Probability approximations using Stein's method

Random graphs, Markov chains and beyond

Fraser Daly*
Heriot–Watt University

LMS Undergraduate Summer School, August 2022

Contents

*f.daly@hw.ac.uk

1 Introduction

One of the main goals of applied probability is to calculate probabilities of interest arising from statistical models of reality. If we are lucky, our models are not too far from reality. If we are very, very lucky, we can calculate those probabilities without too much effort.

Usually we are not so lucky. We can write down a model to help us understand the number of tosses of a fair coin we are likely to need before we first observe the pattern 'Heads Tails Heads' in three consecutive coin tosses, but calculating the probability that this pattern first occurs on the seventh, eighth and ninth coin tosses is too tedious to be practical. One solution to this issue is to use computer simulations to estimate these probabilities. Another solution (and the one we will focus on here) is to turn to limit theorems and approximations.

We can understand a limit theorem as telling us that asymptotically (e.g., as some underlying parameter of the system, or sample size, tends to infinity) our probability of interest tends to a certain limit. 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 certain binomial distributions to a Poisson limit was first established by Poisson in 1837. Generalizations and extensions of these prototypical examples continue to find applications in diverse fields.

However, these limit theorems in isolation have some serious drawbacks as tools for understanding how probabilities of interest behave in practical examples: we don't have an 'infinite sample size' in our experiment, we may have 50 observations. Is 50 'close to infinity'? What does that question even mean? Referring to limit theorems without any indication of a corresponding error bound as *naive*, Aldous [1] writes

"*It is hard to give any argument for the relevance of a proof of a naive limit theorem, except as a vague reassurance that your approximation is sensible, and a good heuristic argument seems equally reassuring.*"

We would like to have a bit more certainty that our estimates are reasonable, perhaps in the form of explicit error bounds in our approximations.

For example, while the result that sufficiently well-behaved maximum likelihood estimators in statistics are asymptotically normally distributed is useful, it would be even better if we

could quantify numerically the maximum error we may be making when we use a normal approximation to construct a confidence interval for our parameter of interest.

In this mini-course we will look at techniques which give us approximations for probabilities that are difficult to calculate, accompanied by explicit error bounds for these approximations. We will mainly focus on a technique known as Stein's method (also referred to as the Stein–Chen method). Charles Stein's seminal 1972 paper [41] first introduced this technique in the setting of normal approximation for sums of weakly dependent random variables. This was followed by work of Louis Chen [12], 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.

Compared other techniques for proving error bounds in probability approximations, Stein's method has two principal advantages:

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

(ii) It can handle dependence between underlying random variables in a natural way.

In this course we will look at applications of Stein's method in univariate settings only, and mainly in the setting of discrete probability distributions. This will allow us to focus on the main ideas underlying the technique without additional technical complications associated with continuous or multivariate distributions. The aim is for this course to be accessible to anyone who has taken an undergraduate-level course covering introductory discrete probability. For those who want to go beyond this starting point, some suggestions for further reading are given in Section 5.

We use the remainder of this section to introduce some notation, and to refresh the fundamentals of discrete probability. Most of the material of the course is covered in Section 2, where we study Poisson approximations with a focus on applications to random graphs and Markov chains. Some further topics are covered in Sections 3 and 4 to illustrate the breadth of application of Stein's method.

Of course, Stein's method is not the only technique for probability approximations, though it is a powerful one. The book [10] gives an overview of many probability approximation techniques, and is an excellent starting point for those interested in this general area.

1.1 Notation

We let $\mathbb{P}$ denote 'probability', so $\mathbb{P}(X = 2)$ denotes the probability that the random variable X takes the value 2, and $\mathbb{P}(X \in A)$ is the probability that X takes a value in the set A.

Similarly, $\mathbb{E}$ will denote 'expectation', so that $\mathbb{E}[X]$ is the expected value (or mean) of a random variable X. As a reminder, for a random variable X which takes values in the set $\mathbb{Z}^+$ of non-negative integers, this expectation is defined by $\mathbb{E}[X] = \sum_{k\in\mathbb{Z}^+} k\mathbb{P}(X = k)$.

$\mathrm{Var}(X)$ is the variance of a random variable X, defined by $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

'IID' stands for 'independent and identically distributed'.

2 Poisson approximation

The random variable Z has a Poisson distribution with mean $\lambda > 0$ if

$$\mathbb{P}(Z = k) = \frac{e^{-\lambda}\lambda^k}{k!}\,,$$

for $k = 0, 1, 2, \ldots$. We write $Z \sim \mathrm{Po}(\lambda)$. Throughout this section we will let the random variable Z have this Poisson distribution.

The Poisson distribution is well-studied as a model for rare events, and as a limiting distribution in such settings, dating back to the work of Poisson in the 1830s. As an initial example, consider IID (independent and identically distributed) random variables $X_1, X_2, \ldots, X_n$, each of which takes the value 1 with probability λ/n and the value 0 with probability $1 - \lambda/n$. Their sum $W = X_1 + \cdots + X_n$ has a binomial distribution with parameters n and λ/n, and probability mass function

$$\mathbb{P}(W = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}\,,$$

for $k = 0, 1, \ldots, n$.

As $n \to \infty$, this probability converges to the probability $\mathbb{P}(Z = k)$. To see this, we note that

$$\lim_{n\to\infty}\frac{n!}{n^k(n-k)!} = 1\,, \qquad \lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}\,, \qquad \text{and} \qquad \lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-k} = 1\,.$$

To show that the probability mass function of our binomial random variable converges to the probability mass function of a Poisson random variable is thus not too hard. Thinking about how we can go further than this leads to some interesting (and much trickier!) questions, including

- Can we bound the error in the approximation? For example, how big is

  $$\sup_k \left|\mathbb{P}(W = k) - \mathbb{P}(Z = k)\right| ?$$

  This is a natural question, but one that is awkward to tackle directly: to bound

  $$\left|\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} - \frac{e^{-\lambda}\lambda^k}{k!}\right|$$

  directly in a meaningful way is not straightforward.

- Can we generalize this example? What happens if our indicator random variables X_i have different distributions? What if they are no longer independent? As we will see in this course, many examples and applications of interest involve underlying random variables which are dependent, so this is an important question.

In this course we will answer both of these questions simultaneously, through the study of Stein's method for Poisson approximation, as developed by Chen [12]. This is a powerful technique which allows us to find explicit error bounds in Poisson approximation settings and which can naturally handle dependence in the underlying random variables.

Before studying Stein's method for Poisson approximation, we will study some other coupling techniques for Poisson approximation for sums of independent indicator random variables. These will help us lay the groundwork for the couplings used in Stein's method, and also allow us to illustrate the benefits of results established by Stein's method over other available bounds even in the independent case.

Throughout this section we will let $X_1, \dots, X_n$ be (possibly dependent) indicator random variables (i.e., they may each take the value 0 or 1). We will write $p_i = \mathbb{P}(X_i = 1)$ and $\lambda = \sum_{i=1}^n p_i$. We are interested in the approximation of the sum $W = X_1 + \cdots + X_n$ by a Poisson random variable $Z \sim \text{Po}(\lambda)$. Note that both W and Z have mean λ.

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

$$d_{TV}(W, Z) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)| .$$

There are other, equivalent ways of expressing the total variation distance. For example, we have the following.

Lemma 2.1. *Let W and Z be random variables taking values in the non-negative integers. Then*

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

Proof. Let $A = \{k \in \mathbb{Z}^+ : \mathbb{P}(W = k) \geq \mathbb{P}(Z = k)\}$. We have that

$$\mathbb{P}(W \in A) - \mathbb{P}(Z \in A) = \sum_{j \in A} [\mathbb{P}(W = j) - \mathbb{P}(Z = j)] , \text{and}$$

$$\mathbb{P}(W \in A) - \mathbb{P}(Z \in A) = \mathbb{P}(Z \in A^c) - \mathbb{P}(W \in A^c) = \sum_{j \in A^c} [\mathbb{P}(Z = j) - \mathbb{P}(W = j)] .$$

Noting that each of the terms in the sums is non-negative, adding these expressions gives

$$2[\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)] = \sum_{j=0}^{\infty} |\mathbb{P}(W = j) - \mathbb{P}(Z = j)| .$$

Finally, we note that this choice of A maximizes the supremum in the definition of the total variation distance. □

2.1 Poisson approximation by maximal coupling

Before we discuss Stein's method for Poisson approximation, we give a brief account of a simple coupling bound for Poisson approximation for sums of independent indicator random variables, due to Le Cam.

A coupling is a construction of two (or more) random variables with given (marginal) distributions on the same probability space. That is, given two random variables with fixed distributions, we are constructing a joint distribution function which has these given marginal distributions. Formally, we have the following.

Definition 2.2. *The pair of random variables $(\widehat{X}, \widehat{Y})$ is a* coupling *of the random variables (X, Y) if $\widehat{X}$ has the same distribution as X, and $\widehat{Y}$ has the same distribution as Y.*

Le Cam's results use the notion of maximal coupling, which we define below.

Definition 2.3. *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 2.4. *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 2.5. *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, due to Le Cam.

Theorem 2.6. *Let $X_1, \ldots, X_n$ be independent indicator random variables, with $\mathbb{P}(X_i = 1) = p_i$. 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. We may construct a Poisson random variable Z with mean λ as the sum of independent Poisson random variables whose means sum to λ. Using this, we write $Z = \sum_{i=1}^{n} Z_i$, where $Z_i \sim \text{Po}(p_i)$.

Using Lemma 2.4, we couple X_i and Z_i maximally for each i to get $(\widehat{X}_i, \widehat{Z}_i)$ with

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

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 Le Cam [11] using operator techniques:

$$\begin{aligned} d_{TV}(W, Z) &\leq 4.5 \max_i p_i \,, \text{ and} \\ d_{TV}(W, Z) &\leq 8\lambda^{-1} \sum_{i=1}^{n} p_i^2 \,. \end{aligned} \tag{1}$$

This last inequality is proved under the assumption that $\max_i p_i \leq 1/4$. We are most interested in the second of these inequalities, which can represent a substantial improvement over the bound of Theorem 2.6 when λ is large, achieved by the inclusion of the 'magic factor' of λ^{-1}. We note also that there is no obvious way of extending the argument of Theorem 2.6 to cover sums of dependent indicator random variables.

2.2 Stein's method for Poisson approximation

Stein's method for Poisson approximation has the advantage of being able to handle dependence between the indicator random variables X_i. We will also see that results here include 'magic factors' akin to that in Le Cam's result (1) which were missing in the coupling argument of Theorem 2.6.

Before proceeding further, we need a little more notation. For any function $g : \mathbb{Z}^+ \to \mathbb{R}$ we let Δ be the forward difference operator, so that $\Delta g(j) = g(j+1) - g(j)$. We let $\|\cdot\|_\infty$ denote the supremum norm, so that $\|g\|_\infty = \sup_x |g(x)|$.

The starting point of Stein's method is the observation that if $Z \sim \text{Po}(\lambda)$ then

$$k\mathbb{P}(Z = k) = \lambda\mathbb{P}(Z = k-1)\,,$$

for each $k = 1, 2, \ldots$. Using this we can compute that, for any bounded function $g : \mathbb{Z}^+ \to \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[Zg(Z)] &= \sum_{k=0}^{\infty} kg(k)\mathbb{P}(Z = k) = \lambda \sum_{k=1}^{\infty} g(k)\mathbb{P}(Z = k-1) \\ &= \lambda \sum_{k=0}^{\infty} g(k+1)\mathbb{P}(Z = k) = \lambda\mathbb{E}[g(Z+1)]\,.\end{aligned}$$

We conclude that if $Z \sim \text{Po}(\lambda)$, then $\mathbb{E}[Zg(Z)] = \lambda\mathbb{E}[g(Z+1)]$ for any bounded function g. In fact, the converse is true too, and we have the following characterization of the Poisson distribution.

Lemma 2.7. *A non-negative, integer-valued random variable X has a Poisson distribution with mean λ if and only if*

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

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

This characterization is at the heart of Stein's method. We know that

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

for all well-behaved g if Z is Poisson with mean λ. We may perhaps hope that

$$\lambda\mathbb{E}\left[g(W+1)\right] - \mathbb{E}\left[Wg(W)\right] \approx 0$$

for all well-behaved g if W is approximately Poisson with mean λ. With a clever choice of function g, Stein's method makes this intuition precise and gives us a way of using the quantity $\lambda\mathbb{E}\left[g(W+1)\right] - \mathbb{E}\left[Wg(W)\right]$ to quantify the distance of W from Poisson.

Consider the following equation: For a given function $h : \mathbb{Z}^+ \to \mathbb{R}$, we let $f = f_h$ solve

$$h(j) - \mathbb{E}h(Z) = \lambda f(j+1) - jf(j)\,, \tag{2}$$

with $f(0) = 0$. We will call (2) the *Stein equation* for the Poisson distribution.

Taking our Stein equation (2), replacing j with W and taking expectations we have that

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\,. \tag{3}$$

If W is approximately Poisson, then the LHS should be small for all well-behaved functions h. So, the RHS should also be small. To make this idea precise (for the case of total variation

distance) we note that the supremum of the absolute value of the LHS of (3) over all indicators of subsets of $\mathbb{Z}^+$ is $d_{TV}(W, Z)$. Denoting this class of functions by

$$\mathcal{H} = \{I(\cdot \in A) : A \subseteq \mathbb{Z}^+\},$$

we may therefore write

$$d_{TV}(W, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| = \sup_{h \in \mathcal{H}} |\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]| . \tag{4}$$

It is not (yet!) clear that this is a useful way of representing this total variation distance, but this turns out to give access to techniques and applications which other methods do not. One advantage of bounding the RHS of (4), rather than the total variation distance directly, is that it only depends on the single random variable W. The dependence on our Poisson target Z is implicit in the form of the RHS (which came from our characterization of the Poisson), but Z itself does not appear explicitly on the RHS.

A selection of metrics other than the total variation distance may be treated similarly, by choosing a different class of functions $\mathcal{H}$. For the purposes of this section we concentrate on total variation distance only.

The solution f of our Stein equation (2) with the choice $h(\cdot) = I(\cdot \in A)$ for some $A \subseteq \mathbb{Z}^+$ is given by

$$f(k+1) = \sum_{j \in A} \frac{I(j \leq k) - \mathbb{P}(Z \leq k)}{\lambda \mathbb{P}(Z = k)/\mathbb{P}(Z = j)} .$$

When applying Stein's method we will need bounds on this function f, which we give below without proof. See Lemma 1.1.1 of [6] for proofs and further discussion.

Lemma 2.8. *If $h \in \mathcal{H}$,*

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

Note that these bounds do not depend on h, nor on the random variable W of interest. In particular, they do not need to be computed every time we want to apply Stein's method to a new example. Such bounds are known as *Stein factors*, or magic factors.

To see how we may bound (4) in practice, we first consider the case where W is a sum of independent indicator random variables.

2.3 Sums of independent indicator random variables

Consider again the setting of Theorem 2.6. In the following theorem, note the improvement over the results stated in Section 2.1. We retain the magic factor of λ^{-1} appearing in (1), but without the restriction on the p_i. The argument here, and its extension to the setting of dependent indicator random variables, is due to Chen [12].

Theorem 2.9. *Let $X_1, \ldots, X_n$ be independent indicator random variables, with $\mathbb{P}(X_i = 1) = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. If $Z \sim Po(\lambda)$,*

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

Proof. By writing $\lambda = \sum_{i=1}^{n} p_i$ and $W = \sum_{i=1}^{n} X_i$, we have that

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

For each i we let $W_i = W - X_i$, and then write $\mathbb{E}[X_i f(W)] = p_i \mathbb{E}[f(W_i + 1)]$. Hence,

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

Since

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

(using Lemma 2.8) we have that

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

□

It is worth noting that the bound given by Theorem 2.9 is of the right order; we have the corresponding lower bound

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

due to Barbour and Hall [5].

The argument of Theorem 2.9 may be relatively easily extended to the case of sums of locally dependent indicator random variables (i.e., where there is some relatively weak dependence between some of the X_i); see [12] for details. We do not pursue this here, but instead consider coupling approaches to applying Stein's method for Poisson approximation of sums of indicator random variables.

2.4 A coupling approach for sums of dependent indicators

Here and in the sections that follow we will define a coupling that can be used in conjunction with Stein's method for Poisson approximation of sums of dependent indicator random variables. We will see how the bounds we obtain simplify considerably in the case of some monotonicity within this coupling, and look at several applications to Poisson approximation problems.

Theorem 2.10. *Let $X_1, \dots, X_n$ be indicator random variables with $\mathbb{P}(X_i = 1) = p_i$. Let $\lambda = p_1 + \cdots + p_n$ and $Z \sim Po(\lambda)$. For each $i = 1, \dots, n$, let U_i have the same distribution as W and let V_i have the same distribution as $(W - 1 | X_i = 1)$. Then*

$$d_{TV}(W, Z) \leq \frac{1 - e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i \mathbb{E}|U_i - V_i| \,.$$

Proof. Using our Stein equation for Poisson approximation,

$$\begin{aligned} d_{TV}(W, Z) &\leq \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{E}[\lambda f(W+1) - W f(W)]| \\ &= \sup_{A \subseteq \mathbb{Z}^+} \left| \sum_{i=1}^{n} \mathbb{E}[p_i f(W+1) - X_i f(W)] \right| \\ &= \sup_{A \subseteq \mathbb{Z}^+} \left| \sum_{i=1}^{n} p_i \mathbb{E}[f(U_i + 1) - f(V_i + 1)] \right| \\ &\leq \sup_{A \subseteq \mathbb{Z}^+} \sum_{i=1}^{n} p_i \mathbb{E}|f(U_i + 1) - f(V_i + 1)| \\ &\leq \frac{1 - e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i \mathbb{E}|U_i - V_i| \,, \end{aligned}$$

where we have used Lemma 2.8 in our final inequality. □

The coupling we have used here is closely related to *size biasing*: see, for example, [3] for an extensive discussion of the role of size biasing in applied probability. As we will see in the following examples, it is often relatively natural to construct the random variables U_i and V_i in order to be able to then evaluate the upper bound of Theorem 2.10.

2.5 Example: Visits to a rare set in a Markov chain

We consider an application of Theorem 2.10 taken from Section 8.5 of [6].

Let $\xi_1, \xi_2, \dots$ be a Markov chain with state space $\mathbb{Z}$ and transition matrix P. That is, P is an infinite matrix whose entries are probabilities and whose rows each sum to 1. We denote the

(i,j)th entry of P by $P_{i,j}$. If $\xi_t = i$, we choose ξ_{t+1} according to the probability distribution given by

$$\mathbb{P}(\xi_{t+1} = j | \xi_t = i) = P_{i,j} \, ,$$

for each $j \in \mathbb{Z}$. Note that this does not depend on t.

We let $P_{i,j}^{(m)}$ denote the m-step transition probability defined by $P_{i,j}^{(m)} = \mathbb{P}(\xi_{t+m} = j | \xi_t = i)$.

We will assume that our Markov chain has a stationary distribution, which we denote by μ, so that $\mu(A)$ is the probability given to the set of states $A \subseteq \mathbb{Z}$ by this stationary distribution. For notational simplicity, we write μ_i for $\mu(\{i\})$. Recall that this stationary distribution is defined such that if ξ_s has the distribution μ, then ξ_t also has this distribution for all $t > s$, and can be calculated using $\boldsymbol{\mu} P = \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is a row vector containing the elements μ_i. We will assume that ξ_1 has our stationary distribution μ.

In the theorem below we will give a bound for Poisson approximation for the number of visits of our Markov chain to the set of states $A \subseteq \mathbb{Z}$ during time $1, \ldots, n$. We will think of this set A as a 'rare set' (in the sense that $\mu(A)$ is small) in order for this approximation to be good.

Theorem 2.11. *Let $\xi_1, \xi_2, \ldots$ be a stationary Markov chain on the state space $\mathbb{Z}$, as described above. For a given $A \subseteq \mathbb{Z}$, let W denote the number of visits of our Markov chain to the set A in time $1, \ldots, n$ and $\lambda = \mathbb{E}[W]$. Then*

$$d_{TV}(W, Z) \leq (1 - e^{-\lambda}) \left[\mu(A) + \frac{2}{\mu(A)} \sum_{r,s \in A} \mu_r \sum_{j \geq 1} |P_{r,s}^{(j)} - \mu_s| \right] \, ,$$

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

Proof. We write $W = X_1 + \cdots + X_n$, where $X_i = I(\xi_i \in A)$, and we note that the X_i are identically distributed. With U_i and V_i defined as in Theorem 2.10, our result will follow if we can construct these random variables in such a way that

$$\mathbb{E}|U_i - V_i| \leq \mu(A) + \frac{2}{\mu(A)} \sum_{r,s \in A} \mu_r \sum_{j \geq 1} |P_{r,s}^{(j)} - \mu_s| \, .$$

We now sketch the argument that achieves such a coupling for each fixed i.

We choose (ξ_i', ξ_i'') in such a way that ξ_i' has the distribution μ restricted to A and ξ_i'' has the distribution μ. We then define Markov chains for times earlier and later than i in the following way:

- For $\{(\xi_{i+j}', \xi_{i+j}'') : j \geq 1\}$ we choose a maximal coupling of two copies of our Markov chain started from the states ξ_i' and ξ_i'', respectively. We let T^+ denote the *coupling time*, the first time $j \geq 1$ at which $\xi_{i+j}' = \xi_{i+j}''$.
- For $\{(\xi_{i-j}', \xi_{i-j}'') : j \geq 1\}$ we choose a maximal coupling of two copies of the *time reversal* our Markov chain started from the states ξ_i' and ξ_i'', respectively. This time reversal is a Markov chain with transition matrix whose (r,s)th entry is given by $\mu_s P_{s,r} / \mu_r$. We let T^- denote the coupling time, the minimal $j \geq 1$ at which $\xi_{i-j}' = \xi_{i-j}''$.

We then define

$$U_i = \sum_{j=1}^{n} I(\xi''_j \in A) \qquad \text{and} \qquad V_i + 1 = \sum_{j=1}^{n} I(\xi'_j \in A)\,,$$

which we can check have the required distributions, and satisfy

$$|U_i - V_i| \leq I(\xi''_i \in A) + \sum_{j\geq 1} \big\{ I(T^+ > j)[I(\xi'_{i+j} \in A) + I(\xi''_{i+j} \in A)] \\ + I(T^- > j)[I(\xi'_{i-j} \in A) + I(\xi''_{i-j} \in A)] \big\}\,.$$

This gives us that

$$\mathbb{E}|U_i - V_i| \leq \mu(A) + \sum_{j\geq 1} \big\{ \mathbb{P}(T^+ > j, \xi'_{i+j} \in A) + \mathbb{P}(T^- > j, \xi'_{i-j} \in A) \\ + \mathbb{P}(T^+ > j, \xi''_{i+j} \in A) + \mathbb{P}(T^- > j, \xi''_{i-j} \in A) \big\}\,. \quad (5)$$

Properties of maximal couplings of Markov chains developed by Griffeath [27] then give us, for example, that

$$\begin{aligned} \mathbb{P}(T^+ > j, \xi'_{i+j} \in A) &= \sum_{s\in A} \left[\frac{1}{\mu(A)} \sum_{r\in A} \mu_r P^{(j)}_{r,s} - \mu_s \right]^+ \\ &\leq \frac{1}{\mu(A)} \sum_{r\in A} \mu_r \sum_{s\in A} [P^{(j)}_{r,s} - \mu_s]^+\,, \end{aligned}$$

where $[\cdot]^+$ denotes the positive part of the given expression. Similar inequalities may be written down for the other probabilities in (5), from which the desired result follows. □

There are many other applications of Stein's method to approximation problems for Markov chains available in the literature: see, for example [8, 16, 17, 20, 21, 33, 36].

2.6 Monotonicity in coupling

The upper bound in Theorem 2.10 can often be delicate to estimate, requiring a detailed study of the random variables U_i and V_i. However, the upper bound simplifies considerably if we have some monotonicity in this coupling construction. We see this in the two results in this section, where we obtain upper bounds that depend essentially only on the first two moments of W. As illustrated by the examples that follow, this monotonicity arises in a variety of settings of interest. Many further examples in which we can exploit this monotonicity are considered by Barbour, Holst and Janson [6].

Theorem 2.12. *With notation as in Theorem 2.10, if $U_i \geq V_i$ with probability 1 for each $i = 1, \ldots, n$, then*

$$d_{TV}(W, Z) \leq 1 - \frac{\mathit{Var}(W)}{\mathbb{E}[W]}\,.$$

Proof. Under our monotonicity assumption, Theorem 2.10 gives

$$d_{TV}(W,Z) \leq \frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^{n} p_i \mathbb{E}|U_i - V_i| = \frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^{n} p_i \mathbb{E}[U_i - V_i] \leq \frac{1}{\lambda}\sum_{i=1}^{n} p_i \mathbb{E}[U_i - V_i]\,.$$

We have

$$\sum_{i=1}^{n} p_i \mathbb{E}[U_i] = \lambda^2\,,$$

and

$$\begin{aligned}\sum_{i=1}^{n} p_i \mathbb{E}[V_i] &= \sum_{i=1}^{n} p_i \mathbb{E}[W-1|X_i=1] = \sum_{i=1}^{n} p_i \mathbb{E}[W|X_i=1] - \lambda = \sum_{i=1}^{n} \mathbb{E}[X_i W] - \lambda \\ &= \mathbb{E}[W^2] - \lambda = \mathrm{Var}(W) + \lambda^2 - \lambda\,.\end{aligned}$$

Combining these gives

$$d_{TV}(W,Z) \leq \frac{1}{\lambda}\left(\lambda^2 - \mathrm{Var}(W) - \lambda^2 + \lambda\right) = 1 - \frac{\mathrm{Var}(W)}{\mathbb{E}[W]}\,,$$

as required. □

A similar argument using a monotonicity condition which is (close to) the reverse of that above gives the following result.

Theorem 2.13. *With notation as in Theorem 2.10, if we can construct U_i, V_i and X_i on the same probability space such that $U_i - X_i \leq V_i$ with probability 1 for each $i = 1, \ldots, n$, then*

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

2.7 Examples: Birthdays and random graphs

A birthday problem

Suppose m people each have a birthday on one of n days of the year, where each person's birthday is chosen uniformly at random (independently of everyone else) from the n available days. Let Y_i be the number of these m people born on day i, for $i = 1, \ldots, n$. Let $X_i = I(Y_i = 0)$, and $W = X_1 + \cdots + X_n$ count the number of days on which no-one has a birthday. We consider Poisson approximation for W, beginning by noting that $\lambda = \mathbb{E}[W] = n(1-1/n)^m$.

We construct U_i by taking it equal to W. For each V_i, consider the setting in which the Y_i people born on day i have their birthday reassigned uniformly at random (and independently of all else) to one of the remaining $n-1$ days. With this reassignment, let $1 + V_i$ denote the

number of days on which no-one has a birthday. We have that V_i has the same distribution as $(W-1|X_i=1)$ and that $\mathbb{E}[V_i] = (n-1)(1-1/(n-1))^m$.

Note that with this construction $U_i \geq V_i$ with probability 1, since for each of the $n-1$ days other than day i the reassignment we made in constructing V_i can only reduce the chance that no-one has a birthday on that day. Hence, Theorem 2.12 gives

$$d_{TV}(W,Z) \leq n\left(1-\frac{1}{n}\right)^m - (n-1)\left(1-\frac{1}{n-1}\right)^m .$$

Triangles in an Erdős–Rényi random graph

Let $G = G(n,p)$ be an Erdős–Rényi random graph with n vertices (i.e., nodes) in which each pair of vertices is connected by an edge between them with probability p, independently of all other pairs of vertices. This is a well-known and well-studied model for random networks.

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 each of these random variables X_α has the same distribution, and so for the remainder of this example we may fix $\alpha = (1,2,3)$ without loss of generality.

Let $W = \sum_{\beta\in\Gamma} X_\beta$ count the number of triangles in G. We can construct U_α by letting it be equal to W. We can construct V_α by adding all the possible edges between vertices 1, 2 and 3 into our graph if they are not already present and letting $1+V_\alpha$ denote the number of triangles in our augmented graph.

Since $W - X_\alpha$ counts the number of triangles in our original graph (except possibly for the triangle involving the vertices $(1,2,3)$), and V_α counts the same quantity for a graph which has at least as many edges as our original graph, it is clear that $V_\alpha \geq W - X_\alpha$ with probability 1. We may therefore apply Theorem 2.13.

We note that $\mathbb{E}X_\beta = p^3$ for all $\beta \in \Gamma$, and so $\lambda = \mathbb{E}W = \binom{n}{3}p^3$. It can be shown that

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

and so Theorem 2.13 gives

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

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

2.8 Extensions and generalizations of Poisson approximation results

We conclude this section by giving some brief remarks on generalizations and extensions of the Poisson approximation results using the Stein's method that we have discussed.

1. **Translated Poisson approximation:** One main disadvantage of Poisson approximation (compared to, for example, normal approximation) is that the Poisson distribution has only one parameter we are able to choose. Röllin [37, 38] 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 added flexibility offered by a second parameter allows us to get closer approximations than we are able to with a simple Poisson approximation, and to get reasonable error bounds in settings where a simple Poisson approximation is not useful.

2. **Compound Poisson approximation:** A natural generalization 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 independent and identically distributed positive random variables and N has a Poisson distribution. This allows a much greater range of applications in which reasonable approximations can 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 assume 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 [4], 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 generalization of the Stein equation for Poisson approximation. 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 2.8) are available. This includes when $k\lambda_k \geq (k+1)\lambda_{k+1}$ for all k [4], or when $\sum_j j(j-1)\lambda_j < \frac{1}{2}\sum_j j\lambda_j$ [7]. 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. Extending this range of applications remains an active area of research.

3. **Poisson process approximation:** Many of the techniques we have considered here can be extended to the setting of approximation by a Poisson process (i.e., approximation of the path of a stochastic process by a Poisson process). A stochastic process is a Poisson process if (i) the count of events in a given set has a Poisson distribution, and (ii)

counts in disjoint sets are independent random variables. In this setting, we can define a suitable coupling in terms of the *Palm process*, and we can characterize a Poisson process based on the fact that a Poisson process has the same distribution as its reduced Palm version. This can be used to define a Stein equation for the Poisson process, which then yields explicit error bounds in approximation of a point process by a Poisson process in appropriate metrics. See [15] for a discussion of this approach to Poisson process approximation via Stein's method.

3 Normal approximation by Stein's method

As we have already noted, normal approximation was the first setting in which Stein's method was applied [41], and was also the main focus of Stein's 1986 monograph [42] that amply demonstrated the power and versatility of this technique. We will use this section to outline Stein's method for normal approximation, and in particular we will see that the core ideas are the same as those we have just discussed for Poisson approximation. We will look at how this approach can be applied in a variety of settings using various different ideas, again with a focus on coupling techniques. Much of our discussion here is based upon the book [13] of Chen, Goldstein and Shao, which is an excellent starting point for a much deeper discussion of Stein's method for normal approximation.

We recall that a normal random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ has density function given by

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\},$$

for $x \in \mathbb{R}$. The normal distribution with mean 0 and variance 1 is described as a *standard* normal distribution, and has a density function which we denote by φ. The corresponding distribution function is denoted by $\Phi(x) = \int_{-\infty}^{x} \varphi(y)\, dy = \mathbb{P}(Z \leq x)$, where here and throughout this section we let Z have a standard normal distribution, which we denote by $Z \sim \mathrm{N}(0,1)$.

3.1 The Stein equation and its solution

Our starting point is the relatively simple observation that a random variable Z has a standard normal distribution if and only if

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

for all absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}$ for which these expectations exist. As in the Poisson case above, this characterization gives us a starting point from which we can define a Stein equation and hence bound the distance of a given random variable from a normal: Given a real-valued random variable W (with mean zero and variance one) which we think of as *approximately* normal, we hope that

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

for all f as above.

Suppose we have a given function $h : \mathbb{R} \to \mathbb{R}$, and let $f = f_h$ be the function satisfying $f(0) = 0$ and which solves the following *Stein equation*:

$$h(x) - \mathbb{E}h(Z) = f'(x) - xf(x)\,, \tag{6}$$

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

We will measure the distance between W and Z by using 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. Note that the total variation distance we have used for error bounds in Poisson approximation may be written in this form. There are two principal metrics that are most commonly used in the setting of normal approximation:

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

- If we take $\mathcal{H} = \mathcal{H}_K = \{I_{\{\cdot\leq y\}} : y \in \mathbb{R}\}$ to be the set of indicator functions of semi-infinite intervals, 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 (6), replacing x by W, taking expectations, and then taking the supremum over the class of functions $\mathcal{H}$, we have

$$d_{\mathcal{H}}(W,Z) = \sup_{h\in\mathcal{H}} |\mathbb{E}[f'(W) - Wf(W)]|\,. \tag{7}$$

This equation is the essence of Stein's method for normal approximation.

As in the Poisson case, one big 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 (7) 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 normal.

Before we consider properties of f, we first state formally the characterization of the standard normal distribution from which this work springs; the proof of this result is deferred until Section 3.4.

Lemma 3.1. *If X has a standard normal distribution, then*

$$\mathbb{E}f'(X) = \mathbb{E}[Xf(X)] \tag{8}$$

for all absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}|f'(Z)| < \infty$. Conversely, if (8) holds for all bounded, continuous and piecewise continuously differentiable functions f with $\mathbb{E}|gf'(Z)| < \infty$, then X has a standard normal distribution.

The solution of the Stein equation (6) is given by the following lemma.

Lemma 3.2. *Let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function with $\mathbb{E}|h(Z)| < \infty$. The unique bounded solution to the Stein equation (6) is given by*

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

There are many bounds available on the solution f to the Stein equation; see Section 2.2 of [13] for a wide selection. We state here only one such bound, whose proof we will sketch in Section 3.4.

Lemma 3.3. *Let f be the function defined in Lemma 3.2. If h is absolutely continuous then*

$$\|f''\|_\infty \leq 2\|h'\|_\infty\,. \tag{9}$$

3.2 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 whose proof dates back to the pioneering work of Stein [41].

Theorem 3.4. *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 (7), we need to bound

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

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,

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

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 (9), 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 normal setting to prove approximation results in 'smooth metrics' (such as the Wasserstein distance) where the functions $h \in \mathcal{H}$ satisfy some differentiability or 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 [14]), whose proof we give in Section 3.4.

Lemma 3.5. *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 $d_K(W, Z) \leq 2\sqrt{\delta}$.*

Unfortunately results such as these typically do not give the best possible error bounds and rates of convergence in many examples and applications, including in the case of sums of independent random variables as we consider here.

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

3.3 Couplings and other approaches

In this section we will briefly describe a selection of approaches that have been successfully applied in conjunction with Stein's method for normal approximation to yield explicit error bounds in normal approximation in settings more exotic than sums of independent random variables. While we will not discuss applications in detail, or provide proofs for many of the results we state, we will give references indicating where many further details can be found.

3.3.1 Local dependence

The argument of Theorem 3.4 may be extended to the case where $X_1, \ldots, X_n$ 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 [14] and assume the following:

$$\begin{aligned} &\text{For each } i \in \mathcal{J}, \text{ there exist } A_i \subseteq B_i \subseteq \mathcal{J} \text{ such that } X_i \text{ is} \\ &\text{independent of } X_{A_i^c} \text{ and } X_{A_i} \text{ is independent of } X_{B_i^c}. \end{aligned} \tag{10}$$

Under this assumption, Chen and Shao [14] give the following bound.

Theorem 3.6. *Let $\{X_i : i \in \mathcal{J}\}$ be such that $\mathbb{E}X_i = 0$ for each i and (10) 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*

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

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

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 3.6 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 3.4.

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 (10) is satisfied with the choices $A_i = \{j \in \mathcal{V} : d(i,j) \leq 2\}$ and $B_i = \cup_{j\in A_i} A_j = \{j \in \mathcal{V} : d(i,j) \leq 4\}$.

3.3.2 Exchangeable pairs

Historically, one of the earliest approaches to Stein's method for normal 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. Recall that (W, W') are said to be exchangeable if the bivariate distributions of (W, W') and (W', W) are identical. In addition to exchangeability, we will assume the following 'linear regression' condition:

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

for some $\lambda \in (0,1)$. See [42] for an extensive discussion of this approach. Subsequent work allows for a relaxation of this condition, for example in permitting a remainder term to appear.

Under this condition, the following theorem may be established. See Section 4.5 of [13] for a proof.

Theorem 3.7. *Let* W *satisfy* $\mathbb{E}W = 0$ *and* $\mathrm{Var}(W) = 1$. *Let* (W, W') *be exchangeable and such that* $\mathbb{E}[W'|W] = (1-\lambda)W$ *for some* $\lambda \in (0,1)$. *Then*

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

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

As an illustration of the construction of an exchangeable pair satisfying the conditions of Theorem 3.7, suppose $W = X_1 + \cdots + X_n$ is a sum of IID random variables, each with mean zero, and with $\mathrm{Var}(W) = 1$. 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 and it can be shown that $\mathbb{E}[W'|W] = \left(1 - \frac{1}{n}\right) W$.

Exchangeable pairs satisfying the conditions of Theorem 3.7 can also be constructed as successive states of a reversible Markov chain in stationarity.

3.3.3 The zero-biased coupling

The zero-biased transformation of a distribution was first introduced by Goldstein and Reinert [24]:

Definition 3.8. *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*

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

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

It was shown in [24] 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

$$p^z(x) = \sigma^{-2} \mathbb{E}[W 1_{\{W > x\}}] \,.$$

The definition of this zero-biasing transformation is motivated by the fact that the mean-zero normal distribution is the unique fixed point of this transformation: Stein's characterization of the normal distribution can be equivalently stated by saying that Z has a mean-zero normal distribution if and only if Z and Z^z have the same distribution. We might therefore expect that we can use a measure of distance between W and W^z to quantify how far W is from normal. The following result (proved by Goldstein and Reinert [24]) does this.

Theorem 3.9. *Let $\mathbb{E}W = 0$ and $\mathrm{Var}(W) = 1$. 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 (6) to write

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

where the final inequality uses (9). The result follows by taking the supremum over $h \in \mathcal{H}_W$. □

This result can be applied by constructing W and W^z on the same probability space, and thus bounding the expected difference between them. There has been much work on how this can be done in various settings: see [13] for several approaches and applications.

3.4 Selected proofs

Proof of Lemma 3.1

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

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

Using Fubini's theorem, we then have

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

Sufficiency: Fix $z \in \mathbb{R}$. Let f be the function defined by Lemma 3.2, with the choice $h(x) = I_{\{x \leq z\}}$. This function f is continuous and piecewise continuously differentiable, and we also know 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.

Sketch proof of Lemma 3.3

Differentiating the Stein equation (6), we have

$$f''(x) = (1 + x^2)f(x) + x[h(x) - \mathbb{E}h(Z)] + h'(x)\,. \tag{11}$$

We can show that

$$h(x) - \mathbb{E}h(Z) = \int_{-\infty}^{x} h'(z)\Phi(z)\,dz - \int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,, \tag{12}$$

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 (12) to show that

$$-\varphi(x)f(x) = [1 - \Phi(x)]\int_{-\infty}^{x} h'(z)\Phi(z)\,dz + \Phi(x)\int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,. \tag{13}$$

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:

$$\theta''(x) = x + (1 + x^2)\theta(x)\,, \tag{14}$$

$$\theta''(-x) = \frac{1 + x^2}{\varphi(x)} - \theta''(x)\,. \tag{15}$$

Combining (11)–(15) we can show that

$$f''(x) = h(x) - \theta''(-x)\int_{-\infty}^{x} h'(z)\Phi(z)\,dz - \theta''(x)\int_{x}^{\infty} h'(z)[1 - \Phi(z)]\,dz\,. \tag{16}$$

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 (16).

Proof of Lemma 3.5

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 $x \leq z$, be 0 for $x \geq z + \alpha$, 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

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

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

4 Additional topics

In this section we give brief, high-level overviews of some further topics that we could have covered in a longer version of this course and some further topics of current research interest related to the material we have discussed.

4.1 Other distributional approximations

There are numerous other topics we could have also considered in this course, even in the setting of univariate approximation, including

- binomial approximation [19],
- geometric approximation [33],
- negative binomial approximation [40],
- exponential approximation [34],
- beta approximation [25],
- chi-square approximation [23],
- Laplace approximation [32], and
- variance-gamma approximation [22],

among many others. Most of these topics have been considered by a number of researchers; the references given here are intended to represent an entry point into the literature, and not an exhaustive bibliography on each topic.

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 [9] have considered the general setting of approximation by the equilibrium distribution of a birth–death process,
- Eichelsbacher and Reinert [18] studied approximation by discrete Gibbs measures, and
- Arras and Houdré [2] have worked on approximation by infinitely divisible distributions with finite first moment, where we recall that a random variable Z is infinitely divisible if, for each $n \geq 1$, there are IID random variables $Z_{1,n}, \ldots, Z_{n,n}$ such that $Z_{1,n} + \cdots + Z_{n,n}$ has the same disribution as Z.

Beyond the univariate setting, see, for example [35] and [15] for starting points in the large amount of literature available on multivariate normal approximation and Poisson process approximation, respectively.

4.2 The Malliavin-Stein method

We note relatively recent work combining Stein's method with the tools of Malliavin calculus, which remains an active area of research. The characterization of the normal distribution which we use as the starting point of Stein's method can be thought of as an integration-by-parts formula with respect to the normal density function (writing an integral of a given function as an integral of the derivative of that function multiplied by another term). There is also an integration-by-parts formula at the heart of the Malliavin calculus, involving the Malliavin derivative and the generator of the Ornstein–Uhlenbeck semigroup. This formula may be combined with the techniques of Stein's method for normal approximation.

One important result in this area is the well-known 'fourth moment theorem' which states that distributions of a (suitably standardised) sequence of elements of a Wiener chaos converge to the normal distribution if and only if the corresponding sequence of fourth moments converge to 3, the fourth moment of a standard normal distribution. This is a considerable simplification of the usual method of moments, which needs all moments to converge in order to obtain this convergence of distributions. See the book of Nourdin and Peccati [31] for a starting point in the study of these techniques.

4.3 Applications in data analysis and machine learning

It is quite often relatively straightforward to write down the kind of characterization you need as a starting point for Stein's method. For example, for a continuous random variable Z with differentiable density function p which is positive on the whole real line, it can easily be checked that

$$\mathbb{E}\left[f'(Z) - \frac{p'(Z)}{p(Z)} f(Z)\right] = 0\,,$$

for all functions f for which this expectation is defined. It is typically only when we want to solve and apply the corresponding Stein equation that we have to focus our attention on a particular target distribution Z rather than a large class of such distributions.

With access to a characterization of our random variable Z (i.e., an operator $\mathcal{A}$ such that $\mathbb{E}[\mathcal{A}f(Z)] = 0$ for all functions f for which the expectation is defined), we can define the *Stein discrepancy* between Z and another random variable W by

$$\sup_{f \in \mathcal{F}} |\mathbb{E}[\mathcal{A}f(W)]|\,,$$

which will depend on the particular class of functions $\mathcal{F}$ that we choose. This is difficult to compute in practice; much of what we have looked at in these notes are ways to find reasonable upper bounds for these kinds of quantities. In 2015, Gorham and Mackey [26] introduced a new, efficiently computable Stein discrepancy between a given data sample and target expectations, and used this as a tool to compare different sampling techniques. For example, in the setting of Markov chain Monte Carlo estimators, we may compare techniques which trade off bias against a reduced variance. Subsequent developments in this direction have included

the combination of Stein discrepancies with the theory of reproducing kernel Hilbert spaces to give a tool that can be used in goodness-of-fit tests and model evaluation, with the benefit that it does not depend on the normalizing constants of the unknown distributions [29]. Since these normalizing constants can be very difficult to calculate, this is very useful in practical applications.

Another current research area at the interface of Stein's method and machine learning is the development of the Stein variational gradient descent algorithm, which can be used to select a representative sample of points from an unknown probability distribution. It does this by exploiting an interesting connection between the Stein discrepancy and the Kullback–Leibler divergence, which gives a measure of the distance between two distributions. This algorithm reduces the Kullback–Leibler divergence between our sample and the target distribution using a gradient-based approach [30].

There are many further interesting recent developments related to the application of ideas from Stein's method in data analysis and machine learning; this is a very active area of current research.

5 Further reading

For those who want to go beyond the introduction to probability approximations in general, and Stein's method in particular, that we have covered in this mini-course, the following books and papers give good starting points for the vast literature in this area:

- An introduction limit theorems in probability at an undergraduate level is given by Lesigne [28];
- At a more advanced level, Čekanavičius [10] gives a survey of numerous methods available for probability approximations;
- The paper by Ross [39] gives a survey of Stein's method that makes an excellent starting point for the study of this technique;
- Barbour, Holst and Janson [6] is dedicated to Poisson approximation by Stein's method, with an emphasis on coupling-based approaches;
- An extensive treatment of Stein's method for normal approximation is given by Chen, Goldstein and Shao [13]; and
- Nourdin and Peccati [31] give a book-length treatment of the use of Malliavin calculus in conjunction with Stein's method for normal approximation.

References

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

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

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

[4] 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.

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

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

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

[8] G. Bresler and D. Nagaraj (2019). Stein's method for stationary distributions of Markov chains and application to Ising models. *Ann. Appl. Probab.* 29(5): 3230–3265.

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

[10] V. Čekanavičius (2016). *Approximation Methods in Probability Theory*. Springer, Switzerland.

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

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

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

[14] 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.

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

[16] F. Daly (2010). Stein's method for compound geometric approximation. *J. Appl. Probab.*

47(1): 146–156.

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

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

[19] W. Ehm (1991). Binomial approximation to the Poisson binomial distribution. *Statist. Prob. Lett.* 11: 7–16.

[20] T. Erhardsson (1999). Compound Poisson approximation for Markov chains using Stein's method. *Ann. Probab.* 27: 565–596.

[21] T. Erhardsson (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. *Ann. Appl. Probab.* 10: 573–591.

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

[23] 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.

[24] 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.

[25] 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.

[26] J. Gorham and L. Mackey (2015). Measuring sample quality with Stein's method. *NIPS 2015*: 226–234.

[27] D. Griffeath (1975). A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitstheorie und ver. Gebiete* 31: 95–106.

[28] E. Lesigne (2005). *An Introduction to Limit Theorems in Probability*. American Mathematical Society, Providence, Rhode Island.

[29] Q. Liu, J. D. Lee and M. I. Jordan (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. Preprint, available at `https://arxiv.org/abs/1602.03253`.

[30] Q. Liu and D. Wang (2016). Stein variational gradient descent: a general purpose Bayesian inference algorithm. Preprint, available at `https://arxiv.org/abs/1608.04471`.

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

[32] J. Pike and H. Ren (2014). Stein's method and the Laplace distribution. *ALEA Lat. Am.*

J. Probab. Math. Stat. 11: 571–587.

[33] E. Peköz (1996). Stein's method for geometric approximation. *J. Appl. Probab.* 33: 707–713.

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

[35] 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.

[36] G. Reinert and N. Ross (2019). Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *Ann. Appl. Probab.* 29(5): 3201–3229.

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

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

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

[40] 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.

[41] 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.

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