Probability Surveys
Vol. 8 (2011) 210–293
ISSN: 1549-5787
DOI: 10.1214/11-PS182

Fundamentals of Stein's method[*]

Nathan Ross

University of California
367 Evans Hall #3860
Berkeley, CA 94720-3860
e-mail: ross@stat.berkeley.edu

Abstract: This survey article discusses the main concepts and techniques of Stein's method for distributional approximation by the normal, Poisson, exponential, and geometric distributions, and also its relation to concentration of measure inequalities. The material is presented at a level accessible to beginning graduate students studying probability with the main emphasis on the themes that are common to these topics and also to much of the Stein's method literature.

AMS 2000 subject classifications: Primary 60F05, 60C05; secondary 05C80.
Keywords and phrases: Stein's method, distributional approximation, concentration of measure inequalities, size-bias distribution, exchangeable pairs.

Received September 2011.

Contents

1 Introduction . . . 211
1.1 Background and notation . . . 213
1.2 Probability metrics . . . 214
2 Normal approximation . . . 215
2.1 The general setup . . . 217
3 Bounding the error . . . 218
3.1 Sum of independent random variables . . . 219
3.2 Dependency neighborhoods . . . 220
3.2.1 Application: Triangles in Erdős-Rényi random graphs . . . 224
3.3 Exchangeable pairs . . . 224
3.3.1 Application: Anti-voter model . . . 228
3.4 Size-bias coupling . . . 231
3.4.1 Coupling construction . . . 232
3.4.2 Applications . . . 234
3.5 Zero-bias coupling . . . 236
3.5.1 Coupling construction . . . 237
3.5.2 Lindeberg-Feller condition . . . 238
3.6 Normal approximation in the Kolmogorov metric . . . 241
3.6.1 Zero-bias transformation . . . 242

[*]This is an original survey paper.

3.6.2 Exchangeable pairs . . . 243
4 Poisson approximation . . . 245
4.1 Law of small numbers . . . 247
4.2 Dependency neighborhoods . . . 248
4.2.1 Application: Head runs . . . 249
4.3 Size-bias coupling . . . 252
4.3.1 Increasing size-bias couplings . . . 255
4.3.2 Application: Subgraph counts . . . 255
4.3.3 Implicit coupling . . . 258
4.3.4 Decreasing size-bias couplings . . . 259
4.4 Exchangeable pairs . . . 262
5 Exponential approximation . . . 265
5.1 Equilibrium coupling . . . 267
5.2 Application: Geometric sums . . . 268
5.3 Application: Critical branching process . . . 270
6 Geometric approximation . . . 275
6.1 Main theorem . . . 275
6.2 Discrete equilibrium coupling . . . 276
6.3 Application: Uniform attachment graph model . . . 277
7 Concentration of measure inequalities . . . 278
7.1 Concentration of measure using exchangeable pairs . . . 280
7.2 Application: Magnetization in the Curie-Weiss model . . . 284
7.3 Concentration of measure using size-bias couplings . . . 287
Acknowledgments . . . 289
References . . . 289

1. Introduction

The fundamental example of the type of result we address in this article is the following version of the classical Berry-Esseen bound for the central limit theorem.

Theorem 1.1. *[56] Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^3 < \infty$, $\mathbb{E}[X_1] = 0$, and $\mathrm{Var}(X_1) = 1$. If Φ denotes the c.d.f. of a standard normal distribution and $W_n = \sum_{i=1}^n X_i/\sqrt{n}$, then*

$$|\mathbb{P}(W_n \leq x) - \Phi(x)| \leq .4785\frac{\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

The theorem quantifies the error in the central limit theorem and has many related embellishments such as assuming independent, but not identically distributed variables, or allowing a specified dependence structure. The proofs of such results typically rely on characteristic function (Fourier) analysis, whereby showing convergence is significantly easier than obtaining error bounds.

More generally, a central theme of probability theory is proving distributional limit theorems, and for the purpose of approximation it is of interest to quantify the rate of convergence in such results. However, many of the methods commonly employed to show distributional convergence (e.g. Fourier analyisis and method of moments) only possibly yield an error rate after serious added effort. Stein's method is a technique that can quantify the error in the approximation of one distribution by another in a variety of metrics. Note that the implications of such an approximation can fall outside of the discussion above which relates only to convergence.

Stein's method was initially conceived by Charles Stein in the seminal paper [54] to provide errors in the approximation by the normal distribution of the distribution of the sum of dependent random variables of a certain structure. However, the ideas presented in [54] are sufficiently abstract and powerful to be able to work well beyond that intended purpose, applying to approximation of more general random variables by distributions other than the normal (such as the Poisson, exponential, etc).

Broadly speaking, Stein's method has two components: the first is a framework to convert the problem of bounding the error in the approximation of one distribution of interest by another, well understood distribution (e.g. the normal) into a problem of bounding the expectation of a certain functional of the random variable of interest (see (2.5) for the normal distribution and (4.4) for the Poisson). The second component of Stein's method is a collection of techniques to bound the expectation appearing in the first component; Stein appropriately refers to this step as "auxiliary randomization." With this in mind, it is no surprise that Stein's monograph [55], which reformulates the method in a more coherent form than [54], is titled "Approximate Computation of Expectations."

There are now hundreds of papers expanding and applying this basic framework above. For the first component, converting to a problem of bounding a certain expectation involving the distribution of interest has been achieved for many well-known distributions. Moreover, canonical methods (which are not guaranteed to be fruitful or easy to implement) have been established for achieving this conversion for new distributions [22, 45].

For the second component, there is now an array of coupling techniques available to bound these functionals for various distributions. Moreover, these coupling techniques can be used in other types of problems which can be distilled into bounding expectations of a function of a distribution of interest. Two examples of the types of problems where this program has succeeded are concentration of measure inequalities [18, 28, 29] (using the well known Proposition 7.1 below), and local limit theorems [50]. We cover the former example in this article.

The purpose of this document is to attempt to elucidate the workings of these two components at a basic level, in order to help make Stein's method more accessible to the uninitiated. There are numerous other introductions to Stein's method which this document draws from, mainly [23, 24] for Normal approximation, [12, 20] for Poisson approximation, and an amalgamation of related topics in the collections [11, 26]. Most of these references focus on one distribution or variation of Stein's method in order to achieve depth, so there are themes and ideas that appear throughout the method which can be difficult to glean from these references. We hope to capture these fundamental concepts in uniform language to give easier entrance to the vast literature on Stein's method and applications. A similar undertaking but with smaller scope can be found in Chapter 2 of [51], which also serves as a nice introduction to the basics of Stein's method.

Of course the purpose of Stein's method is to prove approximation results, so we illustrate concepts in examples and applications, many of which are combinatorial in nature. In order to facilitate exposition, we typically work out examples and applications only in the most straightforward way and provide pointers to the literature where variations of the arguments produce more thorough results.

The layout of this document is as follows. In Section 2, we discuss the basic framework of the first component above in the context of Stein's method for normal approximation, since this setting is the most studied and contains many of the concepts we need later. In Section 3 we discuss the commonly employed couplings used in normal approximation to achieve the second component above. We follow the paradigm of these two sections in discussing Stein's method for Poisson approximation in Section 4, exponential approximation in Section 5, and geometric approximation in Section 6. In the final Section 7 we discuss how to use some of the coupling constructions of Section 3 to prove concentration of measure inequalities.

We conclude this section with a discussion of necessary background and notation.

1.1. Background and notation

This is a document based on a graduate course given at U.C. Berkeley in the Spring semester of 2011 and is aimed at an audience having seen probability theory at the level of [35]. That is, we do not rely heavily on measure theoretic concepts, but exposure at a heuristic level to concepts such as sigma-fields is useful. Also, basic Markov chain theory concepts such as reversibility are assumed along with the notion of coupling random variables which is used frequently in what follows.

Many of our applications concern various statistics of Erdős-Rényi random graphs. We say $G = G(n, p)$ is an Erdős-Rényi random graph on n vertices with edge probability p if, for each of the $\binom{n}{2}$ pairs of vertices, there is an edge connecting the vertices with probability p (and no edge connecting them with probability $1 - p$), independent of all other connections between other pairs of vertices. These objects are a simple and classical model of networks that are well studied; see [14, 37] for book length treatments.

For a set A, we write $\mathbb{I}[\cdot \in A]$ to denote the function which is one on A and 0 otherwise. We write $g(n) \asymp f(n)$ if $g(n)/f(n)$ tends to a positive constant as $n \to \infty$, and $g(n) = \mathrm{O}(f(n))$ if $g(n)/f(n)$ is bounded as $n \to \infty$. For a function f with domain D, we write $\|f\| = \sup_{x \in D} |f(x)|$.

Since Stein's method is mainly concerned with bounding the distance between probability distributions in a given metric, we now discuss the metrics that we use.

1.2. Probability metrics

For two probability measures μ and ν, the probability metrics we use have the form

$$d_{\mathcal{H}}(\mu,\nu)=\sup_{h\in\mathcal{H}}\left|\int h(x)d\mu(x)-\int h(x)d\nu(x)\right|, \tag{1.1}$$

where $\mathcal{H}$ is some family of "test" functions. For random variables X and Y with respective laws μ and ν, we abuse notation and write $d_{\mathcal{H}}(X,Y)$ in place of $d_{\mathcal{H}}(\mu,\nu)$.[1]

We now detail examples of metrics of this form along with some useful properties and relations.

1. By taking $\mathcal{H}=\{\mathbb{I}[\cdot\leq x]:x\in\mathbb{R}\}$ in (1.1), we obtain the *Kolmogorov* metric, which we denote d_{K}. The Kolmogorov metric is the maximum distance between distribution functions, so a sequence of distributions converging to a fixed distribution in this metric implies weak convergence (although the converse is not true since weak convergence only implies pointwise convergence of distribution functions at continuity points of the limiting distribution function).
2. By taking $\mathcal{H}=\{h:\mathbb{R}\to\mathbb{R}:|h(x)-h(y)|\leq|x-y|\}$ in (1.1), we obtain the *Wasserstein* metric, which we denote d_{W}. The Wasserstein metric is a common metric occurring in many contexts and is the main metric we use for approximation by continuous distributions.
3. By taking $\mathcal{H}=\{\mathbb{I}[\cdot\in A]:A\in\mathrm{Borel}(\mathbb{R})\}$ in (1.1), we obtain the *total variation* metric, which we denote d_{TV}. We use the total variation metric for approximation by discrete distributions.

Proposition 1.2. *Retaining the notation for the metrics above, we have the following.*

1. *For random variables W and Z, $d_{\mathrm{K}}(W,Z)\leq d_{\mathrm{TV}}(W,Z)$.*
2. *If the random variable Z has Lebesgue density bounded by C, then for any random variable W,*

$$d_{\mathrm{K}}(W,Z)\leq\sqrt{2C\,d_{\mathrm{W}}(W,Z)}.$$

3. *For W and Z random variables taking values in a discrete space Ω,*

$$d_{\mathrm{TV}}(W,Z)=\frac{1}{2}\sum_{\omega\in\Omega}|\mathbb{P}(W=\omega)-\mathbb{P}(Z=\omega)|.$$

Proof. The first item follows from the fact that the supremum on the right side of the inequality is over a larger set, and the third item is left as an exercise. For the second item, consider the functions $h_x(w)=\mathbb{I}[w\leq x]$, and the 'smoothed'

[1] In some contexts this abuse could cause some confusion, but our use of these metrics is largely outside of such issues.

$h_{x,\varepsilon}(w)$ defined to be one for $w \le x$, zero for $w > x + \varepsilon$, and linear between. Then we have

$$\begin{aligned}\mathbb{E}h_x(W) - \mathbb{E}h_x(Z) &= \mathbb{E}h_x(W) - \mathbb{E}h_{x,\varepsilon}(Z) + \mathbb{E}h_{x,\varepsilon}(Z) - \mathbb{E}h_x(Z)\\ &\le \mathbb{E}h_{x,\varepsilon}(W) - \mathbb{E}h_{x,\varepsilon}(Z) + C\varepsilon/2\\ &\le d_{\mathrm{W}}(W,Z)/\varepsilon + C\varepsilon/2.\end{aligned}$$

Taking $\varepsilon = \sqrt{2\, d_{\mathrm{W}}(W,Z)/C}$ shows half of the desired inequality and a similar argument yields the other half. □

Due to its importance in our framework, we reiterate the implication of Item 2 of the proposition that a bound on the Wasserstein metric between a given distribution and the normal or exponential distribution immediately yields a bound on the Kolmogorov metric.

2. Normal approximation

The main idea behind Stein's method of distributional approximation is to replace the characteristic function typically used to show distributional convergence with a *characterizing operator.*

Lemma 2.1 (Stein's Lemma)**.** *Define the functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(x) = f'(x) - xf(x).$$

1. *If Z has the standard normal distribution, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all absolutely continuous f with $\mathbb{E}|f'(Z)| < \infty$.*
2. *If for some random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all absolutely continuous functions f with $\|f'\| < \infty$, then W has the standard normal distribution.*

The operator $\mathcal{A}$ is referred to as a characterizing operator of the standard normal distribution.

Before proving Lemma 2.1, we record the following lemma and then observe a consequence.

Lemma 2.2. *If $\Phi(x)$ is the c.d.f. of the standard normal distribution, then the unique bounded solution f_x of the differential equation*

$$f_x'(w) - wf_x(w) = \mathbb{I}[w \le x] - \Phi(x) \tag{2.1}$$

is given by

$$\begin{aligned}f_x(w) &= e^{w^2/2}\int_w^\infty e^{-t^2/2}\left(\Phi(x) - \mathbb{I}[t \le x]\right)dt\\ &= -e^{w^2/2}\int_{-\infty}^w e^{-t^2/2}\left(\Phi(x) - \mathbb{I}[t \le x]\right)dt.\end{aligned}$$

Lemmas 2.1 and 2.2 are at the heart of Stein's method; observe the following corollary.

Corollary 2.3. *If f_x is as defined in Lemma 2.2, then for any random variable W,*

$$|\mathbb{P}(W \leq x) - \Phi(x)| = |\mathbb{E}[f_x'(W) - Wf_x(W)]|. \tag{2.2}$$

Although Corollary 2.3 follows directly from Lemma 2.2, it is important to note that Lemma 2.1 suggests that (2.2) may be a fruitful equality. That is, the left hand side of (2.2) is zero for all $x \in \mathbb{R}$ if and only if W has the standard normal distribution. Lemma 2.1 indicates that the right hand side of (2.2) also has this property.

Proof of Lemma 2.2. The method of integrating factors shows that

$$\frac{d}{dw}\left(e^{-w^2/2}f_x(w)\right) = e^{-w^2/2}\left(\mathbb{I}[w \leq x] - \Phi(x)\right),$$

which after integrating and considering the homogeneous solution implies that

$$f_x(w) = e^{w^2/2}\int_w^\infty e^{-t^2/2}\left(\Phi(x) - \mathbb{I}[t \leq x]\right)dt + Ce^{w^2/2} \tag{2.3}$$

is the general solution of (2.1) for any constant C. To show that (2.3) is bounded for $C = 0$ (and then clearly unbounded for other values of C) we use

$$1 - \Phi(w) \leq \min\left\{\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right\}e^{-w^2/2}, \quad w > 0,$$

which follows by considering derivatives. From this point we use the representation

$$f_x(w) = \begin{cases} \sqrt{2\pi}e^{w^2/2}\Phi(w)(1-\Phi(x)), & w \leq x \\ \sqrt{2\pi}e^{w^2/2}\Phi(x)(1-\Phi(w)), & w > x \end{cases}$$

to obtain that $\|f_x\| \leq \sqrt{\frac{\pi}{2}}$. □

Proof of Lemma 2.1. We first prove Item 1 of the lemma. Let Z be a standard normal random variable and let f be absolutely continuous such that $\mathbb{E}|f'(Z)| < \infty$. Then we have the following formal calculation (justified by Fubini's Theorem) which is essentially integration by parts.

$$\begin{aligned}
\mathbb{E}f'(Z) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-t^2/2}f'(t)dt \\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty f'(t)\int_t^\infty we^{-w^2/2}dwdt + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^0 f'(t)\int_{-\infty}^t we^{-w^2/2}dwdt \\
&= \frac{1}{\sqrt{2\pi}}\int_0^\infty we^{-w^2/2}\left[\int_0^w f'(t)dt\right]dw + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^0 we^{-w^2/2}\left[\int_w^0 f'(t)dt\right]dw \\
&= \mathbb{E}[Zf(Z)].
\end{aligned}$$

For the second item of the Lemma, assume that W is a random variable such that $\mathbb{E}[f'(W) - Wf(W)] = 0$ for all absolutely continuous f such that $\|f'\| < \infty$. The function f_x satisfying (2.1) is such a function, so that for all $x \in \mathbb{R}$,

$$0 = \mathbb{E}[f_x'(W) - Wf_x(W)] = \mathbb{P}(W \le x) - \Phi(x),$$

which implies that W has a standard normal distribution. □

Our strategy for bounding the maximum distance between the distribution function of a random variable W and that of the standard normal is now fairly obvious: we want to bound $\mathbb{E}[f_x(W) - Wf_x(W)]$ for f_x solving (2.1). This setup can work, but it turns out that it is easier to work in the Wasserstein metric. Since the critical property of the Kolmogorov metric that we use in the discussion above is the representation (1.1), which the Wasserstein metric shares, extending in this direction comes without great effort.[2]

2.1. The general setup

For two random variables X and Y and some family of functions $\mathcal{H}$, recall the metric

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |\mathbb{E}h(X) - \mathbb{E}h(Y)|, \tag{2.4}$$

and note that such a metric only depends on the marginal laws of X and Y. For $h \in \mathcal{H}$, let f_h solve

$$f_h'(w) - wf_h(w) = h(w) - \Phi(h)$$

where $\Phi(h)$ is the expectation of h with respect to a standard normal distribution. We have the following result which easily follows from the discussion above.

Proposition 2.4. *If W is a random variable and Z has the standard normal distribution, then*

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}[f_h'(W) - Wf_h(W)]|. \tag{2.5}$$

The main idea at this point is to bound the right side of (2.5) by using the structure of W and properties of the solutions f_h. The latter issue is handled by the following lemma.

Lemma 2.5. *Let f_h be the solution of the differential equation*

$$f_h'(w) - wf_h(w) = h(w) - \Phi(h) \tag{2.6}$$

[2] It is important to note, however, that this change typically comes at the cost of sharp rates in the Kolmogorov metric which can be difficult to obtain, even in reasonable problems.

which is given by

$$\begin{aligned} f_h(w) &= e^{w^2/2}\int_w^\infty e^{-t^2/2}\left(\Phi(h)-h(t)\right)dt \\ &= -e^{w^2/2}\int_{-\infty}^w e^{-t^2/2}\left(\Phi(h)-h(t)\right)dt. \end{aligned}$$

1. *If h is bounded, then*

$$\|f_h\| \leq \sqrt{\frac{\pi}{2}}\|h(\cdot)-\Phi(h)\|, \quad \textit{and} \quad \|f_h'\| \leq 2\|h(\cdot)-\Phi(h)\|.$$

2. *If h is absolutely continuous, then*

$$\|f_h\| \leq 2\|h'\|, \qquad \|f_h'\| \leq \sqrt{\frac{2}{\pi}}\|h'\|, \quad \textit{and} \quad \|f_h''\| \leq 2\|h'\|.$$

The proof of Lemma 2.5 is similar to but more technical than the proof of Lemma 2.2. We refer to [24] (Lemma 2.4) for the proof.

3. Bounding the error

We focus mainly on the Wasserstein metric when approximating by continuous distributions. This is not a terrible concession as firstly the Wasserstein metric is a commonly used metric, and also by Proposition 1.2, for Z a standard normal random variable and W any random variable we have

$$d_{\mathrm{K}}(W,Z) \leq (2/\pi)^{1/4}\sqrt{d_{\mathrm{W}}(W,Z)},$$

where d_{K} is the maximum difference between distribution functions (the Kolmogorov metric); d_{K} is an intuitive and standard metric to work with.

The reason for using the Wasserstein metric is that it has the form (2.4) for $\mathcal{H}$ the set of functions with Lipschitz constant equal to one. In particular, if h is a test function for the Wasserstein metric, then $\|h'\| \leq 1$ so that we know the solution f_h of equation (2.6) is bounded with two bounded derivatives by Item 2 of Proposition 2.5. Contrast this to the set of test functions for the Kolmogorov metric where the solution f_h of equation (2.6) is bounded with one bounded derivative (by Item 1 of Proposition 2.5) but is not twice differentiable.

To summarize our progress to this point, we state the following result which is a corollary of Proposition 2.4 and Lemma 2.5. The theorem is the kernel of Stein's method for normal approximation.

Theorem 3.1. *If W is a random variable and Z has the standard normal distribution, and we define the family of functions $\mathcal{F} = \{f : \|f\|, \|f''\| \leq 2, \|f'\| \leq \sqrt{2/\pi}\}$, then*

$$d_{\mathrm{W}}(W,Z) \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - Wf(W)]|. \tag{3.1}$$

In the remainder of this section, we discuss methods which use the structure of W to bound $|\mathbb{E}[f'(W) - Wf(W)]|$. We proceed by identifying general structures that are amenable to this task (for other, more general structures see [49]), but first we illustrate the type of result we are looking for in the following standard example.

3.1. Sum of independent random variables

We have the following result which follows from Theorem 3.1 and Lemma 3.4 below.

Theorem 3.2. *Let $X_1, \ldots, X_n$ be independent mean zero random variables such that $\mathbb{E}|X_i|^4 < \infty$ and $\mathbb{E}X_i^2 = 1$. If $W = (\sum_{i=1}^n X_i)/\sqrt{n}$ and Z has the standard normal distribution, then*

$$d_{\mathrm{W}}(W, Z) \leq \frac{1}{n^{3/2}} \sum_{i=1}^n \mathbb{E}|X_i|^3 + \frac{\sqrt{2}}{\sqrt{\pi} n} \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^4]}.$$

Before the proof we remark that if the X_i of the theorem also have common distribution, then the rate of convergence is order $n^{-1/2}$, which is the best possible. It is also useful to compare this result to Theorem 1.1 which is in a different metric (neither result is recoverable from the other in full strength) and only assumes third moments. A small modification in the argument below yields a similar theorem assuming only third moments (see Lecture 3 of [23]), but the structure of proof for the theorem as stated is one that we copy in what follows.

In order to prepare for arguments to come, we break the proof into a series of lemmas. Since our strategy is to apply Theorem 3.1 by estimating the right side of (3.1) for bounded f with bounded first and second derivative, the first lemma shows an expansion of the right side of (3.1) using the structure of W as defined in Theorem 3.2.

Lemma 3.3. *In the notation of Theorem 3.2, if $W_i = (\sum_{j \neq i} X_i)/\sqrt{n}$ then*

$$\mathbb{E}[Wf(W)] = \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W)\right)\right] \tag{3.2}$$

$$+ \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W)\right]. \tag{3.3}$$

Proof. After noting that the negative of (3.3) is contained in (3.2) and removing these terms from consideration, the lemma is equivalent to

$$\mathbb{E}[Wf(W)] = \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(X_i f(W) - X_i f(W_i)\right)\right]. \tag{3.4}$$

Equation (3.4) follows easily from the fact that W_i is independent of X_i so that $\mathbb{E}[X_i f(W_i)] = 0$. □

The theorem follows after we show that (3.2) is small and that (3.3) is close to $\mathbb{E}f'(W)$; we see this strategy appear frequently in what follows.

Lemma 3.4. *If f is a bounded function with bounded first and second derivative, then in the notation of Theorem 3.2,*

$$|\mathbb{E}[f'(W) - Wf(W)]| \leq \frac{\|f''\|}{2n^{3/2}} \sum_{i=1}^{n} \mathbb{E}|X_i|^3 + \frac{\|f'\|}{n} \sqrt{\sum_{i=1}^{n} \mathbb{E}[X_i^4]}. \tag{3.5}$$

Proof. Using the notation and results of Lemma 3.3, we obtain

$$\begin{aligned}
&|\mathbb{E}[f'(W) - Wf(W)]| \\
&\quad \leq \left| \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W) \right) \right] \right| &(3.6)\\
&\quad + \left| \mathbb{E}\left[f'(W) \left(1 - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(W - W_i) \right) \right] \right|. &(3.7)
\end{aligned}$$

By Taylor expansion, the triangle inequality, and after pushing the absolute value inside the expectation, we obtain that (3.6) is bounded above by

$$\frac{\|f''\|}{2\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}|X_i(W - W_i)^2|.$$

Since $(W - W_i) = X_i/\sqrt{n}$, we obtain the first term in the bound (3.5), and we also find that (3.7) is bounded above by

$$\frac{\|f'\|}{n} \mathbb{E}\left| \sum_{i=1}^{n} (1 - X_i^2) \right| \leq \frac{\|f'\|}{n} \sqrt{\mathrm{Var}\left(\sum_{i=1}^{n} X_i^2 \right)},$$

where we have used the Cauchy-Schwarz inequality. By independence and the fact that $\mathrm{Var}(X_i^2) \leq \mathbb{E}[X_i^4]$, we obtain the second term in the bound (3.5). □

The work above shows that the strategy to bound $\mathbb{E}[f'(W) - Wf(W)]$ is to use the structure of W to rewrite $\mathbb{E}[Wf(W)]$ so that it is seen to be close to $\mathbb{E}[f'(W)]$. Rather than attempt this program anew in each application that arises, we develop ready to use theorems that provide error terms for various canonical structures that arise in many applications.

3.2. Dependency neighborhoods

We generalize Theorem 3.2 to sums of random variables with local dependence.

Definition 3.5. We say that a collection of random variables $(X_1, \ldots, X_n)$ has dependency neighborhoods $N_i \subseteq \{1, \ldots, n\}$, $i = 1, \ldots, n$, if $i \in N_i$ and X_i is independent of $\{X_j\}_{j \notin N_i}$.

Dependency neighborhoods are also referred to as dependency graphs since we can represent their structure in the form of a graph with vertices $\{1, \dots, n\}$ where i is connected to $j \neq i$ if $j \in N_i$. Using the Stein's method framework and a modification of the argument for sums of independent random variables we prove the following theorem, some version of which can be read from the main result of [9].

Theorem 3.6. *Let $X_1, \dots, X_n$ be random variables such that $\mathbb{E}[X_i^4] < \infty$, $\mathbb{E}[X_i] = 0$, $\sigma^2 = \mathrm{Var}\left(\sum_i X_i\right)$, and define $W = \sum_i X_i/\sigma$. Let the collection $(X_1, \dots, X_n)$ have dependency neighborhoods N_i, $i = 1, \dots, n$, and also define $D := \max_{1 \leq i \leq n} |N_i|$. Then for Z a standard normal random variable,*

$$d_{\mathrm{W}}(W, Z) \leq \frac{D^2}{\sigma^3} \sum_{i=1}^{n} \mathbb{E}|X_i|^3 + \frac{\sqrt{28} D^{3/2}}{\sqrt{\pi}\sigma^2} \sqrt{\sum_{i=1}^{n} \mathbb{E}[X_i^4]}. \tag{3.8}$$

Note that this theorem quantifies the heuristic that a sum of many locally dependent random variables are approximately normal. When viewed as an asymptotic result, it is clear that under some conditions a CLT holds even with D growing with n. It is also possible to prove similar theorems using further information about the dependence structure of the variables; see [25].

The proof of the theorem is analogous to the case of sums of independent random variables (a special case of this theorem), but the analysis is a little more complicated due to the dependence.

Proof. From Theorem 3.1, to upper bound $d_{\mathrm{W}}(W, Z)$ it is enough to bound $|\mathbb{E}[f'(W) - Wf(W)]|$, where $\|f\|, \|f''\| \leq 2$ and $\|f'\| \leq \sqrt{2/\pi}$. Let $W_i = \sum_{j \notin N_i} X_j$ and note that X_i is independent of W_i. As in the proof of Theorem 3.2, we can now write

$$\begin{aligned}
&|\mathbb{E}[f'(W) - Wf(W)]| \\
&\qquad \leq \left| \mathbb{E}\left[\frac{1}{\sigma} \sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i) f'(W) \right) \right] \right| \qquad (3.9) \\
&\qquad + \left| \mathbb{E}\left[f'(W) \left(1 - \frac{1}{\sigma} \sum_{i=1}^{n} X_i (W - W_i) \right) \right] \right|. \qquad (3.10)
\end{aligned}$$

We now proceed by showing that (3.9) is bounded above by the first term in (3.8) and (3.10) is bounded above by the second.

By Taylor expansion, the triangle inequality, and after pushing the absolute value inside the expectation, we obtain that (3.9) is bounded above by

$$\begin{aligned}
\frac{\|f''\|}{2\sigma} \sum_{i=1}^{n} \mathbb{E}|X_i (W - W_i)^2| &\leq \frac{1}{\sigma^3} \sum_{i=1}^{n} \mathbb{E} \left| X_i \left(\sum_{j \in N_i} X_j \right)^2 \right| \\
&\leq \frac{1}{\sigma^3} \sum_{i=1}^{n} \sum_{j,k \in N_i} \mathbb{E} \left| X_i X_j X_k \right|. \qquad (3.11)
\end{aligned}$$

The arithmetic-geometric mean inequality implies that

$$\mathbb{E}\left|X_iX_jX_k\right| \leq \frac{1}{3}\left(\mathbb{E}|X_i|^3+\mathbb{E}|X_j|^3+\mathbb{E}|X_k|^3\right),$$

so that (3.9) is bounded above by the first term in the bound (3.8), where we use for example that

$$\sum_{i=1}^n\sum_{j,k\in N_i}\mathbb{E}|X_j|^3 \leq D^2\sum_{j=1}^n\mathbb{E}|X_j|^3.$$

Similar consideration implies that (3.10) is bounded above by

$$\frac{\|f'\|}{\sigma^2}\mathbb{E}\left|\sigma^2-\sum_{i=1}^n X_i\sum_{j\in N_i}X_j\right| \leq \frac{\sqrt{2}}{\sqrt{\pi}\sigma^2}\sqrt{\operatorname{Var}\left(\sum_{i=1}^n\sum_{j\in N_i}X_iX_j\right)}, \tag{3.12}$$

where the inequality follows from the Cauchy-Schwarz inequality coupled with the representation

$$\sigma^2=\mathbb{E}\left[\sum_{i=1}^n X_i\sum_{j\in N_i}X_j\right].$$

The remainder of the proof consists of analysis on (3.12), but note that in practice it may be possible to bound this term directly. In order to bound the variance under the square root in (3.12), we first compute

$$\mathbb{E}\left[\left(\sum_{i=1}^n\sum_{j\in N_i}X_iX_j\right)^2\right]=\sum_{i\neq j}\sum_{k\in N_i}\sum_{l\in N_j}\mathbb{E}[X_iX_jX_kX_l] \tag{3.13}$$

$$+\sum_{i=1}^n\sum_{j\in N_i}\mathbb{E}[X_i^2X_j^2]+\sum_{i=1}^n\sum_{j\in N_i}\sum_{k\in N_i/\{j\}}\mathbb{E}[X_i^2X_jX_k]. \tag{3.14}$$

Using the arithmetic-geometric mean inequality, the first term of the expression (3.14) is bounded above by

$$\frac{1}{2}\sum_{i=1}^n\sum_{j\in N_i}\left(\mathbb{E}[X_i^4]+\mathbb{E}[X_j^4]\right)\leq D\sum_{i=1}^n\mathbb{E}[X_i^4],$$

and the second by

$$\frac{1}{4}\sum_{i=1}^n\sum_{j\in N_i}\sum_{k\in N_i/\{j\}}\left(2\mathbb{E}[X_i^4]+\mathbb{E}[X_j^4]+\mathbb{E}[X_k^4]\right)\leq D(D-1)\sum_{i=1}^n\mathbb{E}[X_i^4].$$

We decompose the term (3.13) into two components;

$$\sum_{i\neq j}\sum_{k\in N_i}\sum_{l\in N_j}\mathbb{E}[X_iX_jX_kX_l] \\ = \sum_{\{i,k\},\{j,l\}}\mathbb{E}[X_iX_k]\mathbb{E}[X_jX_l] + \sum_{\{i,k,j,l\}}\mathbb{E}[X_iX_jX_kX_l], \tag{3.15}$$

where the first sum denotes the indices in which $\{X_i, X_k\}$ are independent of $\{X_j, X_l\}$, and the second term consists of those remaining. Note that by the arithmetic-geometric mean inequality, the second term of (3.15) is bounded above by

$$6D^3\sum_{i=1}^n\mathbb{E}[X_i^4],$$

since the number of "connected components" with at most four vertices of the dependency graph induced by the neighborhoods, is no more than $D\times 2D\times 3D$. Using a decomposition of σ^4 similar to that provided by (3.13) and (3.14), we find the first term of (3.15) is bounded above by

$$\sigma^4 - \sum_{\{i,k,j,l\}}\mathbb{E}[X_iX_k]\mathbb{E}[X_jX_l] - \sum_{i=1}^n\sum_{j\in N_i}\sum_{k\in N_i/\{j\}}\mathbb{E}[X_iX_j]\mathbb{E}[X_iX_k],$$

and a couple applications of the arithmetic-geometric mean inequality yields

$$\begin{aligned}-\mathbb{E}[X_iX_k]\mathbb{E}[X_jX_l] &\leq \frac{1}{2}\left(\mathbb{E}[X_iX_k]^2 + \mathbb{E}[X_jX_l]^2\right)\\ &\leq \frac{1}{2}\left(\mathbb{E}[X_i^2X_k^2] + \mathbb{E}[X_j^2X_l^2]\right)\\ &\leq \frac{1}{4}\left(\mathbb{E}[X_i^4] + \mathbb{E}[X_j^4] + \mathbb{E}[X_k^4] + \mathbb{E}[X_l^4]\right).\end{aligned}$$

Putting everything together, we obtain that

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n\sum_{j\in N_i}X_iX_j\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n\sum_{j\in N_i}X_iX_j\right)^2\right] - \sigma^4\\ &\leq (12D^3 + 2D^2)\sum_{i=1}^n\mathbb{E}[X_i^4] \leq 14D^3\sum_{i=1}^n\mathbb{E}[X_i^4],\end{aligned}$$

which yields the theorem. □

Note that much of the proof of Theorem 3.6 consists of bounding the error in a simple form. However, an upper bound for $d_{\mathrm{W}}(W, Z)$ is obtained by adding the intermediate terms (3.11) and (3.12) which in many applications may be directly bounded (and produce better bounds).

Theorem 3.6 is an intuitively pleasing result that has many applications; a notable example is [8] where CLTs for statistics of various random geometric graphs are shown. We apply it in the following setting.

3.2.1. Application: Triangles in Erdős-Rényi random graphs

Let $G = G(n,p)$ be an Erdős-Rényi random graph on n vertices with edge probability p and let T be the number of triangles in G. We have the decomposition $T = \sum_{i=1}^{N} Y_i$, where $N = \binom{n}{3}$, and Y_i is the indicator that a triangle is formed at the "ith" set of three vertices, in some arbitrary but fixed order. We can take the set $N_i/\{i\}$ to be the indices which share exactly two vertices with those indexed by i, since due to the independence of edges, Y_i is independent of the collection of triangles with which it shares no edges. With this definition, $|N_i| = 3(n-3)+1$, and we can apply Theorem 3.6 with $X_i = Y_i - p^3$ and $D = 3n-8$. Since

$$\mathbb{E}|X_i|^k = p^3(1-p^3)[(1-p^3)^{k-1} + p^{3(k-1)}], \qquad k = 1, 2, \ldots$$

we now only have to compute $\mathrm{Var}(T)$ to apply the theorem. A simple calculation using the decomposition of T into the sum of the indicators Y_i shows that

$$\sigma^2 := \mathrm{Var}(T) = \binom{n}{3} p^3[1 - p^3 + 3(n-3)p^2(1-p)],$$

and Theorem 3.6 implies that for $W = (T - \mathbb{E}[T])/\sigma$ and Z a standard normal random variable

$$\begin{aligned} d_{\mathrm{K}}(W,Z) \leq{}& \frac{(3n-8)^2}{\sigma^3}\binom{n}{3}p^3(1-p^3)[(1-p^3)^2 + p^6] \\ &+ \frac{\sqrt{26}(3n-8)^{3/2}}{\sqrt{\pi}\sigma^2}\sqrt{\binom{n}{3}p^3(1-p^3)[(1-p^3)^3 + p^9]}. \end{aligned}$$

This bound holds for all $n \geq 3$ and $0 \leq p \leq 1$, but some asymptotic analysis shows that if, for example, $p \sim n^{-\alpha}$ for some $0 \leq \alpha < 1$ (so that $\mathrm{Var}(T) \to \infty$), then the number of triangles satisfies a CLT for $0 \leq \alpha < 2/9$, which is only a subset of the regime where normal convergence holds [52]. It is possible that starting from (3.11) and (3.12) would yield better rates in a wider regime, and considering finer structure yields better results [13].

3.3. Exchangeable pairs

We begin with a definition.

Definition 3.7. The ordered pair (W, W') of random variables is called an *exchangeable pair* if $(W, W') \overset{d}{=} (W', W)$. If for some $0 < a \leq 1$, the exchangeable pair (W, W') satisfies the relation

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

then we call (W, W') an a-Stein pair.

The next proposition contains some easy facts related to Stein pairs.

Proposition 3.8. *Let (W, W') an exchangeable pair.*

1. *If $F : \mathbb{R}^2 \to \mathbb{R}$ is an anti-symmetric function; that is $F(x, y) = -F(y, x)$, then $\mathbb{E}[F(W, W')] = 0$.*

If (W, W') is an a-Stein pair with $\operatorname{Var}(W) = \sigma^2$, then

2. *$\mathbb{E}[W] = 0$ and $\mathbb{E}[(W' - W)^2] = 2a\sigma^2$.*

Proof. Item 1 follows by the following equalities, the first by exchangeability and the second by anti-symmetry of F.

$$\mathbb{E}[F(W, W')] = \mathbb{E}[F(W', W)] = -\mathbb{E}[F(W, W')].$$

The first assertion of Item 2 follows from the fact that

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

and the second by calculating

$$\begin{aligned}\mathbb{E}[(W' - W)^2] &= \mathbb{E}[(W')^2] + \mathbb{E}[W^2] - 2\mathbb{E}[W\mathbb{E}[W'|W]] \\ &= 2\sigma^2 - 2(1 - a)\sigma^2 = 2a\sigma^2.\end{aligned}$$

□

From this point we illustrate the use of the exchangeable pair in the following theorem.

Theorem 3.9. *If (W, W') is an a-Stein pair with $\mathbb{E}[W^2] = 1$ and Z has the standard normal distribution, then*

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

Before the proof comes a few remarks.

Remark 3.10. The strategy for using Theorem 3.9 to obtain an error in the approximation of the distribution of a random variable W by the standard normal is to construct W' on the same space as W, such that (W, W') is an a-Stein pair. How can we achieve this construction?

Typically $W = W(\omega)$ is a random variable on some space Ω with probability measure μ. For example, Ω is the set of sequences of zeros and ones of length n where coordinate i is one with probability p, independent of all the other coordinates and W is the number of ones in a sequence (i.e. W has the binomial distribution). It is not too difficult to see that in some generality, if $X_0, X_1, \ldots$ is a stationary Markov chain on Ω which is reversible with respect to μ, then $(W(X_0), W(X_1))$ is an exchangeable pair. In the example above, a Markov chain on Ω which is reversible with respect to the measure defined there follows the rule of independently resampling a randomly chosen coordinate. By only considering the number of ones in this stationary chain, we find an exchangeable pair with binomial marginals.

Since there is much effort put into constructing reversible Markov chains (e.g. Gibbs sampler), this is a useful method to construct exchangeable pairs. However, the linearity condition is not as easily abstractly constructed and must be verified.

Remark 3.11. It is also useful to note that

$$\operatorname{Var}\left(\mathbb{E}[(W'-W)^2|W]\right) \leq \operatorname{Var}\left(\mathbb{E}[(W'-W)^2|\mathcal{F}]\right),$$

for any sigma-field $\mathcal{F}$ which is larger than the sigma-field generated by W. In the setting of the previous remark where $W := W(X_0)$ and $W' := W(X_1)$, it can be helpful to condition on X_0 rather than W when computing the error bound from Theorem 3.9.

Remark 3.12. A heuristic explanation for the form of the error terms appearing in Theorem 3.9 arises by considering an Ornstein-Uhlenbeck (O-U) diffusion process. Define the diffusion process $(D(t))_{t\geq 0}$ by the following properties.

1. $\mathbb{E}[D(t+a) - D(t)|D(t) = x] = -ax + \mathrm{o}(a)$.
2. $\mathbb{E}[(D(t+a) - D(t))^2|D(t) = x] = 2a + \mathrm{o}(a)$.
3. For all $\varepsilon > 0$, $\mathbb{P}[|D(t+a) - D(t)| > \varepsilon | D(t) = x] = \mathrm{o}(a)$.

Here the function $g(a)$ is $\mathrm{o}(a)$ if $g(a)/a$ tends to zero as a tends to zero. These three properties determine the O-U diffusion process, and this process is reversible with the standard normal distribution as its stationary distribution. What does this have to do with Theorem 3.16? Roughly, if we think of W as $D(t)$ and W' as $D(t+a)$ for some small a, then Item 1 corresponds to the a-Stein pair linearity condition. Item 2 implies that the first term of the error in Theorem 3.9 is small since $\mathbb{E}[(D(t+a) - D(t))^2] = 2a + \mathrm{o}(a)$ (compare to Item 2 of Proposition 3.8) so that the variance appearing in the first term of the error will be $\mathrm{o}(a^2)$. Finally, Item 3 relates to the second term in the error.

Proof of Theorem 3.9. The strategy of the proof is to use the exchangeable pair to rewrite $\mathbb{E}[Wf(W)]$ so that it is seen to be close to $\mathbb{E}[f'(W)]$. To this end, let f be bounded with bounded first and second derivative and let $F(w) := \int_0^w f(t)dt$. Now, exchangeability and Taylor expansion imply that

$$\begin{aligned} 0 &= \mathbb{E}[F(W') - F(W)] \\ &= \mathbb{E}\left[(W'-W)f(W) + \frac{1}{2}(W'-W)^2 f'(W) + \frac{1}{6}(W'-W)^3 f''(W^*)\right], \quad (3.16) \end{aligned}$$

where W^* is a random quantity in the interval with endpoints W and W'. Now, the linearity condition on the Stein pair yields

$$\mathbb{E}\left[(W'-W)f(W)\right] = \mathbb{E}[f(W)\mathbb{E}[(W'-W)|W]] = -a\mathbb{E}[Wf(W)]. \quad (3.17)$$

Combining (3.16) and (3.17) we obtain

$$\mathbb{E}[Wf(W)] = \mathbb{E}\left[\frac{(W'-W)^2 f'(W)}{2a} + \frac{(W'-W)^3 f''(W^*)}{6a}\right].$$

From this point we can easily see

$$
\begin{aligned}
&|\mathbb{E}[f'(W) - Wf(W)]| \\
&\qquad \leq \|f'\|\mathbb{E}\left|1 - \frac{\mathbb{E}[(W'-W)^2|W]}{2a}\right| + \|f''\|\frac{\mathbb{E}|W'-W|^3}{6a},
\end{aligned}
\tag{3.18}
$$

and the theorem follows after noting that we are only considering functions f such that $\|f'\| \leq \sqrt{2/\pi}$, and $\|f''\| \leq 2$, and that from Item 2 of Proposition 3.8 (using the assumption $\mathrm{Var}(W) = 1$), we have $\mathbb{E}[\mathbb{E}[(W'-W)^2|W]] = 2a$ so that an application of the Cauchy-Schwarz inequality yields the variance term in the bound. □

Before moving to a heavier application, we consider the canonical example of a sum of independent random variables.

Example 3.13. Let $X_1, \ldots, X_n$ independent with $\mathbb{E}[X_i^4] < \infty$, $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) = 1$, and $W = n^{-1/2}\sum_{i=1}^n X_i$. We construct our exchangeable pair by choosing an index uniformly at random and replacing it by an independent copy. Formally, let I uniform on $\{1, \ldots, n\}$, $(X_1', \ldots, X_n')$ be an independent copy of $(X_1, \ldots, X_n)$, and define

$$W' = W - \frac{X_I}{\sqrt{n}} + \frac{X_I'}{\sqrt{n}}.$$

It is a simple exercise to show that (W, W') is exchangeable, and we now verify that is also a $1/n$-Stein pair. The calculation below is straightforward; in the penultimate equality we use the independence of X_i and X_i' and the fact that $\mathbb{E}[X_i'] = 0$.

$$
\begin{aligned}
\mathbb{E}[W' - W|(X_1, \ldots, X_n)] &= \frac{1}{\sqrt{n}}\mathbb{E}[X_I' - X_I|(X_1, \ldots, X_n)] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{n}\mathbb{E}[X_i' - X_i|(X_1, \ldots, X_n)] \\
&= -\frac{1}{n}\sum_{i=1}^n \frac{X_i}{\sqrt{n}} = -\frac{W}{n}.
\end{aligned}
$$

Since the conditioning on the larger sigma-field only depends on W, we have that $\mathbb{E}[W' - W|W] = -W/n$, as desired.

We can now apply Theorem 3.9. We first bound

$$
\begin{aligned}
\mathbb{E}|W' - W|^3 &= \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}|X_i - X_i'|^3 \\
&\leq \frac{8}{n^{3/2}}\sum_{i=1}^n \mathbb{E}|X_i|^3,
\end{aligned}
$$

where we used the arithmetic-geometric mean inequality for the cross terms of the expansion of the cube of the difference (we could also express the error in terms of these lower moments by independence). Next we compute

$$\begin{aligned}\mathbb{E}[(W'-W)^2|(X_1,\ldots,X_n)] &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[(X_i'-X_i)^2|X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^n 1+X_i^2.\end{aligned}$$

Taking the variance we see that

$$\operatorname{Var}\left(\mathbb{E}[(W'-W)^2|W]\right) \leq \frac{1}{n^4}\sum_{i=1}^n \mathbb{E}[X_i^4].$$

Combining the estimates above we have

$$d_{\mathrm{W}}(W,Z) \leq \sqrt{\frac{2}{\pi}}\frac{\sqrt{\sum_{i=1}^n \mathbb{E}[X_i^4]}}{2n} + \frac{2}{3n}\sum_{i=1}^n \mathbb{E}|X_i|^3.$$

Note that if the X_i are i.i.d. then this term is of order $n^{-1/2}$, which is best possible. Finally, we could probably get away with only assuming three moments for the X_i if we use the intermediate term (3.18) in the proof of Theorem 3.16.

3.3.1. Application: Anti-voter model

In this section we consider an application of Theorem 3.9 found in [47]; we closely follow their treatment. Let G be an r-regular[3] graph with vertex set V and edge set E. We define the *anti-voter* Markov chain on the space $\{-1,1\}^V$ of labelings of the vertices of V by $+1$ and -1. Given the chain is at a state $\mathbf{x}=(x_u)_{u\in V}\in\{-1,1\}^V$, the chain follows the rule of first uniformly choosing a vertex $v\in V$, then uniformly choosing a vertex w from the set of r vertices connected to v and finally obtaining $\mathbf{X}'=(X_u')_{u\in V}$, the next step in the chain, by changing the sign of the label of v to the opposite of this second vertex:

$$X_u' = \begin{cases} x_u & u\neq v, \\ -x_w & u=v. \end{cases} \tag{3.19}$$

The model gets its name from thinking of the vertices as people in a town full of curmudgeons where a positive (negative) labeling corresponding to a yes (no) vote for some measure. At each time unit a random person talks to a random neighbor and decides to switch votes to the opposite of that neighbor.

It is known (Chapter 14, Section 4 of [1]) that if the underlying graph G is not bipartite or a cycle, then the anti-voter chain is irreducible and aperiodic

[3] The term r-regular means that every vertex has degree r.

and has a unique stationary distribution. This distribution can be difficult to describe, but we can use Theorem 3.9 to obtain an error in the Wasserstein distance to the standard normal distribution for the sum of the labels of the vertices. We now state the theorem and postpone discussion of computing the relevant quantities in the error until after the proof.

Theorem 3.14. *Let G be an r-regular graph with n vertices which is not bipartite or a cycle. Let $\boldsymbol{X} = (X_i)_{i=1}^n \in \{-1, 1\}^n$ have the stationary distribution of the anti-voter chain and let $\boldsymbol{X}' = (X_i')_{i=1}^n$ be one step in the chain from $\boldsymbol{X}$ as in the description leading up to (3.19). Let $\sigma_n^2 = \mathrm{Var}(\sum_i X_i)$, $W = \sigma_n^{-1} \sum_i X_i$, and $W' = \sigma_n^{-1} \sum_i X_i'$. Then (W, W') is a $2/n$-Stein pair, and if Z has the standard normal distribution, then*

$$d_{\mathrm{W}}(W, Z) \leq \frac{4n}{3\sigma_n^3} + \frac{\sqrt{\mathrm{Var}(Q)}}{r\sigma_n^2\sqrt{2\pi}},$$

where

$$Q = \sum_{i=1}^n \sum_{j\in N_i} X_i X_j,$$

and N_i denotes the neighbors of i.

Part of the first assertion of the theorem is that (W, W') is exchangeable, which is non-trivial to verify since the anti-voter chain is not necessarily reversible. However, we can apply the following lemma - the proof here appears in [48].

Lemma 3.15. *If W and W' are identically distributed integer-valued random variables defined on the same space such that $\mathbb{P}(|W' - W| \leq 1) = 1$, then (W, W') is an exchangeable pair.*

Proof. The fact that W and W' only differ by at most one almost surely implies

$$\mathbb{P}(W' \leq k) = \mathbb{P}(W < k) + \mathbb{P}(W = k, W' \leq k) + \mathbb{P}(W = k + 1, W' = k),$$

for any k, while we also have

$$\mathbb{P}(W \leq k) = \mathbb{P}(W < k) + \mathbb{P}(W = k, W' \leq k) + \mathbb{P}(W = k, W' = k + 1).$$

Since W and W' have the same distribution, the left hand sides of the equations above are equal, and equating the right hand sides yields

$$\mathbb{P}(W = k + 1, W' = k) = \mathbb{P}(W = k, W' = k + 1),$$

which is the lemma. □

Proof of Theorem 3.14. For the proof below let $\sigma := \sigma_n$ so that $\sigma W = \sum_{i=1}^n X_i$. The exchangeability of (W, W') follows by Lemma 3.15 since

$$\mathbb{P}(\sigma(W' - W)/2 \in \{-1, 0, 1\}) = 1.$$

To show the linearity condition for the Stein pair, we define some auxiliary quantities related to $\mathbf{X}$. Let $a_1 = a_1(\mathbf{X})$ be the number of edges in G which have a one at each end vertex when labeled by $\mathbf{X}$. Similarly, let a_{-1} be the analogous quantity with negative ones at each end vertex and a_0 be the number of edges with a different labal at each end vertex. Due to the fact that G is r-regular, the number of ones in $\mathbf{X}$ is $(2a_1 + a_0)/r$ and the number of negative ones in $\mathbf{X}$ is $(2a_{-1} + a_0)/r$. Note that these two observation imply

$$\sigma W = \frac{2}{r}\left(a_1 - a_{-1}\right). \tag{3.20}$$

Now, since conditional on $\mathbf{X}$ the event $\sigma W' = \sigma W + 2$ is equal to the event that the chain moves to $\mathbf{X}'$ by choosing a vertex labeled -1 and then choosing a neighbor with label -1, we have

$$\mathbb{P}(\sigma(W' - W) = 2|\mathbf{X}) = \frac{2a_{-1}}{nr} \tag{3.21}$$

and similarly

$$\mathbb{P}(\sigma(W' - W) = -2|\mathbf{X}) = \frac{2a_1}{nr}. \tag{3.22}$$

Using these last two formulas and (3.20), we obtain

$$\mathbb{E}[\sigma(W' - W)|\mathbf{X}] = \frac{2}{nr}\left(a_{-1} - a_1\right) = -\frac{2\sigma W}{n},$$

as desired.

From this point we compute the error terms from Theorem 3.9. The first thing to note is that $|W' - W| \leq 2/\sigma$ implies

$$\frac{\mathbb{E}|W' - W|^3}{3a} \leq \frac{4n}{3\sigma^3},$$

which contributes the first part of the error from the Theorem. Now, (3.21) and (3.22) imply

$$\mathbb{E}[(W' - W)^2|\mathbf{X}] = \frac{8}{\sigma^2 nr}\left(a_{-1} + a_1\right), \tag{3.23}$$

and since

$$\begin{aligned} 2a_1 + 2a_{-1} + 2a_0 &= \sum_{i=1}^{n}\sum_{j\in N_i} 1 = nr, \\ 2a_1 + 2a_{-1} - 2a_0 &= \sum_{i=1}^{n}\sum_{j\in N_i} X_iX_j = Q, \end{aligned}$$

we have

$$Q = 4(a_{-1} + a_1) - rn,$$

which combining with (3.23) and a small calculation yields the second error term of the theorem. □

In order for Theorem 3.14 to be useful for a given graph G, we need lower bounds on σ_n^2 and upper bounds on $\mathrm{Var}(Q)$. The former item can be accomplished by the following result of [1] (Chapter 14, Section 4).

Lemma 3.16. *[1] Let G be an r-regular graph and let $\kappa = \kappa(G)$ be the minimum over subsets of vertices A of the number of edges that either have both ends in A or else have both ends in A^c. If σ^2 is the variance of the stationary distribution of the anti-voter model on G, then*

$$\frac{2\kappa}{r} \leq \sigma^2 \leq n.$$

The strategy to upper bound $\mathrm{Var}(Q)$ is to associate the anti-voter model to a so-called "dual process" from interacting particle system theory. This discussion is outside the scope of our work, but see [1, 27] and also Section 2 of [47] where concrete applications of Theorem 3.14 are worked out.

3.4. Size-bias coupling

Our next method of rewriting $\mathbb{E}[Wf(W)]$ to be compared to $\mathbb{E}[f'(W)]$ is through the size-bias coupling which first appeared in the context of Stein's method for normal approximation in [34].

Definition 3.17. For a random variable $X \geq 0$ with $\mathbb{E}[X] = \mu < \infty$, we say the random variable X^s has the *size-bias* distribution with respect to X if for all f such that $\mathbb{E}|Xf(X)| < \infty$ we have

$$\mathbb{E}[Xf(X)] = \mu\mathbb{E}[f(X^s)].$$

Before discussing existence of the size-bias distribution, we remark that our use of X^s is a bit more transparent than the use of the exchangeable pair above. To wit, if $\mathrm{Var}(X) = \sigma^2 < \infty$ and $W = (X - \mu)/\sigma$, then

$$\begin{aligned} \mathbb{E}[Wf(W)] &= \mathbb{E}\left[\frac{X-\mu}{\sigma} f\left(\frac{X-\mu}{\sigma}\right)\right] \\ &= \frac{\mu}{\sigma}\left[f\left(\frac{X^s-\mu}{\sigma}\right) - f\left(\frac{X-\mu}{\sigma}\right)\right], \end{aligned} \tag{3.24}$$

so that if f is differentiable, then the Taylor expansion of (3.24) about W allows us to compare $\mathbb{E}[Wf(W)]$ to $\mathbb{E}[f'(W)]$. We make this precise shortly, but first we tie up a loose end.

Proposition 3.18. *If $X \geq 0$ is a random variable with $\mathbb{E}[X] = \mu < \infty$ and distribution function F, then the size-bias distribution of X is absolutely continuous with respect to the measure of X with density read from*

$$dF^s(x) = \frac{x}{\mu} dF(x).$$

Corollary 3.19. *If $X \geq 0$ is an integer-valued random variable having finite mean μ, then the random variable X^s with the size-bias distribution of X is such that*

$$\mathbb{P}(X^s = k) = \frac{k\mathbb{P}(X = k)}{\mu}.$$

The size-bias distribution arises in other contexts such as the waiting time paradox and sampling theory [3]. We now record our main Stein's method size-bias normal approximation theorem.

Theorem 3.20. *Let $X \geq 0$ be a random variable such that $\mu := \mathbb{E}[X] < \infty$ and $\mathrm{Var}(X) = \sigma^2$. Let X^s be defined on the same space as X and have the size-bias distribution with respect to X. If $W = (X - \mu)/\sigma$ and $Z \sim N(0,1)$, then*

$$d_{\mathrm{W}}(W, Z) \leq \frac{\mu}{\sigma^2}\sqrt{\frac{2}{\pi}}\sqrt{\mathrm{Var}(\mathbb{E}[X^s - X|X])} + \frac{\mu}{\sigma^3}\mathbb{E}[(X^s - X)^2].$$

Proof. Our strategy (as usual) is to bound $|\mathbb{E}[f'(W) - Wf(W)]|$ for f bounded with two bounded derivatives. Starting from (3.24), a Taylor expansion yields

$$\mathbb{E}[Wf(W)] = \frac{\mu}{\sigma}\mathbb{E}\left[\frac{X^s - X}{\sigma}f'\left(\frac{X-\mu}{\sigma}\right) + \frac{(X^s - X)^2}{2\sigma^2}f''\left(\frac{X^* - \mu}{\sigma}\right)\right],$$

for some X^* in the interval with endpoints X and X^s. Using the definition of W in terms of X in the previous expression, we obtain

$$|\mathbb{E}[f'(W) - Wf(W)]| \leq \left|\mathbb{E}\left[f'(W)\left(1 - \frac{\mu}{\sigma^2}(X^s - X)\right)\right]\right| \tag{3.25}$$

$$+ \frac{\mu}{2\sigma^3}\left|\mathbb{E}\left[f''\left(\frac{X^* - \mu}{\sigma}\right)(X^s - X)^2\right]\right|. \tag{3.26}$$

Since we are taking the supremum over functions f with $\|f'\| \leq \sqrt{2/\pi}$ and $\|f''\| \leq 2$, it is clear that (3.26) is bounded above by the second term of the error stated in the theorem and (3.25) is bounded above by

$$\sqrt{\frac{2}{\pi}}\mathbb{E}\left|1 - \frac{\mu}{\sigma^2}\mathbb{E}[X^s - X|X]\right| \leq \frac{\mu}{\sigma^2}\sqrt{\frac{2}{\pi}}\sqrt{\mathrm{Var}(\mathbb{E}[X^s - X|X])};$$

here we use the Cauchy-Schwarz inequality after noting that by the definition of X^s, $\mathbb{E}[X^s] = (\sigma^2 + \mu^2)/\mu$. □

3.4.1. Coupling construction

At this point it is appropriate to discuss methods to couple a random variable X to a size-bias version X^s. In the case that $X = \sum_{i=1}^n X_i$, where $X_i \geq 0$ and $\mathbb{E}[X_i] = \mu_i$, we have the following recipe to construct a size-bias version of X.

1. For each $i = 1, \dots, n$, let X_i^s have the size-bias distribution of X_i independent of $(X_j)_{j \neq i}$ and $(X_j^s)_{j \neq i}$. Given $X_i^s = x$, define the vector $(X_j^{(i)})_{j \neq i}$ to have the distribution of $(X_j)_{j \neq i}$ conditional on $X_i = x$.
2. Choose a random summand X_I, where the index I is chosen proportional to μ_i and independent of all else. Specifically, $\mathbb{P}(I = i) = \mu_i/\mu$, where $\mu = \mathbb{E}[X]$.
3. Define $X^s = \sum_{j \neq I} X_j^{(I)} + X_I^s$.

Proposition 3.21. *Let $X = \sum_{i=1}^n X_i$, with $X_i \geq 0$, $\mathbb{E}[X_i] = \mu_i$, and also $\mu = \mathbb{E}[X] = \sum_i \mu_i$. If X^s is constructed by Items 1 - 3 above, then X^s has the size-bias distribution of X.*

Remark 3.22. Due to the form of the error term in Theorem 3.20 we would like to closely couple X and X^s. In terms of the construction above, it is advantageous to have $X_j^{(i)}$ closely coupled to X_j for $j \neq i$.

Proof. Let $\mathbf{X} = (X_1, \dots, X_n)$ and for $i = 1, \dots, n$, let $\mathbf{X}^i$ be a vector with coordinate j equal to $X_j^{(i)}$ for $j \neq i$ and coordinate i equal to X_i^s as in item 1 above. In order to prove the result, it is enough to show

$$\mathbb{E}[Wf(\mathbf{X})] = \mu \mathbb{E}[f(\mathbf{X}^I)], \tag{3.27}$$

for $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}|Wf(\mathbf{X})| < \infty$. Equation (3.27) follows easily after we show that for all $i = 1, \dots, n$,

$$\mathbb{E}[X_i f(\mathbf{X})] = \mu_i \mathbb{E}[f(\mathbf{X}^i)]. \tag{3.28}$$

To see (3.28), note that for $h(X_i) = \mathbb{E}[f(\mathbf{X})|X_i]$,

$$\begin{aligned} \mathbb{E}[X_i f(\mathbf{X})] &= \mathbb{E}[X_i h(X_i)] \\ &= \mu_i \mathbb{E}[h(X_i^s)], \end{aligned}$$

which is the right hand side of (3.28). □

Note the following special cases of Proposition 3.21.

Corollary 3.23. *Let $X_1, \dots, X_n$ be non-negative independent random variables with $\mathbb{E}[X_i] = \mu_i$, and for each $i = 1, \dots, n$, let X_i^s have the size-bias distribution of X_i independent of $(X_j)_{j \neq i}$ and $(X_j^s)_{j \neq i}$. If $X = \sum_{i=1}^n X_i$, $\mu = \mathbb{E}[X]$, and I is chosen independent of all else with $\mathbb{P}(I = i) = \mu_i/\mu$, then $X^s = X - X_I + X_I^s$ has the size-bias distribution of X.*

Corollary 3.24. *Let $X_1, \dots, X_n$ be zero-one random variables and also let $p_i := \mathbb{P}(X_i = 1)$. For each $i = 1, \dots, n$, let $(X_j^{(i)})_{j \neq i}$ have the distribution of $(X_j)_{j \neq i}$ conditional on $X_i = 1$. If $X = \sum_{i=1}^n X_i$, $\mu = \mathbb{E}[X]$, and I is chosen independent of all else with $\mathbb{P}(I = i) = p_i/\mu$, then $X^s = \sum_{j \neq I} X_j^{(I)} + 1$ has the size-bias distribution of X.*

Proof. Corollary 3.23 is obvious since due to independence, the conditioning in the construction has no effect. Corollary 3.24 follows after noting that for X_i a zero-one random variable, $X_i^s = 1$. □

3.4.2. Applications

Example 3.25. We can use Corollary 3.23 in Theorem 3.20 to bound the Wasserstein distance between the normalized sum of independent variables with finite third moment and the normal distribution; we leave this as an exercise.

Example 3.26. Let $G = G(n,p)$ be an Erdős-Rényi graph and for $i = 1, \dots, n$, let X_i be the indicator that vertex v_i (under some arbitrary but fixed labeling) has degree zero so that $X = \sum_{i=1}^n X_i$ is the number of isolated vertices of G. We use Theorem 3.20 to obtain an upper bound on the Wasserstein metric between the normal distribution and the distribution of $W = (X - \mu)/\sigma$ where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X)$.

We can use Corollary 3.24 to construct X^s, a size-bias version of X. Since X is a sum of identically distributed indicators, Corollary 3.24 states that in order to size-bias X, we first choose an index I uniformly at random from the set $\{1, \dots, n\}$, then size-bias X_I by setting it equal to one, and finally adjust the remaining summands conditional on $X_I = 1$ (the new size-bias value). We can realize $X_I^s = 1$ by erasing any edges connected to vertex v_I. Given that $X_I = 1$ (v_I is isolated), the graph G is just an Erdős-Rényi graph on the remaining $n-1$ vertices. Thus X^s can be realized as the number of isolated vertices in G after erasing all the edges connected to v_I.

In order to apply Theorem 3.20 using this construction, we need to compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, $\mathrm{Var}(\mathbb{E}[X^s - X|X])$, and $\mathbb{E}[(X^s - X)^2]$. Since the chance that a given vertex is isolated is $(1-p)^{n-1}$, we have

$$\mu := \mathbb{E}[X] = n(1-p)^{n-1},$$

and also that

$$\begin{aligned} \sigma^2 := \mathrm{Var}(X) &= \mu\left(1 - (1-p)^{n-1}\right) + n(n-1)\,\mathrm{Cov}(X_1, X_2) \\ &= \mu[1 + (np-1)(1-p)^{n-2}], \end{aligned} \tag{3.29}$$

since $\mathbb{E}[X_1 X_2] = (1-p)^{2n-3}$. Let d_i be the degree of v_i in G and let D_i be the number of vertices connected to v_i which have degree one. Then it is clear that

$$X^s - X = D_I + \mathbb{I}[d_I > 0],$$

so that

$$\begin{aligned} \mathrm{Var}(\mathbb{E}[X^s - X|G]) &= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^n (D_i + \mathbb{I}[d_i > 0])\right) && (3.30) \\ &\le \frac{2}{n^2}\left[\mathrm{Var}\left(\sum_{i=1}^n D_i\right) + \mathrm{Var}\left(\sum_{i=1}^n \mathbb{I}[d_i > 0]\right)\right]. && (3.31) \end{aligned}$$

Since $\sum_{i=1}^n \mathbb{I}[d_i > 0] = n - X$, the second variance term of (3.31) is given by (3.29). Now, $\sum_{i=1}^n D_i$ is the number of vertices in G with degree one which can be expressed as $\sum_{i=1}^{n} Y_i$, where Y_i is the indicator that v_i has degree one in G. Thus,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} D_i\right) &= n(n-1)p(1-p)^{n-2}\left(1-(n-1)p(1-p)^{n-2}\right) \\ &\qquad + n(n-1)\operatorname{Cov}(Y_1, Y_2) \\ &= n(n-1)p(1-p)^{n-2}\big[1-(n-1)p(1-p)^{n-2} \\ &\qquad + (1-p)^{n-2} + (n-1)^2p^2(1-p)^{n-3}\big],\end{aligned}$$

since $\mathbb{E}[Y_1Y_2] = p(1-p)^{2n-4} + (n-1)^2p^2(1-p)^{2n-5}$ (the first term corresponds to v_1 and v_2 being joined).

The final term we need to bound is

$$\begin{aligned}\mathbb{E}[(X^s-X)^2] &= \mathbb{E}\left[\mathbb{E}[(X^s-X)^2|X]\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(D_i+\mathbb{I}[d_i>0])^2] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(D_i+1)^2] \\ &= \mathbb{E}[D_1^2] + 2\mathbb{E}[D_1] + 1.\end{aligned}$$

Expressing D_1 as a sum of indicators, it is not difficult to show

$$\mathbb{E}[D_1^2] = (n-1)p(1-p)^{n-2} + (n-1)(n-2)p^2(1-p)^{2n-5},$$

and after noting that $D_1 \leq D_1^2$ almost surely, we can combine the estimates above with Theorem 3.20 to obtain an explicit upper bound between the distribution of W and the standard normal in the Wasserstein metric. In particular, we can read the following result from our work above.

Theorem 3.27. *If X is the number of isolated vertices in an Erdős-Rényi graph $G(n,p)$, $W = (X-\mu)/\sigma$, and for some $1 \leq \alpha < 2$, $\lim_{n\to\infty} n^\alpha p = c \in (0,\infty)$, then*

$$d_{\mathrm{W}}(W,Z) \leq \frac{C}{\sigma},$$

for some constant C.

Proof. The asymptotic hypothesis $\lim_{n\to\infty} n^\alpha p = c \in (0,\infty)$ for some $1 \leq \alpha < 2$ implies that $(1-p)^n$ tends to a finite positive constant. Thus we can see that $\mu \asymp n$, $\sigma^2 \asymp n^{2-\alpha}$, $\operatorname{Var}(\mathbb{E}[X^s - X|X]) \asymp \sigma^2/n^2$, and $\mathbb{E}[(X^s-X)^2] \asymp n^{1-\alpha}$, from which the result follows from Theorem 3.20. □

Example 3.25 can be generalized to counts of vertices of a given degree d at some computational expense [32, 34]; related results pertain to subgraphs counts in an Erdős-Rényi graph [10]. We examine such constructions in greater detail in our treatment of Stein's method for Poisson approximation where the size-bias coupling plays a large role.

3.5. Zero-bias coupling

Our next method of rewriting $\mathbb{E}[Wf(W)]$ to be compared to $\mathbb{E}[f'(W)]$ is through the zero-bias coupling first introduced in [33].

Definition 3.28. For a random variable W with $\mathbb{E}[W] = 0$ and $\mathrm{Var}(W) = \sigma^2 < \infty$, we say the random variable W^z has the *zero-bias* distribution with respect to W if for all absolutely continuous f such that $\mathbb{E}|Wf(W)| < \infty$ we have

$$\mathbb{E}[Wf(W)] = \sigma^2 \mathbb{E}[f'(W^z)].$$

Before discussing existence and properties of the zero-bias distribution, we note that it is appropriate to view the zero-biasing as a distributional transform which has the normal distribution as its unique fixed point. Also note that zero-biasing is our most transparent effort to compare $\mathbb{E}[Wf(W)]$ to $\mathbb{E}[f'(W)]$, culminating in the following result.

Theorem 3.29. *Let W be a mean zero, variance one random variable and let W^z be defined on the same space as W and have the zero-bias distribution with respect to W. If $Z \sim N(0,1)$, then*

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

Proof. Let $\mathcal{F}$ be the set of functions such that$\|f'\| \leq \sqrt{2/\pi}$ and $\|f\|, \|f''\| \leq 2$. Then

$$\begin{aligned} d_{\mathrm{W}}(W,Z) &\leq \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - Wf(W)]| \\ &= \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - f'(W^z)]| \\ &\leq \sup_{f\in\mathcal{F}} \|f''\| \mathbb{E}\left|W - W^z\right|. \end{aligned}$$

□

Before proceeding further, we discuss some fundamental properties of the zero-bias distribution.

Proposition 3.30. *Let W be a random variable such that $\mathbb{E}[W] = 0$ and $\mathrm{Var}(W) = \sigma^2 < \infty$.*

1. *There is a unique probability distribution for a random variable W^z satisfying*

$$\mathbb{E}[Wf(W)] = \sigma^2 \mathbb{E}[f'(W^z)] \tag{3.32}$$

   for all absolutely continuous f such that $\mathbb{E}|Wf(W)| < \infty$.

2. *The distribution of W^z as defined by (3.32) is absolutely continuous with respect to Lebesgue measure with density*

$$p^z(w) = \sigma^{-2}\mathbb{E}\left[W\mathbb{I}[W > w]\right] = -\sigma^{-2}\mathbb{E}\left[W\mathbb{I}[W \leq w]\right]. \tag{3.33}$$

Proof. Assume that $\sigma^2 = 1$; the proof for general σ is similar. We show Items 1 and 2 simultaneously by showing that p^z defined by (3.33) is a probability density which defines a distribution satisfying (3.32).

Let $f(x) = \int_0^x g(t)dt$ for a non-negative function g integrable on compact domains. Then

$$\begin{aligned}\int_0^\infty f'(u)\mathbb{E}[W\mathbb{I}[W > u]]du &= \int_0^\infty g(u)\mathbb{E}[W\mathbb{I}[W > u]]du \\ &= \mathbb{E}[W \int_0^{\max\{0,W\}} g(u)du = \mathbb{E}[Wf(W)\mathbb{I}[W \geq 0]].\end{aligned}$$

and similarly $\int_{-\infty}^0 f'(u)p^z(u)du = \mathbb{E}\left[Wf(W)\mathbb{I}[W \leq 0]\right]$, which implies that

$$\int_{\mathbb{R}} f'(u)p^z(u)du = \mathbb{E}[Wf(W)] \tag{3.34}$$

for all f as above. However, (3.34) extends to all absolutely continuous f such that $\mathbb{E}|Wf(W)| < \infty$ by routine analytic considerations (e.g. considering the positive and negative part of g).

We now show that p^z is a probability density. That p^z is non-negative follows by considering the two representations in (3.33) - note that these representations are equal since $\mathbb{E}[W] = 0$. We also have

$$\int_0^\infty p^z(u)du = \mathbb{E}[W^2\mathbb{I}[W > 0] \text{ and } \int_{-\infty}^0 p^z(u)du = \mathbb{E}[W^2\mathbb{I}[W < 0],$$

so that $\int_{\mathbb{R}} p^z(u)du = \mathbb{E}[W^2] = 1$.

Finally, uniqueness follows since for random variables X and Y such that $\mathbb{E}[f'(X)] = \mathbb{E}[f'(Y)]$ for all continuously differentiable f with compact support (say), then $X \overset{d}{=} Y$. □

The next result shows that little generality is lost in only considering W with $\mathrm{Var}(W) = 1$ as we have done in Theorem 3.29. The result can be read from the density formula above or by a direct computation.

Proposition 3.31. *If W has mean zero and finite variance then $(aW)^z \overset{d}{=} aW^z$.*

3.5.1. Coupling construction

In order to apply Theorem 3.29 to a random variable W, we need to couple W to a random variable having the zero-bias distribution of W. In general, achieving this coupling can be difficult, but we discuss the nicest case where W is a sum of independent random variables and also work out a neat theoretical application using the construction. Another canonical method of construction that is useful in practice can be derived from a Stein pair - see [33].

Let $X_1, \ldots, X_n$ independent random variables having zero mean and such that $\mathrm{Var}(X_i) = \sigma_i^2$, $\sum_{i=1}^n \sigma_i^2 = 1$, and define $W = \sum_{i=1}^n X_i$. We have the following recipe for constructing a zero-bias version of W.

1. For each $i = 1, \dots, n$, let X_i^z have the zero-bias distribution of X_i independent of $(X_j)_{j \neq i}$ and $(X_j^z)_{j \neq i}$.
2. Choose a random summand X_I, where the index I satisfies $\mathbb{P}(I = i) = \sigma_i^2$ and is independent of all else.
3. Define $W^z = \sum_{j \neq I} X_j + X_I^z$.

Proposition 3.32. *Let $W = \sum_{i=1}^n X_i$ be defined as above. If W^z is constructed as per Items 1 - 3 above, then W^z has the zero-bias distribution of W.*

Remark 3.33. Proposition 3.32 only moves the problem of zero-biasing a sum of random variables to zero-biasing its summands. However, in the setting where we expect a CLT to hold, these summands and their zero-bias versions will be small, so that the error of Theorem 3.29 will also be small.

Proof. We must show that $\mathbb{E}[Wf(W)] = \mathbb{E}[f'(W^z)]$ for all appropriate f. Using the definition of zero-biasing in the coordinate X_i and the fact that $W - X_i$ is independent of X_i, we have

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

Since $\sum_{j \neq I} X_j + X_I^z = W - X_I + X_I^z$, the proof is complete. □

3.5.2. Lindeberg-Feller condition

We now discuss the way in which zero-biasing appears naturally in the proof of the Lindeberg-Feller CLT. Our treatment closely follows [31].

Let $(X_{i,n})_{1 \leq n, 1 \leq i \leq n}$ be a triangular array of random variables[4] such that $\mathrm{Var}(X_{i,n}) = \sigma_{i,n}^2 < \infty$. Let $W_n = \sum_{i=1}^n X_{i,n}$, and assume that $\mathrm{Var}(W_n) = 1$. A sufficient condition for W_n to satisfy a CLT as $n \to \infty$ is the Lindeberg condition: for all $\varepsilon > 0$,

$$\sum_{i=1}^n \mathbb{E}[X_{i,n}^2 \mathbb{I}[|X_{i,n}| > \varepsilon]] \to 0, \text{ as } n \to \infty. \tag{3.35}$$

The condition ensures that no single term dominates in the sum so that the limit is not altered by the distribution of a summand. Note that the condition is not necessary as we could take $X_{1,n}$ to be standard normal and the rest of the terms zero. We now have the following result.

Theorem 3.34. *Let $(X_{i,n})_{n \geq 1, 1 \leq i \leq n}$ be the triangular array defined above and let I_n be a random variable independent of the $X_{i,n}$ with $\mathbb{P}(I_n = i) = \sigma_{i,n}^2$. For*

[4] That is, for each n, $(X_{i,n})_{1 \leq i \leq n}$ is a collection of independent random variables.

each $1 \leq i \leq n$, let $X_{i,n}^z$ have the zero-bias distribution of $X_{i,n}$ independent of all else. Then the Lindeberg condition (3.35) holds if and only if

$$X_{I_n,n}^z \xrightarrow{p} 0 \text{ as } n \to \infty. \tag{3.36}$$

From this point, we can use a modification of Theorem 3.29 to prove the following result which also follows from Theorem 3.34 and the classical Lindeberg-Feller CLT mentioned above.

Theorem 3.35. *In the notation of Theorem 3.34 and the remarks directly preceding it, if $X_{I_n,n}^z \to 0$ in probability as $n \to \infty$, then W_n satisfies a CLT.*

Before proving these two results, we note that Theorem 3.35 is heuristically explained by Theorem 3.29 and the zero-bias construction of W_n. Specifically, $|W_n^z - W_n| = |X_{I_n,n}^z - X_{I_n,n}|$ and Theorem 3.29 implies that W_n is approximately normal if this latter quantity is small (in expectation). The proof of Theorem 3.35 uses a modification of the error in Theorem 3.29 and the (non-trivial) fact that $X_{I_n,n}^z \to 0$ in probability implies that $X_{I_n,n} \to 0$ in probability. Finally, the quantity $|X_{i,n}^z - X_{i,n}|$ is also small if $X_{i,n}$ is approximately normal, which indicates that the zero-bias approach applies to show convergence in the CLT for the sum of independent random variables when such a result holds.

Proof of Theorem 3.34. We first perform a preliminary calculation to relate the Lindeberg-Feller condition to the zero-bias quantity of interest. For some fixed $\varepsilon > 0$, let $f'(x) = \mathbb{I}[|x| \geq \varepsilon]$ and $f(0) = 0$. Using that $xf(x) = (x^2 - \varepsilon|x|)\mathbb{I}[|x| \geq \varepsilon]$ and the definition of the zero-bias transform, we find

$$\begin{aligned}
\mathbb{P}(|X_{I_n,n}^z| \geq \varepsilon) &= \sum_{i=1}^n \sigma_{i,n}^2 \mathbb{P}(|X_{i,n}^z| \geq \varepsilon) \\
&= \sum_{i=1}^n \sigma_{i,n}^2 \mathbb{E}[f'(X_{i,n}^z)] \\
&= \sum_{i=1}^n \mathbb{E}\left[(X_{i,n}^2 - \varepsilon|X_{i,n}|)\mathbb{I}[|X_{i,n}| \geq \varepsilon]\right].
\end{aligned}$$

Now we note that

$$\frac{x^2}{2}\mathbb{I}[|x| \geq 2\varepsilon] \leq (x^2 - \varepsilon|x|)\mathbb{I}[|x| \geq \varepsilon] \leq x^2\mathbb{I}[|x| \geq \varepsilon]$$

which implies that for all $\varepsilon > 0$,

$$\frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[X_{i,n}^2 \mathbb{I}[|X_{i,n}| \geq 2\varepsilon]\right] \leq \mathbb{P}(|X_{I_n,n}^z| \geq \varepsilon) \leq \sum_{i=1}^n \mathbb{E}\left[X_{i,n}^2 \mathbb{I}[|X_{i,n}| \geq \varepsilon]\right],$$

so that (3.35) and (3.36) are equivalent. □

Proof of Theorem 3.35. According to the proof of Theorem 3.29, it is enough to show that

$$|\mathbb{E}[f'(W_n) - f'(W_n^z)]| \to 0 \text{ as } n \to \infty \tag{3.37}$$

for all bounded f with two bounded derivatives. We show that $|W_n^z - W_n| \to 0$ in probability which implies (3.37) by the following calculation.

$$\begin{aligned}|\mathbb{E}[f'(W_n) - f'(W_n^z)]| &\leq \mathbb{E}|f'(W_n) - f'(W_n^z)| \\ &= \int_0^\infty \mathbb{P}(|f'(W_n) - f'(W_n^z)| \geq t)dt \\ &= \int_0^{2\|f'\|} \mathbb{P}(|f'(W_n) - f'(W_n^z)| \geq t)dt \\ &\leq \int_0^{2\|f'\|} \mathbb{P}(\|f''\||W_n - W_n^z| \geq t)dt \\ &\leq \int_0^{2\|f'\|} \mathbb{P}(|W_n - W_n^z| \geq t/\|f''\|)dt,\end{aligned}$$

which tends to zero by dominated convergence.

We now must show that $|W_n^z - W_n| \to 0$ in probability. Since we are assuming that $X_{I_n,n}^z \to 0$ in probability, and $|W_n^z - W_n| = |X_{I_n,n}^z - X_{I_n,n}|$, it is enough to show that $X_{I_n,n} \to 0$ in probability. For $\varepsilon > 0$, and $m_n := \max_{1\leq i\leq n} \sigma_{i,n}^2$,

$$\begin{aligned}\mathbb{P}(|X_{I_n,n}| \geq \varepsilon) &\leq \frac{\operatorname{Var}(X_{I_n,n})}{\varepsilon^2} \\ &= \frac{1}{\varepsilon^2}\sum_{i=1}^n \sigma_{i,n}^4 \\ &\leq \frac{m_n}{\varepsilon^2}\sum_{i=1}^n \sigma_{i,n}^2 = \frac{m_n}{\varepsilon^2}.\end{aligned}$$

From this point we show $m_n \to 0$, which completes the proof. For any $\delta > 0$, we have

$$\begin{aligned}\sigma_{i,n}^2 &= \mathbb{E}[X_{i,n}^2\mathbb{I}[|X_{i,n}| \leq \delta]] + \mathbb{E}[X_{i,n}^2\mathbb{I}[|X_{i,n}| > \delta]] \\ &\leq \delta^2 + \mathbb{E}[X_{i,n}^2\mathbb{I}[|X_{i,n}| > \delta]].\end{aligned} \tag{3.38}$$

Using the calculations in the proof of Theorem 3.34 based on the assumption that $X_{I_n,n}^z \to 0$ in probability, it follows that

$$\sum_{i=1}^n \mathbb{E}[X_{i,n}^2\mathbb{I}[|X_{i,n}| > \delta]] \to 0 \text{ as } n \to \infty,$$

so that the second term of (3.38) goes to zero as n goes to infinity uniformly in i. Thus we have that $\limsup_n m_n \leq \delta^2$ for all $\delta > 0$ which implies that $m_n \to 0$ since $m_n > 0$. □

3.6. Normal approximation in the Kolmogorov metric

Our previous work has been to develop bounds on the Wasserstein metric between a distribution of interest and the normal distribution. For W a random variable and Z standard normal, we have the inequality

$$d_{\mathrm{K}}(W,Z) \leq (2/\pi)^{1/4}\sqrt{d_{\mathrm{W}}(W,Z)},$$

so that our previous effort implies bounds for the Kolmogorov metric. However, it is often the case that this inequality is suboptimal - for example if W is a standardized binomial random variable with parameters n and p, then both $d_{\mathrm{K}}(W,Z)$ and $d_{\mathrm{W}}(W,Z)$ are of order $n^{-1/2}$. In this section we develop Stein's method for normal approximation in the Kolmogorov metric in hopes of reconciling this discrepancy.[5] We follow [24] in our exposition below but similar results using related methods appear elsewhere [43, 47, 53].

Recall the following restatement of Corollary 2.3.

Theorem 3.36. *Let Φ denote the standard normal distribution function and let $f_x(w)$ be the unique bounded solution of*

$$f_x'(w) - wf_x(w) = \mathbb{I}[w \leq x] - \Phi(x). \tag{3.39}$$

If W is a random variable with finite mean and Z is standard normal, then

$$d_{\mathrm{K}}(W,Z) = \sup_{x\in\mathbb{R}} |\mathbb{E}[f_x'(W) - Wf_x(W)]|.$$

Moreover, we have the following lemma, which can be read from [24], Lemma 2.3.

Lemma 3.37. *If f_x is the unique bounded solution to (3.39), then*

$$\|f_x\| \leq \sqrt{\frac{\pi}{2}}, \qquad\qquad \|f_x'\| \leq 2,$$

and for all $u, v, w \in \mathbb{R}$,

$$|(w+u)f_x(w+u) - (w+v)f_x(w+v)| \leq (|w| + \sqrt{2\pi}/4)(|u|+|v|).$$

Our program can be summed up in the following corollary to the results above.

Corollary 3.38. *If $\mathcal{F}$ is the set of functions satisfying the bounds of Lemma 3.37 and W is a random variable with finite mean and Z is standard normal, then*

$$d_{\mathrm{K}}(W,Z) \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - Wf(W)]|.$$

[5] Of course improved rates come at the cost of additional hypotheses, but we will see that the theorems are still useful in application.

3.6.1. Zero-bias transformation

To get better ready to use rates using the zero-bias transform, we must assume a boundedness condition.

Theorem 3.39. *Let W be a mean zero, variance one random variable and suppose there is W^z having the zero-bias distribution of W on the same space as W such that $|W^z - W| \leq \delta$ almost surely. If Z is standard normal, then*

$$d_{\mathrm{K}}(W, Z) \leq \left(1 + \frac{1}{\sqrt{2\pi}} + \frac{\sqrt{2\pi}}{4}\right)\delta.$$

Proof. Our strategy of proof is to show that the condition $|W^z - W| \leq \delta$ implies that $|\, d_{\mathrm{K}}(W, Z) - d_{\mathrm{K}}(W^z, Z)|$ is bounded by a constant times δ. From this point we only need to show that $d_{\mathrm{K}}(W^z, Z)$ is of order δ, which is not as difficult due heuristically to the fact that the zero-bias transform is smooth (absolutely continuous with respect to Lebesgue measure).

We implement the first part of the program. For $z \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z) \\
&\qquad \leq \mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z + \delta) + \mathbb{P}(Z \leq z + \delta) - \mathbb{P}(Z \leq z) \\
&\qquad \leq \mathbb{P}(W^z \leq z + \delta) - \mathbb{P}(Z \leq z + \delta) + \frac{\delta}{\sqrt{2\pi}} \\
&\qquad \leq d_{\mathrm{K}}(W^z, Z) + \frac{\delta}{\sqrt{2\pi}}, \qquad (3.40)
\end{aligned}$$

where the second inequality follows since $\{W \leq z\} \subseteq \{W^z \leq z + \delta\}$ and since Z has density bounded by $(2\pi)^{-1/2}$. Similarly,

$$\begin{aligned}
&\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z) \\
&\qquad \geq \mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z - \delta) + \mathbb{P}(Z \leq z - \delta) - \mathbb{P}(Z \leq z) \\
&\qquad \geq \mathbb{P}(W^z \leq z - \delta) - \mathbb{P}(Z \leq z - \delta) - \frac{\delta}{\sqrt{2\pi}},
\end{aligned}$$

which after taking the supremum over z and combining with (3.40) implies that

$$|\, d_{\mathrm{K}}(W, Z) - d_{\mathrm{K}}(W^z, Z)| \leq \frac{\delta}{\sqrt{2\pi}}. \qquad (3.41)$$

Now, by Corollary 3.38 (and using the notation there), we have

$$d_{\mathrm{K}}(W^z, Z) \leq \sup_{f \in \mathcal{F}} \left|\mathbb{E}[f'(W^z) - W^z f(W^z)]\right|, \qquad (3.42)$$

and for $f \in \mathcal{F}$, we find after using the definition of the zero-bias transform and Lemma 3.37

$$\begin{aligned}
|\mathbb{E}[f'(W^z) - W^z f(W^z)]| &= |\mathbb{E}[Wf(W) - W^z f(W^z)]| \\
&\leq \mathbb{E}\left[\left(|W| + \frac{\sqrt{2\pi}}{4}\right)|W^z - W|\right] \\
&\leq \delta\left(1 + \frac{\sqrt{2\pi}}{4}\right). \qquad (3.43)
\end{aligned}$$

Combining (3.41), (3.42), and (3.43) yields the theorem. □

Theorem 3.36 can be applied to sums of independent random variables which are almost surely bounded (note that W bounded implies W^z bounded), and can also be used to derive a bound in Hoeffding's combinatorial CLT under some boundedness assumption [30].

3.6.2. Exchangeable pairs

To get better rates using exchangeable pairs, we again assume a boundedness condition. A slightly more general version of this theorem appears in [53].

Theorem 3.40. *If* (W, W') *is an* a*-Stein pair with* $\mathrm{Var}(W) = 1$ *and such that* $|W' - W| \leq \delta$*, then*

$$d_{\mathrm{K}}(W, Z) \leq \frac{\sqrt{\mathrm{Var}\left(\mathbb{E}[(W'-W)^2|W]\right)}}{2a} + \frac{\delta^3}{2a} + \frac{3\delta}{2}.$$

Proof. Let f_x the bounded solution of (3.39). Using exchangeability and the linearity condition of the a-Stein pair, a calculation which is similar to that used in (3.17) implies

$$\mathbb{E}[Wf_x(W)] = \frac{1}{2a}\mathbb{E}[(W'-W)(f_x(W') - f_x(W))],$$

so that we can see

$$\begin{aligned}
\mathbb{E}[f_x'(W) - Wf_x(W)] &= \mathbb{E}\left[f_x'(W)\left(1 - \frac{(W'-W)^2}{2a}\right)\right] \qquad (3.44) \\
&\quad + \mathbb{E}\left[\frac{W'-W}{2a}\int_0^{W'-W}[f_x'(W) - f_x'(W+t)]\,dt\right]. \qquad (3.45)
\end{aligned}$$

Exactly as in the proof of Theorem 3.9, (the result analogous to Theorem 3.40 but for the Wasserstein metric) the term (3.44) contributes the first error term from the theorem (using the bounds of Lemma 3.37). Now, since f_x satisfies (3.39), we can rewrite (3.45) as

$$\mathbb{E}\left[\frac{W'-W}{2a}\int_0^{W'-W}[Wf_x(W)-(W+t)f_x(W+t)]\,dt\right] \tag{3.46}$$

$$+\,\mathbb{E}\left[\frac{W'-W}{2a}\int_0^{W'-W}[\mathbb{I}[W\le x]-\mathbb{I}[W+t\le x]\,dt\right], \tag{3.47}$$

and we can apply Lemma 3.37 to find that the absolute value of (3.46) is bounded above by

$$\mathbb{E}\left[\frac{|W'-W|}{2a}\int_0^{W'-W}\left(|W|+\frac{\sqrt{2\pi}}{4}\right)|t|dt\right]$$
$$\le \mathbb{E}\left[\frac{|W'-W|^3}{4a}\left(|W|+\frac{\sqrt{2\pi}}{4}\right)\right]\le\frac{\delta^3}{2a}.$$

In order to bound the absolute value of (3.47), we consider separately the cases $W'-W$ positive and negative. For example,

$$\left|\mathbb{E}\left[\frac{(W'-W)\mathbb{I}[W'<W]}{2a}\int_{W'-W}^{0}\mathbb{I}[x<W\le x-t]dt\right]\right|$$
$$\le\frac{1}{2a}\mathbb{E}\left[(W'-W)^2\mathbb{I}[W'<W]\mathbb{I}[x<W\le x+\delta]\right],$$

where we have used that $|W'-W|\le\delta$. A similar inequality can be obtained for $W'>W$ and combining these terms implies that the absolute value of (3.47) is bounded above

$$\frac{1}{2a}\mathbb{E}\left[(W'-W)^2\mathbb{I}[x<W\le x+\delta]\right]. \tag{3.48}$$

Lemma 3.41 below shows (3.48) is bounded above by $3\delta/2$, which proves the theorem. □

Lemma 3.41. *If (W,W') is an a-Stein pair with* $\mathrm{Var}(W)=1$ *and such that $|W'-W|\le\delta$, then for all $x\in\mathbb{R}$*

$$\mathbb{E}\left[(W'-W)^2\mathbb{I}[x<W\le x+\delta]\right]\le 3\delta a.$$

Proof. Let $g'(w)=\mathbb{I}[x-\delta<w\le x+2\delta]$ and $g(x+\delta/2)=0$. Using that $\|g\|\le 3\delta/2$ in the first inequality below, we have

$$\begin{aligned}3\delta a &\ge 2a\mathbb{E}[Wg(W)]\\ &=\mathbb{E}\left[(W'-W)(g(W')-g(W))\right]\\ &=\mathbb{E}\left[(W'-W)\int_0^{W'-W}g'(W+t)dt\right]\\ &\ge\mathbb{E}\left[(W'-W)\int_0^{W'-W}\mathbb{I}[x-\delta<W+t\le x+2\delta]\mathbb{I}[x<W\le x+\delta]dt\right]\\ &=\mathbb{E}\left[(W'-W)^2\mathbb{I}[x<W\le x+\delta]\right],\end{aligned}$$

as desired. □

Theorem 3.40 can be applied to sums of independent random variables which are almost surely bounded, and can also be applied to the anti-voter model to yield rates in the Kolmogorov metric that are comparable to those we obtained in the Wasserstein metric in Section 3.3.1.

4. Poisson approximation

One great advantage of Stein's method is that it can easily be adapted to various distributions and metrics. In this section we develop Stein's method for bounding the total variation distance (see Section 1.2) between a distribution of interest and the Poisson distribution. We move quickly through the material analogous to that of Section 2 for normal approximation, as the general framework is similar. We follow the exposition of [12].

Lemma 4.1. *For $\lambda > 0$, define the functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(k) = \lambda f(k+1) - kf(k).$$

1. *If the random variable Z has the Poisson distribution with mean λ, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all bounded f.*
2. *If for some non-negative integer-valued random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all bounded functions f, then W has the Poisson distribution with mean λ.*

The operator $\mathcal{A}$ is referred to as a characterizing operator of the Poisson distribution.

Before proving the lemma, we state one more result and then its consequence.

Lemma 4.2. *Let $\mathcal{P}_\lambda$ denote probability with respect to a Poisson distribution with mean λ and $A \subseteq \mathbb{N} \cup \{0\}$. The unique solution f_A of*

$$\lambda f_A(k+1) - kf_A(k) = \mathbb{I}[k \in A] - \mathcal{P}_\lambda(A) \tag{4.1}$$

with $f_A(0) = 0$ is given by

$$f_A(k) = \lambda^{-k} e^{\lambda}(k-1)!\left[\mathcal{P}_\lambda(A \cap U_k) - \mathcal{P}_\lambda(A)\mathcal{P}_\lambda(U_k)\right],$$

where $U_k = \{0, 1, \ldots, k-1\}$.

Analogous to normal approximation, this setup immediately yields the following promising result.

Corollary 4.3. *If $W \geq 0$ is an integer-valued random variable with mean λ, then*

$$|\mathbb{P}(W \in A) - \mathcal{P}_\lambda(A)| = |\mathbb{E}[\lambda f_A(W+1) - Wf_A(W)]|\,.$$

Proof of Lemma 4.2. The relation (4.1) defines f_A recursively, so it is obvious that the solution is unique under the boundary condition $f_A(0) = 0$. The fact that the solution is as claimed can be easily verified by substitution into the recursion (4.1). □

Proof of Lemma 4.1. Item 1 follows easily by direct calculation: if $Z \sim \mathrm{Po}(\lambda)$ and f is bounded, then

$$\begin{aligned}\lambda \mathbb{E}[f(Z+1)] &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} f(k+1) \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} (k+1) f(k+1) \\ &= \mathbb{E}[Zf(Z)].\end{aligned}$$

For Item 2, let $\mathbb{E}\mathcal{A}f(W) = 0$ for all bounded functions f. Lemma 4.4 below shows that $f_k \equiv f_{\{k\}}$ is bounded, and then $\mathbb{E}\mathcal{A}f_k(W) = 0$ implies that W has Poisson point probabilities. Alternatively, for $j \in \mathbb{N} \cup \{0\}$, we could take $f(k) = \mathbb{I}[k = j]$ so that the $\mathbb{E}\mathcal{A}f(W) = 0$ implies that

$$\lambda \mathbb{P}(W = j - 1) = j \mathbb{P}(W = j),$$

which is defining since W is a non-negative integer-valued random variable. A third proof can be obtained by taking $f(k) = e^{-uk}$, from which the Laplace transform of W can be derived. □

We now derive useful properties of the solutions f_A of (4.1).

Lemma 4.4. *If f_A solves (4.1), then*

$$\|f_A\| \leq \min\left\{1, \lambda^{-1/2}\right\} \text{ and } \|\Delta f_A\| \leq \frac{1 - e^{-\lambda}}{\lambda} \leq \min\left\{1, \lambda^{-1}\right\}, \tag{4.2}$$

where $\Delta f(k) := f(k+1) - f(k)$.

Proof. The proof of Lemma 4.4 follows from careful analysis. We prove the second assertion and refer to [12] for further details. Upon rewriting

$$f_A(k) = \lambda^{-k}(k-1)! e^{\lambda} \left[\mathcal{P}_\lambda(A \cap U_k)\mathcal{P}_\lambda(U_k^c) - \mathcal{P}_\lambda(A \cap U_k^c)\mathcal{P}_\lambda(U_k)\right],$$

some consideration leads us to observe that for $j \geq 1$, $f_j := f_{\{j\}}$ satisfies

- $f_j(k) \leq 0$ for $k \leq j$ and $f_j(k) \geq 0$ for $k > j$,
- $\Delta f_j(k) \leq 0$ for $k \neq j$, and $\Delta f_j(j) \geq 0$,
- $\Delta f_j(j) \leq \min\left\{j^{-1}, (1 - e^{-\lambda})/\lambda\right\}$.

And also $\Delta f_0(k) < 0$. Since

$$\Delta f_A(k) = \sum_{j \in A} f_j(k)$$

is a sum of terms which are all negative except for at most one, we find

$$\Delta f_A(k) \leq \frac{1 - e^{-\lambda}}{\lambda}. \tag{4.3}$$

Since $f_{A^c} = -f_A$, (4.3) yields the second assertion. □

We can now state our main Poisson approximation theorem which follows from Corollary 4.3 and Lemma 4.4.

Theorem 4.5. *Let $\mathcal{F}$ be the set of functions satisfying (4.2). If $W \geq 0$ is an integer-valued random variable with mean λ and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \sup_{f \in \mathcal{F}} \left| \mathbb{E}[\lambda f(W+1) - W f(W)] \right|. \tag{4.4}$$

We are ready to apply Theorem 4.5 to some examples, but first some remarks. Recall that our main strategy for normal approximation was to find some structure in W, the random variable of interest, that allows us to compare $\mathbb{E}[Wf(W)]$ to $\mathbb{E}[f'(W)]$ for appropriate f. The canonical such structures were

1. Sums of independent random variables.
2. Sums of locally dependent random variables.
3. Exchangeable pairs.
4. Size-biasing.
5. Zero-biasing.

Note that each of these structures essentially provided a way to break down $\mathbb{E}[Wf(W)]$ into a functional of f and some auxiliary random variables. Also, from the form of the Poisson characterizing operator, we want to find some structure in W (the random variable of interest) that allows us to compare $\mathbb{E}[Wf(W)]$ to $\lambda\mathbb{E}[f(W+1)]$ for appropriate f. These two observations imply that the first four items on the list above may be germane to Poisson approximation, which is exactly the program we pursue (since zero-biasing involves f', we won't find use for it in our discrete setting).

4.1. Law of small numbers

It is well known that if $W_n \sim \mathrm{Bi}(n, \lambda/n)$ and $Z \sim \mathrm{Po}(\lambda)$ then $d_{\mathrm{TV}}(W_n, Z) \to 0$ as $n \to \infty$, and it is not difficult to obtain a rate of this convergence. From this fact, it is easy to believe that if $X_1, \ldots, X_n$ are independent indicators with $\mathbb{P}(X_i = 1) = p_i$, then $W = \sum_{i=1}^n X_i$ is approximately Poisson if $\max_i p_i$ is small. In fact, we show the following result.

Theorem 4.6. *Let $X_1, \ldots, X_n$ independent indicators with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_i p_i$. If $Z \sim \mathrm{Po}(\lambda)$, then*

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

Proof. The second inequality is clear and is only included to address the discussion preceding the theorem. For the first inequality, we apply Theorem 4.5.

Let f satisfy (4.2) and note that

$$\begin{aligned}\mathbb{E}[Wf(W)] &= \sum_{i=1}^{n} \mathbb{E}[X_i f(W)] \\ &= \sum_{i=1}^{n} \mathbb{E}[f(W)|X_i = 1]\mathbb{P}[X_i = 1] \\ &= \sum_{i=1}^{n} p_i \mathbb{E}[f(W_i + 1)], \qquad (4.5)\end{aligned}$$

where $W_i = W - X_i$ and (4.5) follows since X_i is independent of W_i. Since $\lambda f(W+1) = \sum_i p_i f(W+1)$, we obtain

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

where the inequality is by rewriting $f(W+1) - f(W_i+1)$ as a telescoping sum of $|W - W_i|$ first differences of f. Combining this last calculation with Theorem 4.5 yields the desired result. □

4.2. Dependency neighborhoods

Analogous to normal approximation, we can generalize Theorem 4.6 to sums of locally dependent variables [4, 5].

Theorem 4.7. *Let $X_1, \ldots, X_n$ indicator variables with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_i p_i$. For each $i = 1, \ldots, n$, let $N_i \subseteq \{1, \ldots, n\}$ such that $i \in N_i$ and X_i is independent of $\{X_j : j \notin N_i\}$. If $p_{ij} := \mathbb{E}[X_i X_j]$ and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda^{-1}\} \left(\sum_{i=1}^{n} \sum_{j \in N_i} p_i p_j + \sum_{i=1}^{n} \sum_{j \in N_i/\{i\}} p_{ij} \right).$$

Remark 4.8. The neighborhoods N_i can be defined with greater flexibility (i.e. dropping the assumption that X_i is independent of the variables not indexed by N_i) at the cost of an additional error term that (roughly) measures dependence (see [4, 5]).

Proof. We want to mimic the proof of Theorem 4.6 up to (4.5), the point where the hypothesis of independence is used. Let f satisfy (4.2), $W_i = W - X_i$, and $V_i = \sum_{j \notin N_i} X_j$. Since $X_i f(W) = X_i f(W_i + 1)$ almost surely, we find

$$\mathbb{E}[\lambda f(W+1) - W f(W)] = \sum_{i=1}^{n} p_i \mathbb{E}[f(W+1) - f(W_i+1)] \tag{4.6}$$

$$+ \sum_{i=1}^{n} \mathbb{E}[(p_i - X_i) f(W_i + 1)] \tag{4.7}$$

As in the proof of Theorem 4.6, the absolute value of (4.6) is bounded above by $\|\Delta f\| \sum_i p_i^2$. Due to the independence of X_i and V_i, and the fact that $\mathbb{E}[X_i] = p_i$, we find that (4.7) is equal to

$$\sum_{i=1}^{n} \mathbb{E}[(p_i - X_i)(f(W_i+1) - f(V_i+1))],$$

so that the absolute value of (4.7) is bounded above by

$$\begin{aligned} \|\Delta f\| \sum_{i=1}^{n} \mathbb{E}\Big[\, |p_i - X_i|\, |W_i - V_i| \,\Big] &\leq \|\Delta f\| \sum_{i=1}^{n} \mathbb{E}\Big[(p_i + X_i) \sum_{j \in N_i / \{i\}} X_j \Big] \\ &= \|\Delta f\| \sum_{i=1}^{n} \sum_{j \in N_i/\{i\}} (p_i p_j + p_{ij}) \,. \end{aligned}$$

Combining these bounds for (4.6) and (4.7) yields the theorem. □

4.2.1. Application: Head runs

In this section we consider an example that arises in an application from biology, that of DNA comparison. We postpone discussion of the details of this relation until the end of the section.

In a sequence of zeros and ones we call a given occurrence of the pattern $\cdots 011 \cdots 10$ (or $11 \cdots 10 \cdots$ or $\cdots 011 \cdots 1$ at the boundaries of the sequence) with exactly k ones a head run of length k. Let W be the number of head runs of length at least k in a sequence of n independent tosses of a coin with head probability p. More precisely, let $Y_1, \ldots, Y_n$ be i.i.d. indicator variables with $\mathbb{P}(Y_i = 1) = p$ and let

$$X_1 = \prod_{j=1}^{k} Y_j,$$

and for $i = 2, \ldots, n-k+1$ let

$$X_i = (1 - Y_{i-1}) \prod_{j=0}^{k-1} Y_{i+j}.$$

Then X_i is the indicator that a run of ones of length at least k begins at position i in the sequence $(Y_1, \ldots, Y_n)$ so that we set $W = \sum_{i=1}^{n-k+1} X_i$. Note that the factor $1-Y_{i-1}$ is used to "de-clump" the runs of length greater than k so that we do not count the same run more than once. At this point we can apply Theorem 4.7 with only a little effort to obtain the following result.

Theorem 4.9. *Let W be the number of head runs of at least length k in a sequence of n independent tosses of a coin with head probability p as defined above. If $\lambda = \mathbb{E}[W] = p^k((n-k)(1-p)+1)$ and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \lambda^2 \frac{2k+1}{n-k+1} + 2\lambda p^k. \tag{4.8}$$

Remark 4.10. Although Theorem 4.9 provides an error for all n, p, k, it can also be interpreted asymptotically as $n \to \infty$ and λ bounded away from zero and infinity. Roughly, if

$$k = \frac{\log(n(1-p))}{\log(1/p)} + c$$

for some constant c, then for fixed p, $\lim_{n\to\infty} \lambda = p^c$. In this case the bound (4.8) is of order $\log(n)/n$.

Proof of Theorem 4.9. As discussed in the remarks preceding the theorem, W has representation as a sum of indicators: $W = \sum_{i=1}^{n-k+1} X_i$. The fact that λ is as stated follows from this representation using that $\mathbb{E}[X_1] = p^k$ and $\mathbb{E}[X_i] = (1-p)p^k$ for $i \neq 1$.

We apply Theorem 4.7 with $N_i = \{1 \leq j \leq n-k+1 : |i-j| \leq k\}$ which clearly has the property that X_i is independent of $\{X_j : j \notin N_i\}$. Moreover, if $j \in N_i/\{i\}$, then $\mathbb{E}[X_i X_j] = 0$ since two runs of length at least k cannot begin within k positions of each other. Theorem 4.7 now implies

$$d_{\mathrm{TV}}(W, Z) \leq \sum_{i=1}^{n} \sum_{j \in N_i} \mathbb{E}[X_i]\mathbb{E}[X_j].$$

It only remains to show that this quantity is bounded above by (4.8) which follows by grouping and counting the terms of the sum into those that contain $\mathbb{E}[X_1]$ and those that do not. □

A related quantity which is of interest in the biological application below is R_n, the length of the longest head run in n independent coin tosses. Due to the equality of events, we have $\mathbb{P}(W=0) = \mathbb{P}(R_n < k)$, so that we can use Remark 4.10 to roughly state

$$\left| \mathbb{P}\left(R_n - \frac{\log(n(1-p))}{\log(1/p)} < x \right) - e^{-p^x} \right| \leq C\left(\frac{\log(n)}{n} \right).$$

The inequality above needs some qualification due to the fact that R_n is integer-valued, but it can be made precise - see [4, 5, 6] for more details.

Theorem 4.9 was relatively simple to derive, but many embellishments are possible which can also be handled similarly, but with more technicalities. For example, for $0 < a \leq 1$, we can define a "quality a" run of length j to be a run of length j with at least aj heads. We could then take W to be the number of quality a runs of length at least k and R_n to be the longest quality a run in n independent coin tosses. A story analogous to that above emerges.

These particular results can also be viewed as elaborations of the classical theorem:

Theorem 4.11 (Erdős-Rényi Law)**.** *If R_n is the longest quality a head run in a sequence of n independent tosses of a coin with head probability p as defined above, then almost surely,*

$$\frac{R_n}{\log(n)} \to \frac{1}{H(a,p)},$$

where for $0 < a < 1$, $H(a,p) = a\log(a/p) + (1-a)\log((1-a)/(1-p))$, and $H(1,p) = \log(1/p)$.

Remark 4.12. Some of the impetus for the results above and especially their embellishments stems from an application in computational biology - see [4, 5, 6, 58] for an entry into this literature. We briefly describe this application here.

DNA is made up of long sequences of the letters A, G, C, and T which stand for certain nucleotides. Frequently it is desirable to know how closely[6] two sequences of DNA are related.

Assume for simplicity that the two sequences of DNA to be compared both have length n. One possible measure of closeness between these sequences is the length of the longest run where the sequences agree when compared coordinatewise. More precisely, if sequence $\mathbf{A}$ is $A_1A_2\cdots A_n$, sequence $\mathbf{B}$ is $B_1B_2\cdots B_n$, and we define $Y_i = \mathbb{I}[A_i = B_i]$, then the measure of closeness between the sequences $\mathbf{A}$ and $\mathbf{B}$ would be the length of the longest run of ones in $(Y_1, \ldots, Y_n)$.

Now, given the sequences $\mathbf{A}$ and $\mathbf{B}$, how long should the longest run be in order to consider them close? The usual statistical setup to handle this question is to assume a probability model under the hypothesis that the sequences are not related, and then compute the probability of the event "at least as long a run" as the observed run. If this probability is low enough, then it is likely that the sequences are closely related (assuming the model is accurate).

We make the simplifying assumption that letters in a sequence of DNA are independently chosen from the alphabet $\{A, G, C, T\}$ under some probability distribution with frequencies p_A, p_G, p_C, and p_T (a more realistic assumption is that letters are generated under some local dependence of nearby letters in the sequence). The hypothesis that the sequences are unrelated corresponds to the sequences being generated *independently* of each other.

[6] For example, whether one sequence could be transformed to the other by few mutations, or whether the two sequences have similar biological function.

In this framework, the distribution of the longest run between two unrelated sequences of DNA of length n is exactly R_n above with

$$p := \mathbb{P}(Y_i = 1) = p_A^2 + p_G^2 + p_C^2 + p_T^2.$$

Thus the work above can be used to approximate tail probabilities of the longest run length under the assumption that two sequences of DNA are unrelated, which can then be used to determine the likelihood of the observed longest run lengths.

In practice, there are different methods of obtaining approximate tail probabilities or "p-values" to determine the likelihood that two sequences are closely related. Those most related to our work above are called *alignment free* [46, 57] and basically count the number of common words of various lengths occurring at different positions in the two sequences. Our Poisson approximation result above is not directly relevant since we counted the number of common words of a fixed length starting from the same position in the two sequences. However, embellishments of the Stein's method argument above can be fruitfully applied to these statistics in some settings [39].

Another method of DNA sequence comparison aligns the two sequences by optimizing a "score" function on alignments. In this setting, the main issues are in determining algorithms that align two sequences of DNA in an optimal way under various score functions and deriving the associated tail probabilities (p-values) for such alignments. The results above and those related to them do not rigorously apply to the typical methods used for such issues (e.g. BLAST), but they have provided heuristic guidance. For a more thorough discussion of the interplay between theory and practice in this setting, see Chapters 8, 9, and 11 of [58].

4.3. Size-bias coupling

The most powerful method of rewriting $\mathbb{E}[Wf(W)]$ so that it can be usefully compared to $\mathbb{E}[W]\mathbb{E}[f(W+1)]$ is through the size-bias coupling already defined in Section 3.4 - recall the relevant definitions and properties there. The book [12] is almost entirely devoted to Poisson approximation through the size-bias coupling (although that terminology is not used), so we spend some time fleshing out their powerful and general results.

Theorem 4.13. *Let $W \geq 0$ be an integer-valued random variable such that $\mathbb{E}[W] = \lambda > 0$ and let W^s be a size-bias coupling of W. If $Z \sim \mathrm{Po}(\lambda)$, then*

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

Proof. Let f bounded and $\|\Delta f\| \leq \min\{1, \lambda^{-1}\}$. Then

$$\begin{aligned} |\mathbb{E}[\lambda f(W+1) - Wf(W)]| &= \lambda\,|\mathbb{E}[f(W+1) - f(W^s)]| \\ &\leq \lambda\|\Delta f\|\mathbb{E}|W + 1 - W^s|, \end{aligned}$$

where we have used the definition of the size-bias distribution and rewritten $f(W+1) - f(W^s)$ as a telescoping sum of $|W + 1 - W^s|$ terms. □

Due to the canonical "law of small numbers" for Poisson approximation, we are mostly concerned with approximating a sum of indicators by a Poisson distribution. Recall the following construction of a size-bias coupling from Section 3.4, and useful special case.

Corollary 4.14. *Let $X_1, \ldots, X_n$ be indicator variables with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_i p_i$. If for each $i = 1, \ldots, n$, $(X_j^{(i)})_{j \neq i}$ has the distribution of $(X_j)_{j \neq i}$ conditional on $X_i = 1$ and I is a random variable independent of all else such that $\mathbb{P}(I = i) = p_i/\lambda$, then $W^s = \sum_{j \neq I} X_j^{(I)} + 1$ has the size-bias distribution of X.*

Corollary 4.15. *Let $X_1, \ldots, X_n$ be exchangeable indicator variables and let $(X_j^{(1)})_{j \neq 1}$ have the distribution of $(X_j)_{j \neq 1}$ conditional on $X_1 = 1$. If we define $X = \sum_{i=1}^n X_i$, then $X^s = \sum_{j \neq 1} X_j^{(1)} + 1$ has the size-bias distribution of X.*

Proof. Corollary 4.14 was proved in Section 3.4 and Corollary 4.15 follows from the fact that exchangeability first implies that I is uniform and also that $\sum_{j \neq i} X_j^{(i)} + X_i^s \overset{d}{=} \sum_{j \neq 1} X_j^{(1)} + X_1^s$. □

Example 4.16 (Law of small numbers)**.** Let $W = \sum_{i=1}^n X_i$ where the X_i are independent indicators with $\mathbb{P}(X_i = 1) = p_i$. According to Corollary 4.14, in order to size-bias W, we first choose an index I with $\mathbb{P}(I = i) = p_i/\lambda$, where $\lambda = \mathbb{E}[W] = \sum_i p_i$. Given $I = i$ we construct $X_j^{(i)}$ having the distribution of X_j conditional on $X_i = 1$. However, by independence, $(X_j^{(i)})_{j \neq i}$ has the same distribution as $(X_j)_{j \neq i}$ so that we can take $W^s = \sum_{j \neq I} X_j + 1$. Applying Theorem 4.13 we find that for $Z \sim \mathrm{Po}(\lambda)$,

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda\} \mathbb{E}[X_I] = \min\{1, \lambda\} \sum_{i=1}^n \frac{p_i}{\lambda} \mathbb{E}[X_i] = \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i^2,$$

which agrees with our previous bound for this example.

Example 4.17 (Isolated Vertices)**.** Let W be the number of isolated vertices in an Erdős-Rényi random graph on n vertices with edge probabilities p. Note that $W = \sum_{i=1}^n X_i$, where X_i is the indicator that vertex v_i (in some arbitrary but fixed labeling) is isolated. We constructed a size-bias coupling of W in Section 3.4 using Corollary 4.14, and we can simplify this coupling by using Corollary 4.15[7] as follows.

We first generate an Erdős-Rényi random graph G, and then erase all edges connected to vertex v_1. Then take $X_j^{(1)}$ be the indicator that vertex v_j is isolated in this new graph. By the independence of the edges in the graph, it is clear that $(X_j^{(1)})_{j \neq 1}$ has the distribution of $(X_j)_{j \neq 1}$ conditional on $X_1 = 1$, so that by Corollary 4.15, we can take $W^s = \sum_{j \neq 1} X_j^{(1)} + 1$ and of course we take W to be the number of isolated vertices in G.

[7] This simplification would not have yielded a useful error bound in Section 3.4 since the size-bias normal approximation theorem contains a variance term; there the randomization provides an extra factor of $1/n$.

In order to apply Theorem 4.13, we only need to compute $\lambda = \mathbb{E}[W]$ and $\mathbb{E}|W+1-W^s|$. From Example 3.26 in Section 3.4, $\lambda = n(1-p)^{n-1}$ and from the construction above

$$\begin{aligned}\mathbb{E}|W+1-W^s| &= \mathbb{E}\left|X_1 + \sum_{j=2}^{n} X_j - X_j^{(1)}\right| \\ &= \mathbb{E}[X_1] + \sum_{j=2}^{n} \mathbb{E}\left[X_j^{(1)} - X_j\right],\end{aligned}$$

where we use the fact that $X_j^{(1)} \geq X_j$ which follows since we can only increase the number of isolated vertices by erasing edges. Thus, $X_j^{(1)} - X_j$ is equal to zero or one and the latter happens only if vertex v_j has degree one and is connected to v_1 which occurs with probability $p(1-p)^{n-2}$. Putting this all together in Theorem 4.13, we obtain the following.

Proposition 4.18. *Let W the number of isolated vertices in an Erdős-Rényi random graph and $\lambda = \mathbb{E}[W]$. If $Z \sim \mathrm{Po}(\lambda)$, then*

$$\begin{aligned}d_{\mathrm{TV}}(W,Z) &\leq \min\{1,\lambda\}\left((n-1)p(1-p)^{n-2} + (1-p)^{n-1}\right) \\ &\leq \min\{\lambda,\lambda^2\}\left(\frac{p}{1-p} + \frac{1}{n}\right).\end{aligned}$$

To interpret this result asymptotically, if λ is to stay away from zero and infinity as n gets large, p must be of order $\log(n)/n$, in which case the error above is of order $\log(n)/n$.

Example 4.19 (Degree d vertices)**.** We can generalize Example 4.17 by taking W to be the number of degree $d \geq 0$ vertices in an Erdős-Rényi random graph on n vertices with edge probabilities p. Note that $W = \sum_{i=1}^{n} X_i$, where X_i is the indicator that vertex v_i (in some arbitrary but fixed labeling) has degree d. We can construct a size-bias coupling of W by using Corollary 4.15 as follows. Let G be an Erdős-Rényi random graph.

- If the degree of vertex v_1 is $d_1 \geq d$, then erase $d_1 - d$ edges chosen uniformly at random from the d_1 edges connected to v_1.
- If the degree of vertex v_1 is $d_1 < d$, then add edges from v_1 to the $d - d_1$ vertices not connected to v_1 chosen uniformly at random from the $n - d_1 - 1$ vertices unconnected to v_1.

Let $X_j^{(1)}$ be the indicator that vertex v_j has degree d in this new graph. By the independence of the edges in the graph, it is clear that $(X_j^{(1)})_{j\neq 1}$ has the distribution of $(X_j)_{j\neq 1}$ conditional on $X_1 = 1$, so that by Corollary 4.15, we can take $W^s = \sum_{j\neq 1} X_j^{(1)} + 1$ and of course we take W to be the number of isolated vertices in G.

Armed with this coupling, we could apply Theorem 4.13 to yield a bound in the variation distance between W and a Poisson distribution. However, the analysis for this particular example is a bit technical, so we refer to Section 5.2 of [12] for the details.

4.3.1. Increasing size-bias couplings

A crucial simplification occurred in Example 4.17 because the size-bias coupling was increasing in a certain sense. The following result quantifies this simplification.

Theorem 4.20. *Let $X_1, \ldots, X_n$ be indicator variables with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_i p_i$. For each $i = 1, \ldots, n$, let $(X_j^{(i)})_{j \neq i}$ have the distribution of $(X_j)_{j \neq i}$ conditional on $X_i = 1$ and let I be a random variable independent of all else, such that $\mathbb{P}(I = i) = p_i/\lambda$ so that $W^s = \sum_{j \neq I} X_j^{(I)} + 1$ has the size-bias distribution of W. If $X_j^{(i)} \geq X_j$ for all $i \neq j$, and $Z \sim \mathrm{Po}(\lambda)$, then*

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

Proof. Let $W_i = \sum_{j \neq i} X_j^{(i)} + 1$. From Theorem 4.13 and the size-bias construction of Corollary 4.14, we have

$$\begin{aligned} d_{\mathrm{TV}}(W, Z) &\leq \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}|W + 1 - W_i| \\ &= \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}\left[\sum_{j \neq i} \left(X_j^{(i)} - X_j \right) + X_i \right] \\ &= \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}\left[W_i - W - 1 + 2X_i \right], \end{aligned} \tag{4.9}$$

where the penultimate equality uses the monotonicity of the size-bias coupling. Using again the construction of the size-bias coupling we obtain that (4.9) is equal to

$$\min\{1, \lambda^{-1}\} \left(\lambda \mathbb{E}[W^s] - \lambda^2 - \lambda + 2 \sum_{i=1}^n p_i^2 \right),$$

which yields the desired inequality by the definition of the size-bias distribution. □

4.3.2. Application: Subgraph counts

Let $G = G(n, p)$ be an Erdős-Rényi random graph on n vertices with edge probability p and let H be a graph on $0 < v_H \leq n$ vertices with e_H edges and no isolated vertices. We want to analyze the number of copies of H in G; that is, the number of subgraphs of the complete graph on n vertices which are isomorphic to H which appear in G. For example, we could take H to be a triangle so that $v_H = e_H = 3$.

Let Γ be the set of all copies of H in K_n, the complete graph on n vertices and for $\alpha \in \Gamma$, let X_α be the indicator that there is a copy of H in G at α and set $W = \sum_{\alpha\in\Gamma} X_\alpha$. We now have the following result.

Theorem 4.21. *Let H be a graph with no isolated vertices and W be the number of copies H in G as defined above and let $\lambda = \mathbb{E}[W]$. If H has e_H edges and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda^{-1}\} \left(\mathrm{Var}(W) - \lambda + 2\lambda p^{e_H}\right).$$

Proof. We show that Theorem 4.20 applies to W. Since $W = \sum_{\alpha\in\Gamma} X_\alpha$ is a sum of exchangeable indicators, we can apply Corollary 4.15 to construct a size-bias coupling of W. To this end, for a fixed $\alpha \in \Gamma$, let $X_\beta^{(\alpha)}$ be the indicator that there is a copy of H in $G \cup \{\alpha\}$ at β. Here, $G \cup \{\alpha\}$ means we add the minimum edges necessary to G to have a copy of H at α. The following three evident facts now imply the theorem:

1. $(X_\beta^{(\alpha)})_{\beta\neq\alpha}$ has the distribution of $(X_\beta)_{\beta\neq\alpha}$ given that $X_\alpha = 1$.
2. For all $\beta \in \Gamma/\{\alpha\}$, $X_\beta^{(\alpha)} \geq X_\beta$.
3. $\mathbb{E}[X_\alpha] = p^{e_H}$.

□

Theorem 4.21 is a very general result, but it can be difficult to interpret. That is, what properties of a subgraph H make W approximately Poisson? We can begin to answer that question by expressing the mean and variance of W in terms of properties of H which yields the following.

Corollary 4.22. *Let W be the number of copies of a graph H with no isolated vertices in G as defined above and let $\lambda = \mathbb{E}[W]$. For fixed $\alpha \in \Gamma$, let $\Gamma_\alpha^t \subseteq \Gamma$ be the set of subgraphs of K_n isomorphic to H with exactly t edges not in α. If H has e_H edges and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda\} \left(p^{e_H} + \sum_{t=1}^{e_H-1} |\Gamma_\alpha^t| \left(p^t - p^{e_H}\right)\right).$$

Proof. The corollary follows after deriving the mean and variance of W. The terms $|\Gamma_\alpha^t|$ account for the number of covariance terms for different types of pairs of indicators. In detail,

$$\begin{aligned}\mathrm{Var}(W) &= \sum_{\alpha\in\Gamma} \mathrm{Var}(X_\alpha) + \sum_{\alpha\in\Gamma}\sum_{\beta\neq\alpha} \mathrm{Cov}(X_\alpha, X_\beta) \\ &= \lambda(1 - p^{e_H}) + \sum_{\alpha\in\Gamma}\sum_{t=1}^{e_H}\sum_{\beta\in\Gamma_\alpha^t} \mathrm{Cov}(X_\alpha, X_\beta)\end{aligned}$$

$$= \lambda(1 - p^{e_H}) + \sum_{\alpha \in \Gamma} p^{e_H} \sum_{t=1}^{e_H} \sum_{\beta \in \Gamma_\alpha^t} \left(\mathbb{E}[X_\beta | X_\alpha = 1] - p^{e_H}\right)$$
$$= \lambda \left(1 - p^{e_H} + \sum_{t=1}^{e_H - 1} |\Gamma_\alpha^t| \left(p^t - p^{e_H}\right) \right),$$

since $\lambda = \sum_{\alpha \in \Gamma} p^{e_H}$ and for $\beta \in \Gamma_\alpha^t$, $\mathbb{E}[X_\beta | X_\alpha = 1] = p^t$. □

It is possible to rewrite the error in other forms which can be used to make some general statements (see [12], Chapter 5), but we content ourselves with some examples.

Example 4.23 (Triangles)**.** Let H be a triangle. In this case, $e_H = 3$, $|\Gamma_\alpha^2| = 3(n-3)$, and $|\Gamma_\alpha^1| = 0$ since triangles either share one edge or all three edges (corresponding to $t = 2$ and $t = 0$). Thus Corollary 4.22 implies that for W the number of triangles in G and Z an appropriate Poisson variable,

$$d_{\mathrm{TV}}(W, Z) \le \min\{1, \lambda\} \left(p^3 + 3(n-3)p^2(1-p)\right). \tag{4.10}$$

Since $\lambda = \binom{n}{3}p^3$ we can view (4.10) as an asymptotic result with p of order $1/n$. In this case, (4.10) is of order $1/n$.

Example 4.24 (k-cycles)**.** More generally, we can let H be a k-cycle (a triangle is 3-cycle). Now note that for some constants c_t and C_k,

$$|\Gamma_\alpha^t| \le \binom{k}{k-t} c_t n^{t-1} \le C_k n^{t-1},$$

since we choose the $k - t$ edges shared in the k-cycle α, and then we have order n^{t-1} sequences of vertices to create a cycle with t edges outside of the $k - t$ edges shared with α. The second equality follows by maximizing $\binom{k}{k-t}c_t$ over the possible values of t. We can now find for W the number of k-cycles in G and Z an appropriate Poisson variable,

$$d_{\mathrm{TV}}(W, Z) \le \min\{1, \lambda\} \left(p^k + C_k p \sum_{t=1}^{k-1} (np)^{t-1} \right). \tag{4.11}$$

To interpret this bound asymptotically, we note that $\lambda = |\Gamma| p^k$ and

$$|\Gamma| = \binom{n}{k} \frac{(k-1)!}{2} p^k,$$

since the number of non-isomorphic k-cycles on K_k is $k!/(2k)$ (since $k!$ is the number of permutations of the vertices, which over counts by a factor of $2k$ due to reflections and rotations). Thus λ is of order $(np)^k$ for fixed k so that we take p to be of order $1/n$ and in this regime (4.11) is of order $1/n$.

Similar results can be derived for *induced* and *isolated* subgraph counts - again we refer to [12], Chapter 5.

4.3.3. Implicit coupling

In this section we show that it can be possible to apply Theorem 4.20 without constructing the size-bias coupling explicitly. We first need some terminology.

Definition 4.25. We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is increasing (decreasing) if for all $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ such that $x_i \leq y_i$ for all $i = 1, \ldots, n$, we have $f(\mathbf{x}) \leq f(\mathbf{y})$ $(f(\mathbf{x}) \geq f(\mathbf{y}))$.

Theorem 4.26. *Let $\boldsymbol{Y} = (Y_i)_{j=1}^N$ be a finite collection of independent indicators and assume $X_1, \ldots, X_n$ are increasing or decreasing functions from $\{0,1\}^N$ into $\{0,1\}$. If $W = \sum_{i=1}^n X_i(\boldsymbol{Y})$ and $\mathbb{E}[W] = \lambda$, then*

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

Proof. We show that there exists a size-bias coupling of W satisfying the hypotheses of Theorem 4.20, which implies the result. From Lemma 4.27 below, it is enough to show that $\mathrm{Cov}(X_i(\mathbf{Y}), \phi \circ \mathbf{X}(\mathbf{Y})) \geq 0$ for all increasing indicator functions ϕ. However, since each $X_i(\mathbf{Y})$ is an increasing or decreasing function applied to independent indicators, then so is $\phi \circ \mathbf{X}(\mathbf{Y})$. Thus we may apply the FKG inequality (see Chapter 2 of [38]) which in this case states

$$\mathbb{E}[X_i(\mathbf{Y})\phi \circ \mathbf{X}(\mathbf{Y})] \geq \mathbb{E}[X_i(\mathbf{Y})]\mathbb{E}[\phi \circ \mathbf{X}(\mathbf{Y})],$$

as desired. □

Lemma 4.27. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of indicator variables and let $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)}) \stackrel{d}{=} \boldsymbol{X} | X_i = 1$. Then the following are equivalent:*

1. *There exists a coupling such that $X_j^{(i)} \geq X_j$.*
2. *For all increasing indicator functions ϕ, $\mathbb{E}[\phi(\boldsymbol{X}^{(i)})] \geq \mathbb{E}[\phi(\boldsymbol{X})]$.*
3. *For all increasing indicator functions ϕ, $\mathrm{Cov}(X_i, \phi(\boldsymbol{X})) \geq 0$.*

Proof. The equivalence 1⇔2 follows from a general version of Strassen's theorem which can be found in [38]. In one dimension, Strassen's theorem says that there exists a coupling of random variables X and Y such that $X \geq Y$ if and only if $F_X(z) \leq F_Y(z)$ for all $z \in \mathbb{R}$ where F_X and F_Y are distribution functions.

The equivalence 2⇔3 follows from the following calculation.

$$\mathbb{E}[\phi(\mathbf{X}^{(i)})] = \mathbb{E}[\phi(\mathbf{X}) | X_i = 1] = \mathbb{E}[X_i \phi(\mathbf{X}) | X_i = 1] = \frac{\mathbb{E}[X_i \phi(\mathbf{X})]}{\mathbb{P}(X_i = 1)}.$$

□

Example 4.28 (Subgraph counts)**.** Theorem 4.26 applies to the example of Section 4.3.2, since the indicator of a copy of H at a given location is an increasing function of the edge indicators of the graph G.

Example 4.29 (Large degree vertices)**.** Let $d \geq 0$ and let W be the number of vertices with degree at least d. Clearly W is a sum of indicators that are increasing functions of the edge indicators of the graph G so that Theorem 4.26 can be applied. After some technical analysis, we arrive at the following result - see Section 5.2 of [12] for details. If $q_1 = \sum_{k \geq d} \binom{n-1}{k} p^k (1-p)^{n-k-1}$ and $Z \sim \mathrm{Po}(nq_1)$, then

$$d_{\mathrm{TV}}(W, Z) \leq q_1 + \frac{d^2(1-p)\left[\binom{n-1}{d} p^d (1-p)^{n-d-1}\right]^2}{(n-1)pq_1}.$$

Example 4.30 (Small degree vertices)**.** Let $d \geq 0$ and let W be the number of vertices with degree at most d. Clearly W is a sum of indicators that are decreasing functions of the edge indicators of the graph G so that Theorem 4.26 can be applied. After some technical analysis, we arrive at the following result - see Section 5.2 of [12] for details. If $q_2 = \sum_{k \leq d} \binom{n-1}{k} p^k (1-p)^{n-k-1}$ and $Z \sim \mathrm{Po}(nq_2)$, then

$$d_{\mathrm{TV}}(W, Z) \leq q_2 + \frac{(n-d-1)^2 p\left[\binom{n-1}{d} p^d (1-p)^{n-d-1}\right]^2}{(n-1)(1-p)q_2}.$$

4.3.4. Decreasing size-bias couplings

In this section we prove and apply a result complementary to Theorem 4.20.

Theorem 4.31. *Let $X_1, \ldots, X_n$ be indicator variables with $\mathbb{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^n X_i$, and $\lambda = \mathbb{E}[W] = \sum_i p_i$. For each $i = 1, \ldots, n$, let $(X_j^{(i)})_{j \neq i}$ have the distribution of $(X_j)_{j \neq i}$ conditional on $X_i = 1$ and let I be a random variable independent of all else, such that $\mathbb{P}(I = i) = p_i/\lambda$ so that $W^s = \sum_{j \neq I} X_j^{(I)} + 1$ has the size-bias distribution of W. If $X_j^{(i)} \leq X_j$ for all $i \neq j$, and $Z \sim \mathrm{Po}(\lambda)$, then*

$$d_{\mathrm{TV}}(W, Z) \leq \min\{1, \lambda\} \left(1 - \frac{\mathrm{Var}(W)}{\lambda}\right).$$

Proof. Let $W_i = \sum_{j \neq i} X_j^{(i)} + 1$. From Theorem 4.13 and the size-bias construction of Corollary 4.14, we have

$$\begin{aligned} d_{\mathrm{TV}}(W, Z) &\leq \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}|W + 1 - W_i| \\ &= \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}\left[\sum_{j \neq i} \left(X_j - X_j^{(i)}\right) + X_i\right] \\ &= \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i \mathbb{E}\left[W - W_i + 1\right], \qquad (4.12) \end{aligned}$$

where the penultimate equality uses the monotonicity of the size-bias coupling. Using again the construction of the size-bias coupling we obtain that (4.12) is equal to

$$\min\{1, \lambda^{-1}\} \left(\lambda^2 - \lambda\mathbb{E}[W^s] + \lambda\right),$$

which yields the desired inequality by the definition of the size-bias distribution. □

Example 4.32 (Hypergeometric distribution)**.** Suppose we have N balls in an urn in which $1 \le n \le N$ are colored red and we draw $1 \le m \le N$ balls uniformly at random without replacement so that each of the $\binom{N}{m}$ subsets of balls is equally likely. Let W be the number of red balls. It is well known that if N is large and m/N is small, then W has approximately a binomial distribution since the dependence diminishes. Thus, we would also expect W to be approximately Poisson distributed if in addition, n/N is small. We can use Theorem 4.31 to make this heuristic precise.

Label the red balls in the urn arbitrarily and let X_i be the indicator that ball i is chosen in the m-sample so that we have the representation $W = \sum_{i=1}^n X_i$. Since the X_i are exchangeable, we can use Corollary 4.15 to size-bias W. If the ball labelled one already appears in the m-sample then do nothing, otherwise, we force $X_1^s = 1$ by adding the ball labelled one to the sample and putting back a ball chosen uniformly at random from the initial m-sample. If we let $X_i^{(1)}$ be the indicator that ball i is in the sample after this procedure, then it is clear that $(X_i^{(1)})_{i\ge 2}$ has the distribution of $(X_i)_{i\ge 2}$ conditional on $X_1 = 1$ so that we can take $W^s = \sum_{i\ge 2} X_i^{(1)} + 1$.

In the construction of W^s above, no additional red balls labeled $2, \ldots, n$ can be added to the m-sample. Thus $X_i^{(1)} \le X_i$ for $i \ge 2$, and we can apply Theorem 4.31. A simple calculation yields

$$\mathbb{E}[W] = \frac{nm}{N} \quad \text{and} \quad \mathrm{Var}(W) = \frac{nm(N-n)(n-m)}{N^2(N-1)},$$

so that for $Z \sim \mathrm{Po}(nm/N)$,

$$d_{\mathrm{TV}}(W, Z) \le \min\left\{1, \frac{nm}{N}\right\} \left(\frac{n}{N-1} + \frac{m}{N-1} - \frac{nm}{N(N-1)} - \frac{1}{N-1}\right).$$

Remark 4.33. This same bound can be recovered after noting the well known fact that W has the representation as a sum of *independent* indicators (see [44]) which implies that Theorem 4.31 can be applied.

Example 4.34 (Coupon collecting)**.** Assume that a certain brand of cereal puts a toy in each carton. There are n distinct types of toys and each week you pick up a carton of this cereal from the grocery store in such a way as to receive a uniformly random type of toy, independent of the toys received previously. The classical coupon collecting problem asks the question of how many cartons of cereal you must pick up in order to have received all n types of toys.

We formulate the problem as follows. Assume you have n boxes and k balls are tossed independently into these boxes uniformly at random. Let W be the number of empty boxes after tossing all k balls into the boxes. Viewing the n boxes as types of toys and the k balls as cartons of cereal, it is easy to see that the event $\{W = 0\}$ corresponds to the event that k cartons of cereal are sufficient to receive all n types of toys. We use Theorem 4.31 to show that W is approximately Poisson which yields an estimate with error for the probability of this event.

Let X_i be the indicator that box i (under some arbitrary labeling) is empty after tossing the k balls so that $W = \sum_{i=1}^n X_i$. Since the X_i are exchangeable, we can use Corollary 4.15 to size-bias W by first setting $X_1^s = 1$ by emptying box 1 (if it is not already empty) and then redistributing the balls in box 1 uniformly among boxes 2 through n. If we let $X_i^{(1)}$ be the indicator that box i is empty after this procedure, then it is clear that $(X_i^{(1)})_{i\geq 2}$ has the distribution of $(X_i)_{i\geq 2}$ conditional on $X_1 = 1$ so that we can take $W^s = \sum_{i\geq 2} X_i^{(1)} + 1$.

In the construction of W^s above we can only add balls to boxes 2 through n, which implies $X_i^{(1)} \leq X_i$ for $i \geq 2$ so that we can apply Theorem 4.31. In order to apply the theorem we only need to compute the mean and variance of W. First note that $\mathbb{P}(X_i = 1) = ((n-1)/n)^k$ so that

$$\lambda := \mathbb{E}[W] = n\left(1 - \frac{1}{n}\right)^k,$$

and also that for $i \neq j$, $\mathbb{P}(X_i = 1, X_j = 1) = ((n-2)/n)^k$ so that

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

Using these calculations in Theorem 4.31 yields that for $Z \sim \mathrm{Po}(\lambda)$,

$$\begin{aligned} &d_{\mathrm{TV}}(W, Z) \\ &\leq \min\{1, \lambda\}\left(\left(1 - \frac{1}{n}\right)^k + (n-1)\left[\left(1 - \frac{1}{n}\right)^k - \left(1 - \frac{1}{n-1}\right)^k\right]\right). \end{aligned} \tag{4.13}$$

In order to interpret this result asymptotically, let $k = n\log(n) - cn$ for some constant c so that $\lambda = e^c$ is bounded away from zero and infinity as $n \to \infty$. In this case (4.13) is asymptotically of order

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

and since for $0 < a \leq 1$, $1 - a^x \leq -\log(a)x$ and also $\log(1+x) \leq x$, we find

$$1 - \left(\frac{n(n-2)}{(n-1)^2}\right)^k \leq k\log\left(\frac{(n-1)^2}{n(n-2)}\right) \leq \frac{k}{n(n-2)} = \frac{\log(n) - c}{(n-2)},$$

which implies

$$d_{\mathrm{TV}}(W,Z) \le C\frac{\log(n)}{n}.$$

Example 4.35 (Coupon collecting continued)**.** We can embellish the coupon collecting problem of Example 4.34 in a number of ways; recall the notation and setup there. For example, rather than distribute the balls into the boxes uniformly and independently, we could distribute each ball independently according to some probability distribution, say p_i is the chance that a ball goes into box i with $\sum_{i=1}^n p_i = 1$. Note that Example 4.34 had $p_i = 1/n$ for all i.

Let X_i be the indicator that box i (under some arbitrary labeling) is empty after tossing k balls and $W = \sum_{i=1}^n X_i$. In this setting, the X_i are not necessarily exchangeable so that we use Corollary 4.14 to construct W^s. First we compute

$$\lambda := \mathbb{E}[W] = \sum_{i=1}^{n}(1-p_i)^k$$

and let I be a random variable such that $\mathbb{P}(I=i) = (1-p_i)^k/\lambda$. Corollary 4.14 now states that in order to construct W^s, we empty box I (forcing $X_I^s = 1$) and then redistribute the balls that were removed into the remaining boxes independently and with chance of landing in box j equal to $p_j/(1-p_I)$. If we let $X_j^{(I)}$ be the indicator that box j is empty after this procedure, then it is clear that $(X_j^{(I)})_{j\neq I}$ has the distribution of $(X_j)_{j\neq I}$ conditional on $X_I = 1$ so that we can take $W^s = \sum_{j\neq I} X_j^{(I)} + 1$.

Analogous to Example 4.34, $X_j^{(i)} \le X_j$ for $j \neq i$ so that we can apply Theorem 4.31. The mean and variance are easily computed, but not as easily interpreted. A little analysis (see [12] Section 6.2) yields that for $Z \sim \mathrm{Po}(\lambda)$,

$$d_{\mathrm{TV}}(W,Z) \le \min\{1,\lambda\}\left[\max_i (1-p_i)^k + \frac{k}{\lambda}\left(\frac{\lambda\log(k)}{k-\log(\lambda)} + \frac{4}{k}\right)^2\right].$$

Remark 4.36. We could also consider the number of boxes with at most $m \ge 0$ balls; we just studied the case $m = 0$. The coupling is still decreasing because it can be constructed by first choosing a box randomly and if it has greater than m balls in it, redistributing a random number of them among the other boxes. Thus Theorem 4.31 can be applied and an error in the Poisson approximation can be obtained in terms of the mean and the variance. We again refer to Chapter 6 of [12] for the details.

Finally, one could also consider the number of boxes containing exactly $m \ge 0$ balls, or containing at least $m \ge 1$ balls; Theorem 4.20 applies to the latter problem. See Chapter 6 of [12].

4.4. Exchangeable pairs

In this section, we develop Stein's method for Poisson approximation using exchangeable pairs as detailed in [20]. The applications for this theory are not as developed as that of dependency neighborhoods and size-biasing, but the method fits well into our framework and the ideas here prove useful elsewhere [50].

As we have done for dependency neighborhoods and size-biasing, we could develop exchangeable pairs for Poisson approximation by following our development for normal approximation which involved rewriting $\mathbb{E}[Wf(W)]$ as the expectation of a term involving the exchangeable pair and f. However, this approach is not as useful as a different one which has the added advantage of removing the a-Stein pair linearity condition.

Theorem 4.37. *Let W be a non-negative integer valued random variable and let (W, W') be an exchangeable pair. If $\mathcal{F}$ is a sigma-field with $\sigma(W) \subseteq \mathcal{F}$, and $Z \sim \mathrm{Po}(\lambda)$, then for all $c \in \mathbb{R}$,*

$$\begin{aligned}&d_{\mathrm{TV}}(W, Z)\\&\le \min\{1, \lambda^{-1/2}\}\Big(\mathbb{E}\big|\lambda - c\mathbb{P}(W' = W+1|\mathcal{F})\big| + \mathbb{E}\big|W - c\mathbb{P}(W' = W-1|\mathcal{F})\big|\Big).\end{aligned}$$

Before the proof, a few remarks.

Remark 4.38. Typically c is chosen to be approximately equal to $\lambda/\mathbb{P}(W' = W+1) = \lambda/\mathbb{P}(W' = W-1)$ so that the terms in absolute value have a small mean.

Remark 4.39. As with exchangeable pairs for normal approximation, there is a stochastic interpretation for the terms appearing in the error of Theorem 4.37. We can define a birth-death process on $\mathbb{N} \cup \{0\}$ where the birth rate at state k is $\alpha(k) = \lambda$ and death rate at state k is $\beta(k) = k$. This birth-death process has a $\mathrm{Po}(\lambda)$ stationary distribution so that the theorem says that if there is a reversible Markov chain with stationary distribution equal to the distribution of W such that the chance of increasing by one is approximately proportional to some constant λ and the chance of decreasing by one is approximately proportional to the current state, then W is approximately Poisson.

Proof. As usual, we want to bound $|\mathbb{E}[\lambda f(W+1) - Wf(W)]|$ for functions f such that $\|f\| \le \min\{1, \lambda^{-1/2}\}$ and $\|\Delta f\| \le \{1, \lambda^{-1}\}$. Now, the function

$$F(w, w') = \mathbb{I}[w' = w+1]f(w') - \mathbb{I}[w' = w-1]f(w)$$

is anti-symmetric, so that $\mathbb{E}[F(W, W')] = 0$. Evaluating the expectation by conditioning on $\mathcal{F}$, we obtain that

$$\mathbb{E}\left[\mathbb{P}(W' = W+1|\mathcal{F})f(W+1) - \mathbb{P}(W' = W-1|\mathcal{F})f(W)\right] = 0.$$

Hence, for any $c \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{E}[\lambda f(W+1) - Wf(W)]\\&\quad = \mathbb{E}\left[(\lambda - c\mathbb{P}(W' = W+1|\mathcal{F}))\, f(W+1) - (W - c\mathbb{P}(W' = W-1|\mathcal{F}))\, f(W)\right].\end{aligned} \tag{4.14}$$

Taking the absolute value and applying the triangle inequality yields the theorem. □

Example 4.40 (Law of small numbers)**.** Let $W = \sum_{i=1}^{n} X_i$ where the X_i are independent indicators with $\mathbb{P}(X_i = 1) = p_i$ and define $W' = W - X_I + X'_I$, where I is uniform on $\{1, \ldots, n\}$ independent of W and $X'_1, \ldots, X'_n$ are independent copies of the X_i independent of each other and all else. It is easy to see that (W, W') is an exchangeable pair and that

$$\begin{aligned}
\mathbb{P}(W' = W + 1 | (X_i)_{i \geq 1}) &= \frac{1}{n} \sum_{i=1}^{n} (1 - X_i) p_i, \\
\mathbb{P}(W' = W - 1 | (X_i)_{i \geq 1}) &= \frac{1}{n} \sum_{i=1}^{n} X_i (1 - p_i),
\end{aligned}$$

so that Theorem 4.37 with $c = n$ yields

$$\begin{aligned}
&d_{\mathrm{TV}}(W, Z) \\
&\leq \min\{1, \lambda^{-1/2}\} \left(\mathbb{E} \left| \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} (1 - X_i) p_i \right| + \mathbb{E} \left| \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i (1 - p_i) \right| \right) \\
&= 2 \min\{1, \lambda^{-1/2}\} \mathbb{E} \left[\sum_{i=1}^{n} p_i X_i \right] = 2 \min\{1, \lambda^{-1/2}\} \sum_{i=1}^{n} p_i^2.
\end{aligned}$$

This bound is not as good as that obtained using size-biasing (for example), but the better result can be recovered with exchangeable pairs by bounding the absolute value of (4.14) directly.

Example 4.41 (Fixed points of permutations)**.** Let π be a permutation of $\{1, \ldots, n\}$ chosen uniformly at random, let τ be a uniformly chosen random transposition, and let $\pi' = \pi\tau$. Let W be the number of fixed points of π and W' be the number of fixed points of π'. Then (W, W') is exchangeable and if W_2 is the number of transpositions when π is written as a product of disjoint cycles, then

$$\begin{aligned}
\mathbb{P}(W' = W + 1 | \pi) &= \frac{n - W - 2W_2}{\binom{n}{2}}, \\
\mathbb{P}(W' = W - 1 | \pi) &= \frac{W(n - W)}{\binom{n}{2}}.
\end{aligned}$$

To see why these expressions are true, note that in order for the number of fixed points of a permutation to increase by exactly one after multiplication by a transposition, a letter must be fixed that is not already and is not in a transposition (else the number of fixed points would increase by two). Similar considerations lead to the second expression.

By considering W as a sum of indicators, it is easy to see that $\mathbb{E}[W] = 1$ so that applying Theorem 4.37 with $c = (n-1)/2$ yields

$$\begin{aligned} d_{\mathrm{TV}}(W, Z) &\le \mathbb{E}\left|1 - \frac{n - W - 2W_2}{n}\right| + \mathbb{E}\left|W - \frac{W(n-W)}{n}\right| \\ &= \frac{1}{n}\mathbb{E}\left[W + 2W_2\right] + \frac{1}{n}\mathbb{E}[W^2] \\ &= 4/n, \end{aligned}$$

where the final inequality follows by considering W and W_2 as a sum of indicators which leads to $\mathbb{E}[W^2] = 2$ and $\mathbb{E}[W_2] = 1/2$. As is well known, the true rate of convergence is much better than order $1/n$; it is not clear how to get a better rate with this method.

We could also handle the number of i-cycles in a random permutation, but the analysis is a bit more tedious and is not worth pursuing due to the fact that so much more is known in this example; see Section 1.1 of [2] for a thorough account.

5. Exponential approximation

In this section we develop Stein's method for bounding the Wasserstein distance (see Section 1.2) between a distribution of interest and the Exponential distribution. We move quickly through the material analogous to that of Section 2 for normal approximation, as the general framework is similar. Our treatment follows [43] closely; alternative approaches can be found in [21].

Definition 5.1. We say that a random variable Z has the exponential distribution with rate λ, or $Z \sim \mathrm{Exp}(\lambda)$, if Z has density $\lambda e^{-\lambda z}$ for $z > 0$.

Lemma 5.2. *Define the functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(x) = f'(x) - f(x) + f(0).$$

1. *If $Z \sim \mathrm{Exp}(1)$, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all absolutely continuous f with $\mathbb{E}|f'(Z)| < \infty$.*
2. *If for some non-negative random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all absolutely continuous f with $\mathbb{E}|f'(Z)| < \infty$, then $W \sim \mathrm{Exp}(1)$.*

The operator $\mathcal{A}$ is referred to as a characterizing operator of the exponential distribution.

Before proving the lemma, we state one more result and then its consequence.

Lemma 5.3. *Let $Z \sim \mathrm{Exp}(1)$ and for some function h let f_h be the unique solution of*

$$f_h'(x) - f_h(x) = h(x) - \mathbb{E}[h(Z)] \tag{5.1}$$

such that $f_h(0) = 0$.

1. *If h is non-negative and bounded, then*
$$\|f_h\| \leq \|h\| \text{ and } \|f_h'\| \leq 2\|h\|.$$
2. *If h is absolutely continuous, then*
$$\|f_h'\| \leq \|h'\| \text{ and } \|f_h''\| \leq 2\|h'\|.$$

Analogous to normal approximation, this setup immediately yields the following promising result.

Theorem 5.4. *Let $W \geq 0$ be a random variable with $\mathbb{E}W < \infty$ and also let $Z \sim \mathrm{Exp}(1)$.*

1. *If $\mathcal{F}_W$ is the set of functions with $\|f'\| \leq 1$, $\|f''\| \leq 2$, and $f(0) = 0$, then*
$$d_{\mathrm{W}}(W, Z) \leq \sup_{f \in \mathcal{F}_W} |\mathbb{E}[f'(W) - f(W)]|.$$
2. *If $\mathcal{F}_K$ is the set of functions with $\|f\| \leq 1$, $\|f'\| \leq 2$, and $f(0) = 0$, then*
$$d_{\mathrm{K}}(W, Z) \leq \sup_{f \in \mathcal{F}_K} |\mathbb{E}[f'(W) - f(W)]|.$$

Proof of Lemma 5.3. Writing $\tilde{h}(t) := h(t) - \mathbb{E}[h(Z)]$, the relation (5.1) can easily be solved to yield
$$f_h(x) = -e^x \int_x^\infty \tilde{h}(t)e^{-t}dt. \tag{5.2}$$

1. If h is non-negative and bounded, then (5.2) implies that
$$|f_h(x)| \leq e^x \int_x^\infty |\tilde{h}(t)|e^{-t}dt \leq \|h\|.$$
Since f_h solves (5.1), we have
$$|f_h'(x)| = |f_h(x) + \tilde{h}(x)| \leq \|f_h\| + \|\tilde{h}\| \leq 2\|h\|,$$
where we have used the bound on $\|f_h\|$ above and that h is non-negative.
2. If h is absolutely continuous, then by the form of (5.2) it is clear that f_h is twice differentiable. Thus we have that f_h satisfies
$$f_h''(x) - f_h'(x) = h'(x).$$
Comparing this equation to (5.1), we can use the arguments of the proof of the previous item to establish the bounds on $\|f_h'\|$ and $\|f_h''\|$. □

Proof of Lemma 5.2. Item 1 essentially follows by integration by parts. More formally, for f absolutely continuous, we have
$$\begin{aligned}\mathbb{E}[f'(Z)] &= \int_0^\infty f'(t)e^{-t}dt = \int_0^\infty f'(t)\int_t^\infty e^{-x}dxdt \\ &= \int_0^\infty e^{-x}\int_0^x f'(t)dtdx = \mathbb{E}[f(Z)] - f(0),\end{aligned}$$

as desired. For the second item, assume that the random variable $W \geq 0$ satisfies $\mathbb{E}[f'(W)] = \mathbb{E}[f(W)] - f(0)$ for all absolutely continuous f with $\mathbb{E}|f'(Z)| < \infty$. The functions $f(x) = x^k$ are in this family, so that

$$k\mathbb{E}[W^{k-1}] = \mathbb{E}[W^k],$$

and this relation determines the moments of W as those of an exponential distribution with rate one, which satisfy Carleman's condition (using Stirling's approximation). Alternatively, the hypothesis on W also determines the Laplace transform as that of an exponential variable. □

It is clear from the form of the error in Theorem 5.4 that we want to find some structure in W, the random variable of interest, that allows us to compare $\mathbb{E}[f(W)]$ to $\mathbb{E}[f'(W)]$ for appropriate f. Unfortunately the tools we have previously developed for the analogous task in Poisson and Normal approximation do not help us directly here. However, we can define a transformation amenable to the form of the exponential characterizing operator which will prove fruitful. An alternative approach (followed in [21, 22]) is to use exchangeable pairs with a modified a-Stein condition.

5.1. Equilibrium coupling

We begin with a definition.

Definition 5.5. Let $W \geq 0$ a random variable with $\mathbb{E}[W] = \mu$. We say that W^e has the *equilibrium* distribution with respect to W if

$$\mathbb{E}[f(W)] - f(0) = \mu\mathbb{E}[f'(W^e)] \tag{5.3}$$

for all Lipschitz functions f.

Proposition 5.7 below implies that the equilibrium distribution is absolutely continuous with respect to Lebesgue measure, so that the right hand side of (5.3) is well defined. Before discussing that result, we note the following consequence of this definition.

Theorem 5.6. *Let $W \geq 0$ a random variable with $\mathbb{E}[W] = 1$ and $\mathbb{E}[W^2] < \infty$. If W^e has the equilibrium distribution with respect to W and is coupled to W, then*

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

Proof. Note that if $f(0) = 0$ and $\mathbb{E}[W] = 1$, then $\mathbb{E}[f(W)] = \mathbb{E}[f'(W^e)]$ so that for f with bounded first and second derivative and such that $f(0) = 0$, we have

$$|\mathbb{E}[f'(W) - f(W)]| = |\mathbb{E}[f'(W) - f'(W^e)]| \leq \|f''\|\mathbb{E}|W^e - W|.$$

Applying Theorem 5.4 now proves the desired conclusion. □

We now state a constructive definition of the equilibrium distribution which we also use later.

Proposition 5.7. *Let $W \geq 0$ be a random variable with $\mathbb{E}[W] = \mu$ and let W^s have the size-bias distribution of W. If U is uniform on the interval $(0,1)$ and independent of W^s, then UW^s has the equilibrium distribution of W.*

Proof. Let f be Lipschitz with $f(0) = 0$. Then

$$\mathbb{E}[f'(UW^s)] = \mathbb{E}\left[\int_0^1 f'(uW^s)du\right] = \mathbb{E}\left[\frac{f(W^s)}{W^s}\right] = \mu^{-1}\mathbb{E}[f(W)],$$

where in the final equality we use the definition of the size-bias distribution. □

Remark 5.8. This proposition shows that the equilibrium distribution is the same as that from renewal theory. That is, a stationary renewal process with increments distributed as a random variable Y is given by $Y^e + Y_1 + \cdots + Y_n$, where Y^e has the equilibrium distribution of Y and is independent of the i.i.d. sequence $Y_1, Y_2, \ldots$

Below we apply Theorem 5.6 in some non-trivial applications, but first we handle a canonical exponential approximation result.

Example 5.9 (Geometric distribution)**.** Let N be geometric with parameter p with positive support (denoted $N \sim \mathrm{Ge}(p)$), specifically, $\mathbb{P}(N = k) = (1-p)^{k-1}p$ for $k \geq 1$. It is well known that as $p \to 0$, pN converges weakly to an exponential distribution; this fact is not surprising as a simple calculation shows that if $Z \sim \mathrm{Exp}(\lambda)$, then the smallest integer no greater than Z is geometrically distributed. We can use Theorem 5.6 above to obtain an error in this approximation.

A little calculation shows that N has the same distribution as a variable which is uniform on $\{1, \ldots, N^s\}$, where N^s has the size-bias distribution of N (heuristically this is due to the memoryless property of the geometric distribution - see Remark 5.8). Thus Proposition 5.7 implies that for U uniform on $(0,1)$ independent of N, $N - U$ has the equilibrium distribution of N.

It is easy to verify that for a constant c and a non-negative variable X, $(cX)^e \overset{d}{=} cX^e$, so that if we define $W = pN$ our remarks above imply that $W^e := W - pU$ has the equilibrium distribution with respect to W. We now apply Theorem 5.6 to find that for $Z \sim \mathrm{Exp}(1)$,

$$d_{\mathrm{W}}(W, Z) \leq 2\mathbb{E}[pU] = p.$$

5.2. Application: Geometric sums

Our first application is a generalization of the following classical result which in turn generalizes Example 5.9.

Theorem 5.10 (Rényi's Theorem)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of non-negative random variables with $\mathbb{E}[X_i] = 1$ and let $N \sim \mathrm{Ge}(p)$ independent of* *the* X_i. *If* $W = p\sum_{i=1}^{N} X_i$ *and* $Z \sim \mathrm{Exp}(1)$, *then*

$$\lim_{p\to 0} d_{\mathrm{K}}(W, Z) = 0.$$

The special case where $X_i \equiv 1$ is handled in Example 5.9; intuitively, the example can be generalized because for p small, N is large so that the law of large numbers implies that $\sum_{i=1}^{N} X_i$ is approximately equal to N. We show the following result which implies Rényi's Theorem.

Theorem 5.11. *Let* $X_1, X_2, \ldots$ *be square integrable, non-negative, and independent random variables with* $\mathbb{E}[X_i] = 1$. *Let* $N > 0$ *be an integer valued random variable with* $\mathbb{E}[N] = 1/p$ *for some* $0 < p \leq 1$ *and let* M *be defined on the same space as* N *such that*

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

If $W = p\sum_{i=1}^{N} X_i$, $Z \sim \mathrm{Exp}(1)$, *and* X_i^e *is an equilibrium coupling of* X_i *independent of* N, M, *and* $(X_j)_{j\neq i}$, *then*

$$\begin{aligned} d_{\mathrm{W}}(W, Z) &\leq 2p\left(\mathbb{E}|X_M - X_M^e| + \mathbb{E}|N - M|\right) && (5.4)\\ &\leq 2p\left(1 + \frac{\mu_2}{2} + \mathbb{E}|N - M|\right), && (5.5)\end{aligned}$$

where $\mu_2 := \sup_i \mathbb{E}[X_i^2]$.

Before the proof of the theorem, we make a few remarks.

Remark 5.12. The theorem can be a little difficult to parse on first glance, so we make a few comments to interpret the error. The random variable M is a discrete version of the equilibrium transform which we have already seen above in Example 5.9. More specifically, it is easy to verify that if N^s has the size-bias distribution of N, then M is distributed uniformly on the set $\{1, \ldots, N^s\}$. If $N \sim \mathrm{Ge}(p)$, then we can take $M \equiv N$ so that the final term of the error of (5.4) and (5.5) is zero. Thus $\mathbb{E}|N - M|$ quantifies the proximity of the distribution of N to a geometric distribution. We formalize this more precisely when we cover Stein's method for geometric approximation below.

The first term of the error in (5.5) can be interpreted as a term measuring the regularity of the X_i. The heuristic is that the law of large numbers needs to kick in so that $\sum_{i=1}^{N} X_i \approx N$, and the theorem shows that in order for this to occur, it is enough that the X_i's have uniformly bounded second moments. Moreover, the first term of the error (5.4) shows that the approximation also benefits from having the X_i be close to exponentially distributed. In particular, if all of the X_i are exponential and N is geometric, then the theorem shows $d_{\mathrm{W}}(W, Z) = 0$ which can also be easily verified using Laplace transforms.

Remark 5.13. A more general theorem can be proved with a little added technicality which allows for the X_i to have different means and for the X_i to have a certain dependence - see [43].

Proof. We show that

$$W^e = p\left[\sum_{i=1}^{M-1} X_i + X_M^e\right] \tag{5.6}$$

is an equilibrium coupling of W. From this point Theorem 5.6 implies that

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

which proves (5.4). The second bound (5.5) follows from (5.4) after noting that

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

where the equality follows from the definition of the equilibrium coupling.

It only remains to show (5.6). Let f be Lipschitz with $f(0) = 0$ and define

$$g(m) = f\left(p\sum_{i=1}^{m} X_i\right).$$

On one hand, using independence and the defining relation of X_m^e, we obtain

$$\mathbb{E}\left[f'\left(p\sum_{i=1}^{M-1} X_i + pX_M^e\right)\Big|M\right] = p^{-1}\mathbb{E}[g(M) - g(M-1)|M],$$

and on the other, the definition of M implies that

$$p^{-1}\mathbb{E}[g(M) - g(M-1)|(X_i)_{i\geq 1}] = \mathbb{E}[g(N)|(X_i)_{i\geq 1}],$$

so that altogether we obtain $\mathbb{E}[f'(W^e)] = \mathbb{E}[g(N)] = \mathbb{E}[f(W)]$, as desired. □

5.3. Application: Critical branching process

In this section we obtain an error in a classical theorem of Yaglom pertaining to the generation size of a critical Galton-Watson branching process conditioned on non-extinction. Although we attempt to have this section be self-contained, it is useful to have some exposure to the elementary properties and definitions of branching processes, found for example in the first chapter of [7].

Let $Z_0 = 1$, and Z_1 be a non-negative integer valued random variable with finite mean. For $i, j \geq 1$, let $Z_{i,j}$ be i.i.d. copies of Z_1 and define for $n \geq 1$

$$Z_{n+1} = \sum_{i=1}^{Z_n} Z_{n,i}.$$

We think of Z_n as the generation size of a population that initially has one individual and where each individual in a generation has a Z_1 distributed number of offspring (or children) independently of the other individuals in the generation. We also assume that all individuals in a generation have offspring at the same time (creating the next generation) and die immediately after reproducing.

It is a basic fact (see Theorem 1 on page 7 of [7]) that if $\mathbb{E}[Z_1] \leq 1$ and $\mathbb{P}(Z_1 = 1 < 1)$ then the population almost surely dies out, whereas if $\mathbb{E}[Z_1] > 1$, then the probability that the population lives forever is strictly positive. Thus the case where $\mathbb{E}[Z_1] = 1$ is referred to as the critical case and a fundamental result of the behavior in this case is the following.

Theorem 5.14 (Yaglom's Theorem)**.** *Let $1 = Z_0, Z_1, \ldots$ be the generation sizes of a Galton-Watson branching process where $\mathbb{E}[Z_1] = 1$ and $\mathrm{Var}(Z_1) = \sigma^2 < \infty$. If $Y_n \overset{d}{=} (Z_n | Z_n > 0)$ and $Z \sim \mathrm{Exp}(1)$, then*

$$\lim_{n\to\infty} d_{\mathrm{K}}\left(\frac{2Y_n}{n\sigma^2}, Z\right) = 0.$$

We provide a rate of convergence in this theorem under a stricter moment assumption.

Theorem 5.15. *Let $Z_0, Z_1, \ldots$ as in Yaglom's Theorem above and assume also that $\mathbb{E}|Z_1|^3 < \infty$. If $Y_n \overset{d}{=} (Z_n | Z_n > 0)$ and $Z \sim \mathrm{Exp}(1)$, then for some constant C,*

$$d_{\mathrm{W}}\left(\frac{2Y_n}{n\sigma^2}, Z\right) \leq C\frac{\log(n)}{n}.$$

Proof. On the same probability space, we construct copies of Y_n and of Y_n^e, where the latter has the equilibrium distribution, and then show that $\mathbb{E}|Y_n - Y_n^e| \leq C\log(n)$. Once this is established, the result is proved by Theorem 5.6 and the fact that $(cY_n)^e \overset{d}{=} cY_n^e$ for any constant c.

In order to couple Y_n and Y_n^e, we construct a "size-bias" tree and then find copies of the variables we need in it. The clever construction we use is due to [40], and implicit in their work is the fact that $\mathbb{E}|Y_n - Y_n^e|/n \to 0$ (used to show that $n\mathbb{P}(Z_n > 0) \to 2/\sigma^2$), but we extract a rate from their analysis.

We view the size-bias tree as labeled and ordered, in the sense that, if w and v are vertices in the tree from the same generation and w is to the left of v, then the offspring of w is to the left of the offspring of v. Start in generation 0 with one vertex v_0 and let it have a number of offspring distributed according to the size-bias distribution of Z_1. Pick one of the offspring of v_0 uniformly at random and call it v_1. To each of the siblings of v_1 attach an independent Galton-Watson branching process having the offspring distribution of Z_1. For v_1 proceed as for v_0, i.e., give it a size-bias number of offspring, pick one uniformly at random, call it v_2, attach independent Galton-Watson branching process to the siblings of v_2 and so on. It is clear that this process always yields an infinite tree as the "spine" $v_0, v_1, v_2, \ldots$ is infinite. See Figure 1 of [40] for an illustration of this tree.

Now, for a fixed tree t, let $G_n(t)$ be the chance that the original branching process driven by Z_1 agrees with t up to generation n, let $G_n^s(t)$ be the chance that the size-bias tree just described agrees with t up to generation n, and for v an individual of t in generation n, let $G_n^s(t, v)$ be the chance that the size-bias tree agrees with t up to generation n and has the vertex v as the distinguished vertex v_n in generation n. We claim that

$$G_n^s(t, v) = G_n(t). \tag{5.7}$$

Before proving this claim, we note some immediate consequences which imply that our size-bias tree naturally contains a copy of Y_n^e. Let S_n be the size of generation n in the size-bias tree.

1. S_n has the size-bias distribution of Z_n.
2. If Y_n^s has the size-bias distribution of Y_n, then $S_n \overset{d}{=} Y_n^s$.
3. Given S_n, v_n is uniformly distributed among the individuals of generation n.
4. If R_n is the number of individuals to the right (inclusive) of v_n and U is uniform on $(0, 1)$, independent of all else, then $Y_n^e := R_n - U$ has the equilibrium distribution of Y_n.

To show the first item, note that (5.7) implies

$$G_n^s(t) = t_n G_n(t), \tag{5.8}$$

where t_n is the number of individuals in the nth generation of t. Now, $\mathbb{P}(S_n = k)$ is obtained by integrating the left hand side of (5.8) over trees t with $t_n = k$, and performing the same integral on the right hand side of (5.8) yields $k\mathbb{P}(Z_n = k)$. The second item follows from the more general fact that conditioning a non-negative random variable to be positive does not change the size-bias distribution. Item 3 can be read from the right hand side of (5.7), since it does not depend on v. For Item 4, Item 3 implies that R_n is uniform on $\{1, \ldots, S_n\}$ so that $R_n - U \overset{d}{=} US_n$, from which the result follows from Item 2 and Proposition 5.7.

At this point, we would like to find a copy of Y_n in the size-bias tree, but before proceeding further we prove (5.7). Since trees formed below distinct vertices in a given generation are independent, we prove the formula by writing down a recursion. To this end, for a given planar tree t with k individuals in the first generation, label the subtrees with these k individuals as a root from left to right by $t_1, t_2, \ldots, t_k$. Now, a vertex v in generation $n+1$ of t lies in exactly one of the subtrees $t_1, \ldots, t_k$, say t_i. With this setup, we have

$$G_{n+1}^s(t, v) = G_n(t_i, v) \prod_{j \neq i} G_n(t_j)[k\mathbb{P}(Z_1 = k)]\frac{1}{k}.$$

The first factor corresponds to the chance of seeing the tree t_i up to generation $n+1$ below the distinguished vertex v_1 and choosing v as the distinguished vertex in generation $n + 1$. The second factor is the chance of seeing the remaining subtrees up to generation $n+1$, and the remaining factors correspond to having

k offspring initially (with the size-bias distribution of Z_1) and choosing vertex v_1 (the root of t_i) as the distinguished vertex initially. With this formula in hand, it is enough to verify that (5.7) follows this recursion.

We must now find a copy of Y_n in our size-bias tree. If L_n is the number of individuals to the left of v_n (exclusive, so $S_n = L_n + R_n$), then we claim that

$$S_n \big| \{L_n = 0\} \overset{d}{=} Y_n. \tag{5.9}$$

Indeed, we have

$$\begin{aligned}\mathbb{P}(S_n = k | L_n = 0) &= \frac{\mathbb{P}(L_n = 0 | S_n = k)\mathbb{P}(S_n = k)}{\mathbb{P}(L_n = 0)} \\ &= \frac{\mathbb{P}(S_n = k)}{k\mathbb{P}(L_n = 0)} = \frac{\mathbb{P}(Z_n = k)}{\mathbb{P}(L_n = 0)},\end{aligned}$$

where we have used Items 2 and 3 above and the claim now follows since

$$\mathbb{P}(L_n = 0) = \sum_{k\geq 1} \mathbb{P}(L_n = 0 | S_n = k)\mathbb{P}(S_n = k) = \sum_{k\geq 1} \frac{\mathbb{P}(S_n = k)}{k} = \mathbb{P}(Z_n > 0).$$

We are only part of the way to finding a copy of Y_n in the size-bias tree since we still need to realize S_n given the event $L_n = 0$. Denote by $S_{n,j}$ the number of particles in generation n that stem from any of the siblings of v_j (but not v_j itself). Clearly, $S_n = 1 + \sum_{j=1}^n S_{n,j}$, where the summands are independent. Likewise, let $L_{n,j}$ and $R_{n,j}$, be the number of particles in generation n that stem from the siblings to the left and right of v_j (exclusive) and note that $L_{n,n}$ and $R_{n,n}$ are just the number of siblings of v_n to the left and to the right, respectively. We have the relations $L_n = \sum_{j=1}^n L_{n,j}$ and $R_n = 1 + \sum_{j=1}^n R_{n,j}$. Note that for fixed j, $L_{n,j}$ and $R_{n,j}$ are in general not independent, as they are linked through the offspring size of v_{j-1}.

Let now $R'_{n,j}$ be independent random variables such that

$$R'_{n,j} \overset{d}{=} R_{n,j} \big| \{L_{n,j} = 0\}.$$

and

$$R^*_{n,j} = R_{n,j}\mathbb{I}[L_{n,j} = 0] + R'_{n,j}\mathbb{I}[L_{n,j} > 0] = R_{n,j} + (R'_{n,j} - R_{n,j})\mathbb{I}[L_{n,j} > 0].$$

Finally, if $R^*_n = 1 + \sum_{j=1}^n R^*_{n,j}$, then (5.9) implies that we can take $Y_n := R^*_n$.

Having coupled Y_n and Y^e_n, we can now proceed to show $\mathbb{E}|Y^e_n - Y_n| = \mathrm{O}(\log(n))$. By Item 4 above,

$$\begin{aligned}|Y_n - Y^e_n| &= \left|1 - U + \sum_{j=1}^n (R'_{n,j} - R_{n,j})\mathbb{I}[L_{n,j} > 0]\right| \\ &\leq |1 - U| + \sum_{j=1}^n R'_{n,j}\mathbb{I}[L_{n,j} > 0] + \sum_{j=1}^n R_{n,j}\mathbb{I}[L_{n,j} > 0].\end{aligned}$$

Taking expectation in the inequality above, our result follows after we show that

$$\begin{aligned}
&(i)\ \ \mathbb{E}\left[R'_{n,j}\mathbb{I}[L_{n,j}>0]\right] \leq \sigma^2\mathbb{P}(L_{n,j}>0),\\
&(ii)\ \ \mathbb{E}\left[R_{n,j}\mathbb{I}[L_{n,j}>0]\right] \leq \mathbb{E}[Z_1^3]\mathbb{P}(Z_{n-j}>0),\\
&(iii)\ \ \mathbb{P}(L_{n,j}>0) \leq \sigma^2\mathbb{P}(Z_{n-j}>0) \leq C(n-j+1)^{-1} \text{ for some } C>0.
\end{aligned}$$

For part (i), independence implies that

$$\mathbb{E}\left[R'_{n,j}\mathbb{I}[L_{n,j}>0]\right] = \mathbb{E}[R'_{n,j}]\mathbb{P}(L_{n,j}>0),$$

and using that $S_{n,j}$ and $\mathbb{I}[L_{n,j}=0]$ are negatively correlated (in the second inequality) below, we find

$$\begin{aligned}
\mathbb{E}[R'_{n,j}] &= \mathbb{E}[R_{n,j}|L_{n,j}=0]\\
&\leq \mathbb{E}[S_{n,j}-1|L_{n,j}=0]\\
&\leq \mathbb{E}[S_{n,j}]-1\\
&\leq \mathbb{E}[S_n]-1=\sigma^2.
\end{aligned}$$

For part (ii), if X_j denotes the number of siblings of v_j, having the size-bias distribution of Z_1 minus 1, we have

$$\begin{aligned}
\mathbb{E}\left[R_{n,j}\mathbb{I}[L_{n,j}>0]\right] &\leq \mathbb{E}[X_j\mathbb{I}[L_{n,j}>0]]\\
&\leq \sum_k k\mathbb{P}(X_j=k, L_{n,j}>0)\\
&\leq \sum_k k\mathbb{P}(X_j=k)\mathbb{P}(L_{n,j}>0|X_j=k)\\
&\leq \sum_k k^2\mathbb{P}(X_j=k)\mathbb{P}(Z_{n-j}>0)\\
&\leq \mathbb{E}[Z_1^3]\mathbb{P}(Z_{n-j}>0),
\end{aligned}$$

where we have used that $\mathbb{E}[R_{n,j}\mathbb{I}[L_{n,j}>0]|X_j] \leq X_j\mathbb{I}[L_{n,j}>0]$ in the first inequality and that $\mathbb{P}(L_{n,j}>0|X_j=k) \leq k\mathbb{P}(Z_{n-j}>0)$ in the penultimate inequality.

Finally, we have

$$\begin{aligned}
\mathbb{P}(L_{n,j}>0) &= \mathbb{E}\left[\mathbb{P}(L_{n,j}>0|X_j)\right]\\
&\leq \mathbb{E}\left[X_j\mathbb{P}(Z_{n-j}>0\right]\\
&\leq \sigma^2\mathbb{P}(Z_{n-j}>0).
\end{aligned}$$

Using Kolmogorov's estimate (see Chapter 1, Section 9 of [7]), we have

$$\lim_{n\to\infty} n\mathbb{P}(Z_n>0)=2/\sigma^2,$$

which implies the final statement of (iii). □

6. Geometric approximation

Due to Example 5.9, if $W > 0$ is integer-valued such that $W/\mathbb{E}[W]$ is approximately exponential and $\mathbb{E}[W]$ is large, then we expect W to be approximately geometrically distributed. In fact, if we write $\mathbb{E}[W] = 1/p$, and let $X \sim \mathrm{Ge}(p)$ and $Z \sim \mathrm{Exp}(1)$, then the triangle inequality implies that

$$|d_{\mathrm{W}}(pW, Z) - d_{\mathrm{W}}(X, W)| \leq p.$$

However, if we want to bound $d_{\mathrm{TV}}(W, X)$, then the inequality above is not useful. For example, if $W \stackrel{d}{=} kX$ for some positive integer k, then $d_{\mathrm{TV}}(W, X) \approx (k-1)/k$ since the support of W and X do not match. This issue of support mismatch is typical in bounding the total variation distance between integer-valued random variables and can be handled by introducing a term into the bound that quantifies the "smoothness" of the random variable of interest.

The version of Stein's method for geometric approximation which we discuss below can be used to handle these types of technicalities [42], but the arguments can be a bit technical. Thus, we develop a simplified version of the method and apply it to an example where these technicalities do not arise and where exponential approximation is not useful.

We parallel the development of Stein's method for exponential approximation above, so we move quickly through the initial theoretical framework; our work below follows [42].

6.1. Main theorem

A typical issue when discussing the geometric distribution is whether to have the support begin at zero or one. In our work below we focus on the geometric distribution which puts mass at zero; that is $N \sim \mathrm{Ge}^0(p)$ if for $k = 0, 1, \ldots$, we have $\mathbb{P}(N = k) = (1-p)^k p$. Developing the theory below for the geometric distribution with positive support is similar in flavor, but different in detail; see [42].

As usual we begin by defining the characterizing operator that we use.

Lemma 6.1. *Define the functional operator $\mathcal{A}$ by*

$$\mathcal{A}f(k) = (1-p)\Delta f(k) - pf(k) + pf(0).$$

1. *If $Z \sim \mathrm{Ge}^0(p)$, then $\mathbb{E}\mathcal{A}f(Z) = 0$ for all bounded f.*
2. *If for some non-negative random variable W, $\mathbb{E}\mathcal{A}f(W) = 0$ for all bounded f, then $W \sim \mathrm{Ge}^0(p)$.*

The operator $\mathcal{A}$ is referred to as a characterizing operator of the geometric distribution.

We now state the properties of the solution to the Stein equation that we need.

Lemma 6.2. *If $Z \sim \mathrm{Ge}^0(p)$, $A \subseteq \mathbb{N} \cup \{0\}$, and f_A is the unique solution with $f_A(0) = 0$ of*

$$(1-p)\Delta f_A(k) - p f_A(k) = \mathbb{I}[k \in A] - \mathbb{P}(Z \in A),$$

then $-1 \leq \Delta f(k) \leq 1$.

These two lemmas lead easily to the following result.

Theorem 6.3. *Let $\mathcal{F}$ be the set of functions with $f(0) = 0$ and $\|\Delta f\| \leq 1$ and let $W \geq 0$ be an integer-valued random variable with $\mathbb{E}[W] = (1-p)/p$ for some $0 < p \leq 1$. If $N \sim \mathrm{Ge}^0(p)$, then*

$$d_{\mathrm{TV}}(W, N) \leq \sup_{f \in \mathcal{F}} \left|\mathbb{E}[(1-p)\Delta f(W) - p f(W)]\right|.$$

Before proceeding further, we briefly indicate the proofs of Lemmas 6.1 and 6.2.

Proof of Lemma 6.1. The first assertion is a simple computation while the second can be verified by choosing $f(k) = \mathbb{I}[k = j]$ for each $j = 0, 1, \ldots$ which yields a recursion for the point probabilities for W. □

Proof of Lemma 6.2. After noting that

$$f_A(k) = \sum_{i \in A} (1-p)^i - \sum_{i \in A, i \geq k} (1-p)^{i-k},$$

we easily see that

$$\Delta f_A(k) = \mathbb{I}[k \in A] - p \sum_{i \in A, i \geq k+1} (1-p)^{i-k-1},$$

which is the difference of two non-negative terms, each of which is bounded above by one. □

It is clear from the form of the error of Theorem 6.3 that it may be fruitful to attempt to define a discrete version of the equilibrium distribution used in the exponential approximation formulation above, which is the program we follow. An alternative coupling is used in [41].

6.2. Discrete equilibrium coupling

We begin with a definition.

Definition 6.4. Let $W \geq 0$ be a random variable such that $\mathbb{E}[W] = (1-p)/p$ for some $0 < p \leq 1$. We say that W^e has the *discrete equilibrium* distribution with respect to W if for all functions f with $\|\Delta f\| < \infty$,

$$p\mathbb{E}[f(W)] - p f(0) = (1-p)\mathbb{E}[\Delta f(W^e)]. \tag{6.1}$$

The following result provides a constructive definition of the discrete equilibrium distribution, so that the right hand side of (6.1) defines a probability distribution.

Proposition 6.5. *Let $W \geq 0$ be an integer-valued random variable such that $\mathbb{E}[W] = (1-p)/p$ for some $0 < p \leq 1$ and let W^s have the size-bias distribution of W. If conditional on W^s, W^e is uniform on $\{0, 1, \ldots, W^s - 1\}$, then W^e has the discrete equilibrium distribution with respect to W.*

Proof. Let f be such that $\|\Delta f\| < \infty$ and $f(0) = 0$. If W^e is uniform on $\{0, 1, \ldots, W^s - 1\}$ as dictated by the proposition, then

$$\mathbb{E}[\Delta f(W^e)] = \mathbb{E}\left[\frac{1}{W^s}\sum_{i=0}^{W^s-1}\Delta f(i)\right] = \mathbb{E}\left[\frac{f(W^s)}{W^s}\right] = \frac{p}{1-p}\mathbb{E}[f(W)],$$

where in the final equality we use the definition of the size-bias distribution. □

Remark 6.6. This proposition shows that the equilibrium distribution is the same as that from renewal theory - see Remark 5.8.

Theorem 6.7. *Let $N \sim \mathrm{Ge}^0(p)$ and $W \geq 0$ be an integer-valued random variable with $\mathbb{E}[W] = (1-p)/p$ for some $0 < p \leq 1$ such that $\mathbb{E}[W^2] < \infty$. If W^e has the equilibrium distribution with respect to W and is coupled to W, then*

$$d_{\mathrm{TV}}(W, N) \leq 2(1-p)\mathbb{E}|W^e - W|.$$

Proof. If $f(0) = 0$ and $\|\Delta f\| \leq 1$, then

$$\begin{aligned}\left|\mathbb{E}[(1-p)\Delta f(W) - pf(W)]\right| &= (1-p)|\mathbb{E}[\Delta f(W) - \Delta f(W^e)]| \\ &\leq 2(1-p)\mathbb{E}|W^e - W|,\end{aligned}$$

where the inequality follows after noting that $\Delta f(W) - \Delta f(W^e)$ can be written as a sum of $|W^e - W|$ terms each of size at most

$$|\Delta f(W + i + 1) - \Delta f(W + i)| \leq 2\|\Delta f\|.$$

Applying Theorem 6.3 now proves the desired conclusion. □

6.3. Application: Uniform attachment graph model

Let G_n be a directed random graph on n nodes defined by the following recursive construction. Initially the graph starts with one node with a single loop where one end of the loop contributes to the "in-degree" and the other to the "out-degree." Now, for $2 \leq m \leq n$, given the graph with $m - 1$ nodes, add node m along with an edge directed from m to a node chosen uniformly at random among the m nodes present. Note that this model allows edges connecting a node with itself. This random graph model is referred to as uniform attachment. We prove the following geometric approximation result (convergence was shown without rate in [15]), which is weaker than the result of [42] but has a slightly simpler proof.

Theorem 6.8. *If W is the in-degree of a node chosen uniformly at random from the random graph G_n which is generated according to uniform attachment and $N \sim \mathrm{Ge}^0(1/2)$, then*

$$d_{\mathrm{TV}}(W, N) \leq \frac{\log(n) + 1}{n}. \tag{6.2}$$

Proof. For $j = 1, \ldots, n$, let X_j have a Bernoulli distribution, independent of all else, with parameter $\mu_j := (n - j + 1)^{-1}$, and let I be an independent random variable that is uniform on the integers $1, 2, \ldots, n$. If we imagine that node $n + 1 - I$ is the randomly selected node, then it is easy to see that we can write $W := \sum_{j=1}^{I} X_j$.

We show that

$$W^e = \sum_{j=1}^{I-1} X_j \tag{6.3}$$

is an equilibrium coupling of W. From this point Theorem 6.7 implies that

$$\begin{aligned} d_{\mathrm{TV}}(W, N) &\leq \mathbb{E}|W^e - W| = \mathbb{E}\,|X_I| \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|X_i| = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{j} \leq \frac{\log(n) + 1}{n}, \end{aligned}$$

which is (6.2).

It only remains to show (6.3) has the equilibrium distribution with respect to W. Let f be bounded with $f(0) = 0$ and define

$$g(i) = f\left(\sum_{j=1}^{i} X_j\right).$$

On one hand, using independence and the fact that $X_i^e \equiv 0$ for $1 \leq i \leq n$,

$$\mathbb{E}\left[\Delta f\left(\sum_{j=1}^{I-1} X_j\right)\middle| I\right] = \frac{1}{\mu_I}\mathbb{E}[g(I) - g(I-1)|I],$$

and on the other, the definition of I implies that

$$\mathbb{E}\left[\frac{g(I) - g(I-1)}{\mu_I}\middle|(X_j)_{j\geq 1}\right] = \mathbb{E}[g(I)|(X_j)_{j\geq 1}],$$

so that altogether we obtain $\mathbb{E}[\Delta f(W^e)] = \mathbb{E}[g(I)] = \mathbb{E}[f(W)]$, as desired. □

7. Concentration of measure inequalities

The techniques we have developed for estimating expectations of characterizing operators (e.g. exchangeable pairs) can also be used to prove concentration of measure inequalities (or large deviations). By concentration of measure inequalities, we mean estimates of $\mathbb{P}(W \geq t)$ and $\mathbb{P}(W \leq -t)$, for $t > 0$ and some centered random variable W. Of course our previous work was concerned with such estimates, but here we are after the rate that these quantities tend to zero as t tends to infinity - in the tails of the distribution. Distributional error terms are maximized in the body of the distribution and so typically do not provide information about the tails.

The study of concentration of measure inequalities has a long history and has also found recent use in machine learning and analysis of algorithms - see [16] and references therein for a flavor of the modern considerations of these types of problems. Our results hinge on the following fundamental observation.

Proposition 7.1. *If W is random variable and there is a $\delta > 0$ such that $\mathbb{E}[e^{\theta W}] < \infty$ for all $\theta \in (-\delta, \delta)$, then for all $t > 0$ and $0 < \theta < \delta$,*

$$\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}[e^{\theta W}]}{e^{\theta t}} \text{ and } \mathbb{P}(W \leq -t) \leq \frac{\mathbb{E}[e^{-\theta W}]}{e^{\theta t}}.$$

Proof. Using first that e^x is an increasing function, and then Markov's inequality,

$$\mathbb{P}(W > t) = \mathbb{P}\left(e^{\theta W} > e^{\theta t}\right) \leq \frac{\mathbb{E}[e^{\theta W}]}{e^{\theta t}},$$

which proves the first assertion. The second assertion follows similarly. □

Before discussing the use of Proposition 7.1 in Stein's method, we first work out a couple of easy examples.

Example 7.2 (Normal distribution)**.** Let Z have the standard normal distribution and recall that for $t > 0$ we have the Mills ratio bound

$$\mathbb{P}(Z \geq t) \leq \frac{e^{-t^2/2}}{t\sqrt{2\pi}}. \tag{7.1}$$

A simple calculation implies $\mathbb{E}[e^{\theta Z}] = e^{\theta^2/2}$ for all $\theta \in \mathbb{R}$, so that for $\theta, t > 0$ Proposition 7.1 implies

$$\mathbb{P}(Z \geq t) \leq e^{\theta^2/2 - \theta t},$$

and choosing the minimizer $\theta = t$ yields

$$\mathbb{P}(Z \geq t) \leq e^{-t^2/2}, \tag{7.2}$$

which implies that this is the best behavior we can hope for using Proposition 7.1 in examples where the random variable is approximately normal (such as sums of independent variables). Comparing (7.2) to the Mills ratio rate given by (7.1), we expect that our rates obtained using Proposition 7.1 in such applications will be suboptimal (by a factor of t) for large t.

Example 7.3 (Poisson distribution)**.** Let Z have the Poisson distribution with mean λ. A simple calculation implies $\mathbb{E}[e^{\theta Z}] = \exp\{\lambda(e^{\theta} - 1)\}$ for all $\theta \in \mathbb{R}$, so that for $\theta, t > 0$ Proposition 7.1 implies

$$\mathbb{P}(Z - \lambda \geq t) \leq \exp\{\lambda(e^{\theta} - 1) - \theta(t + \lambda)\},$$

and choosing the minimizer $\theta = \log(1 + t/\lambda)$ yields

$$\mathbb{P}(Z - \lambda \geq t) \leq \exp\left\{-t\left(\log\left(1 + \frac{t}{\lambda}\right) - 1\right) - \lambda \log\left(1 + \frac{t}{\lambda}\right)\right\},$$

which is of smaller order than e^{-ct} for t large and fixed $c > 0$, but of bigger order than $e^{-t\log(t)}$. This is the best behavior we can hope for using Proposition 7.1 in examples where the random variable is approximately Poisson (such as sums of independent indicators, each with a small probability of being one).

How does Proposition 7.1 help us to use the techniques from Stein's method to obtain concentration of measure inequalities? If W is random variable and there is a $\delta > 0$ such that $\mathbb{E}[e^{\theta W}] < \infty$ for all $\theta \in (-\delta, \delta)$, then we can define $m(\theta) = \mathbb{E}[e^{\theta W}]$ for $0 < \theta < \delta$, and we also have that $m'(\theta) = \mathbb{E}[We^{\theta W}]$. Thus $m'(\theta)$ is of the form $\mathbb{E}[Wf(W)]$, where $f(W) = e^{\theta W}$ so that we can use the techniques that we developed to bound the characterizing operator for the normal and Poisson distribution to obtain a differential inequality for $m(\theta)$. Such an inequality leads to bounds on $m(\theta)$ so that we can apply Proposition 7.1 to obtain bounds on the tail probabilities of W. This observation was first made in [17].

7.1. Concentration of measure using exchangeable pairs

Our first formulation using the couplings of Sections 3 and 4 for concentration of measure inequalities uses exchangeable pairs. We follow the development of [18].

Theorem 7.4. *Let (W, W') an a-Stein pair with $\mathrm{Var}(W) = \sigma^2 < \infty$. If $\mathbb{E}[e^{\theta W}|W' - W|] < \infty$ for all $\theta \in \mathbb{R}$ and for some sigma-algebra $\mathcal{F} \supseteq \sigma(W)$ there are non-negative constants B and C such that*

$$\frac{\mathbb{E}[(W' - W)^2|\mathcal{F}]}{2a} \leq BW + C, \tag{7.3}$$

then for all $t > 0$,

$$\mathbb{P}(W \geq t) \leq \exp\left\{\frac{-t^2}{2C + 2Bt}\right\} \text{ and } \mathbb{P}(W \leq -t) \leq \exp\left\{\frac{-t^2}{2C}\right\}.$$

Remark 7.5. The reason the left tail has a better bound stems from condition (7.3) which implies that $BW + C \geq 0$. Thus, the condition essentially forces the centered variable W to be bounded from below whereas there is no such requirement for large positive values. As can be understood from the proof of the theorem, conditions other than (7.3) may be substituted to yield different bounds; see [19].

Before proving the theorem, we apply it in a simple example.

Example 7.6 (Sum of independent variables)**.** Let $X_1, \dots, X_n$ be independent random variables with $\mu_i := \mathbb{E}[X_i]$, $\sigma_i^2 := \mathrm{Var}(X_i) < \infty$ and define $W = \sum_i X_i - \mu_i$. Let $X_1', \dots, X_n'$ be an independent copy of the X_i and for I independent of all else and uniform on $\{1, \dots, n\}$, let $W' = W - X_I + X_I'$ so that as usual, (W, W') is a $1/n$-Stein pair. We consider two special cases of this setup.

1. For $i = 1, \dots, n$, assume $|X_i - \mu_i| \le C_i$. Then clearly the moment generating function condition of Theorem 7.4 is satisfied and we also have

$$\begin{aligned}\mathbb{E}[(W' - W)^2 | (X_j)_{j\ge1}] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i' - X_i)^2 | (X_j)_{j\ge1}] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i' - \mu_i)^2] + \frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)^2 \\ &\le \frac{1}{n}\sum_{i=1}^n (\sigma_i^2 + C_i^2),\end{aligned}$$

so that we can apply Theorem 7.4 with $B = 0$ and $2C = \sum_{i=1}^n (\sigma_i^2 + C_i^2)$. We have shown that for $t > 0$,

$$\mathbb{P}\left(|W - \mathbb{E}[W]| \ge t\right) \le 2\exp\left\{-\frac{t^2}{\sum_{i=1}^n (\sigma_i^2 + C_i^2)}\right\},$$

which is some version of Hoeffding's inequality [36].

2. For $i = 1, \dots, n$, assume $0 \le X_i \le 1$. Then the moment generating function condition of the theorem is satisfied and

$$\begin{aligned}\mathbb{E}[(W' - W)^2 | (X_j)_{j\ge1}] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i' - X_i)^2 | (X_j)_{j\ge1}] \\ &= \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}[(X_i')^2] - 2\mu_i X_i + X_i^2\right) \\ &\le \frac{1}{n}\sum_{i=1}^n (\mu_i + X_i) = \frac{1}{n}(2\mu + W),\end{aligned}$$

where $\mu := \mathbb{E}[W]$ and we have used that $-2\mu_i X_i \le 0$ and $X_i^2 \le X_i$. We can now apply Theorem 7.4 with $B = 1/2$ and $C = \mu$ to find that for $t > 0$,

$$\mathbb{P}\left(|W - \mu| \ge t\right) \le 2\exp\left\{-\frac{t^2}{2\mu + t}\right\}. \tag{7.4}$$

Note that if μ is constant in n, then (7.4) is of the order e^{-t} for large t, which according to Example 7.3 is similar to the order of Poisson tails.

However, if μ and $\sigma^2 := \mathrm{Var}(W)$ both tend to infinity at the same rate, then (7.4) implies

$$\mathbb{P}\left(\frac{|W-\mu|}{\sigma} \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\frac{\mu}{\sigma^2}+\frac{t}{\sigma}}\right\},$$

so that for σ^2 large, the tails are of order e^{-ct^2}, which according to Example 7.2 is similar to the order of Gaussian tails.

Proof of Theorem 7.4. Let $m(\theta) = \mathbb{E}[e^{\theta W}]$ and note that $m'(\theta) = \mathbb{E}[We^{\theta W}]$. Since (W, W') is an a-Stein pair, we can use (3.17) to find that for all f such that $\mathbb{E}|Wf(W)| < \infty$,

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

In particular,

$$m'(\theta) = \frac{\mathbb{E}[(W'-W)(e^{\theta W'}-e^{\theta W})]}{2a}. \tag{7.5}$$

In order to bound this term, we use the convexity of the exponential function to obtain for $x > y$

$$\frac{e^x-e^y}{x-y} = \int_0^1 \exp\{tx+(1-t)y\}dt \leq \int_0^1 [te^x+(1-t)e^y]dt = \frac{e^x+e^y}{2}. \tag{7.6}$$

Combining (7.5) and (7.6), we find that for all $\theta \in \mathbb{R}$,

$$\begin{aligned} |m'(\theta)| &\leq |\theta|\frac{\mathbb{E}[(W'-W)^2(e^{\theta W'}+e^{\theta W})]}{4a} \\ &= |\theta|\frac{\mathbb{E}[(W'-W)^2e^{\theta W}]}{2a} \\ &\leq |\theta|\,\mathbb{E}[(BW+C)e^{\theta W}] \\ &\leq B|\theta|m'(\theta)+C|\theta|m(\theta), \end{aligned} \tag{7.7}$$

where the equality is by exchangeability and the penultimate inequality follow from the hypothesis (7.3). Now, since m is convex and $m'(0) = 0$, we find that $m'(\theta)/\theta > 0$ for $\theta \neq 0$. We now break the proof into two cases, corresponding to the positive and negative tails of the distribution of W.

$\theta > 0$. In this case, our calculation above implies that for $0 < \theta < 1/B$,

$$\frac{d}{d\theta}\log(m(\theta)) = \frac{m'(\theta)}{m(\theta)} \leq \frac{C\theta}{1-B\theta},$$

which yields that

$$\log(m(\theta)) \leq \int_0^\theta \frac{Cu}{1-Bu}du \leq \frac{C\theta^2}{2(1-B\theta)},$$

and from this point we easily find

$$m(\theta) \leq \exp\left\{\frac{C\theta^2}{2(1-B\theta)}\right\}.$$

According to Proposition 7.1 we now have for $t > 0$ and $0 < \theta < 1/B$,

$$\mathbb{P}(W \geq t) \leq \exp\left\{\frac{C\theta^2}{2(1-B\theta)} - \theta t\right\},$$

and choosing $\theta = t/(C + Bt)$ proves the first assertion of the theorem.

$\theta < 0$. In this case, since $m'(\theta) < 0$, (7.7) is bounded above by $-C\theta m(\theta)$ which implies

$$C\theta \leq \frac{d}{d\theta}\log(m(\theta)) < 0.$$

From this equation, some minor consideration shows that

$$\log(m(\theta)) \leq \frac{C\theta^2}{2}$$

According to Proposition 7.1 we now have for $t > 0$ and $\theta < 0$,

$$\mathbb{P}(W \leq -t) \leq \exp\left\{\frac{C\theta^2}{2} + \theta t\right\},$$

and choosing $\theta = -t/C$ proves the second assertion of the theorem.

□

Example 7.7 (Hoeffding's combinatorial CLT)**.** Let $(a_{ij})_{1\leq i,j\leq n}$ be an array of real numbers and let σ be a uniformly chosen random permutation of $\{1, \ldots, n\}$. Let

$$W = \sum_{i=1}^{n} a_{i\sigma_i} - \frac{1}{n}\sum_{i,j} a_{ij},$$

and define $\sigma' = \sigma\tau$, where τ is a uniformly chosen transposition and

$$W' = \sum_{i=1}^{n} a_{i\sigma'_i} - \frac{1}{n}\sum_{i,j} a_{ij}.$$

It is a straightforward exercise to show that that $\mathbb{E}[W] = 0$ and (W, W') is a $2/(n-1)$-Stein pair so that it may be possible to apply Theorem 7.4. In fact, we have the following result.

Proposition 7.8. *If W is defined as above with $0 \leq a_{ij} \leq 1$, then for all $t > 0$,*

$$\mathbb{P}(|W| \geq t) \leq 2\exp\left\{\frac{-t^2}{\frac{4}{n}\sum_{i,j} a_{ij} + 2t}\right\}.$$

Proof. Let (W, W') be the $2/(n-1)$-Stein pair as defined in the remarks preceding the statement of the proposition. We now have

$$\begin{aligned}\mathbb{E}[(W'-W)^2|\sigma] &= \frac{1}{n(n-1)}\sum_{i,j}\left(a_{i\sigma_i}+a_{j\sigma_j}-a_{i\sigma_j}-a_{j\sigma_i}\right)^2 \\ &\le \frac{2}{n(n-1)}\sum_{i,j}\left(a_{i\sigma_i}+a_{j\sigma_j}+a_{i\sigma_j}+a_{j\sigma_i}\right) \\ &= \frac{4}{n-1}W + \frac{8}{n(n-1)}\sum_{i,j}a_{ij}, \end{aligned} \tag{7.8}$$

where in the final equality we use that

$$\sum_{i,j}a_{i\sigma_j} = \sum_{i,j}a_{j\sigma_i} = \sum_{i,j}a_{ij}$$

and

$$W = \frac{1}{n}\sum_{i,j}a_{i\sigma_i} - \frac{1}{n}\sum_{i,j}a_{i\sigma_j} = \frac{1}{n}\sum_{i,j}a_{j\sigma_j} - \frac{1}{n}\sum_{i,j}a_{j\sigma_i}.$$

Using the expression (7.8), the result is proved by applying Theorem 7.4 with $B=1$ and $C=(2/n)\sum_{i,j}a_{ij}$. □

7.2. Application: Magnetization in the Curie-Weiss model

Let $\beta > 0$, $h \in \mathbb{R}$ and for $\sigma \in \{-1,1\}^n$ define the Gibbs measure

$$\mathbb{P}(\sigma) = Z^{-1}\exp\left\{\frac{\beta}{n}\sum_{i<j}\sigma_i\sigma_j + \beta h\sum_i \sigma_i\right\}, \tag{7.9}$$

where Z is the appropriate normalizing constant (the so-called "partition function" of statistical physics).

We think of σ as a configuration of "spins" (± 1) on a system with n sites. The spin of a site depends on those at all other sites since for all $i \neq j$, each of the terms $\sigma_i\sigma_j$ appears in the first sum. Thus, the most likely configurations are those that have many of the spins the same spin ($+1$ if $h>0$ and -1 if $h<0$). This probability model is referred to as the Curie-Weiss model and a quantity of interest is $m = \frac{1}{n}\sum_{i=1}^n \sigma_i$, the "magnetization" of the system. We show the following result found in [18].

Proposition 7.9. *If* $m = \frac{1}{n}\sum_{i=1}^n\sigma_i$*, then for all* $\beta > 0$, $h \in \mathbb{R}$*, and* $t \ge 0$,

$$\mathbb{P}\left(|m - \tanh(\beta m + \beta h)| \ge \frac{\beta}{n} + \frac{t}{\sqrt{n}}\right) \le 2\exp\left\{-\frac{t^2}{4(1+\beta)}\right\},$$

where $\tanh(x) := (e^x - e^{-x})/(e^x + e^{-x})$.

In order to the prove the proposition, we need a more general result than that of Theorem 7.4; the proofs of the two results are very similar.

Theorem 7.10. *Let (X, X') an exchangeable pair of random elements on a Polish space. Let F an antisymmetric function and define*

$$f(X) := \mathbb{E}[F(X, X')|X].$$

If $\mathbb{E}[e^{\theta f(X)}|F(X, X')|] < \infty$ for all $\theta \in \mathbb{R}$ and there are constants $B, C \geq 0$ such that

$$\frac{1}{2}\mathbb{E}\left[|(f(X) - f(X'))F(X, X')|\,\middle|X\right] \leq Bf(X) + C,$$

then for all $t > 0$,

$$\mathbb{P}(f(X) > t) \leq \exp\left\{\frac{-t^2}{2C + 2Bt}\right\} \text{ and } \mathbb{P}(f(X) \leq -t) \leq \exp\left\{\frac{-t^2}{2C}\right\}.$$

In order to recover Theorem 7.4 from the result above, note that if (W, W') is an a-Stein pair, then we can take $F(W, W') = (W - W')/a$, which implies $f(W) := \mathbb{E}[F(W, W')|W] = W$.

Proof of Proposition 7.9. In the notation of Theorem 7.10, we set $X = \sigma$ and $X' = \sigma'$, where σ is chosen according to the Gibbs measure given by (7.9) and σ' is a step from σ in the following reversible Markov chain: at each step of the chain a site from the n possible sites is chosen uniformly at random and then the spin at that site is resampled according to the Gibbs measure (7.9) conditional on the value of the spins at all other sites. This chain is most commonly known as the *Gibbs sampler*. We take $F(\sigma, \sigma') = \sum_{i=1}^{n}(\sigma_i - \sigma'_i)$ apply Theorem 7.10 to study $f(\sigma) = \mathbb{E}[F(\sigma, \sigma')|\sigma]$.

The first thing we need to do is compute the transition probabilities for the Gibbs sampler chain. Suppose the chosen site is site i, then

$$\mathbb{P}(\sigma'_i = 1|(\sigma_j)_{j\neq i}) = \frac{\mathbb{P}(\sigma_i = 1, (\sigma_j)_{j\neq i})}{\mathbb{P}((\sigma_j)_{j\neq i})},$$
$$\mathbb{P}(\sigma'_i = -1|(\sigma_j)_{j\neq i}) = \frac{\mathbb{P}(\sigma_i = -1, (\sigma_j)_{j\neq i})}{\mathbb{P}((\sigma_j)_{j\neq i})}.$$

Note that $\mathbb{P}((\sigma_j)_{j\neq i}) = \mathbb{P}(\sigma_i = 1, (\sigma_j)_{j\neq i}) + \mathbb{P}(\sigma_i = -1, (\sigma_j)_{j\neq i})$, and that

$$\mathbb{P}(\sigma'_i = 1, (\sigma_j)_{j\neq i}) = \frac{1}{Z}\exp\left\{\frac{\beta}{n}\left(\sum_{k<j, j\neq i}\sigma_j\sigma_k + \sum_{j\neq i}\sigma_j\right) + \beta h\sum_{j\neq i}\sigma_j + \beta h\right\},$$
$$\mathbb{P}(\sigma'_i = -1, (\sigma_j)_{j\neq i}) = \frac{1}{Z}\exp\left\{\frac{\beta}{n}\left(\sum_{k<j, j\neq i}\sigma_j\sigma_k - \sum_{j\neq i}\sigma_j\right) + \beta h\sum_{j\neq i}\sigma_j - \beta h\right\}.$$

Thus

$$\mathbb{P}(\sigma_i' = 1|(\sigma_j)_{j\neq i}) = \frac{\exp\{\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta h\}}{\exp\{\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta h\} + \exp\{-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta h\}},$$

$$\mathbb{P}(\sigma_i' = -1|(\sigma_j)_{j\neq i}) = \frac{\exp\{-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta h\}}{\exp\{\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta h\} + \exp\{-\frac{\beta}{n}\sum_{j\neq i}\sigma_j - \beta h\}},$$

and hence

$$\mathbb{E}[F(\sigma,\sigma')|\sigma] = \frac{1}{n}\sum_{i=1}^n \sigma_i - \frac{1}{n}\sum_{i=1}^n \tanh\left(\frac{\beta}{n}\sum_{j\neq i}\sigma_j + \beta h\right), \tag{7.10}$$

where the summation over i and the factor of $1/n$ is due to the fact that the resampled site is chosen uniformly at random (note also that for $j \neq i$, we have $\mathbb{E}[\sigma_j - \sigma_j'|\sigma \text{ and chose site } i] = 0$).

We give a concentration of measure inequality for (7.10) using Theorem 7.10, and then show that the difference between (7.10) and the quantity of interest in the proposition is almost surely bounded by a small quantity, from which the result follows.

If we denote $m_i := \frac{1}{n}\sum_{j\neq i}\sigma_j$, then

$$f(\sigma) := \mathbb{E}[F(\sigma,\sigma')|\sigma] = m - \frac{1}{n}\sum_{i=1}^n \tanh\{\beta m_i + \beta h\},$$

and we need to check the conditions of Theorem 7.10 for $f(\sigma)$. The condition involving the moment generating function is obvious since all quantities involved are finite, so we only need to find constants $B, C > 0$ such that

$$\frac{1}{2}\mathbb{E}\left[|(f(\sigma) - f(\sigma'))\, F(\sigma,\sigma')|\,\middle|\sigma\right] \leq Bf(\sigma) + C. \tag{7.11}$$

Since $F(\sigma,\sigma')$ is the difference of the sum of the spins in one step of the Gibbs sampler and only one spin can change in a step of the chain, we find that $|F(\sigma,\sigma')| \leq 2$.

Also, if we denote $m' := \frac{1}{n}\sum_{i=1}^n \sigma'$, then using that

$$|\tanh(x) - \tanh(y)| \leq |x - y|,$$

we find

$$|f(\sigma) - f(\sigma')| \leq |m - m'| + \frac{\beta}{n}\sum_{i=1}^n |m_i - m_i'| \leq \frac{2(1+\beta)}{n},$$

Hence, (7.11) is satisfied with $B = 0$ and $C = \frac{2(1+\beta)}{n}$ and Theorem 7.10 now yields

$$\mathbb{P}\left(\left|m - \frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i + \beta h)\right| > \frac{t}{\sqrt{n}}\right) \leq 2\exp\left\{-\frac{t^2}{4(1+\beta)}\right\}.$$

To complete the proof we note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\tanh(\beta m_i+\beta h)-\tanh(\beta m+\beta h)\right]\right| \leq \frac{1}{n}\sum_{i=1}^{n}|\beta m_i-\beta m| \leq \frac{\beta}{n},$$

and thus an application of the triangle inequality yields the bound in the proposition. □

7.3. Concentration of measure using size-bias couplings

As previously mentioned, the key step in the proof of Theorem 7.4 was to rewrite $m'(\theta) := \mathbb{E}[We^{\theta W}]$ using exchangeable pairs in order to get a differential inequality for $m(\theta)$. We can follow this same program, but with a size-bias coupling in place of the exchangeable pair. We follow the development of [29].

Theorem 7.11. *Let $X \geq 0$ with $\mathbb{E}[X]=\mu$ and $0<\mathrm{Var}(X)=\sigma^2<\infty$ and let X^s be a size-biased coupling of X such that $|X-X^s|\leq C<\infty$.*

1. *If $X^s \geq X$, then*

$$\mathbb{P}\left(\frac{X-\mu}{\sigma}\leq -t\right)\leq \exp\left\{\frac{-t^2}{2\left(\frac{C\mu}{\sigma^2}\right)}\right\}.$$

2. *If $m(\theta)=\mathbb{E}[e^{\theta X}]<\infty$ for $\theta = 2/C$, then*

$$\mathbb{P}\left(\frac{X-\mu}{\sigma}\geq t\right)\leq \exp\left\{\frac{-t^2}{2\left(\frac{C\mu}{\sigma^2}+\frac{C}{2\sigma}t\right)}\right\}.$$

Proof. To prove the first item, let $\theta \leq 0$ so that $m(\theta) := \mathbb{E}[e^{\theta X}]<\infty$ since $X \geq 0$. As in the proof of Theorem 7.4, we need the inequality (7.6): for all $x, y \in \mathbb{R}$,

$$\left|\frac{e^x-e^y}{x-y}\right| \leq \frac{e^x+e^y}{2}.$$

Using this fact and that $X^s \geq X$, we find

$$\mathbb{E}[e^{\theta X}-e^{\theta X^s}] \leq \frac{C|\theta|}{2}\left(\mathbb{E}[e^{\theta X}]+\mathbb{E}[e^{\theta X^s}]\right) \leq C|\theta|\mathbb{E}[e^{\theta X}]. \tag{7.12}$$

The definition of the size-bias distribution implies that $m'(\theta)=\mu\mathbb{E}[e^{\theta X^s}]$ so that (7.12) yields the differential inequality $m'(\theta)\geq \mu(1+C\theta)m(\theta)$, or put otherwise

$$\frac{d}{d\theta}\left[\log(m(\theta))-\mu\theta\right] \geq \mu C\theta. \tag{7.13}$$

Setting $\tilde{m}(\theta) = \log(m(\theta)) - \mu\theta$, (7.13) implies $\tilde{m}(\theta) \leq \mu C\theta^2/2$, and it follows that

$$\begin{aligned} \mathbb{E}\left[\exp\left\{\theta\left(\frac{X-\mu}{\sigma}\right)\right\}\right] &= m\left(\frac{\theta}{\sigma}\right)\exp\left\{-\frac{\mu\theta}{\sigma}\right\} \\ &= \exp\left\{\tilde{m}\left(\frac{\theta}{\sigma}\right)\right\} \leq \exp\left(\frac{\mu C\theta^2}{2\sigma^2}\right). \end{aligned}$$

We can now apply Proposition 7.1 to find that

$$\mathbb{P}\left(\frac{X-\mu}{\sigma} < -t\right) \leq \exp\left(\frac{\mu C\theta^2}{2\sigma^2} + \theta t\right). \tag{7.14}$$

The right hand side of this (7.14) is minimized at $\theta = -\sigma^2 t/\mu C$, and substituting this value into (7.14) yields the first item of the theorem.

For the second assertion of the theorem, suppose that $0 \leq \theta < 2/C$. A calculation similar to (7.12) above shows that

$$\frac{m'(\theta)}{\mu} - m(\theta) \leq \frac{C\theta}{2}\left(\frac{m'(\theta)}{\mu} + m(\theta)\right),$$

so that we can write

$$m'(\theta) \leq \frac{\mu\left(1 + \frac{C\theta}{2}\right)}{1 - \frac{C\theta}{2}} m(\theta).$$

Again defining $\tilde{m}(\theta) = \log(m(\theta)) - \mu\theta$, we have $\tilde{m}'(\theta) \leq C\mu\theta/(1 - \frac{C\theta}{2})$ so that

$$\tilde{m}\left(\frac{\theta}{\sigma}\right) \leq \frac{C\mu\theta^2}{\sigma^2\left(2 - \frac{C\theta}{\sigma}\right)} \quad \text{for} \quad 0 \leq \theta < \min\{2/C, 2\sigma/C\}.$$

We can now apply Proposition 7.1 to find that

$$\mathbb{P}\left(\frac{X-\mu}{\sigma} \geq t\right) \leq \exp\left(\frac{\mu C\theta^2}{\sigma^2\left(2 - \frac{C\theta}{\sigma}\right)} - \theta t\right). \tag{7.15}$$

The right hand side of this (7.15) is minimized at

$$\theta = t\left(\frac{C\mu}{\sigma^2} + \frac{Ct}{2\sigma}\right)^{-1},$$

and substituting this value into (7.15) yields the second item of the theorem. □

Theorem 7.11 can be applied in many of the examples we have discussed in the context of the size-bias transform for normal and Poisson approximation and others. We content ourselves with a short example and refer to [28] for many more applications.

Example 7.12 (Head runs)**.** We apply Theorem 7.11 to a variation of the random variable studied in the application of Section 4.2.1. Let $Y_1, \ldots, Y_n$ be i.i.d. indicator variables with $\mathbb{P}(Y_i = 1) = p$ and for $i = 1, \ldots, n$ and $k < n/2$, let

$$X_i = \prod_{j=1}^{k} Y_j,$$

where we interpret the bounds in the product "modularly," for example $X_n = Y_n Y_1 \cdots Y_{k-1}$. We say that X_i is the indicator that there is a head run of length k at position i and we set $X = \sum_{i=1}^n X_i$, the number of head runs of length k.

In order to apply Theorem 7.11, we must find a size-bias coupling of X that satisfies the hypotheses of the theorem. Since X is a sum of identically distributed indicators, we can apply the size-bias construction of Corollary 4.14. For this construction, we first realize the variables $Y_1, \ldots, Y_n$ as above. We then choose an index I uniformly from the set $\{1, \ldots, n\}$ and "condition" X_I to be one by setting $Y_I = Y_{I+1} = \cdots = Y_{I+k-1} = 1$. By defining $X_j^{(I)}$ to be the indicator that there is a head run of length k at position j after this procedure, Corollary 4.14 implies that

$$X^s = 1 + \sum_{j \neq I} X_j^{(I)}$$

is a size-bias coupling of X.

It is easy to see that $X^s \geq X$ and $|X^s - X| \leq 2k - 1$ so that an application of Theorem 7.11 yields

$$\mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2(2k-1)\left(\frac{\mu}{\sigma^2} + \frac{t}{2\sigma}\right)}\right),$$

where $\mu = np^k$ and $\sigma^2 = \mu\left(1 - p^k + \sum_{i=1}^{k-1}(p^i - p^k)\right)$.

Acknowledgments

The author thanks the students who sat in on the course on which this document is based. Special thanks go to those students who contributed notes, parts of which may appear in some form above: Josh Abramson, Miklos Racz, Douglas Rizzolo, and Rachel Wang. We also thank Haiyan Huang for help with the discussion of the biological application in Remark 4.12. Finally, we are grateful to an anonymous referee for helpful and detailed comments.

References

[1] D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs. http://www.stat.berkeley.edu/~aldous/RWG/book.html, 2010.

[2] R. Arratia, A. D. Barbour, and S. Tavaré. *Logarithmic combinatorial structures: a probabilistic approach.* EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2003. MR2032426

[3] R. Arratia and L. Goldstein. Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent? http://arxiv.org/abs/1007.3910, 2011.

[4] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989. MR0972770

[5] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statist. Sci.*, 5(4):403–434, 1990. With comments and a rejoinder by the authors. MR1092983

[6] R. Arratia, L. Gordon, and M. S. Waterman. The Erdős-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.*, 18(2):539–570, 1990. MR1056326

[7] K. B. Athreya and P. E. Ney. *Branching processes*. Springer-Verlag, New York, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196. MR0373040

[8] F. Avram and D. Bertsimas. On central limit theorems in geometrical probability. *Ann. Appl. Probab.*, 3(4):1033–1046, 1993. MR1241033

[9] P. Baldi, Y. Rinott, and C. Stein. A normal approximation for the number of local maxima of a random function on a graph. In *Probability, statistics, and mathematics*, pages 59–81. Academic Press, Boston, MA, 1989. MR1031278

[10] A. D. Barbour. Poisson convergence and random graphs. *Math. Proc. Cambridge Philos. Soc.*, 92(2):349–359, 1982. MR0671189

[11] A. D. Barbour and L. H. Y. Chen, editors. *An introduction to Stein's method*, volume 4 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Singapore University Press, Singapore, 2005. Lectures from the Meeting on Stein's Method and Applications: a Program in Honor of Charles Stein held at the National University of Singapore, Singapore, July 28–August 31, 2003.

[12] A. D. Barbour, L. Holst, and S. Janson. *Poisson approximation*, volume 2 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1992. Oxford Science Publications. MR1163825

[13] A. D. Barbour, M. Karoński, and A. Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *J. Combin. Theory Ser. B*, 47(2):125–145, 1989. MR1047781

[14] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001. MR1864966

[15] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures Algorithms*, 18(3):279–290, 2001. MR1824277

[16] O. Bousquet, S. Boucheron, and G. Lugosi. Concentration inequalities. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003*, Lecture Notes in Artificial Intelligence, pages 208–240. Springer, 2004.

[17] S. Chatterjee. Concentration inequalities with exchangeable pairs. http://arxiv.org/abs/math/0507526, 2005. Ph.D. dissertation, Stanford University. MR2707160

[18] S. Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138(1-2):305–321, 2007. MR2288072

[19] S. Chatterjee and P. S. Dey. Applications of Stein's method for concentration inequalities. *Ann. Probab.*, 38(6):2443–2485, 2010. MR2683635

[20] S. Chatterjee, P. Diaconis, and E. Meckes. Exchangeable pairs and Poisson approximation. *Probab. Surv.*, 2:64–106 (electronic), 2005. MR2121796

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

[22] S. Chatterjee and Q.-M. Shao. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. Appl. Probab.*, 21(2):464–483, 2011. MR2807964

[23] S. Chatterjee and Students. Stein's method course notes. http://www.stat.berkeley.edu/~sourav/stat206Afall07.html, 2007.

[24] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011. MR2732624

[25] L. H. Y. Chen and Q.-M. Shao. Normal approximation under local dependence. *Ann. Probab.*, 32(3A):1985–2028, 2004. MR2073183

[26] P. Diaconis and S. Holmes, editors. *Stein's method: expository lectures and applications*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 46. Institute of Mathematical Statistics, Beachwood, OH, 2004. Papers from the Workshop on Stein's Method held at Stanford University, Stanford, CA, 1998. MR2118599

[27] P. Donnelly and D. Welsh. The antivoter problem: random 2-colourings of graphs. In *Graph theory and combinatorics (Cambridge, 1983)*, pages 133–144. Academic Press, London, 1984. MR0777170

[28] S. Ghosh and L. Goldstein. Applications of size biased couplings for concentration of measures. *Electronic Communications in Probability*, 16:70–83, 2011. MR2763529

[29] S. Ghosh and L. Goldstein. Concentration of measures via size-biased couplings. *Probability Theory and Related Fields*, 149:271–278, 2011. 10.1007/s00440-009-0253-3. MR2773032

[30] L. Goldstein. Berry-Esseen bounds for combinatorial central limit theorems and pattern occurrences, using zero and size biasing. *J. Appl. Probab.*, 42(3):661–683, 2005. MR2157512

[31] L. Goldstein. A probabilistic proof of the Lindeberg-Feller central limit theorem. *Amer. Math. Monthly*, 116(1):45–60, 2009. MR2478752

[32] L. Goldstein. A Berry-Esseen bound with applications to counts in the Erdös-Rényi random graph. http://arxiv.org/abs/1005.4390, 2010.

[33] L. Goldstein and G. Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.*, 7(4):935–952, 1997. MR1484792

[34] L. Goldstein and Y. Rinott. Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.*, 33(1):1–17, 1996. MR1371949

[35] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001. MR2059709

[36] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. MR0144363

[37] S. Janson, T. Łuczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000. MR1782847

[38] T. M. Liggett. *Interacting particle systems*. Classics in Mathematics. Springer-Verlag, Berlin, 2005. Reprint of the 1985 original. MR2108619

[39] R. A. Lippert, H. Huang, and M. S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA*, 99(22):13980–13989 (electronic), 2002. MR1944413

[40] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann. Probab.*, 23(3):1125–1138, 1995. MR1349164

[41] E. Peköz. Stein's method for geometric approximation. *J. Appl. Probab.*, 33(3):707–713, 1996. MR1401468

[42] E. Peköz, A. Röllin, and N. Ross. Total variation and local limit error bounds for geometric approximation. `http://arxiv.org/abs/1005.2774`, 2010. To appear in *Bernoulli*.

[43] E. A. Peköz and A. Röllin. New rates for exponential approximation and the theorems of Rényi and Yaglom. *Ann. Probab.*, 39(2):587–608, 2011. MR2789507

[44] J. Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros. *J. Combin. Theory Ser. A*, 77(2):279–303, 1997. MR1429082

[45] G. Reinert. Three general approaches to Stein's method. In *An introduction to Stein's method*, volume 4 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 183–221. Singapore Univ. Press, Singapore, 2005. MR2235451

[46] G. Reinert, D. Chew, F. Sun, and M. S. Waterman. Alignment-free sequence comparison. I. Statistics and power. *J. Comput. Biol.*, 16(12):1615–1634, 2009. MR2578699

[47] Y. Rinott and V. Rotar. On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics. *Ann. Appl. Probab.*, 7(4):1080–1105, 1997. MR1484798

[48] A. Röllin. A note on the exchangeability condition in Stein's method. `http://arxiv.org/abs/math/0611050v1`, 2006.

[49] L. H. Y. Chen and A. Röllin. Stein couplings for normal approximation. http://arxiv.org/abs/1003.6039, 2010.

[50] A. Röllin and N. Ross. A probabilistic approach to local limit theorems with applications to random graphs. http://arxiv.org/abs/1011.3100, 2010.

[51] S. Ross and E. Peköz. *A second course in probability*. www.ProbabilityBookstore.com, Boston, 2007.

[52] A. Ruciński. When are small subgraphs of a random graph normally distributed? *Probab. Theory Related Fields*, 78(1):1–10, 1988. MR0940863

[53] Q.-M. Shao and Z.-G. Su. The Berry-Esseen bound for character ratios. *Proc. Amer. Math. Soc.*, 134(7):2153–2159 (electronic), 2006. MR2215787

[54] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, Berkeley, Calif., 1972. Univ. California Press. MR0402873

[55] C. Stein. *Approximate computation of expectations*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7. Institute of Mathematical Statistics, Hayward, CA, 1986. MR0882007

[56] I. S. Tyurin. Sharpening the upper bounds for constants in Lyapunov's theorem. *Uspekhi Mat. Nauk*, 65(3(393)):201–201, 2010. MR2682728

[57] L. Wan, G. Reinert, F. Sun, and M. S. Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J. Comput. Biol.*, 17(11):1467–1490, 2010. MR2739743

[58] M. Waterman. *Introduction to computational biology*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 1995.