 2. In Section 3 we will consider exponential approximation, and then in Section 4 we will move from a continuous to a discrete setting, looking at Poisson approximation. Again in a discrete context, Section 5 looks at approximation by geometric sums. Finally, in Section 6 we conclude with a brief look at one case in which multiple distributions may be considered simultaneously, the case of approximation by an infinitely divisible distribution with finite mean.

There are numerous other topics we could have also considered, even in the setting of univariate approximation, including binomial approximation [34], negative binomial approximation [56], beta approximation [41], chi-square approximation [38], approximation by the Laplace distribution [49] and variance-gamma approximation [37], among many others. We indicate these references here as an entry point into the relevant literature.

Beyond the univariate setting, see, for example [52] and [21] for starting points in the large amount of literature available on multivariate Gaussian approximation and Poisson process approximation, respectively.

2 Gaussian approximation and related topics

2.1 Introduction

As we have already noted, Gaussian approximation was the first setting in which Stein's method was applied [59], and was also the main focus of Stein's 1986 monograph [60] that amply demonstrated the power and versatility of this technique. We will use this first part of this section to outline the basic idea behind Stein's method, before stating and proving the lemmas necessary to make this approach rigorous. Following this, we will look at how this approach can be applied in a variety of settings using various different ideas. Our focus here will be on coupling techniques, though we will indicate briefly some of the other ideas which have also been successfully applied in conjunction with Stein's method for Gaussian approximation. Much of our discussion here is based upon the book [19] of Chen, Goldstein and Shao.

The starting point for Stein's method for Gaussian approximation is the relatively simple observation that a random variable Z has a standard Gaussian distribution if and only if

$$\mathbb{E}f'(Z) = \mathbb{E}[Zf(Z)]\,,$$

for all absolutely continuous functions $f : \mathbb{R} \mapsto \mathbb{R}$ for which these expectations exist.

Given a real-valued random variable W (with mean zero and variance one) which we think of (in some sense) as *approximately* Gaussian, we may therefore hope that

$$\mathbb{E}f'(W) \approx \mathbb{E}[Wf(W)]\,,$$

for all f as above. Roughly speaking, Stein's method quantifies the proximity of W to Gaussian by bounding the difference between the LHS and RHS of the above, over a wide enough class of functions f to give useful information.

To make this idea precise, suppose we have a given test function $h : \mathbb{R} \mapsto \mathbb{R}$, and let $f = f_h$ be the unique bounded solution to the following *Stein equation*:

$$h(x) - \mathbb{E}h(Z) = f'(x) - xf(x)\,, \tag{1}$$

for all $x \in \mathbb{R}$, where $Z \sim \mathrm{N}(0, 1)$.

We may usefully quantify the proximity of W and Z by a variety of metrics of the form

$$d_{\mathcal{H}}(W, Z) = \sup_{h\in\mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)|\,,$$

where $\mathcal{H}$ is a suitably rich class of functions. For example, if we take $\mathcal{H} = \mathcal{H}_W = \{h : |h(x) - h(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}$ to be the set of all Lipschitz functions on $\mathbb{R}$ with Lipschitz constant 1, then we obtain the Wasserstein distance:

$$d_W(W, Z) = \sup_{h\in\mathcal{H}_W} |\mathbb{E}h(W) - \mathbb{E}h(Z)|\,.$$

Similarly, if we take $\mathcal{H} = \mathcal{H}_K = \{I_{\{\cdot\leq y\}} : y \in \mathbb{R}\}$, we obtain the Kolmogorov distance:

$$d_K(W, Z) = \sup_{y\in\mathbb{R}} |\mathbb{P}(W \leq y) - \mathbb{P}(Z \leq y)|\,.$$

So, beginning with the Stein equation (1), replacing x by W, taking expectations, and then taking the supremum over the class of functions $\mathcal{H}$, we have

$$d_{\mathcal{H}}(W, Z) = \sup_{h\in\mathcal{H}} |\mathbb{E}[f'(W) - Wf(W)]|\,. \tag{2}$$

This equation is the essence of Stein's method for Gaussian approximation. The advantage here is that it turns out to be considerably simpler to bound the RHS of this equation than to bound the LHS directly. Partly, this is due to the fact that the random variable $Z \sim \mathrm{N}(0, 1)$ does not appear directly on the RHS, but only implicitly in the form of the equation itself. Thus, we no longer have to work with two random variables (W and Z), but only with the random variable W.

There are several techniques available for bounding the RHS of (2) which we will discuss later in this section. All of these will require some estimates of the boundedness or smoothness of the function $f = f_h$, which we discuss below. It is worth emphasising that the Stein equation we use here, and bounds on the corresponding solution f do not depend on the random variable W we wish to approximate, only on the fact that our target distribution is Gaussian.

Before we consider properties of f, we first state formally the characterisation of the standard Gaussian distribution from which this work springs; the proof of this result is deferred until later in this section. Throughout the remainder of this section, we let $Z \sim \mathrm{N}(0,1)$.

Lemma 2.1. *If X has a standard Gaussian distribution, then*

$$\mathbb{E}f'(X) = \mathbb{E}[Xf(X)] \tag{3}$$

for all absolutely continuous functions $f : \mathbb{R} \mapsto \mathbb{R}$ with $\mathbb{E}|f'(Z)| < \infty$. Conversely, if (3) holds for all bounded, continuous and piecewise continuously differentiable functions f with $\mathbb{E}|f'(Z)| < \infty$, then X has a standard Gaussian distribution.

The solution of the Stein equation (1) is given by the following lemma.

Lemma 2.2. *Let $h : \mathbb{R} \mapsto \mathbb{R}$ be a measurable function with $\mathbb{E}|h(Z)| < \infty$. The unique bounded solution to the Stein equation (1) is given by*

$$\begin{aligned} f(x) &= e^{x^2/2} \int_{-\infty}^{x} (h(y) - \mathbb{E}h(Z))\, e^{-y^2/2}\, dy \\ &= -e^{x^2/2} \int_{x}^{\infty} (h(y) - \mathbb{E}h(Z))\, e^{-y^2/2}\, dy \,. \end{aligned}$$

Exercise: Use the integrating factor $e^{-x^2/2}$ to solve the differential equation (1) and hence prove Lemma 2.2.

Bounds on the solution to this function f will be expressed in terms of the supremum norm: for any function g, let

$$\|g\|_\infty = \sup_x |g(x)| \,.$$

There are many bounds available on the solution f to the Stein equation; see Section 2.2 of [19] for a wide selection. We state here only the bounds that we will need for the particular applications we will discuss later on.

Lemma 2.3. *Let f be the function defined in Lemma 2.2.*

(i) If h is absolutely continuous then

$$\|f''\|_\infty \le 2\|h'\|_\infty \,. \tag{4}$$

(ii) For any $u, v, w \in \mathbb{R}$ and any h with $\|h\|_\infty \le 1$,

$$|(w+u)f(w+u) - (w+v)f(w+v)| \le \left(|w| + \frac{\sqrt{2\pi}}{4} \right) (|u| + |v|) \,. \tag{5}$$

Proof. We sketch the proof of (4) below. For a proof of (5), see Lemma 2.3 of [19].

Differentiating the Stein equation (1), we have

$$f''(x) = (1 + x^2)f(x) + x[h(x) - \mathbb{E}h(Z)] + h'(x)\,. \tag{6}$$

We can show that

$$h(x) - \mathbb{E}h(Z) = \int_{-\infty}^{x} h'(z)\Phi(z)\,dz - \int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,, \tag{7}$$

where $\Phi(z) = \mathbb{P}(Z \leq z)$. Letting $\varphi(z)$ be the corresponding density function, we can use the fact that $\varphi(x)f(x) = \int_{-\infty}^{x}[h(y) - \mathbb{E}h(Z)]\varphi(y)\,dy$ together with (7) to show that

$$-\varphi(x)f(x) = [1 - \Phi(x)]\int_{-\infty}^{x} h'(z)\Phi(z)\,dz + \Phi(x)\int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,. \tag{8}$$

Now, define $\theta(x) = \frac{\Phi(x)}{\varphi(x)}$. It can shown that $\theta''(x) \geq 0$ for all x, and that the following equations hold:

$$\theta''(x) = x + (1 + x^2)\theta(x)\,, \tag{9}$$

$$\theta''(-x) = \frac{1 + x^2}{\varphi(x)} - \theta''(x)\,. \tag{10}$$

Combining (6)–(10) we can show that

$$f''(x) = h'(x) - \theta''(-x)\int_{-\infty}^{x} h'(z)\Phi(z)\,dz - \theta''(x)\int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,. \tag{11}$$

After showing that

$$\theta''(-x)\int_{-\infty}^{x} \Phi(z)\,dz + \theta''(x)\int_{x}^{\infty} [1 - \Phi(z)]\,dz = 1\,,$$

for all x, the proof is completed by using the triangle inequality in (11). □

Proof of Lemma 2.1

Necessity: We, essentially, use an integration by parts approach here. Let f be an absolutely continuous function such that $\mathbb{E}|f'(Z)| < \infty$. If $X \sim \mathrm{N}(0, 1)$ then

$$\begin{aligned}\mathbb{E}f'(X) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(x)e^{-x^2/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{0} f'(x)\left(\int_{-\infty}^{x} -ye^{-y^2/2}\,dy\right)\,dx + \int_{0}^{\infty} f'(x)\left(\int_{x}^{\infty} ye^{-y^2/2}\,dy\right)\,dx\right].\end{aligned}$$

Using Fubini's theorem, we then have

$$\begin{aligned}\mathbb{E}f'(X) &= \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{0}\left(\int_{y}^{0} f'(x)\,dx\right)(-y)e^{-y^2/2}\,dy + \int_{0}^{\infty}\left(\int_{0}^{x} f'(x)\,dx\right)ye^{-y^2/2}\,dy\right]\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}[f(y)-f(0)]ye^{-y^2/2}\,dy\\ &= \mathbb{E}[Xf(X)]\,.\end{aligned}$$

Sufficiency: Fix $z \in \mathbb{R}$. Let f be the function defined by Lemma 2.2, with the choice $h(x) = I_{\{x\leq z\}}$. This function f is continuous and piecewise continuously differentiable, and we know from Lemma 2.3 that f is bounded. Hence, by assumption

$$0 = \mathbb{E}[f'(X) - Xf(X)] = \mathbb{E}[I_{\{X\leq z\}}] - \mathbb{P}(Z \leq z) = \mathbb{P}(X \leq z) - \mathbb{P}(Z \leq z)\,,$$

so that X has a standard normal distribution.

2.2 Sums of independent random variables

To illustrate the application of the above framework, we prove a central limit theorem for sums of independent random variables.

Theorem 2.4. *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}X_i = 0$ and Var$(X_i) = \sigma_i^2$ for each i. Suppose that $\sigma_1^2 + \cdots + \sigma_n^2 = 1$ and let $W = X_1 + \cdots + X_n$. Then*

$$d_W(W, Z) \leq 4\sum_{i=1}^{n}\mathbb{E}|X_i^3|\,,$$

where $Z \sim N(0,1)$.

Proof. We write $W_i = W - X_i$ for each i. Using (2), we need to bound

$$\sup_{h\in\mathcal{H}_W} |\mathbb{E}\,[f'(W) - Wf(W)]\,|\,.$$

To that end, we firstly note that $\mathbb{E}[Wf(W)] = \mathbb{E}\sum_{i=1}^{n} X_i f(W_i + X_i)$. Then, using a Taylor expansion,

$$X_i f(W_i + X_i) = X_i f(W_i) + X_i^2 \int_0^1 f'(W_i + uX_i)\,du\,.$$

By independence, the first term vanishes on taking expectations. Hence,

$$\mathbb{E}[Wf(W)] = \mathbb{E}\sum_{i=1}^{n} X_i^2 \int_0^1 f'(W_i + uX_i)\,du\,.$$

Also,

$$\begin{aligned}\mathbb{E}[f'(W)] &= \mathbb{E}\sum_{i=1}^{n}\sigma_i^2 f'(W)\\ &= \mathbb{E}\sum_{i=1}^{n}\sigma_i^2 f'(W_i) + \mathbb{E}\sum_{i=1}^{n}\sigma_i^2\left(f'(W)-f'(W_i)\right)\\ &= \mathbb{E}\sum_{i=1}^{n}X_i^2 f'(W_i) + \mathbb{E}\sum_{i=1}^{n}\sigma_i^2\left(f'(W)-f'(W_i)\right)\,.\end{aligned}$$

Combining these,

$$\begin{aligned}&\mathbb{E}[f'(W)-Wf(W)]\\ &\quad= \mathbb{E}\sum_{i=1}^{n}X_i^2\int_0^1\left(f'(W_i)-f'(W_i+uX_i)\right)\,du + \mathbb{E}\sum_{i=1}^{n}\sigma_i^2\left(f'(W)-f'(W_i)\right)\,.\end{aligned}$$

By the mean value theorem, $|f'(W_i)-f'(W_i+uX_i)| \leq |X_i|\|f''\|_\infty$. The same bound may also be applied in the second term of the above (with $u=1$). Hence

$$|\mathbb{E}[f'(W)-Wf(W)]| \leq \|f''\|_\infty\sum_{i=1}^{n}\left(\mathbb{E}|X_i^3| + \sigma_i^2\mathbb{E}|X_i|\right) \leq 2\|f''\|_\infty\sum_{i=1}^{n}\mathbb{E}|X_i^3|\,.$$

Applying the bound (4), we then have

$$|\mathbb{E}[f'(W)-Wf(W)]| \leq 4\|h'\|_\infty\sum_{i=1}^{n}\mathbb{E}|X_i^3|\,.$$

The proof is completed by noting that $\|h'\|_\infty \leq 1$ for each $h\in\mathcal{H}_W$. □

As (perhaps) suggested by the above application, it is often significantly easier in the Gaussian setting to prove approximation results in ‘smooth metrics’ (such as the Wasserstein distance) where the functions $h\in\mathcal{H}$ satisfy some differentiability of other smoothness conditions. However, these are certainly not the only metrics of interest. For example, the Kolmogorov distance is of practical importance, but relies on (non-smooth) indicator test functions h. One solution to this problem is to replace h by a smoothed version, and then to control the difference between the original test function h and its smoothed version. This typically works well, though leads to additional technical complications compared to proving results in smooth metrics.

Somewhat weaker bounds in non-smooth distances can also be established by exploiting general inequalities such as that in the lemma below (see page 13 of [8]).

Lemma 2.5. *Suppose that there exists $\delta>0$ such that, for any $h\in\mathcal{H}_W$, $|\mathbb{E}h(W)-\mathbb{E}h(Z)| \leq \delta\|h'\|_\infty$. Then $d_K(W,Z)\leq 2\sqrt{\delta}$.*

Proof. We may assume that $\delta < 1/4$, otherwise the result is trivial. Now, let $\alpha = \delta^{1/2}(2\pi)^{1/4}$. For a fixed z, let $h_\alpha(x)$ be 1 for $x \le z$, be 0 for $x \ge z + \alpha$, and interpolate linearly between these two values for $z < x < z + \alpha$. It is clear that $\|h'\|_\infty = 1/\alpha$, and by the assumptions of the lemma we have

$$\begin{aligned}\mathbb{P}(W \le z) - \mathbb{P}(Z \le z) &\le \mathbb{E}h_\alpha(W) - \mathbb{E}h_\alpha(Z) + \mathbb{E}h_\alpha(Z) - \mathbb{P}(Z \le z)\\ &\le \frac{\delta}{\alpha} + \mathbb{P}(z \le Z \le z + \alpha) \le \frac{\delta}{\alpha} + \frac{\alpha}{\sqrt{2\pi}} \le 2(2\pi)^{-1/4}\delta^{1/2} \le 2\delta^{1/2}\,.\end{aligned}$$

A similar argument gives $\mathbb{P}(W \le z) - \mathbb{P}(Z \le z) \ge -2\delta^{1/2}$ and completes the proof. □

Note that these difficulties do not arise in the settings where the underlying distributions are discrete.

2.3 Coupling and other approaches to Stein's method for Gaussian approximation

In this section we will describe a selection of approaches that have been successfully applied in conjunction with Stein's method for Gaussian approximation to yield explicit error bounds in Gaussian approximation in settings more exotic than sums of independent random variables.

2.3.1 Local dependence

The argument of Theorem 2.4 may be extended to the case where $X_1, \dots, X_n$ satisfy a *local dependence* assumption. We let $W = \sum_{j\in\mathcal{J}} X_j$, where $\mathcal{J}$ is a fixed index set with n elements. We assume that $\mathbb{E}X_j = 0$ for all j and that $\mathrm{Var}(W) = 1$. For any subset $A \subseteq \mathcal{J}$, we define $X_A = \{X_j : j \in A\}$. There are numerous ways that assumptions of local dependence can be made, here we will follow [20] and assume the following:

$$\begin{aligned}&\text{For each } i \in \mathcal{J}\text{, there exist } A_i \subseteq B_i \subseteq \mathcal{J} \text{ such that } X_i \text{ is}\\ &\text{independent of } X_{A_i^c} \text{ and } X_{A_i} \text{ is independent of } X_{B_i^c}.\end{aligned} \qquad (12)$$

Under this assumption, we have the following bound.

Theorem 2.6. *Let $\{X_i : i \in \mathcal{J}\}$ be such that $\mathbb{E}X_i = 0$ for each i and (12) holds. Let $W = \sum_{j\in\mathcal{J}} X_j$, and assume that $\mathrm{Var}(W) = 1$. Let $\eta_i = \sum_{j\in A_i} X_j$ and $\tau_i = \sum_{j\in B_i} X_j$. Then*

$$d_W(W, Z) \le 2\sum_{i\in\mathcal{J}} \{\mathbb{E}|X_i\eta_i\tau_i| + |\mathbb{E}[X_i\eta_i]|\mathbb{E}|\tau_i|\} + \sum_{i\in\mathcal{J}} \mathbb{E}|X_i\eta_i^2|\,,$$

where $Z \sim N(0, 1)$.

Proof. Let $h \in \mathcal{H}_W$, and f be the corresponding solution to the Stein equation. We have

$$\mathbb{E}[Wf(W)] = \sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i f(W)\right] = \sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i \left\{f(W) - f(W - \eta_i)\right\}\right] ,$$

using the independence of X_i and $W - \eta_i$. Thus, we may write

$$\mathbb{E}[Wf(W)] = \sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i \left\{f(W) - f(W - \eta_i) - \eta_i f'(W)\right\}\right] + \mathbb{E}\left[\left(\sum_{i \in \mathcal{J}} X_i \eta_i\right) f'(W)\right] .$$

Now, since $\mathbb{E}X_i = 0$ for each i, and by the independence of X_i and $X_{A_i^c}$,

$$1 = \mathbb{E}[W^2] = \sum_{i \in \mathcal{J}} \sum_{j \in \mathcal{J}} \mathbb{E}[X_i X_j] = \sum_{i \in \mathcal{J}} \mathbb{E}[X_i \eta_i] ,$$

giving

$$\begin{aligned}\mathbb{E}[f'(W) - Wf(W)] = -\mathbb{E}&\left[\sum_{i \in \mathcal{J}} \left(X_i \eta_i - \mathbb{E}[X_i \eta_i]\right) f'(W)\right] \\ &- \sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i \left\{f(W) - f(W - \eta_i) - \eta_i f'(W)\right\}\right] .\end{aligned}$$

Since $W - \tau_i$ and $X_i \eta_i$ are independent, we may further write

$$\begin{aligned}\mathbb{E}[f'(W) - Wf(W)] = -\mathbb{E}&\left[\sum_{i \in \mathcal{J}} \left(X_i \eta_i - \mathbb{E}[X_i \eta_i]\right) \left(f'(W) - f'(W - \tau_i)\right)\right] \\ &- \sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i \left\{f(W) - f(W - \eta_i) - \eta_i f'(W)\right\}\right] .\end{aligned}$$

Exercise: Use suitable Taylor expansions to show the following:

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{i \in \mathcal{J}} \left(X_i \eta_i - \mathbb{E}[X_i \eta_i]\right) \left(f'(W) - f'(W - \tau_i)\right)\right]\right| &\leq \|f''\|_\infty \sum_{i \in \mathcal{J}} \left\{\mathbb{E}|X_i \eta_i \tau_i| + |\mathbb{E}[X_i \eta_i]| \mathbb{E}|\tau_i|\right\} , \\ \left|\sum_{i \in \mathcal{J}} \mathbb{E}\left[X_i \left\{f(W) - f(W - \eta_i) - \eta_i f'(W)\right\}\right]\right| &\leq \frac{1}{2} \|f''\|_\infty \sum_{i \in \mathcal{J}} \mathbb{E}|X_i \eta_i^2| .\end{aligned}$$

The proof is complete on applying (4). □

Note that if the random variables X_i were independent, then we could choose $A_i = B_i = \{i\}$, so that $\eta_i = \tau_i = X_i$. We can then use Theorem 2.6 to obtain $d_W(W, Z) \leq 5 \sum_{i \in \mathcal{J}} \mathbb{E}|X_i^3|$, which is only slightly worse than the bound we obtained directly in this case in Theorem 2.4.

Example: Local maxima on a graph

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and IID continuous random variables $\{\xi_i : i \in \mathcal{V}\}$. For each vertex i we let $N_i = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ be the set of vertices neighbouring i. Define the indicator random variables

$$Y_i = \begin{cases} 1, & \text{if } \xi_i > \xi_j \text{ for all } j \in N_i\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

that show when vertex i is a local maximum. Then $Y = \sum_{i \in \mathcal{V}} Y_i$ counts the number of local maxima on the graph.

If we write $d(i, j)$ for the distance between vertices i and j in the graph (that is, the minimum number of edges that need to be used to move from i to j), then (12) is satisfied with the choices $A_i = \{j \in \mathcal{V} : d(i, j) \leq 2\}$ and $B_i = \cup_{j \in A_i} A_j = \{j \in \mathcal{V} : d(i, j) \leq 4\}$.

Exercise: Show that if $\mathcal{G}$ is a regular graph with degree d (i.e., all vertices have d edges), then $\mathbb{E}Y = |\mathcal{V}|/(d + 1)$ and (more challenging!)

$$\text{Var}(Y) = \sum_{i,j \in \mathcal{V}, d(i,j)=2} s(i, j)(2d + 2 - s(i, j))^{-1}(d + 1)^{-2}\,,$$

where $s(i, j) = |N_i \cap N_j|$.

2.3.2 Exchangeable pairs

Historically, one of the earliest approaches to Stein's method for Gaussian approximation for a random variable W (with mean zero and variance 1, say) relied on the construction of a pair (W, W') of exchangeable random variables (i.e., for which (W, W') has the same distribution as (W', W)) satisfying the following 'linear regression' condition:

$$\mathbb{E}[W'|W] = (1 - \lambda)W\,,$$

for some $\lambda \in (0, 1)$. Subsequent work allows for a relaxation of this condition, for example in permitting a remainder term to appear.

This condition appears, in some sense, as an analogue of the fact that if (W, W') were bivariate normal with correlation ρ, then $\mathbb{E}[W'|W] = \rho W$.

Under this condition, the following theorem (which we give without proof; see Section 4.5 of [19]) may be established.

Theorem 2.7. *Let W have distribution function F, and satisfy $\mathbb{E}W = 0$ and Var$(W) = 1$. Let (W, W') be exchangeable and such that $\mathbb{E}[W'|W] = (1 - \lambda)W$ for some $\lambda \in (0, 1)$. Then*

$$d_W(W, Z) \leq \frac{1}{2\lambda}\left[\sqrt{\frac{2}{\pi}}\sqrt{Var\left(\mathbb{E}[(W' - W)^2|W]\right)} + \mathbb{E}[|W' - W|^3]\right]\,,$$

where $Z \sim N(0, 1)$.

The starting point of the proof of this result is the observation that for (W, W') exchangeable, $\mathbb{E}g(W, W') = 0$ for all antisymmetric functions g for which this expectation exists. We apply this with the choice

$$g(x, y) = (x - y)[f(y) + f(x)] \, ,$$

where we will take f to be the solution of our Stein equation, which is sufficiently bounded to guarantee existence of the expectation. Hence,

$$\begin{aligned} 0 &= \mathbb{E}\left[(W - W')(f(W') + f(W))\right] \\ &= \mathbb{E}\left[(W - W')(f(W') - f(W))\right] + 2\mathbb{E}\left[f(W)(W - W')\right] \\ &= \mathbb{E}\left[(W - W')(f(W') - f(W))\right] + 2\mathbb{E}\left[f(W)\mathbb{E}[W - W'|W]\right] \\ &= \mathbb{E}\left[(W - W')(f(W') - f(W))\right] + 2\lambda\mathbb{E}\left[Wf(W)\right] \, , \end{aligned}$$

so that

$$\mathbb{E}\left[Wf(W)\right] = \frac{1}{2\lambda}\mathbb{E}\left[(W - W')(f(W') - f(W))\right] \, .$$

As an illustration of the construction of an exchangeable pair satisfying the conditions of Theorem 2.7, suppose $W = X_1 + \cdots + X_n$ is a sum of IID random variables, each with mean zero, and with $\text{Var}(W) = 1$. Letting I be uniformly distributed on $\{1, \ldots, n\}$ (independent of all else) and $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$, we may write $W' = W - X_I + X_I'$. Then W and W' are exchangeable.

Exercise: In this setting, show that $\mathbb{E}[W'|W] = \left(1 - \frac{1}{n}\right) W$.

Exchangeable pairs satisfying the conditions of Theorem 2.7 can also be constructed as successive states of a reversible Markov chain in stationarity with stationary distribution F. For example, let W be the sum of a simple random sample (without replacement) of size n from a set of numbers $\mathcal{S} = \{a_1, \ldots, a_N\}$ of size $N > n$. We assume appropriate normalisation to give $\mathbb{E}W = 0$ and $\text{Var}(W) = 1$. We may construct a Markov chain whose state space consists of subsets of $\mathcal{S}$ of size n. This Markov chain evolves according to the following rule: at each time step, choose uniformly at random an element of $\mathcal{S}$ in your current sample, and an element of $\mathcal{S}$ not in your current sample, and interchange them. We start the chain in stationarity, so that it remains stationary, and it is clearly reversible. Letting W and W' be two consecutive states of this chain, these random variables are exchangeable, and we may prove the linearity condition required by Theorem 2.7 in a similar way to the IID example above.

2.3.3 The zero-biased coupling

The zero-biased transformation of a distribution was first introduced by Goldstein and Reinert [40]:

Definition 2.8. *Let W be a random variable with mean zero and finite variance σ^2. The random variable W^z has the W-zero-biased distribution if*

$$\mathbb{E}[Wg(W)] = \sigma^2\mathbb{E}g'(W^z) \, ,$$

for all absolutely continuous functions $g : \mathbb{R} \mapsto \mathbb{R}$ for which these expectations exist.

It was shown in [40] that W^z exists for all W with zero mean and finite variance. This random variable W^z is always continuous (regardless of whether W is discrete or continuous), and has density function given by

$$p^z(x) = \sigma^{-2}\mathbb{E}[W1_{\{W>x\}}] \, .$$

Exercise: Check that this is indeed the density function of W^z.

Exercise: Verify that for any constant $c \neq 0$, $(cW)^z$ is equal in distribution to cW^z.

It is clear from the definition and Stein's characterisation of the Gaussian distribution that the Gaussian (with mean zero and variance σ^2) is the unique fixed point of the zero-bias transformation. We may quantify how close W is to Gaussian with the following result, which follows almost immediately from the definition.

Theorem 2.9. *Let $\mathbb{E}W = 0$ and Var$(W) = 1$. Then*

$$d_W(W, Z) \leq 2\mathbb{E}|W - W^z| \, ,$$

where $Z \sim$ N$(0,1)$ and W^z has the W-zero-biased distribution.

Proof. For any $h \in \mathcal{H}_W$, we use the Stein equation (1) to write

$$\begin{aligned} |\mathbb{E}h(W) - \mathbb{E}h(Z)| = |\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)]| &= |\mathbb{E}f'(W) - \mathbb{E}f'(W^z)| \\ &\leq \|f''\|_\infty \mathbb{E}|W - W^z| \leq 2\|h'\|_\infty \mathbb{E}|W - W^z| \, , \end{aligned}$$

where the final inequality uses (4). The result follows by taking the supremum over $h \in \mathcal{H}_W$. □

We illustrate this result with an application to a sum of independent random variables. For this, we will need the following lemma, which allows us to construct the zero-biased distribution of the sum in terms of the zero-biased distribution of the individual summands.

Lemma 2.10. *Let $X_1, \dots, X_n$ be independent, mean zero random variables with Var$(X_i) = \sigma_i^2$. Let $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2 > 0$ and let $W = X_1 + \cdots + X_n$. Define the random index I, independent of $X_1, \dots, X_n$, such that*

$$\mathbb{P}(I = i) = \frac{\sigma_i^2}{\sigma^2} \, , \qquad i = 1, \dots, n \, .$$

If X_i^z has the X_i-zero-biased distribution (independent of I and X_j for $j \neq i$), then the random variable

$$W^z = W - X_I + X_I^z$$

has the W-zero-biased distribution.

Proof. For all absolutely continuous functions g (either directly or by using a suitable limiting argument), we use independence of the underlying random variables to write

$$\begin{aligned}\mathbb{E}[Wg(W)] &= \sum_{i=1}^{n}\mathbb{E}[X_i g(W)] = \sum_{i=1}^{n}\mathbb{E}\left[X_i g\left(X_i + \sum_{j\neq i} X_j\right)\right] \\ &= \sum_{i=1}^{n}\sigma_i^2\mathbb{E}\left[g'\left(X_i^z + \sum_{j\neq i} X_j\right)\right] = \sigma^2\sum_{i=1}^{n}\frac{\sigma_i^2}{\sigma^2}\mathbb{E}\left[g'\left(W - X_i + X_i^z\right)\right] \\ &= \sigma^2\mathbb{E}g'(W - X_I + X_I^z) = \sigma^2\mathbb{E}g'(W^z)\,.\end{aligned}$$

□

Suppose now that $X_1, \ldots, X_n$ are independent, mean zero random variables whose variances sum to 1. Letting $W = X_1 + \cdots + X_n$ and $Z \sim \mathrm{N}(0,1)$, we may combine the above results to get that

$$d_W(W,Z) \leq 2\mathbb{E}|X_I - X_I^z| = 2\sum_{i=1}^{n}\sigma_i^2\mathbb{E}|X_i - X_i^z|\,,$$

using the notation of Lemma 2.10. Taking $g(x) = x^2\mathrm{sgn}(x)$, so that $g'(x) = 2|x|$, in the definition of X_i^z gives us that

$$\sigma_i^2\mathbb{E}|X_i^z| \leq \frac{1}{2}\mathbb{E}|X_i^3|\,,$$

so that by the triangle inequality

$$d_W(W,Z) \leq 2\sum_{i=1}^{n}\left(\sigma_i^2\mathbb{E}|X_i| + \frac{1}{2}\mathbb{E}|X_i^3|\right)\,.$$

Many other situations may also be treated by the bound in Theorem 2.9; see Chapter 4 of [19] for several applications of this result.

In proving bounds in normal approximation in Kolmogorov distance using zero-biased couplings, assumptions of boundedness of the underlying coupling are often very useful, as illustrated by the following theorem (which is given as Theorem 5.1 of [19]).

Theorem 2.11. *Let W be a random variable with mean zero and variance 1. Letting W^z have the W-zero-biased distribution, suppose we may construct W and W^z on the same space such that $|W - W^z| \leq \delta$ almost surely. Then*

$$d_K(W,Z) \leq 2.03\delta\,,$$

where $Z \sim N(0,1)$.

Proof. For any $z \in \mathbb{R}$, we let $\Phi(z) = \mathbb{P}(Z \leq z)$ and apply a well-known inequality for increments of Φ to write

$$\begin{aligned}\mathbb{P}(W \leq z) - \Phi(z) &= [\Phi(z-\delta) - \Phi(z)] + \mathbb{P}(W \leq z) - \Phi(z-\delta) \\ &\geq -\frac{\delta}{\sqrt{2\pi}} + \mathbb{P}(W \leq z) - \Phi(z-\delta)\,. \end{aligned} \tag{13}$$

Now, let f be the solution to the Stein equation (1) for the test function $h(x) = I_{\{x \leq z-\delta\}}$. Then we have that

$$\begin{aligned} f'(W^z) &= I_{\{W^z \leq z-\delta\}} - \Phi(z-\delta) + W^z f(W^z) \\ &\leq I_{\{W \leq z\}} - \Phi(z-\delta) + W^z f(W^z)\,. \end{aligned} \tag{14}$$

Taking expectations in (14) and applying this in (13) we obtain

$$\begin{aligned}\mathbb{P}(W \leq z) - \Phi(z) &\geq -\frac{\delta}{\sqrt{2\pi}} + \mathbb{E}\left[f'(W^z) - W^z f(W^z)\right] \\ &= -\frac{\delta}{\sqrt{2\pi}} + \mathbb{E}\left[W f(W) - W^z f(W^z)\right]\,. \end{aligned} \tag{15}$$

Letting $\Delta = W^z - W$ and using (5) we have

$$\begin{aligned}\left|\mathbb{E}\left[W f(W) - W^z f(W^z)\right]\right| &= \left|\mathbb{E}\left[W f(W) - (W+\Delta) f(W+\Delta)\right]\right| \\ &\leq \mathbb{E}\left[\left(|W| + \frac{\sqrt{2\pi}}{4}\right)|\Delta|\right] \\ &\leq \delta\left(1 + \frac{\sqrt{2\pi}}{4}\right)\,. \end{aligned}$$

Combining this with (15) we have

$$\mathbb{P}(W \leq z) - \Phi(z) \geq -\delta\left(\frac{1}{\sqrt{2\pi}} + 1 + \frac{\sqrt{2\pi}}{4}\right) \geq -2.03\delta\,.$$

A similar argument gives the reverse inequality, and the result follows by taking the supremum over $z \in \mathbb{R}$. □

Example: Combinatorial central limit theorem

Let $(a_{i,j})_{i,j=1}^n$ be an array of real numbers and π a uniformly chosen permutation of $\{1, \ldots, n\}$. Let $Y = \sum_{i=1}^n a_{i,\pi(i)}$. We further define

$$a_{\bullet\bullet} = \frac{1}{n^2}\sum_{i,j=1}^n a_{i,j}\,, \quad a_{i\bullet} = \frac{1}{n}\sum_{j=1}^n a_{i,j}\,, \quad \text{and} \quad a_{\bullet j} = \frac{1}{n}\sum_{i=1}^n a_{i,j}\,.$$

Note that $\mathbb{E}Y = na_{\bullet\bullet}$ and

$$\text{Var}(Y) = \sigma^2 = \frac{1}{n-1}\sum_{i,j=1}^{n} (a_{i,j} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet})^2 \, .$$

See, for example, Section 4.4 of [19]. Letting $W = \sigma^{-1}(Y - na_{\bullet\bullet})$ and

$$C = \max_{1\le i,j\le n} |a_{i,j} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet}| \, ,$$

the proof of Theorem 6.1 of [19] shows that $\mathbb{E}|W - W^z| \le 8C\sigma^{-1}$, and so we have from Theorem 2.11 that

$$d_K(W, Z) \le \frac{16.3C}{\sigma} \, .$$

2.3.4 The size-biased coupling

Given a non-negative random variable Y with $\mathbb{E}Y = \mu > 0$, we may define $Y^\star$, the size-biased version of Y, by letting

$$\mu\mathbb{E}[g(Y^\star)] = \mathbb{E}[Yg(Y)] \, ,$$

for all functions $g : \mathbb{R}^+ \mapsto \mathbb{R}$ for which these expectations exist. The strongest links between size-biasing and Stein's method come when considering Poisson approximation, and we will discuss the construction and properties of size-biased random variables in detail in Section 4. For now, we just note that this construction can also be used to yield Gaussian approximation results. As in the zero-biasing case above, assumptions of boundedness of size-biased couplings may be used to help yield bounds in Kolmogorov distance. We state one typical result here (for a proof, see Section 5.3 of [19]), but defer any further discussion of size-biasing until Section 4.

Theorem 2.12. *Let Y be a non-negative random variable with mean $0 < \mu < \infty$ and variance $0 < \sigma^2 < \infty$. Suppose $Y^\star$ may be coupled to Y in such a way that $|Y^\star - Y| \le A$ almost surely, for some A. Let $W = (Y - \mu)/\sigma$. Then*

$$d_K(W, Z) \le \frac{6\mu A^2}{\sigma^3} + \frac{2\mu}{\sigma^2}\sqrt{\mathit{Var}(\mathbb{E}[Y^\star - Y|Y])} \, ,$$

where $Z \sim N(0, 1)$.

2.3.5 The Malliavin–Stein method

Finally in this section, we note relatively recent work combining Stein's method with the tools of Malliavin calculus. This begins with the observation that the Malliavin calculus integration by parts formula may be combined with analogous techniques at the heart of the proof of the characterisation of the Gaussian distribution, Lemma 2.1. See the book of Nourdin and Peccati [47] for a starting point for the study of these techniques.

2.4 Variance bounds using zero biasing

Isoperimetric inequalities, giving upper bounds on the variance of a function of a random variable, have a long and rich history, beginning with the work of Chernoff [22]. Chernoff proved that if $Z \sim \mathrm{N}(0, \sigma^2)$ then

$$\mathrm{Var}(g(Z)) \leq \sigma^2 \mathbb{E}[(g'(Z))^2] \,,$$

for any absolutely continuous function $g : \mathbb{R} \mapsto \mathbb{R}$ such that $g(Z)$ has finite variance. This inequality has since been investigated and generalized by many authors, including Cacoullos [14], Chen [18], Klaassen [44] and Borovkov and Utev [12]. To accompany these upper variance bounds, many of these authors have also established corresponding lower bounds, in the form of generalized Cramér-Rao inequalities. In particular, if $Z \sim \mathrm{N}(0, \sigma^2)$ we have

$$\mathrm{Var}(g(Z)) \geq \sigma^2 \mathbb{E}[g'(Z)]^2,$$

see, for example, [14]. There has since been much further work on such inequalities. Here we present (a special case of) some recent work [28] that allows the proof of such upper and lower variance bounds in the framework of zero-biased couplings.

Let W be a random variable with mean zero and finite, non-zero variance σ^2. Recall the defining relation for the zero-biased distribution:

$$\mathbb{E}[W \alpha(W)] = \sigma^2 \mathbb{E}\alpha'(W^z) \,, \tag{16}$$

for all absolutely continuous functions $\alpha : \mathbb{R} \mapsto \mathbb{R}$ for which these expectations exist.

Let g be a real-valued differentiable function such that all expectations and variances written below exist. We then have that

$$\begin{aligned} \mathrm{Var}(g(W)) \leq \mathbb{E}\left[(g(W) - g(0))^2\right] &= \mathbb{E}\left[\left(\int_0^W g'(u)\, du\right)^2\right] \\ &\leq \mathbb{E}\left[W \int_0^W (g'(u))^2 \, du\right] \,, \end{aligned}$$

where the final equality follows from the Cauchy-Schwarz inequality. Applying (16) and rules for differentiating integrals we deduce that

$$\mathrm{Var}(g(W)) \leq \sigma^2 \mathbb{E}\left[(g'(W^z))^2\right] \,. \tag{17}$$

The definition (16) can also readily be combined with the Cauchy-Schwarz inequality to obtain lower variance bounds:

$$(\mathbb{E}\,[W g(W)])^2 = (\mathbb{E}\,[W(g(W) - \mathbb{E}[g(W)])])^2 \leq \mathbb{E}\left[W^2\right] \mathrm{Var}(g(W)) \,.$$

Then from (16) we have

$$\mathrm{Var}(g(W)) \geq \sigma^2 (\mathbb{E}\,[g'(W^z)])^2 \,.$$

It would, of course, be preferable to have variance bounds in terms of the original random variable W rather than its zero-biased version. Using that, for any twice differentiable function g, we have

$$\left|g'(x+t)^2 - g'(x)^2\right| \leq 2\|g'g''\|_\infty |t| \,,$$

we may immediately deduce from (17) that

$$\text{Var}(g(W)) \leq \sigma^2 \mathbb{E}\left[g'(W)^2\right] + 2\sigma^2 \|g'g''\|_\infty \mathbb{E}|W^z - W| \,. \tag{18}$$

This can be applied in situations in which the zero-biased coupling is bounded, as in the combinatorial central limit theorem considered in Section 2.3.3.

We use the reminder of this section to explore the example in which W is a sum of (suitably normalised) independent and identically distributed Bernoulli random variables, each with parameter $p = 1-q$. That is, $W = X_1 + \cdots + X_n$, where, for each i,

$$\mathbb{P}\left(X_i = \frac{q}{\sqrt{npq}}\right) = p \,, \qquad \text{and} \qquad \mathbb{P}\left(X_i = \frac{-p}{\sqrt{npq}}\right) = q \,,$$

so that $\mathbb{E}X_i = 0$ and $\text{Var}(X_i) = 1/n$.

In this case, using Lemma 2.10 in conjunction with (18) gives the bound

$$\text{Var}(g(W)) \leq \mathbb{E}\left[g'(W)^2\right] + 2\|g'g''\|_\infty \mathbb{E}|X_1^z - X_1| \,.$$

Exercise: Show that we may construct X_1^z as $\frac{U-p}{\sqrt{npq}}$, where $U \sim \text{U}(0,1)$ has a uniform distribution.

Using this fact, Corollary 4.1 of [19] shows that we may couple X_1 and X_1^z such that $\mathbb{E}|X_1^z - X_1| = \frac{p^2+q^2}{2\sqrt{npq}}$, which gives

$$\text{Var}(g(W)) \leq \mathbb{E}\left[g'(W)^2\right] + \|g'g''\|_\infty \frac{p^2+q^2}{\sqrt{npq}} \,.$$

We may derive tighter bounds in this example under restrictions on the class of functions g we consider. Again using the fact that a zero-biased Bernoulli distribution is uniformly distributed, we have that X_1^z is smaller than X_1 in a convex sense; that is, $\mathbb{E}\beta(X_1^z) \leq \mathbb{E}\beta(X_1)$ for all convex functions β. Using Lemma 2.10 together with standard closure properties of convex ordering under mixtures and convolutions (see Section 3.A of [58]), it then follows that W^z is smaller than W in a convex sense.

Using this fact, (17) gives

$$\text{Var}(g(W)) \leq \mathbb{E}\left[g'(W)^2\right] \,,$$

for any function g such that the function $x \mapsto g'(x)^2$ is convex.

3 Exponential approximation

3.1 Introduction

We use this section to illustrate how the same principles we have developed for Gaussian approximation using Stein's method may also be used in many other contexts. We do this by looking at exponential approximation. We will later also study Poisson approximation, among other topics. Recall that $Z \sim \text{Exp}(\nu)$ has an exponential distribution with rate ν if Z has density function $\nu e^{-\nu z}$ for $z > 0$. See [16], [36] and [50] for approaches to Stein's method for exponential approximation; the particular techniques we discuss here are developed by [50].

As in the Gaussian case, we begin with a suitable characterisation of the exponential distribution.

Lemma 3.1. *A random variable X has and exponential distribution with rate 1, written $X \sim$ Exp(1), if and only if*

$$\mathbb{E}f'(X) = \mathbb{E}f(X) - f(0)\,, \tag{19}$$

for all absolutely continuous $f : \mathbb{R}^+ \mapsto \mathbb{R}$ such that $\mathbb{E}|f'(Z)| < \infty$, where $Z \sim$ Exp(1).

Exercise: Check that if $X \sim \text{Exp}(1)$ then (19) holds for f as in the above lemma.

Exercise: Prove the other part of the lemma too: check that the condition (19) shows that the Laplace transform of X is that of an exponential random variable with rate 1.

Using this characterisation, we define the following Stein equation: for a given function $h : \mathbb{R}^+ \mapsto \mathbb{R}$, we let $f : \mathbb{R}^+ \mapsto \mathbb{R}$ be such that $f(0) = 0$ and

$$f'(x) - f(x) = h(x) - \mathbb{E}h(Z)\,, \tag{20}$$

where $Z \sim \text{Exp}(1)$. Note that f defined in this way is unique.

Exercise: Check that f is given by

$$f(x) = -e^x \int_x^\infty [h(t) - \mathbb{E}h(Z)]\, e^{-t}\, dt\,. \tag{21}$$

In order to apply this in practice we will (as in the Gaussian case) need bounds on the solution to this Stein equation. The bound that we will need here is presented in the following lemma; a selection of other analogous bounds may be derived similarly.

Lemma 3.2. *Let $h : \mathbb{R}^+ \mapsto \mathbb{R}$ be absolutely continuous and let f solve (20). Then*

$$\|f''\|_\infty \leq 2\|h'\|_\infty\,. \tag{22}$$

Proof. Differentiating (20), since h is absolutely continuous, we have that f satisfies

$$f''(x) = f'(x) + h'(x)\,,$$

and so it remains only to prove that $|f'(x)| \leq \|h'\|_\infty$. By the same arguments that showed that (21) solves (20) in the exercise above, we may show that

$$f'(x) = -e^x \int_x^\infty h'(t)e^{-t}\,dt\,,$$

from which it follows that

$$|f'(x)| \leq \|h'\|_\infty e^x \int_x^\infty e^{-t}\,dt = \|h'\|_\infty\,,$$

as required. □

We will again focus on relevant coupling techniques which may be applied in conjunction with Stein's method for exponential approximation. Other approaches (e.g., exchangeable pairs [36]) are available. We use the *equilibrium coupling*, defined below in this continuous setting. Discrete analogues of this coupling have also been employed for geometric [51] and negative binomial [56] approximations.

Definition 3.3. *Let $W \geq 0$ be a random variable with mean λ. The random variable W^e has the equilibrium distribution with respect to W if*

$$\mathbb{E}g(W) - g(0) = \lambda \mathbb{E}g'(W^e)\,,$$

for all Lipschitz functions $g : \mathbb{R}^+ \mapsto \mathbb{R}$.

The key to this definition is that the exponential distribution is the unique fixed-point of this transformation; this follows from (a slight generalisation of) our characterisation of the exponential distribution. That is, we can think of our characterisation of the exponential distribution as saying that X and X^e have the same distribution if and only if X is exponentially distributed.

Equivalently, we may define W^e by using the integrated tail of W:

$$\mathbb{P}(W^e \leq x) = \frac{1}{\mathbb{E}W} \int_0^x \mathbb{P}(W > y)\,dy\,,$$

which is familiar from renewal theory and other areas. Again, it is easy to see from this that if X is exponential, then X and X^e have the same distribution.

In what follows we will consider approximation in the Wasserstein distance, defined by

$$d_W(W, Z) = \sup_{h \in \mathcal{H}_W} |\mathbb{E}h(W) - \mathbb{E}h(Z)|\,,$$

where $\mathcal{H}_W = \{h : \mathbb{R} \mapsto \mathbb{R} \mid |h(x) - h(y)| \leq |x - y|\}$ is the set of 1-Lipschitz functions. Now, using (20) we have that

$$d_W(W, Z) \leq \sup_{h \in \mathcal{H}_W} |\mathbb{E}f'(W) - \mathbb{E}f(W)|$$
$$= \sup_{h \in \mathcal{H}_W} |\mathbb{E}f'(W) - \mathbb{E}f'(W^e)| \leq \sup_{h \in \mathcal{H}_W} \|f''\|_\infty \mathbb{E}|W^e - W| \,.$$

We now apply (22) to obtain the following result.

Proposition 3.4. *Let $W \geq 0$ be a random variable with $\mathbb{E}W = 1$ and Var$(W) < \infty$. Then*

$$d_W(W, Z) \leq 2\mathbb{E}|W^e - W| \,,$$

where $Z \sim$ Exp(1).

Note that the second moment condition here ensures that W^e has finite mean. In the next section, we apply this result to the approximation of suitably scaled geometric sums by an exponential distribution.

3.2 Approximation of geometric sums

In this section, we will consider the case where $W = p\sum_{i=1}^N Y_i$, where $Y_1, Y_2, \ldots$ are independent, non-negative, square-integrable random variables with $\mathbb{E}Y_i = 1$, and N is a positive integer-valued random variable independent of the Y_i with mean $1/p$. A classical theorem of Rényi states that if N is geometrically distributed, then W converges in distribution to exponential as $p \to 0$. We will give a proof of the following result, which complements Rényi's theorem with explicit error bounds, and applies in the somewhat more general setting where N is not necessarily geometric. The result here was established by Peköz and Röllin [50]. In fact, they established a more powerful result, allowing some dependence between the Y_i.

Theorem 3.5. *Let $W = p\sum_{i=1}^N Y_i$, where $Y_1, Y_2, \ldots$ are independent, non-negative, square-integrable random variables with $\mathbb{E}Y_i = 1$, and N is a positive integer-valued random variable independent of the Y_i with mean $1/p$. Let M be a positive integer-valued random variable on the same probability space as N, satisfying*

$$\mathbb{P}(M = m) = p\mathbb{P}(N \geq m) \,.$$

Let Y_i^e have the equilibrium coupling of Y_i, and be independent of M, N and Y_j for $j \neq i$. Then

$$d_W(W, Z) \leq 2p\left(\mathbb{E}|Y_M - Y_M^e| + \mathbb{E}|N - M|\right)$$
$$\leq 2p\left(1 + \frac{\mu_2}{2} + \mathbb{E}|N - M|\right) \,,$$

where $Z \sim$ Exp(1) and $\mu_2 = \sup_i \mathbb{E}[Y_i^2]$.

Proof. We begin by constructing the random variable W^e.

Exercise: Let $h : \mathbb{R}^+ \mapsto \mathbb{R}$ be a Lipschitz function with $h(0) = 0$, and define $g(m) = h(p\sum_{i=1}^{m} Y_i)$. Show that

$$\mathbb{E}\left[h'\left(p\sum_{i=1}^{M-1} Y_i + pY_M^e \right) \middle| M \right] = \frac{1}{p}\mathbb{E}\left[g(M) - g(M-1) | M \right] ,$$

and that

$$\frac{1}{p}\mathbb{E}\left[g(M) - g(M-1) | (Y_i)_{i\geq 1} \right] = \mathbb{E}\left[g(N) | (Y_i)_{i\geq 1} \right] .$$

Hence, conclude that we may let

$$W^e = p\left[\sum_{i=1}^{M-1} Y_i + Y_M^e \right] . \tag{23}$$

Now, using (23) and Proposition 3.4, we have that

$$\begin{aligned} d_W(W, Z) &\leq 2p\mathbb{E}\left| Y_M^e - Y_M + \text{sgn}(N - M) \sum_{i=(M\wedge N)+1}^{M\vee N} Y_i \right| \\ &\leq 2p\left[\mathbb{E}|Y_M^e - Y_M| + \mathbb{E}|N - M| \right] , \end{aligned}$$

which gives the first bound of the theorem. For the second inequality, we have

$$\begin{aligned} \mathbb{E}\left[|Y_M^e - Y_M| \mid M \right] &\leq \mathbb{E}[Y_M^e|M] + \mathbb{E}[Y_M|M] \\ &= \frac{1}{2}\mathbb{E}[Y_M^2|M] + 1 \\ &\leq \frac{\mu_2}{2} + 1 , \end{aligned}$$

where the equality uses the definition of the equilibrium coupling. □

Note that M is defined using a discrete version of the equilibrium coupling, for which the geometric distribution (with mean $1/p$) is the unique fixed point; i.e., if N is geometrically distributed, then we may choose $M = N$ in Theorem 3.5. We may think of the term $\mathbb{E}|N - M|$ as measuring how close N is to geometric. See [51] for applications of this discrete equilibrium coupling to geometric approximation.

Hence, note that the upper bound of Theorem 3.5 is zero if N is geometrically distributed and the Y_i are exponentially distributed.

Exercise: Check that (with notation as in Theorem 3.5) if $Y_i \sim \text{Exp}(1)$ and N is geometrically distributed, then M is also geometrically distributed and $W \sim \text{Exp}(1)$.

We will return to the study of geometric sums in Section 5, where we will consider approximation by (rather than for) geometric sums.

4 Poisson approximation and related topics

Throughout this section we will let $X_1, \dots, X_n$ be (possibly dependent) Bernoulli random variables. We will write $p_i = \mathbb{E}X_i$ and $\lambda = \sum_{i=1}^n p_i$. We are interested in the approximation of $W = X_1 + \cdots + X_n$ by a Poisson random variable $Z \sim \text{Po}(\lambda)$ with mass function $\mathbb{P}(Z = j) = \frac{e^{-\lambda}\lambda^j}{j!}$ for $j \in \mathbb{Z}^+$.

For the most part, we will assess closeness of non-negative, integer-valued random variables using the total variation distance:

$$d_{TV}(W, Z) = \frac{1}{2}\sum_{j=0}^{\infty} |\mathbb{P}(W = j) - \mathbb{P}(Z = j)| = \sup_{\|h\|_\infty \leq 1} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \,.$$

4.1 Poisson approximation by maximal coupling

Before we discuss the Stein–Chen method for Poisson approximation, we give a brief account of a simple coupling bound for Poisson approximation for sums of independent Bernoulli random variables, due to Le Cam.

Definition 4.1. *A coupling $(\widehat{X}, \widehat{Y})$ of random variables (X, Y) is* maximal *if*

$$\mathbb{P}(\widehat{X} = \widehat{Y}) = \sup\left\{\mathbb{P}(\widetilde{X} = \widetilde{Y}) : (\widetilde{X}, \widetilde{Y}) \textit{ is a coupling of } (X, Y)\right\} \,.$$

Before we give a Poisson approximation bound, we note some properties of maximal couplings (which we state without proof).

Lemma 4.2. *Let $(\widehat{X}, \widehat{Y})$ be a maximal coupling of the non-negative, integer-valued random variables X and Y. Then*

$$\mathbb{P}(\widehat{X} = \widehat{Y}) = \sum_{j=0}^{\infty} \min\{\mathbb{P}(X = j), \mathbb{P}(Y = j)\} \,.$$

Lemma 4.3. *If $(\widehat{X}, \widehat{Y})$ is a maximal coupling of X and Y*

$$d_{TV}(X, Y) = \mathbb{P}(\widehat{X} \neq \widehat{Y}) \,.$$

We are now in a position to state and prove the following well-known Poisson approximation result.

Theorem 4.4 (Le Cam)**.** *Let $X_1, \dots, X_n$ be independent Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. If $Z \sim Po(\lambda)$,*

$$d_{TV}(W, Z) \leq \sum_{i=1}^{n} p_i^2 \,.$$

Proof. Write $Z = \sum_{i=1}^n Z_i$, where $Z_i \sim \text{Po}(p_i)$. We can couple X_i and Z_i maximally for each i (using Lemma 4.2) to get $(\widehat{X}_i, \widehat{Z}_i)$ with

$$\mathbb{P}(\widehat{X}_i = \widehat{Z}_i) = \sum_{j=0}^{\infty} \min\{\mathbb{P}(X_i = j), \mathbb{P}(Z_i = j)\}$$
$$= \min\{1 - p_i, e^{-p_i}\} + \min\{p_i, p_i e^{-p_i}\} = 1 - p_i + p_i e^{-p_i} \geq 1 - p_i^2 \,.$$

Then, since $\left(\sum_{i=1}^n \widehat{X}_i, \sum_{i=1}^n \widehat{Z}_i\right)$ is a coupling of W and Z,

$$d_{TV}(W, Z) \leq \mathbb{P}\left(\sum_{i=1}^n \widehat{X}_i \neq \sum_{i=1}^n \widehat{Z}_i\right) \leq \mathbb{P}\left(\bigcup_{i=1}^n \left\{\widehat{X}_i \neq \widehat{Z}_i\right\}\right) \leq \sum_{i=1}^n \mathbb{P}(\widehat{X}_i \neq \widehat{Z}_i) \leq \sum_{i=1}^n p_i^2 \,.$$

□

This is an elegant result, but there is much room for improvement. To see this, consider the following results, established for sums of independent Bernoulli random variables by [15] using operator techniques:

$$d_{TV}(W, Z) \leq 4.5 \max_i p_i \,,$$
$$d_{TV}(W, Z) \leq 8\lambda^{-1} \sum_{i=1}^n p_i^2 \,, \tag{24}$$

this last inequality proved under the assumption $\max_i p_i \leq 1/4$. We are most interested in the second of these inequalities, which can represent a substantial improvement over the bound of Theorem 4.4 when λ is large, achieved by the inclusion of the "magic factor" of λ^{-1}

4.2 The Stein–Chen method for Poisson approximation

As in the case of Gaussian approximation, the Stein–Chen method has the advantage of being able to handle dependence between the random variables X_i. We will also see that results here include "magic factors" akin to that in Le Cam's result (24) and that were missing in the coupling argument of Theorem 4.4.

Before proceeding further, we need a little notation. Let $\mathbb{Z}^+$ denote the non-negative integers. For any function $g : \mathbb{Z}^+ \mapsto \mathbb{R}$ we let Δ be the forward difference operator, so that $\Delta g(j) = g(j+1) - g(j)$.

As in the Gaussian case, the starting point of Stein's method is a characterization of the Poisson distribution (with a proof similar to the Gaussian and exponential cases we have already seen).

Lemma 4.5. *A non-negative, integer-valued random variable X has a Poisson distribution with mean λ, written $X \sim Po(\lambda)$, if and only if*

$$\lambda \mathbb{E}\left[g(X+1)\right] = \mathbb{E}\left[X g(X)\right]$$

for all bounded $g : \mathbb{Z}^+ \mapsto \mathbb{R}$.

Exercise: Verify that if $Z \sim$ Po(λ), then $\lambda\mathbb{E}\left[g(Z+1)\right] = \mathbb{E}\left[Zg(Z)\right]$ for all bounded $g : \mathbb{Z}^+ \mapsto \mathbb{R}$.

We may now write down the corresponding Stein equation: For a given function $h : \mathbb{Z}^+ \mapsto \mathbb{R}$, we let $f = f_h$ solve

$$h(j) - \mathbb{E}h(Z) = \lambda f(j+1) - jf(j)\,. \tag{25}$$

This is used in an analogous way to previous Stein equations we have seen. Replacing j with W and taking expectations we have that

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\,.$$

If $W \approx$ Po(λ), then the LHS should be small for a suitably large class of functions h. So, the RHS should also be small. We can 'measure how close W is to Poisson' by looking at how large the RHS can become for h in some suitable class. To make this idea precise (for the case of total variation distance) let

$$\mathcal{H} = \mathcal{H}_{TV} = \{h : \mathbb{Z}^+ \mapsto \mathbb{R} \mid \|h\|_\infty \leq 1\}$$

Then

$$d_{TV}(W, Z) = \sup_{h \in \mathcal{H}_{TV}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| = \sup_{h \in \mathcal{H}_{TV}} |\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]|\,. \tag{26}$$

Other metrics may be treated similarly. For example, defining the Wasserstein distance

$$d_W(W, Z) = \sum_{j=0}^{\infty} |\mathbb{P}(W \leq j) - \mathbb{P}(Z \leq j)| = \sup_{h \in \mathcal{H}_W} |\mathbb{E}h(W) - \mathbb{E}h(Z)|\,,$$

where $\mathcal{H}_W = \{h : \mathbb{Z}^+ \mapsto \mathbb{R} \mid \|\Delta h\| \leq 1\}$ is the set of Lipschitz functions, we have

$$d_W(W, Z) = \sup_{h \in \mathcal{H}_W} |\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]|\,.$$

We will need bounds on the solution f of this Stein equation, which are given below without proof. The interested reader is referred to Lemmas 1.1.1 and 1.1.5 of [10] for proofs and further discussion.

Lemma 4.6. *1. If $h \in \mathcal{H}_{TV}$,*

$$\|f\|_\infty \leq \min\left\{1, \sqrt{\frac{2}{e\lambda}}\right\} \qquad \textit{and} \qquad \|\Delta f\|_\infty \leq \frac{1-e^{-\lambda}}{\lambda}\,.$$

2. If $h \in \mathcal{H}_W$,

$$\|f\|_\infty \leq 3 \qquad \textit{and} \qquad \|\Delta f\|_\infty \leq 3\min\left\{1, \sqrt{\frac{1}{\lambda}}\right\}\,.$$

Note that these bounds do not depend on h, nor on the random variable W of interest. Such bounds are known as *Stein factors*, or magic factors.

To see how we may bound (26) in practice, we first consider the case where W is a sum of independent Bernoulli random variables.

4.2.1 Independent summands

In the following theorem, note the improvement over previous results. We retain the magic factor of λ^{-1} appearing in (24), but without the restriction on the p_i. The argument here, and its extension to the setting of local dependence, is due to Chen [17].

Theorem 4.7. *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. If $Z \sim$ Po(λ),*

$$d_{TV}(W, Z) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \,.$$

Proof. For each i we write $W_i = W - X_i$. We begin by noting that

$$\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] = \sum_{i=1}^{n} \mathbb{E}\left[p_i f(W+1) - X_i f(W)\right] \,.$$

For each i, $\mathbb{E}[X_i f(W)] = p_i \mathbb{E}[f(W_i + 1)]$ and so

$$\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] = \sum_{i=1}^{n} p_i \mathbb{E}\left[f(W+1) - f(W_i+1)\right] \,.$$

Since

$$\begin{aligned}
|\mathbb{E}\left[f(W+1) - f(W_i+1)\right]| &\leq \sup_{h \in \mathcal{H}_{TV}} \|\Delta f\|_\infty \mathbb{E}|W - W_i| \\
&\leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \mathbb{E}X_i \\
&= \left(\frac{1 - e^{-\lambda}}{\lambda}\right) p_i \,,
\end{aligned}$$

(using Lemma 4.6) we have that

$$|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]| \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \,.$$

□

It is worth noting that the bound given by Theorem 4.7 is of the right order. We have the corresponding lower bound

$$d_{TV}(W,Z) \geq \frac{1}{32}\min\left\{1,\frac{1}{\lambda}\right\}\sum_{i=1}^{n} p_i^2\,;$$

See Barbour and Hall [9].

4.2.2 Dependent summands: the local approach

One of the advantages to Stein's approach is that the argument of Theorem 4.7 may be easily adapted to cover the case where the X_i are no longer independent. In the Poisson case, there are two widely used approaches to doing this: the 'local approach' and the 'coupling approach'. We will begin with a brief look at the local approach (due to Chen [17]), but most of our attention will be devoted to the coupling approach.

Theorem 4.8. *Let $X_1,\dots,X_n$ be Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1+\cdots+X_n$ and $\lambda = \mathbb{E}W = p_1+\cdots+p_n$. For each i, divide $\{1,\dots,i-1,i+1,\dots,n\}$ into two subsets Γ_i and Θ_i so that, informally,*

$$\Gamma_i = \{j : X_j \textit{ is strongly dependent on } X_i\}\,.$$

Let $Z_i = \sum_{j\in\Gamma_i} X_j$ and $W_i = \sum_{j\in\Theta_i} X_j$. If $Z \sim Po(\lambda)$,

$$d_{TV}(W,Z) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_{i=1}^{n}\left(p_i\mathbb{E}[X_i+Z_i] + \mathbb{E}[X_iZ_i]\right) + \sqrt{\frac{2}{e\lambda}}\sum_{i=1}^{n}\mathbb{E}\left|p_i - \mathbb{E}[X_i|W_i]\right|\,.$$

Proof. We write

$$\begin{aligned}
\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] &= \sum_{i=1}^{n}\mathbb{E}\left[p_if(W+1) - X_if(W)\right]\\
&= \sum_{i=1}^{n}\mathbb{E}\left[p_if(W+1) - p_if(W_i+1)\right]\\
&+ \sum_{i=1}^{n}\mathbb{E}\left[p_if(W_i+1) - X_if(W_i+1)\right]\\
&+ \sum_{i=1}^{n}\mathbb{E}\left[X_if(W_i+1) - X_if(W)\right]\,.
\end{aligned}$$

For each i we have the bounds

$$\begin{aligned}
|f(W+1) - f(W_i+1)| &\leq \|\Delta f\|_\infty (X_i+Z_i)\,,\\
|X_if(W_i+1) - X_if(W)| &\leq \|\Delta f\|_\infty X_iZ_i\,,\\
\left|\mathbb{E}\left[p_if(W_i+1) - X_if(W_i+1)\right]\right| &\leq \|f\|_\infty\mathbb{E}\left|p_i - \mathbb{E}[X_i|W_i]\right|\,.
\end{aligned}$$

Combining all these we have

$$\begin{aligned}&|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]| \\ &\qquad\qquad \leq \|\Delta f\|_\infty \sum_{i=1}^{n} \left(p_i \mathbb{E}[X_i + Z_i] + \mathbb{E}[X_i Z_i]\right) + \|f\|_\infty \sum_{i=1}^{n} \mathbb{E}\left|p_i - \mathbb{E}[X_i|W_i]\right| ,\end{aligned}$$

from which the result follows using Lemma 4.6. □

In the case where X_i are independent, we choose $\Gamma_i = \emptyset$ for each i and recover the bound of Theorem 4.7.

Example: the birthday problem

This example is taken from [5]. Suppose m balls (people) are thrown independently and equiprobably into d boxes (days of the year). Let W be the number of pairs that go into the same box. How close is W to Poisson?

Let Γ be the set of all 2–subsets of $\{1, \ldots, m\}$. That is, $\Gamma = \{i \subset \{1, \ldots, m\} : |i| = 2\}$. If $i = \{i_1, i_2\}$, we write X_i for the indicator that balls i_1 and i_2 land in the same box. So, $W = \sum_{i \in \Gamma} X_i$.

Note that $\mathbb{E}X_i = d^{-1}$ for all $i \in \Gamma$, and so $\lambda = \mathbb{E}W = \binom{m}{2} d^{-1}$. Also, $\mathbb{E}[X_i X_j] = d^{-2}$ for all $i \neq j$.

We choose $\Gamma_i = \{j \in \Gamma \setminus \{i\} : i \cap j \neq \emptyset\}$. Then X_i is independent of X_j for all $j \notin \Gamma_i \cup \{i\}$ and so the final term of the bound in Theorem 4.8 vanishes.

We obtain

$$\begin{aligned} d_{TV}(W, Z) &\leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i \in \Gamma} \left(p_i \mathbb{E}[X_i + Z_i] + \mathbb{E}[X_i Z_i]\right) \\ &= \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \binom{m}{2} \left(\frac{2(m-1)+1}{d^2} + \frac{2(m-1)}{d^2}\right) \\ &= \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \binom{m}{2} \frac{4m-3}{d^2} \\ &\leq \frac{8\lambda(1 - e^{-\lambda})}{m-1} . \end{aligned}$$

4.3 The size-biased coupling and monotonicity

Definition 4.9. *If W is a non-negative, integer-valued random variable with mean $\lambda > 0$, we let $W^\star$ have the W-size-biased distribution, given by*

$$\mathbb{P}(W^\star = j) = \frac{j\mathbb{P}(W = j)}{\lambda} .$$

Equivalently, we may define $W^\star$ by letting

$$\lambda\mathbb{E}[g(W^\star)] = \mathbb{E}[Wg(W)]\,, \tag{27}$$

for all $g : \mathbb{Z}^+ \mapsto \mathbb{R}$ for which the expectations above exist.

With this definition, it is clear that we may write our Stein equation as

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \lambda\mathbb{E}\left[f(W+1) - f(W^\star)\right]\,. \tag{28}$$

Note that we may also rewrite our characterization of the Poisson distribution by saying that X has a Poisson distribution if and only if $X+1$ is equal in distribution to $X^\star$.

Since the absolute value of the RHS of (28) may be bounded by $\lambda\|\Delta f\|_\infty\mathbb{E}|W+1-W^\star|$, we immediately obtain the following from Lemma 4.6.

Theorem 4.10. *Let W be a non-negative, integer-valued random variable with $\mathbb{E}W = \lambda > 0$. Let $Z \sim$ Po(λ). Then*

$$d_{TV}(W, Z) \leq \min\{1, \lambda\}\mathbb{E}|W+1-W^\star|\,.$$

The size-biased coupling approach to the Stein–Chen method for Poisson approximation is often easiest to apply in the presence of monotonicity, either some form of 'negative dependence' or 'positive dependence'. In the remainder of this section we will focus on results in the presence of such monotonicity (basing our discussion on [31]), but see [10] for a more extensive discussion, including results that hold without any monotonicity.

Definition 4.11. *For any two random variables X and Y, Y is stochastically larger than X (written $X \leq_{st} Y$) if $\mathbb{P}(X > t) \leq \mathbb{P}(Y > t)$ for all t. This is equivalent to having $\mathbb{E}g(X) \leq \mathbb{E}g(Y)$ for all increasing functions g, and to the existence of a coupling $(\widetilde{X}, \widetilde{Y})$ of (X, Y) such that $\widetilde{X} \leq \widetilde{Y}$ almost surely.*

4.3.1 Negative dependence

Our 'negative dependence' assumption is that $W^\star \leq_{st} W+1$. Consider the following cases in which this condition holds:

1. W is *ultra log-concave* (of degree ∞). That is,

   $$\frac{(j+1)\mathbb{P}(W=j+1)}{\mathbb{P}(W=j)} \text{ is increasing in } j\,.$$

   Or, equivalently, $W^\star \leq_{lr} W+1$. Here '$\leq_{lr}$' is the likelihood ratio ordering, which is known to be stronger than the stochastic ordering $\leq_{st}$. The ULC(∞) class was introduced by Liggett [46] to capture negative dependence.

2. $W = X_1 + \cdots + X_n$ is a sum of (dependent) Bernoulli random variables with

$$\text{Cov}(g_1(X_i), g_2(W - X_i)) \leq 0\,,$$

for each $i = 1, \ldots, n$ and all increasing functions g_1 and g_2. That is, $X_1, \ldots, X_n$ are *totally negatively dependent* (TND); see Papadatos and Papathanasiou [48], who show that TND Bernoulli random variables also satisfy the well-known negative dependence property of being negatively related.

Let $W_i = W - X_i$ and g be an increasing function. In the case where $X_1, \ldots, X_n$ are TND we have that, almost surely,

$$\begin{aligned}\mathbb{E}[Wg(W)] = \sum_{i=1}^{n} \mathbb{E}[X_i g(W)] \leq \sum_{i=1}^{n} \mathbb{E}[X_i g(W_i + 1)] &\leq \sum_{i=1}^{n} \mathbb{E}[X_i]\mathbb{E}[g(W_i + 1)] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[X_i]\mathbb{E}[g(W + 1)] = \mathbb{E}[W]\mathbb{E}[g(W + 1)]\,,\end{aligned}$$

so that $W^\star \leq_{st} W + 1$.

Theorem 4.12. *Let W be a non-negative, integer-valued random variable with $\mathbb{E}W = \lambda > 0$. Suppose that $W^* \leq_{st} W + 1$. Then*

$$d_{TV}(W, Z) \leq \frac{1 - e^{-\lambda}}{\lambda}\left[\lambda - \mathit{Var}(W)\right]\,,$$

where $Z \sim Po(\lambda)$.

Proof. We have

$$\begin{aligned}\mathbb{E}f(W + 1) - \mathbb{E}f(W^\star) &= \sum_{j=0}^{\infty} f(j)\big[\mathbb{P}(W + 1 = j) - \mathbb{P}(W^\star = j)\big] \\ &= \sum_{j=0}^{\infty} \Delta f(j)\big[\mathbb{P}(W + 1 > j) - \mathbb{P}(W^\star > j)\big]\,.\end{aligned}$$

Hence, taking absolute values and using our Stein equation,

$$d_{TV}(W, Z) \leq \lambda \sup_{h \in \mathcal{H}_{TV}} \|\Delta f\|_\infty \sum_{j=0}^{\infty} |\mathbb{P}(W + 1 > j) - \mathbb{P}(W^\star > j)|\,.$$

Our negative dependence assumption allows us to remove the absolute values from this expression, so that (applying Lemma 4.6) we obtain

$$d_{TV}(W, Z) \leq (1 - e^{-\lambda})\mathbb{E}[W + 1 - W^\star] = \frac{1 - e^{-\lambda}}{\lambda}\left[\lambda - \text{Var}(W)\right]\,.$$

□

In order to be able to apply this result, we will need to understand how to construct the size-biased random variable $W^\star$, particularly in the case where $W = X_1 + \cdots + X_n$ is a sum of dependent Bernoulli random variables. The following lemma allows us to do this, so that we can verify the monotonicity condition needed for Theorem 4.12. This result, together with an extensive discussion of the role of size biasing in various areas of applied probability, can be found in [6].

Lemma 4.13. *Let $W = X_1 + \cdots + X_n$ be a sum of dependent Bernoulli random variables. Let $p_i = \mathbb{E}X_i$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n > 0$. Let I be a random variable, independent of all else, with $\mathbb{P}(I = i) = p_i/\lambda$. Then*

$$W^\star = 1 + \sum_{j \neq I} X_j^I$$

has the W-size-biased distribution, where $X_j^I \overset{d}{=} (X_j | X_I = 1)$.

Examples: the hypergeometric distribution and coupon collecting

Suppose we have N balls in an urn, of which n are red and the remaining $N - n$ are white. We select m of these N balls, uniformly at random without replacement (so that any of the $\binom{N}{m}$ subsets of m balls are equally likely to be chosen).

Let X_i, $i = 1, \ldots, n$, be an indicator that red ball i is chosen as part of our sample, so that $W = X_1 + \cdots + X_n$ counts the number of red balls chosen in our sample. These X_i are exchangeable, so without loss of generality we may assume (in the notation of Lemma 4.13) that $I = 1$.

Now carry out the following procedure:

- if the red ball labelled 1 is in our sample, do nothing.
- if the red ball labelled 1 is not in our sample, choose uniformly at random one of the m balls in the sample, and replace it with the red ball labelled 1.

Now let $X_i^{(1)}$, $i = 2, \ldots, n$, be an indicator that red ball i is in this (potentially modified) sample of balls. Clearly we have that $X_i^{(1)} \leq X_i$ almost surely for each $i = 2, \ldots, n$. Lemma 4.13 tells us that we may take

$$W^\star = 1 + \sum_{j=2}^{n} X_j^{(1)} \leq_{st} 1 + \sum_{j=2}^{n} X_j \leq_{st} W + 1\,,$$

and so our negative dependence assumption holds.

Then, since $\mathbb{E}W = \lambda = \frac{mn}{N}$ and

$$\text{Var}(W) = \frac{mn(N-n)(n-m)}{N^2(N-1)}\,,$$

Theorem 4.12 gives us that

$$d_{TV}(W,Z) \leq \frac{n}{N-1} + \frac{m}{N-1} - \frac{mn}{N(N-1)} - \frac{1}{N-1}\,.$$

where $Z \sim \text{Po}(\lambda)$.

Exercise: Now suppose that we have n urns and k balls. Each of these k balls is thrown into an urn which is chosen uniformly at random. Let X_i, $i = 1,\dots,n$, be an indicator that urn i remains empty after all k balls are distributed. Then $W = X_1 + \cdots + X_n$ counts the number of empty urns. Note that, as in the hypergeometric example above, these X_i are exchangeable random variables. Use a procedure similar to that of the example above to show that $W^\star \leq_{st} W + 1$. Show also that $\mathbb{E}W = \lambda = n\left(1-\frac{1}{n}\right)^k$, $\mathbb{P}(X_i = 1, X_j = 1) = \left(1-\frac{2}{n}\right)^k$ for $i \neq j$, and that

$$\text{Var}(W) = \lambda\left[1-\left(1-\frac{1}{n}\right)^k\right] + n(n-1)\left[\left(1-\frac{2}{n}\right)^k - \left(1-\frac{1}{n}\right)^{2k}\right]\,.$$

Hence, use Theorem 4.12 to give an explicit bound in the Poisson approximation for W.

4.3.2 Positive dependence

It would be natural to conjecture that our positive dependence condition would be "$W+1 \leq_{st} W^\star$", however this is not the case. (**Exercise:** use the supports of these two random variables to argue that this condition would not make sense.)

We focus on the case where $W = X_1 + \cdots + X_n$ is a sum of dependent Bernoulli random variables, and let the random index I, independent of all else, be defined as in Lemma 4.13. In this case our 'positive dependence' condition is that $W + 1 - X_I \leq_{st} W^\star$. Under this condition we obtain the following.

Theorem 4.14. *Let $W = X_1 + \cdots + X_n$, with X_j, λ and the random index I as in Lemma 4.13. Assume $W + 1 - X_I \leq_{st} W^\star$. Then*

$$d_{TV}(W,Z) \leq \frac{1-e^{-\lambda}}{\lambda}\left[\mathit{Var}(W) - \lambda + 2\sum_{j=1}^{n} p_j^2\right]\,,$$

where $Z \sim Po(\lambda)$.

Exercise: Prove this result by making suitable modifications to the proof of Theorem 4.12.

Example: triangles in an Erdős–Rényi random graph

Let $G = G(n,p)$ be an Erdős–Rényi random graph with n vertices, in which each pair of vertices has an edge between them with probability p, independently of all other pairs of vertices.

Let Γ be the set of all $\binom{n}{3}$ triples (x, y, z) of vertices in G, and for $\alpha \in \Gamma$, let X_α be an indicator that the three corresponding vertices form a triangle (i.e., that all edges between these three vertices are present in the graph). Note that these random variables X_α are exchangeable, so for the remainder of this example we fix $\alpha = (1, 2, 3)$ and (in the notation of Lemma 4.13) take $I = \alpha$.

Let $W = \sum_{\beta\in\Gamma} X_\beta$ count the number of triangles in G. For $\beta \in \Gamma$, we let $X_\beta^{(\alpha)}$ be an indicator that a triangle is present at β in the graph constructed by adding to G any edges missing between vertices in α. Clearly we have that $X_\beta^{(\alpha)} \geq X_\beta$ almost surely, and so using Lemma 4.13 we have

$$W + 1 - X_\alpha = 1 + \sum_{\beta\in\Gamma\backslash\{\alpha\}} X_\beta \leq_{st} 1 + \sum_{\beta\in\Gamma\backslash\{\alpha\}} X_\beta^{(\alpha)} = W^\star ,$$

so that our positive dependence assumption holds. To apply the bound of Theorem 4.14, we note that $\mathbb{E}X_\beta = p^3$ for all $\beta \in \Gamma$, and so $\lambda = \mathbb{E}W = \binom{n}{3}p^3$.

Exercise: Show that

$$\text{Var}(W) = \lambda \left[1 - p^3 + 3(n-3)p^2(1-p)\right] .$$

Combining all these ingredients, Theorem 4.14 gives

$$d_{TV}(W, Z) \leq p^3 + 3(n-3)p^2(1-p) ,$$

where $Z \sim \text{Po}(\lambda)$.

4.4 Further applications of monotone size-biased couplings

In this section we look at some other situations in which the our negative dependence assumption $W^* \leq_{st} W + 1$ may be used to compare the random variable W with a Poisson random variable of the same mean. These comparisons are each in a different sense, but make use of the same monotonicity assumption on the size-biased coupling.

4.4.1 Bounds on the Poincaré constant

For a non-negative, integer-valued random variable W, we define the Poincaré (or inverse spectral gap) constant by

$$R_W = \sup_{g\in\mathcal{G}(W)} \left\{ \frac{\mathbb{E}[g(W)^2]}{\mathbb{E}[\Delta g(W)^2]} \right\} ,$$

where the supremum is taken over the set

$$\mathcal{G}(W) = \{g : \mathbb{Z}^+ \mapsto \mathbb{R} \text{ with } \mathbb{E}[g(W)^2] < \infty \text{ and } \mathbb{E}[g(W)] = 0\} .$$

Exercise: Show that $R_W \geq \text{Var}(W)$.

In general, the problem of finding sharp upper bounds on R_W is a hard one. Note for what follows that $Z \sim \text{Po}(\lambda)$ is characterised by the fact that $R_Z = \lambda$.

Daly and Johnson [29] prove the following result.

Theorem 4.15. *Let W be a non-negative, integer-valued random variable with mean λ, and let $Y \geq 1$ be a random variable defined on the same space as W such that $W^\star \leq_{st} W + Y$. Then for any $g \in \mathcal{G}(W)$,*

$$\mathbb{E}[g(W)^2] \leq \lambda \sum_{j=0}^{\infty} \Delta g(j)^2 \mathbb{P}(j - Y < W \leq j)\,.$$

We have the immediate corollary

Corollary 4.16. *Let W be a non-negative, integer-valued random variable. Suppose that $W^\star \leq_{st} W + 1$. Then $R_W \leq \mathbb{E}W$.*

The proof of Theorem 4.15 uses the kernel function given by equation (2.17) of [44]:

$$\chi(i,j) = I(\lfloor x_0 \rfloor \leq j < i) - I(i \leq j < \lfloor x_0 \rfloor) - (x_0 - \lfloor x_0 \rfloor) I(j = \lfloor x_0 \rfloor)\,, \tag{29}$$

for some $x_0 \in \mathbb{R}$, and uses the following lemma.

Lemma 4.17. *Let W be a non-negative, integer-valued random variable. Then for any $g \in \mathcal{G}(W)$ and any $x_0 \in \mathbb{R}$,*

$$\mathbb{E}[g(W)^2] \leq \sum_{j=0}^{\infty} \Delta g(j)^2 \sum_{i=0}^{\infty} (i - x_0) P(W = i) \chi(i,j)\,. \tag{30}$$

Proof. For any given integer i, we consider the cases $\{i < x_0\}$, $\{i > x_0\}$ and $\{i = x_0\}$ to deduce that, for any function h,

$$\sum_{j=0}^{\infty} \chi(i,j) h(j) = \begin{cases} -\sum_{j=i}^{\lfloor x_0 \rfloor - 1} h(j) + (\lfloor x_0 \rfloor - x_0) h(\lfloor x_0 \rfloor) & \text{for } i < \lfloor x_0 \rfloor, \\ (\lfloor x_0 \rfloor - x_0) h(\lfloor x_0 \rfloor) & \text{for } i = \lfloor x_0 \rfloor, \\ \sum_{\lfloor x_0 \rfloor}^{i-1} h(j) + (\lfloor x_0 \rfloor - x_0) h(\lfloor x_0 \rfloor) & \text{for } i > x_0. \end{cases}$$

Now let $h \equiv \Delta g$ to get that $\sum_{j=0}^{\infty} \chi(i,j) \Delta g(j) = g(i) - g^*$, where

$$g^* = g(\lfloor x_0 \rfloor) + \Delta g(\lfloor x_0 \rfloor)(x_0 - \lfloor x_0 \rfloor)\,.$$

In particular, taking $h(j) \equiv 1$ we deduce that $\sum_{j=0}^{\infty} \chi(i,j) = (i - x_0)$. Observe that by the Cauchy–Schwarz inequality this means that

$$\begin{aligned} (g(i) - g^*)^2 = \left(\sum_{j=0}^{\infty} \chi(i,j) \Delta g(j) \right)^2 &\leq \left(\sum_{j=0}^{\infty} \chi(i,j) \right) \left(\sum_{j=0}^{\infty} \chi(i,j) \Delta g(j)^2 \right) \\ &= (i - x_0) \left(\sum_{j=0}^{\infty} \chi(i,j) \Delta g(j)^2 \right). \end{aligned} \tag{31}$$

Note that although $\chi(i,j)$ is a signed measure on j, the use of the Cauchy–Schwarz inequality is justified since $\chi(i,j)$ has constant sign for any given i. If $i \geq \lfloor x_0 \rfloor$ then $\chi(i,j) \geq 0$ for all j, and otherwise we have that $\chi(i,j) \leq 0$ for all j.

The lemma follows on combining (31) with the fact that, for all $g \in \mathcal{G}(W)$,

$$\mathbb{E}[g(W)^2] \leq \sum_{i=0}^{\infty} P(W=i)\,(g(i)-g^*)^2 \ ,$$

and reversing the order of summation in the resulting expression. □

Theorem 4.15 follows immediately from Lemma 4.17. To see this, choose $x_0 = \mathbb{E}W = \lambda$. Then, using the definition of size-biasing, for a fixed $j \in \mathbb{Z}$ the inner sum in (30) can be expressed as

$$\begin{aligned} \mathbb{E}W\chi(W,j) - \lambda\mathbb{E}\chi(W,j) &= \lambda\left(\mathbb{E}\chi(W^\star,j) - \mathbb{E}\chi(W,j)\right) \\ &\leq \lambda\mathbb{E}\left[\chi(W+Y,j) - \chi(W,j)\right], \end{aligned} \tag{32}$$

using the negative dependence assumption of Theorem 4.15, and the fact that $\chi(i,j)$ is increasing in i for fixed j. Using (29), and assuming that $j \geq \lfloor x_0 \rfloor$ first, we have that, for any $w, z \in \mathbb{Z}$,

$$\chi(w+z,j) - \chi(w,j) = I(j < w+z) - I(j < w) = I(w \leq j < w+z). \tag{33}$$

Similarly, Equation (33) also holds in the case $j < \lfloor x_0 \rfloor$. Substituting this in (32) we obtain $\lambda\mathbb{P}(W \leq j < W+Y) = \lambda\mathbb{P}(j-Y < W \leq j)$ as required to complete the proof of Theorem 4.15.

4.4.2 Entropy bounds

We define the entropy $H(W)$ of a non-negative, integer-valued random variable W in the usual way, although for convenience we take natural logarithms:

$$H(W) = -\sum_{i=0}^{\infty} \mathbb{P}(W=i)\log(\mathbb{P}(W=i))\,.$$

In this section we will state (but not prove) a result which lets us compare the entropy of a random variable W satisfying our negative dependence property with that of a Poisson random variable of the same mean (λ, say). In fact, we will go further: defining a parametric family of random variables which evolves from W to $Z \sim \text{Po}(\lambda)$, we will study the evolution of the entropy along this path.

For any $\alpha \in [0,1]$, we define the thinning operator T_α by letting $T_\alpha W = \sum_{i=1}^{W} \eta_i$, where $\eta_1, \eta_2, \ldots$ are IID Bernoulli random variables (independent of W) with mean α. We will write Z_λ for a Poisson random variable with mean λ.

We will define the required family of random variables using the operator U_α, given by

$$U_\alpha W = T_\alpha W + Z_{(1-\alpha)\lambda}\,, \tag{34}$$

where $Z_{(1-\alpha)\lambda}$ is independent of all else. In what follows we will write W_α for a random variable equal in distribution to $U_\alpha W$ for $\alpha \in [0,1]$. Note that W_1 is equal in distribution to W, and that $W_0 \sim \text{Po}(\lambda)$.

It is easy to see that, for any $\alpha \in [0,1]$, we have $\mathbb{E}[W_\alpha] = \mathbb{E}[W] = \lambda$. We also note that we have a semigroup-like property: for any $\alpha, \beta \in [0,1]$, $U_\beta(U_\alpha W)$ is equal in distribution to $U_{\alpha\beta}W$. Finally, it is useful to note that U_α acts trivially on Poisson distributions. That is, $U_\alpha Z_\lambda$ is equal in distribution to Z_λ for any $\lambda \geq 0$ and $\alpha \in [0,1]$. Further properties of the operators U_α, and their link with the M/M/∞ queue, are discussed by Johnson [42].

Theorem 5.1 of [42] shows that for W within the ultra log-concave class ULC(∞) of random variables, the entropy of W_α is a decreasing and concave function of α. In particular, this implies that the Poisson distribution has maximum entropy in this class.

Based on the arguments given by [42], we can show that the same result applies in the wider class of random variables W satisfying our negative dependence assumption. See [25] for the proof.

Theorem 4.18. *Let W be a non-negative, integer-valued random variable satisfying $W^\star \leq_{st} W+1$. Then*

$$\frac{\partial}{\partial\alpha} H(W_\alpha) \leq 0 \qquad \textit{and} \qquad \frac{\partial^2}{\partial\alpha^2} H(W_\alpha) \leq 0\,, \tag{35}$$

with equality if and only if W has a Poisson distribution.

Note that although no closed-form expression exists for $H(Z_\lambda)$, there are several bounds on this quantity available in the literature. For example, there is the well-known bound

$$H(Z_\lambda) \leq \frac{1}{2}\log\left(2\pi e\left(\lambda + \frac{1}{12}\right)\right).$$

The proof of Theorem 4.18 uses the following closure property:

Lemma 4.19. *Let W be a non-negative, integer-valued random variable with positive mean. If we have $W^\star \leq_{st} W+1$ then $W_\alpha^\star \leq_{st} W_\alpha + 1$ for all $\alpha \in [0,1]$.*

The proof of this lemma uses Lemma 4.13 on the size-biasing of sums, as well as various closure properties of stochastic ordering.

4.5 Concentration via size-biased couplings

As well as monotonicity of size-biased couplings, a simplifying assumption which is both useful and relevant in applications is that of boundedness of size-biased couplings. In the case where $|W - W^\star|$ is bounded almost surely, Gaussian approximation results for W are given in Section 5.3 of [19].

Another situation in which boundedness of size-biased couplings may be applied is in establishing concentration inequalities for W. For example, Ghosh and Goldstein [39] establish the following result.

Theorem 4.20. *Let W be a non-negative, integer-valued random variable with mean and variance λ and σ^2, respectively, both of which are finite and positive. Suppose that there exists a coupling of $W^\star$ to W such that $0 \leq W^\star - W \leq C$ almost surely for some $C > 0$. Then*

$$\mathbb{P}\left(\frac{W-\lambda}{\sigma} \leq -t\right) \leq \exp\left\{-\frac{t^2}{2A}\right\} ,$$

for all $t > 0$, where $A = C\lambda/\sigma^2$.

Proof. The elementary inequality

$$\frac{e^y - e^x}{y - x} \leq \frac{e^y + e^x}{2} ,$$

for all $x \neq y$, gives us that

$$|e^{\theta W^\star} - e^{\theta W}| \leq \frac{1}{2}|\theta(W^\star - W)||e^{\theta W^\star} + e^{\theta W}| \leq \frac{C|\theta|}{2}\left(e^{\theta W^\star} + e^{\theta W}\right) , \tag{36}$$

almost surely, for any $\theta \in \mathbb{R}$.

Letting $m(\theta) = \mathbb{E}[e^{\theta W}]$ be the MGF of W (which exists for all $\theta < 0$), note that

$$m'(\theta) = \mathbb{E}[We^{\theta W}] = \lambda\mathbb{E}[e^{\theta W^\star}] . \tag{37}$$

Letting $\theta < 0$ and using the fact that $e^{\theta W^\star} \leq e^{\theta W}$ almost surely, (36) gives $e^{\theta W} - e^{\theta W^\star} \leq C|\theta|e^{\theta W}$. Hence,

$$\mathbb{E}[e^{\theta W^\star}] \geq (1 - C|\theta|)\mathbb{E}[e^{\theta W}] = (1 + C\theta)\mathbb{E}[e^{\theta W}] ,$$

so that

$$m'(\theta) \geq \lambda(1 + C\theta)m(\theta) , \tag{38}$$

for all $\theta < 0$.

Now let $M(\theta)$ be the MGF of $(W - \lambda)/\sigma$ and note that $M(\theta) = e^{-\lambda\theta/\sigma}m(\theta/\sigma)$. Hence, writing (38) in terms of this standardised random variable gives

$$M'(\theta) \geq \frac{C\lambda\theta}{\sigma^2}M(\theta) ,$$

from which (since $M(0) = 1$) it follows that

$$-\log M(\theta) = \int_\theta^0 \frac{M'(s)}{M(s)}\,ds \geq \int_\theta^0 \frac{C\lambda s}{\sigma^2}\,ds = -\frac{C\lambda\theta^2}{2\sigma^2} .$$

Hence,

$$M(\theta) \leq \exp\left\{\frac{C\lambda\theta^2}{2\sigma^2}\right\} ,$$

and

$$\mathbb{P}\left(\frac{W-\lambda}{\sigma} \leq -t\right) = \mathbb{P}\left(\theta\left(\frac{W-\lambda}{\sigma}\right) \geq -\theta t\right) = \mathbb{P}\left(e^{\theta(\frac{W-\lambda}{\sigma})} \geq e^{-\theta t}\right)$$

$$\leq e^{\theta t}M(\theta) \leq \exp\left\{\theta t + \frac{C\lambda\theta^2}{2\sigma^2}\right\} ,$$

where the first inequality follows from Markov's inequality. We let $\theta = -\frac{t\sigma^2}{C\lambda}$ to complete the proof. □

Note that since $W \leq_{st} W^\star$, there always exists a coupling such that $W \leq W^\star$ almost surely. There is, however, no guarantee in general that this coupling is bounded.

Exercise: Use Lemma 4.13 to check that Theorem 4.20 may be applied to a sum of independent Bernoulli random variables with the choice $C = 1$.

Example: runs

This application is taken from [39], where many further applications are also given. Consider a sequence $\xi_1, \dots, \xi_n$ of IID Bernoulli random variables with $\mathbb{E}\xi_1 = p \in (0,1)$. Let W count the number of runs of m consecutive 1s in these Bernoulli trials, where we assume that $n \geq 2m$. That is, $W = X_1 + \cdots + X_n$, where $X_i = \xi_i \cdots \xi_{i+m-1}$ and we use the convention that $\xi_{n-k} = \xi_k$.

Clearly $\lambda = np^m$, and Goldstein and Ghosh [39] show by direct calculation that

$$\sigma^2 = np^m\left(1 + 2\frac{p-p^m}{1-p} - (2m-1)p^m\right) .$$

It is clear that the random variables $X_1, \dots, X_n$ are $(m-1)$-dependent. That is, if $|i-j| \geq m$, then X_i and X_j are independent (where, as always here, we treat all indices modulo n). Hence, for each i there are $2m-1$ random variables X_j which depend on X_i. An argument analogous to that used in the exercise above for independent Bernoulli random variables then gives us that we may take $C = 2m-1$ in Theorem 4.20, so that

$$A = \frac{2m-1}{1 + 2\frac{p-p^m}{1-p} - (2m-1)p^m} .$$

4.6 Extensions and generalisations of Poisson approximation results

In this section we give some brief remarks on generalisations and extensions of the Poisson approximation results using the Stein–Chen method that we have discussed.

1. **Generalising and relaxing monotonicity assumptions:** Results on monotone couplings may be generalised in a number of directions: Daly, Lefèvre and Utev [31] consider the somewhat more general setting of approximation by the equilibrium distribution of a birth-death process, and Daly and Johnson [30] allow for some relaxation of the rather strict monotonicity we have assumed here.

2. **Translated Poisson approximation:** One disadvantage of Poisson approximation compared to, for example, Gaussian approximation is that the Poisson distribution has only one parameter we are able to choose. Röllin [53, 54] has explored a two-parameter *translated* Poisson approximation, allowing one to (almost) match the first two moments of W with those of the approximating random variable. The techniques employed here include taking advantage of conditional independence or using an exchangeable pairs coupling.

3. **Compound Poisson approximation:** A natural generalisation of Poisson approximation is to consider *compound* Poisson approximation, where the approximating random variable has the form $Y_1 + \cdots + Y_N$, where the Y_i are IID positive random variables and N has a Poisson distribution. This would allow a much greater range of regimes in which limiting results could be obtained: it allows for situations in which rare events can happen in 'clumps'. Consider, for example, the number of observed runs of r Heads in a sequence of independent coin tosses, each coin showing Heads with probability p. The probability of seeing such a run of Heads at a given time is p^r, but having just observed one, we observe another at the next time point with the (relatively large) probability p. Thus, the usefulness of a Poisson approximation may be limited; instead we may want to use a compound Poisson approximation, where we are assuming that the occurrence of'clumps' is rare (i.e., approximately Poisson), and the random variables Y_i take account of the number of events we see in each clump. For more background on this idea, see the book by Aldous [1].

   Stein's method for compound Poisson approximation was first studied by Barbour, Chen and Loh [7], using the Stein equation

$$h(j) - \mathbb{E}h(Z) = \sum_{k=1}^{\infty} k\lambda_k f(j+k) - jf(j)\,,$$

   where $\lambda = \mathbb{E}N$, $\mu_j = \mathbb{P}(Y_i = j)$ and $\lambda_j = \lambda\mu_j$, which is a natural generalisation of the Poisson Stein equation. Unfortunately, the solution of this Stein equation is not as well-behaved as we might hope, with the solution f being bounded only exponentially in λ in the general case. This limits how useful the resulting approximation bounds will be. There are, however, some cases in which bounds of an order comparable to those in the Poisson case (as in Lemma 4.6) are available. This includes when $k\lambda_k \geq (k+1)\lambda_{k+1}$ for all k [7], or when $\sum_j j(j-1)\lambda_j < \frac{1}{2}\sum_j j\lambda_j$ [11]. These conditions each quantify the idea of the approximating compound Poisson distribution being 'not too far from Poisson'. In these cases, useful compound Poisson approximation theorems may be derived in a wide range of applications. Again, the proofs of such results may be approached by either a 'local approach' or a 'coupling approach'.

There has been some work on using monotonicity of the couplings associated with compound Poisson approximation using the Stein–Chen method; see [24].

4. **Poisson process approximation:** Many of the techniques we have considered here can be extended to the setting of approximation by a Poisson process. Recall that a point process Ξ on a space Γ (which we assume to be a locally compact complete separable metric space) with locally finite mean measure $\boldsymbol{\lambda}$ is a Poisson process if (i) for any $B \subseteq \Gamma$, $\Xi(B)$ has a Poisson distribution with mean $\boldsymbol{\lambda}(B)$, and (ii) for disjoint sets $B_1, \dots, B_k \subseteq \Gamma$, the random variables $\Xi(B_1), \dots, \Xi(B_k)$ are independent.

   In this setting, the role of size biasing is played by the Palm distributions $\{P_\alpha : \alpha \in \Gamma\}$ defined by
$$P_\alpha(B) = \frac{\mathbb{E}[I_{\{\Xi \in B\}}\Xi(d\alpha)]}{\boldsymbol{\lambda}(d\alpha)},$$
   or, letting Ξ_α be a process with the Palm distribution P_α of Ξ,
$$\mathbb{E}\int_B f(\alpha, \Xi)\Xi(d\alpha) = \mathbb{E}\int_B f(\alpha, \Xi_\alpha)\boldsymbol{\lambda}(d\alpha).$$
   for any measurable function f. We can characterize a Poisson process based on the fact that Ξ is a Poisson process if and only if Ξ and $\Xi_\alpha - \delta_\alpha$ have the same distribution, where $\Xi_\alpha - \delta_\alpha$ is the *reduced* Palm process. This can be used to define a Stein equation for the Poisson process, which may then be used to yield explicit error bounds in approximation of a point process by a Poisson process in appropriate metrics. See [21] for a a discussion of this approach to Poisson process approximation via Stein's method.

5 Approximation by geometric sums

In this section we will consider some settings in which we wish to approximate a (non-negative, integer-valued) random variable W of interest by a geometric sum Z which may be written as $Y_1 + \cdots + Y_N$, where $Y, Y_1, Y_2, \dots$ are IID positive integer-valued random variables, and $N \sim \text{Geom}(p)$ has a geometric distribution with mass function $\mathbb{P}(N = k) = p(1-p)^k$ for $k = 0, 1, \dots$. Unless stated otherwise, we will choose this parameter p to be such that
$$p = \mathbb{P}(N = 0) = \mathbb{P}(Z = 0) = \mathbb{P}(W = 0),$$
and consider the approximation of W by Z in total variation distance.

Geometric sums such as these appear in various areas of probability, including risk theory (since the infinite-horizon ruin probability in various risk models may be expressed as a tail probability of such a geometric sum), random walks and records processes. See the book by Kalashnikov [43].

We will begin by giving a Stein equation which may be used in this setting, together with bounds on its solution. We will then consider two applications: the approximation of passage times for stationary Markov chains and the approximation of random variables having an increasing failure rate. The material in this section is based on the papers [23], [26] and [27].

5.1 The Stein equation and its solution

In approximating by the geometric sum Z defined above, we may use the Stein equation given by

$$h(j) - \mathbb{E}h(Z) = (1-p)\mathbb{E}f(j+Y) - f(j)\,.$$

Exercise: Check that if Z is the geometric sum defined above, then for all abounded functions f we have that $(1-p)\mathbb{E}f(Z+Y) = \mathbb{E}f(Z)$ (recall that Y is independent of $Y_1, Y_2, \ldots$).

Exercise: Check that we may represent the solution f of this Stein equation in the following way: let $f(0) = 0$ and

$$f(j) = -\mathbb{E}\sum_{r=0}^{\infty}(1-p)^r \widehat{h}(j+S_r)\,, \tag{39}$$

for $j \geq 1$, where $\widehat{h}(j) = h(j) = \mathbb{E}h(Z)$ and $S_r = Y_1 + \cdots + Y_r$.

We will now give the bounds on this function f that we will need in what follows.

Lemma 5.1. *For $h \in \mathcal{H}_{TV}$,*

$$\|\Delta f\|_\infty \leq 1 + (1-p)\mathbb{P}(Y > 1)\,, \tag{40}$$

$$|f(j) - f(k)| \leq \frac{1}{p}\,, \tag{41}$$

for any $j, k \in \mathbb{Z}^+$.

Proof. For (40), we may assume without loss of generality that $h = I_A$, the indicator function for some $A \subseteq \mathbb{Z}^+$. We may then write

$$f(j) = -\sum_{i=0}^{\infty}(1-p)^i \left[\mathbb{P}(j + Y_1 + \cdots + Y_i \in A) - \mathbb{P}(Z + Y_1 + \cdots + Y_i \in A)\right]\,,$$

from which it follows that $|\Delta f(j)|$ may be bounded by

$$\sum_{i=0}^{\infty}(1-p)^i \left|\mathbb{P}(j + 1 + Y_1 + \cdots + Y_i \in A) - \mathbb{P}(j + Y_1 + \cdots + Y_i \in A)\right|\,.$$

To complete the proof of the first inequality, we let $N \sim \text{Geom}(p)$ and write this as

$$\frac{1}{p}|\mathbb{P}(j + 1 + Y_1 + \cdots + Y_N \in A) - \mathbb{P}(j + Y_1 + \cdots + Y_N \in A)|$$
$$\leq \frac{1}{p}d_{TV}(\mathcal{L}(Z), \mathcal{L}(Z+1)) \leq 1 + \frac{(1-p)}{p}\mathbb{P}(Y > 1)\,,$$

where the final inequality uses Theorem 3.1 of [62].

For (41), note that for any $j, k \geq 0$, we have

$$\begin{aligned}|f(j) - f(k)| &= \left| \mathbb{E}\left[\sum_{r=0}^{\infty} (1-p)^r \left[\hat{h}(S_r + j) - \hat{h}(S_r + k) \right] \right] \right| \\ &= \left| \mathbb{E}\left[\sum_{r=0}^{\infty} (1-p)^r \left[h(S_r + j) - h(S_r + k) \right] \right] \right| \\ &\leq \sum_{r=0}^{\infty} (1-p)^r \; = \; \frac{1}{p}\,,\end{aligned}$$

as required. □

Given a random variable W which we wish to approximate, we define the random variable V such that

$$V + Y \overset{d}{=} W | W > 0\,,$$

where W and Y are independent. We may then write

$$\begin{aligned}Eh(W) - Eh(Z) = (1-p)E[f(W+Y)] - (1-p)E[f(W)|W > 0] \\ = (1-p)E[f(W+Y) - f(V+Y)]\,,\end{aligned} \quad (42)$$

since $f(0) = 0$.

In [23], the following general bound is established.

Proposition 5.2. *Let the random variables V, W and Z be as above.*

$$d_{TV}(W, Z) \leq \frac{1-p}{p} d_{TV}(W, V)\,.$$

Proof. From (42) we have that, for any $h \in \mathcal{H}_{TV}$,

$$|Eh(W) - Eh(Z)| \leq (1-p)\mathbb{P}(W \neq V) \sup_{j,k \in \mathbb{Z}^+} |f(j) - f(k)|\,,$$

from which the result follows on applying (41) and taking the maximal coupling of (V, W). □

In the following section we will use this to consider the approximation of passage times of stationary Markov chains.

5.2 Markov chain hitting times

Let $\{\xi_t : t \geq 0\}$ be an irreducible, ergodic, discrete time Markov chain with finite state space S, transition matrix P, and stationary distribution π. We suppose throughout that this Markov chain is started according to its stationary distribution. We will consider the approximation of $W = \inf\{t \geq 0 : \xi_t = x\}$, the first time this Markov chain reaches a given state x. Fill and Lyzinski [35] have shown in their Theorem 4.2 that such a hitting time may, under certain conditions, be expressed exactly as a geometric sum. In particular, this holds if the underlying Markov chain is reversible and $P^t(x, x)$, the t-step transition probability from x to x, is decreasing in t. In this case, the summands Y are strong stationary times for this Markov chain with initial distribution π_x, the stationary distribution restricted to states other than x. See [2] and [32] for background on strong stationary times. We recall some relevant definitions and background below.

Definition 5.3. *A strong stationary time, T, for $\{\xi_t : t \geq 0\}$ is a randomized stopping time such that $\xi_T \sim \pi$ and X_T is independent of T. Note that the distribution of T depends on the initial distribution of the Markov chain, but this dependence is often suppressed in the notation.*

The tails of strong stationary times are closely related to separation, which may be used to estimate how far ξ_t is from stationarity. The separation at time t is defined by

$$s(t) = 1 - \min_{y \in S} \left\{ \frac{\mathbb{P}(\xi_t = y)}{\pi(y)} \right\} .$$

A strong stationary time T is called 'fastest' if it is stochastically smaller than any other strong stationary time. Proposition 3.2 of [2] guarantees the existence of such a fastest strong stationary time for any Markov chain X of the type we consider here.

Lemma 5.4. *There exists a strong stationary time T for $\{\xi_t : t \geq 0\}$ such that $\mathbb{P}(T > t) = s(t)$ for all $t \geq 0$.*

We collect some well-known properties of strong stationary times in the following lemma.

Lemma 5.5. *Let $x \in S$ be such that $\pi(y)\mathbb{P}(\xi_t = x) \leq \pi(x)\mathbb{P}(\xi_t = y)$, for all $t \geq 0$ and $y \in S$, when $\xi_0 \sim \pi_x$. Then $P^t(x, x)$ is decreasing in t and, letting the random variable Y be defined by*

$$\mathbb{P}(Y > t) = \frac{P^t(x, x) - \pi(x)}{1 - \pi(x)} , \qquad t = 0, 1, \dots ,$$

Y has the distribution of a fastest strong stationary time for our Markov chain with initial distribution $\xi_0 \sim \pi_x$. Furthermore, $\mathbb{P}(Y > t, \xi_t = x) = 0$ for all $t \geq 0$.

Proof. Under the given condition on state x, when $\xi_0 \sim \pi_x$,

$$\begin{aligned} s(t) = 1 - \frac{\mathbb{P}(\xi_t = x)}{\pi(x)} &= 1 - \frac{\sum_{s \neq x} P^t(s, x)\pi(s)}{\pi(x)(1 - \pi(x))} \\ &= 1 - \frac{\pi(x) - P^t(x, x)\pi(x)}{\pi(x)(1 - \pi(x))} = \frac{P^t(x, x) - \pi(x)}{1 - \pi(x)} . \end{aligned}$$

Since $s(t)$ is known to be decreasing in t (see Chapter 9 of [3]), we have that $P^t(x,x)$ is decreasing. By Lemma 5.4, Y as defined has the distribution of a fastest strong stationary time. Finally,

$$\begin{aligned}\mathbb{P}(Y > t, \xi_t = x) &= \mathbb{P}(\xi_t = x) - \mathbb{P}(Y \le t, \xi_t = x) \\ &= \mathbb{P}(\xi_t = x) - \mathbb{P}(Y \le t)\pi(x) = \pi(x)\left(1 - s(t) - \mathbb{P}(Y \le t)\right) = 0\,,\end{aligned}$$

where the second equality follows from Lemma 6.9 of [45], the third by the assumption made on the state x, and the final equality follows from Lemma 5.4. □

We are now in a position to state our main result of this section.

Theorem 5.6. *Let $W = \inf\{t \ge 0 : \xi_t = x\}$, where we assume that $\xi_0 \sim \pi$. Let Y be a strong stationary time (independent of W) for this Markov chain with initial distribution π_x. Let $N \sim$ Geom$(\pi(x))$. Then*

$$d_{TV}(W, Z) \le (1 - \pi(x))\mathbb{E}Y - \pi(x)\mathbb{E}W\,,$$

where $Z = Y_1 + \cdots + Y_N$ and $Y, Y_1, Y_2, \ldots$ are IID.

Proof. Write $\widetilde{W} = (W | W > 0)$. We construct $\widetilde{W}$ as the time of the first visit of our Markov chain to the state x when initialized with distribution π_x. This hitting time may come either before the strong stationary time Y for this chain, or not. On the event that $\widetilde{W} < Y$, we set $V = \widetilde{W} - Y$; otherwise we have achieved stationarity at time Y, and the remaining time until we reach state x is distributed as W, and we may set $V = W$. Here we construct W using a Bernoulli random variable $\eta \sim \text{Be}(\pi(x))$ (independent of all else) with $\mathbb{P}(\eta = 1) = 1 - \mathbb{P}(\eta = 0) = \pi(x)$, and set $W = 0$ if $\eta = 1$; otherwise we set $W = \widetilde{W}$.

Letting $\{\widetilde{\xi}_t : t \ge 0\}$ denote a copy of our Markov chain started according to π_x, we therefore have that

$$d_{TV}(W, V) \le \mathbb{P}(\widetilde{W} < Y) \le \sum_{t=1}^{\infty} \mathbb{P}(\widetilde{\xi}_t = x, Y > t)\,,$$

and hence

$$d_{TV}(W, V) \le \sum_{t=1}^{\infty} \left[\mathbb{P}(\widetilde{\xi}_t = x) - \mathbb{P}(Y \le t, \widetilde{\xi}_t = x)\right]\,.$$

Now, Lemma 6.9 of [45] gives us that $\mathbb{P}(Y \le t, \widetilde{\xi}_t = x) = \pi(x)\mathbb{P}(Y \le t)$. Further,

$$\mathbb{P}(\widetilde{\xi}_t = x) = \sum_{s \ne x} P^t(s, x)\frac{\pi(s)}{1 - \pi(x)} = \frac{\pi(x)}{1 - \pi(x)}\left[1 - P^t(x, x)\right]\,,$$

since $\sum_{s\in S}\pi(s)P^t(s,x)=\pi(x)$. Hence,

$$\begin{aligned}
d_{TV}(W,V) &\leq \pi(x)\sum_{t=1}^{\infty}\left[\frac{1-P^t(x,x)}{1-\pi(x)}-\mathbb{P}(Y\leq t)\right]\\
&=\pi(x)\sum_{t=1}^{\infty}\left[\mathbb{P}(Y>t)-\frac{P^t(x,x)-\pi(x)}{1-\pi(x)}\right]\\
&=\pi(x)\mathbb{E}Y-\frac{\pi(x)^2}{1-\pi(x)}\mathbb{E}W\,.
\end{aligned}$$

The final equality follows from the identity $\pi(x)\mathbb{E}W=\sum_{t=0}^{\infty}\left[P^t(x,x)-\pi(x)\right]$ given in Proposition 10.19 of [45]. The result now follows from Proposition 5.2. □

We may use this result, for example, to give conditions in the spirit of [35] under which W is distributed exactly as a geometric sum. If x is such that $\pi(y)\mathbb{P}(\xi_t=x)\leq\pi(x)\mathbb{P}(\xi_t=y)$ for all $t\geq 0$ and $y\in S$, then Lemma 5.5 gives us that we may take the strong stationary time Y to have distribution

$$\mathbb{P}(Y>t)=\frac{P^t(x,x)-\pi(x)}{1-\pi(x)}\,,\qquad t=0,1,\ldots\,,$$

and so the right-hand side of (5.6) is zero (again using Proposition 10.19 of [45]).

Finally in this section, we note that W is stochastically smaller than the approximating geometric sum we have constructed.

Theorem 5.7. *Let $W=\inf\{t\geq 0:\xi_t=x\}$, where we assume that $\xi_0\sim\pi$. Let Y be a strong stationary time (independent of W) for this Markov chain with initial distribution $\xi_0\sim\pi_x$. Let $Z=Y_1+\cdots+Y_N$, where $Y,Y_1,Y_2,\ldots$ are IID and $N\sim$ Geom$(\pi(x))$. Then $W\leq_{st}Z$.*

Proof. Following (42), we write

$$\mathbb{E}h(W)-\mathbb{E}h(Z)=(1-\pi(x))\mathbb{E}\left[f(W+Y)-f(V+Y)\right]\,, \tag{43}$$

where V is the random variable constructed in the proof of Theorem 5.6, $f(0)=0$, and

$$f(x)=-\mathbb{E}\left[\sum_{r=0}^{\infty}(1-\pi(x))^r\{h(x+S_r)-\mathbb{E}h(Z)\}\right]\,, \tag{44}$$

for $x>0$, where $S_r=Y_1+\cdots+Y_r$; see (39).

Exercise: Check that if h is increasing, then f is decreasing.

Now, since we have constructed V in the proof of Theorem 5.6 in such a way that $V\leq W$ almost surely, using (43) gives us the desired inequality. □

5.3 Geometric approximation for IFR random variables

In this section we explore one particular monotonicity condition in the setting of geometric approximation: the case where W has an increasing failure rate (IFR). The failure (or hazard) rate of a non-negative, integer-valued random variable W is defined to be

$$r_W(j) = \frac{\mathbb{P}(W = j)}{\mathbb{P}(W > j)}, \qquad j \in \mathbb{Z}^+ .$$

We may define the failure rate for a continuous random variable analogously, with a density function replacing the mass function. W is said to be IFR if $r_W(j)$ is a non-decreasing function of j. Note that if $\mathbb{P}(W = 0) = p$, then $r_W(0) = p(1 - p)^{-1}$.

We will limit ourselves to considering only geometric approximation here (i.e., the case where $Y = 1$ almost surely) using the following result. This is a special case of a result in approximation by a geometric sum for random variables whose failure rate may be bounded from below established by [26].

Theorem 5.8. *Let W be a non-negative, integer-valued random variable with $\mathbb{P}(W = 0) = p$. If $r_W(j) \geq p(1 - p)^{-1}$ (e.g., if W is IFR) then*

$$d_{TV}(W, Z) \leq 1 - p(1 + \mathbb{E}W),$$

where $Z \sim Geom(p)$.

The proof of this result will need the following lemma. Note that the random variable V we have employed previously is here defined by $V + 1 \overset{d}{=} (W|W > 0)$.

Lemma 5.9. *Let W be a nonnegative, integer-valued random variable with $\mathbb{P}(W = 0) = p$ and $r_W(j) \geq p(1 - p)^{-1}$ for all $j \in \mathbb{Z}^+$. Let V be as above. Then $V \leq_{st} W$.*

Proof. We have that

$$\mathbb{P}(V > j) = \mathbb{P}(W > j + 1|W > 0) = \frac{\mathbb{P}(W > j + 1)}{1 - p} .$$

Hence, writing $\mathbb{P}(W > j) = \mathbb{P}(W > j + 1) + \mathbb{P}(W = j + 1)$, the required conclusion holds if

$$\mathbb{P}(W > j + 1) \leq (1 - p)[\mathbb{P}(W > j + 1) + \mathbb{P}(W = j + 1)],$$

that is, if

$$p\mathbb{P}(W > j + 1) \leq (1 - p)\mathbb{P}(W = j + 1),$$

which holds by assumption. □

We complete the proof of Theorem 5.8 in a similar way to the proof of Theorem 4.12. For $h \in \mathcal{H}_{TV}$, we write

$$\begin{aligned}\mathbb{E}h(W) - \mathbb{E}h(Z) &= (1-p)\mathbb{E}\left[f(W+1) - f(V+1)\right] \\ &= (1-p)\sum_{j=0}^{\infty} \Delta f(j)\left[\mathbb{P}(W+1>j) - \mathbb{P}(V+1>j)\right] .\end{aligned}$$

Now, taking absolute values and using the stochastic ordering of Lemma 5.9 gives us that

$$\begin{aligned}d_{TV}(W,Z) \leq (1-p) \sup_{h\in\mathcal{H}_{TV}} \|\Delta f\|_\infty \mathbb{E}[W-V] &\leq (1-p)\mathbb{E}[W-V] \\ &= p\mathbb{E}[Z-W] = (1-p) - p\mathbb{E}W ,\end{aligned}$$

where the second inequality uses (40), and the first equality uses the definition of V. This completes the proof of Theorem 5.8.

We conclude this section with two short applications of Theorem 5.8.

1. **The Pólya distribution:** Suppose that m balls are distributed randomly among $d \geq 2$ urns, so that all assignments are equally likely. Let W count the number of balls in the first urn. Then $W \sim \text{Pya}(m,d)$ has a Pólya distribution, with

$$\mathbb{P}(W=k) = \frac{\binom{d+m-k-2}{m-k}}{\binom{d+m-1}{m}} , \qquad 0 \leq k \leq m .$$

   It is known that W is IFR, and we may apply Theorem 5.8 to obtain

$$d_{TV}\left(W,Z\right) \leq \frac{m}{d(d+m-1)} ,$$

   where $Z \sim \text{Geom}\left(\frac{d-1}{d+m-1}\right)$.

   A simple corresponding lower bound is given by

$$d_{TV}(W,Z) \geq |\mathbb{P}(W=1) - \mathbb{P}(Z=1)| = \frac{m(d-1)}{(d+m-2)(d+m-1)^2} .$$

   In the case where d is of order $O(m)$, this lower bound is of the same order as our upper bound.

2. **Stopped Poisson process:** Let $\{N(t) : t \geq 0\}$ be a homogeneous Poisson process of rate λ and let T be an IFR random variable independent of $\{N(t) : t \geq 0\}$. By Corollary 5.2 of [57], $N(T)$ is also IFR. Since $\mathbb{P}(N(T)=0) = \mathbb{E}e^{-\lambda T}$ and $\mathbb{E}N(T) = \lambda\mathbb{E}T$, Theorem 5.8 gives

$$d_{TV}\left(N(T),Z\right) \leq 1 - \left(\mathbb{E}e^{-\lambda T}\right)\left(1+\lambda\mathbb{E}T\right) ,$$

   where $Z \sim \text{Geom}\left(\mathbb{E}e^{-\lambda T}\right)$.

6 Infinitely divisible distributions

6.1 Introduction

As (perhaps) suggested by the structure of these notes, it is typically the case that Stein's method is developed and introduced for a single limiting law at a time. There have been, however, several works which present a more unified approach to treating families of random variables simultaneously. For example, Brown and Xia [13] consider the general setting of approximation by the equilibrium distribution of a birth-death process, and Eichelsbacher and Reinert [33] study approximation by discrete Gibbs measures.

In this section we will give a brief account of (a small part of) recent work by Arras and Houdré [4] on approximation by infinitely divisible distributions with finite first moment. We will firstly present some background on infinitely divisible distributions and the important subclass of self-decomposable distributions. Following this, we state the necessary ingredients to apply Stein's method: a characterisation, a Stein equation, and bounds on its solution. These we will state without proof: see [4] for proofs of all results we give here.

Finally, in the next section we will present a concrete application of Stein's method in this setting; approximation of the maximum of a sample of IID exponential data by a Gumbel distribution.

Definition 6.1. *A random variable Z is infinitely divisible if, for each $n \geq 1$, there are IID random variables $Z_{1,n}, \dots, Z_{n,n}$ such that $Z \overset{d}{=} Z_{1,n} + \cdots + Z_{n,n}$.*

Examples of infinitely divisible distributions include the Gaussian, compound Poisson, negative binomial and gamma random variables, so this is a rather wide class, and the work of Arras and Houdré [4] in this setting has potentially many applications. Note, however, that there are random variables (such as the binomial) for which the tools of Stein's method have been developed [34], but which do not fit into this framework.

Exercise: Use properties of the gamma distribution to conclude that it is infinitely divisible (and in particular, so is the exponential distribution).

The famous Lévy–Khintchine representation states that Z is infinitely divisible if and only if its characteristic function $\varphi(t) = \mathbb{E}[e^{itZ}]$ has the form

$$\varphi(t) = \exp\left\{ itb - \frac{\sigma^2 t^2}{2} + \int_{-\infty}^{\infty} \left(e^{itu} - 1 - ituI_{\{|u| \leq 1\}} \right) \nu(du) \right\} ,$$

for all $t \in \mathbb{R}$, where $b \in \mathbb{R}$, $\sigma \geq 0$ and ν is a Borel measure (called the Lévy measure of Z) on $\mathbb{R}$ such that $\nu(0) = 0$ and $\int_{-\infty}^{\infty} (1 \wedge u^2) \nu(du) < \infty$. In this case, we write that $Z \sim \text{ID}(b, \sigma^2, \nu)$.

This representation is the starting point for a Stein-type characterisation of infinitely divisible distributions: see Theorem 3.1 of [4] for a proof of the following.

Lemma 6.2. *Let X be a random variable with $\mathbb{E}|X| < \infty$. Let $b \in \mathbb{R}$, $\sigma \geq 0$ and ν a Borel measure on $\mathbb{R}$ such that $\nu(0) = 0$, $\int_{-\infty}^{\infty}(1 \wedge u^2)\nu(du) < \infty$ and $\int_{|u|>1} |u|\, \nu(du) < \infty$. Then*

$$\mathbb{E}\left[Xf(X) - bf(X) - \sigma^2 f'(X) - \int_{-\infty}^{\infty} \left[f(X+u) - f(X)I_{\{|u|\leq 1\}}\right] u\,\nu(du)\right] = 0$$

for all bounded Lipschitz functions $f : \mathbb{R} \mapsto \mathbb{R}$, if and only if $X \sim ID(b, \sigma^2, \nu)$.

There are other, equivalent forms of this characterisation that also prove useful. For example, in the setting where $\sigma = 0$, we have that if X is infinitely divisible with Lévy measure ν, then

$$\mathbb{E}[Xf(X)] - \left(\mathbb{E}X - \int_{-\infty}^{\infty} u\,\nu(dv)\right)\mathbb{E}f(X) = \mathbb{E}\int_{-\infty}^{\infty} f(X+u)u\,\nu(du)\,. \tag{45}$$

Exercise: Let X have a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, which has Lévy measure $\nu(du) = \alpha u^{-1} I_{\{u>0\}} \exp(-\beta u)\, du$. Use (45) to show that for any function f as in Lemma 6.2,

$$\mathbb{E}[Xf(X)] = \mathbb{E}X\mathbb{E}f(X+Y)\,, \tag{46}$$

where $Y \sim \text{Exp}(\beta)$ and is independent of X.

Special attention is paid by [4] to the setting where $\sigma = 0$ (i.e., Z has no Gaussian part) and Z is self decomposable:

Definition 6.3. *Z is self-decomposable if, for any $0 < c < 1$, there exists a random variable Z_c, independent of Z, such that $Z \overset{d}{=} cZ + Z_c$.*

Note that non-degenerate self-decomposable random variables are infinitely divisible and absolutely continuous, and that the class of self-decomposable random variables is closed under convolution. Self decomposable distributions include stable laws, gamma distributions, log-normal distributions, the Laplace distributions and the logisitc distributions, among many other well-known examples.

In the case where $Z \sim \text{ID}(b, 0, \nu)$ is self-decomposable, the corresponding Stein equation employed by Arras and Houdré [4] is given by

$$h(x) - \mathbb{E}h(Z) = (\mathbb{E}Z - x)\, f'(x) + \int_{-\infty}^{\infty} \left[f'(x+u) - f'(x)\right] u\,\nu(du)\,. \tag{47}$$

The solution $f = f_h$ of this Stein equation is shown to satisfy the following properties (see Chapter 5 of [4]):

- If h is a continuously differentiable function with $\|h\|_\infty \leq 1$ and $\|h'\|_\infty \leq 1$, then f is differentiable on $\mathbb{R}$ and $\|f'\|_\infty \leq 1$.
- If h is a twice continuously differentiable function with $\|h\|_\infty \leq 1$, $\|h'\|_\infty \leq 1$ and $\|h''\|_\infty \leq 1$, then f is twice differentiable on $\mathbb{R}$ and $\|f''\|_\infty \leq \frac{1}{2}$.

6.2 Application: Gumbel approximation for the maximum of exponential data

Let $Y_1, Y_2, \dots$ be a sequence of IID exponentially distributed random variables with rate 1, and let

$$W = \sum_{k=1}^{n} \frac{Y_k}{k} - \log(n)\,,$$

where $\log$ is the natural logarithm.

Well-known properties of the exponential distribution guarantee that we have $\max_{1 \le k \le n} Y_k \overset{d}{=} \sum_{k=1}^{n} \frac{Y_k}{k}$.

Exercise: Check this!

Hence, W converges in distribution to a (standard) Gumbel random variable Z with distribution function $F(z) = \exp\{-\exp(z)\}$ for $z \in \mathbb{R}$ and mean $\mathbb{E}Z = \gamma = 0.5772\dots$ given by Euler's constant.

The Gumbel random variable Z is well-known to be self-decomposable, with $Z \sim \mathrm{ID}(\gamma, 0, \nu)$, where

$$\nu(du) = I_{\{u>0\}} \frac{e^{-u}}{u(1-e^{-u})}\, du\,,$$

so that

$$\int_0^1 u\, \nu(du) = \infty\,, \qquad \text{and} \qquad \int_0^\infty u^2\, \nu(du) = \frac{\pi^2}{6}\,.$$

See Example 11.10 of [61]. Hence, the framework of the preceding section applies.

We give the proof of [4] of an explicit rate of convergence of W to Z in the smooth Wasserstein distance d_{W_2} defined by

$$d_{W_2}(W, Z) = \sup_{h \in \mathcal{H}_2} |\mathbb{E}h(W) - \mathbb{E}h(Z)|\,,$$

where $\mathcal{H}_2$ is the set of twice continuously differentiable functions $h : \mathbb{R} \mapsto \mathbb{R}$ with $\|h\|_\infty \le 1$, $\|h'\|_\infty \le 1$ and $\|h''\|_\infty \le 1$.

Theorem 6.4. *Let W be as defined above and Z have the standard Gumbel distribution. Then*

$$d_{W_2}(W, Z) \le \frac{C}{n}\,,$$

for some $C > 0$ which does not depend on n.

Proof. In line with the discussion of the previous section, we let $h \in \mathcal{H}_2$ and bound

$$\begin{aligned} &\left| \mathbb{E}\left[(\gamma - W)\, f'(W) + \int_0^\infty [f'(W+u) - f'(W)]\, \frac{e^{-u}}{1-e^{-u}}\, du \right] \right| \\ &\quad \le |\gamma - \mathbb{E}W| + \left| \mathbb{E}\left[(\mathbb{E}W - W)\, f'(W) + \int_0^\infty [f'(W+u) - f'(W)]\, \frac{e^{-u}}{1-e^{-u}}\, du \right] \right| \qquad (48) \end{aligned}$$

where f is the solution to the Stein equation (47), which satisfies $\|f'\|_\infty \leq 1$ for $h \in \mathcal{H}_2$. We may bound $d_{W_2}(W, Z)$ by taking the supremum of (48) over $h \in \mathcal{H}_2$.

Now,

$$|\gamma - \mathbb{E}W| = \left|\gamma + \log(n) - \sum_{k=1}^{n} \frac{1}{k}\right| \leq \frac{C_1}{n}\,,$$

for some constant C_1 independent of n. It remains only to deal with the final term in (48).

Let $W_k = W - \frac{Y_k}{k}$. From the definition of W we have

$$\begin{aligned}\mathbb{E}\left[Wf'(W)\right] &= \sum_{k=1}^{n} \frac{1}{k}\mathbb{E}\left[Y_k f'\left(W_k + \frac{Y_k}{k}\right)\right] - \log(n)\mathbb{E}f'(W) \\ &= \sum_{k=1}^{n} \frac{1}{k} \int_0^\infty e^{-u}\mathbb{E}\left[f'\left(W + \frac{u}{k}\right)\right] du - \log(n)\mathbb{E}f'(W)\,,\end{aligned}$$

where the second equality applies (46), using independence of W_k and Y_k.

Hence, and again using independence of W_k and Y_k, we have

$$\begin{aligned}&\mathbb{E}\left[(\mathbb{E}W - W)\,f'(W) + \int_0^\infty \left[f'(W+u) - f'(W)\right] \frac{e^{-u}}{1-e^{-u}}\,du\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{n} \frac{1}{k} \int_0^\infty \left[f'(W) - f'\left(W + \frac{u}{k}\right)\right] e^{-u}\,du + \int_0^\infty \left[f'(W+u) - f'(W)\right] \frac{e^{-u}}{1-e^{-u}}\,du\right]\,,\end{aligned}$$

which in turn is equal to

$$\begin{aligned}&\mathbb{E}\left[\int_0^\infty \left[f'(W) - f'(W+u)\right] \frac{1-e^{-nu}}{e^u - 1}\,du + \int_0^\infty \left[f'(W+u) - f'(W)\right] \frac{1}{e^u - 1}\,du\right] \\ &\qquad = \mathbb{E}\int_0^\infty \left[f'(W+u) - f'(W)\right] \frac{e^{-nu}}{e^u - 1}\,du\,.\end{aligned}$$

The absolute value of this may be bounded by

$$\frac{1}{2}\int_0^\infty \frac{ue^{-nu}}{e^u - 1}\,du = \sum_{k=0}^{\infty} \frac{1}{(k+n+1)^2}\,,$$

using properties of f (since $h \in \mathcal{H}_2$), and where the proof is completed on using an asymptotic expansion of this final expression. □

References

[1] D. Aldous (1989). *Probability Approximations via the Poisson Clumping Heuristic*. Springer, New York.

[2] D. Aldous and P. Diaconis (1987). Strong uniform times and finite random walks. *Adv. in Appl. Math.* 8: 69–97.

[3] D. J. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. Unfinished monograph, available at `https://www.stat.berkeley.edu/~aldous/RWG/book.html`.

[4] B. Arras and C. Houdré (2019). *On Stein's Method for Infinitely Divisible Laws with Finite First Moment*. Springer.

[5] R. Arratia, L. Goldstein and L. Gordon (1989). Two moments suffice for Poisson approximation: the Chen–Stein method. *Ann. Probab.* 17: 9–25.

[6] R. Arratia, L. Goldstein and F. Kochman (2019). Size bias for one and all. *Probab. Surveys* 16: 1–61.

[7] A. D. Barbour, L. H. Y. Chen and W.-L. Loh (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.* 20: 1843–1866.

[8] A. D. Barbour and L. H. Y. Chen (eds.) (2005). *An Introduction to Stein's Method*. Lecture Notes Series 4, Institute for Mathematical Sciences, Singapore University Press, Singapore.

[9] A. D. Barbour and P. Hall (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* 95: 473–480.

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

[11] A. D. Barbour and A. Xia (1999). Poisson perturbations. *ESAIM Probab. Stat.* 3: 131–150.

[12] A. A. Borovkov and S. A. Utev (1984). On an inequality and on the related characterization of the normal distribution. *Theory Probab. Appl.* 28(2): 219–228.

[13] T. C. Brown and A. Xia (2001). Stein's method and birth–death processes. *Ann. Probab.* 29(3): 1373–1403.

[14] T. Cacoullos (1982). On upper and lower bounds for the variance of a function of a random variable. *Ann. Probab.* 10(3): 799-809.

[15] L. Le Cam (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* 10: 1181–1197.

[16] S. Chatterjee, J. Fulman and A. Röllin (2011). Exponential approximation by Stein's method and spectral graph theory. *ALEA Lat. Am. J. Probab. Math. Stat.* 8: 197–223.

[17] L. H. Y. Chen (1975). Poisson approximation for dependent trials. *Ann. Probab.* 3: 534–545.

[18] L. H. Y. Chen (1982). An inequality for the multivariate normal distribution. *J. Multi-*

variate Anal. 12: 306–315.

[19] L. H. Y. Chen, L. Goldstein and Q.-M. Shao (2011). *Normal Approximation by Stein's Method*. Springer, Berlin.

[20] L. H. Y. Chen and Q.-M. Shao (2005). Stein's method for normal approximation. In *An Introduction to Stein's Method*, Eds: A. D. Barbour and L. Y. H. Chen, 1–59, Lecture Notes Series 4, Institute for Mathematical Sciences, Singapore University Press, Singapore.

[21] L. H. Y. Chen and A. Xia (2004). Stein's method, Palm theory and Poisson process approximation. *Ann. Probab.* 32(3B): 2545–2569.

[22] H. Chernoff (1981). A note on an inequality involving the normal distribution. *Ann. Probab.* 9(3): 533-535.

[23] F. Daly (2010). Stein's method for compound geometric approximation. *J. Appl. Probab.* 47(1): 146–156.

[24] F. Daly (2013). Compound Poisson approximation with association or negative association via Stein's method. *Electron. Comm. Probab.* 18(30): 1–12.

[25] F. Daly (2016). Negative dependence and stochastic orderings. *ESAIM: Probab. Stat.* 20: 45–65.

[26] F. Daly (2016). Compound geometric approximation under a failure rate constraint. *J. Appl. Probab.* 53(3): 700–714.

[27] F. Daly (2019). On strong stationary times and approximation of Markov chain hitting times by geometric sums. *Stat. Prob. Lett.* 150: 74–80.

[28] F. Daly, F. Ghaderinezhad, C. Ley and Y. Swan (2019). Simple variance bounds with applications to Bayesian posteriors and intractable distributions. Preprint. Available at `arXiv:1911.03396`.

[29] F. Daly and O. Johnson (2013). Bounds on the Poincaré constant under negative dependence *Stat. Probab. Lett.* 83(2): 511–518.

[30] F. Daly and O. Johnson (2018). Relaxation of monotone coupling conditions: Poisson approximation and beyond *J. Appl. Probab.* 55(3): 742-759.

[31] F. Daly, C. Lefèvre and S. Utev (2012). Stein's method and stochastic orderings. *Adv. Appl. Prob.* 44: 343–372.

[32] P. Diaconis and J. A. Fill (1990). Strong stationary times via a new form of duality. *Ann. Probab.* 18: 1483–1522.

[33] P. Eichelsbacher and G. Reinert (2008). Stein's method for discrete Gibbs measures. *Ann. Appl. Probab.* 18(4): 1588-1618.

[34] W. Ehm (1991). Binomial approximation to the Poisson binomial distribution. *Statist.*

Prob. Lett. 11: 7–16.

[35] J. A. Fill and V. Lyzinski (2014). Hitting times and interlacing eigenvalues: a stochastic approach using intertwinings. *J. Theoret. Probab.* 27: 954–981.

[36] J. Fulman and N. Ross (2013). Exponential approximation and Stein's method of exchangeable pairs. *ALEA, Lat. Am. J. Probab. Math. Stat.* 10(1): 1–13.

[37] R. Gaunt (2014). Variance-Gamma approximation via Stein's method. *Electron. J. Probab.* 19(38): 1–33.

[38] R. Gaunt, A. Pickett and G. Reinert (2017). Chi-square approximation by Stein's method with application to Pearson's statistic. *Ann. Appl. Probab.* 27: 720–756.

[39] S. Ghosh and L. Goldstein (2011). Concentration of measures via size biased couplings. *Probab. Th. Relat. Fields* 149: 271–278.

[40] L. Goldstein and G. Reinert (1997). Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab* 7: 935–952.

[41] L. Goldstein and G. Reinert (2013). Stein's method for the Beta distribution and the Pólya–Eggenberger urn. *J. Appl. Probab.* 50(4): 1187–1205.

[42] O. Johnson (2007). Log-concavity and the maximum entropy property of the Poisson distribution. *Stochastic Processes Appl.* 117: 791–802.

[43] V. Kalashnikov (1997). *Geometric Sums: bounds for rare events with applications*. Kluwer, Dordrecht.

[44] C. Klaassen (1985). On an inequality of Chernoff. *Ann. Probab.* 13(3): 966–974.

[45] D. A. Levin, Y. Peres and E. L. Wilmer (2009). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island.

[46] T. M. Liggett (1997). Ultra logconcave sequences and negative dependence. *J. Combin. Theory Ser. A* 79: 315–325.

[47] I. Nourdin and G. Peccati (2012). *Normal Approximation with Malliavin Calculus: From Stein's Method to Universality*. Cambridge University Press.

[48] N. Papadatos and V. Papathanasiou (2002). Approximation for a sum of dependent indicators: an alternative approach. *Adv. Appl. Prob.* 34: 609–625.

[49] J. Pike and H. Ren (2014). Stein's method and the Laplace distribution. *ALEA Lat. Am. J. Probab. Math. Stat.* 11: 571–587.

[50] E. A. Peköz and A. Röllin (2011). New rates for exponential approximation and the theorems of Rényi and Yaglom. *Ann. Probab.* 39: 587–608.

[51] E. A. Peköz, A. Röllin and N. Ross (2013). Total variation error bounds for geometric approximation. *Bernoulli* 19: 610–632.

[52] G. Reinert and A. Röllin (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* 37(6): 2150–2173.

[53] A. Röllin (2005). Approximation of sums of conditionally independent variables by the translated Poisson distribution. *Bernoulli* 11: 1115–1128.

[54] A. Röllin (2007). Translated Poisson approximation using exchangeable pair couplings. *Ann. Appl. Probab.* 17: 1596–1614.

[55] N. Ross (2011). Fundamentals of Stein's method. *Probab. Surveys* 8: 210–293.

[56] N. Ross (2013). Power laws in preferential attachment graphs and Stein's method for the negative binomial distribution. *Adv. Appl. Probab.* 45(3): 876–893.

[57] S. M. Ross, J. G. Shanthikumar and Z. Zhu (2005). On increasing-failure-rate random variables. *J. Appl. Probab.* 42: 797–809.

[58] M. Shaked and J. G. Shanthikumar (2007). *Stochastic Orders*. Springer, New York.

[59] C. Stein (1972). A bound for the error in normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statis. Probab.* 2:583–602.

[60] C. Stein (1986). *Approximate Computation of Expectations*. IMS, Hayward, California.

[61] F. W. Steuel and K. van Harn. *Infinite Divisibility of Probability Distributions on the Real Line*. CRC Press, New York.

[62] P. Vellaisamy and B. Chaudhuri (1996). Poisson and compound Poisson approximations for random sums of random variables. *J. Appl. Probab.* 33(1): 127–137.