J. Appl. Prob. **49,** 745–757 (2012)
Printed in England
© *Applied Probability Trust* 2012

A PÓLYA APPROXIMATION TO THE POISSON-BINOMIAL LAW

MAX SKIPPER,* *University of Oxford and The University of Sydney*

Abstract

Using Stein's method, we derive explicit upper bounds on the total variation distance between a Poisson-binomial law (the distribution of a sum of independent but not necessarily identically distributed Bernoulli random variables) and a Pólya distribution with the same support, mean, and variance; a nonuniform bound on the pointwise distance between the probability mass functions is also given. A numerical comparison of alternative distributional approximations on a somewhat representative collection of case studies is also exhibited. The evidence proves that no single one is uniformly most accurate, though it suggests that the Pólya approximation might be preferred in several parameter domains encountered in practice.

Keywords: Stein's method; Poisson-binomial approximation; Bernoulli variable

2010 Mathematics Subject Classification: Primary 62E17
Secondary 60F05; 60C99

1. Introduction

The Poisson-binomial distribution is the law of the number of successes in a sequence of independent, nonidentically distributed trials and, as such, has found utility in several modelling applications, including Bayesian heirarchical models [24], generalized linear models [15], and noisy threshold models [21]. What is more, as the convolution product of nonidentical two-point distributions, the Poisson-binomial distribution is also intimately linked to several combinatorial and occupancy problems [26], including the 'birthday paradox' [32].

Naturally, in some contexts it is preferable to compute the point probabilities exactly and for this the interested reader is referred to two quadratic-time (in the worst case) algorithms reviewed in [15]. In other contexts, explicit bounds unrelated to an approximate distribution may be the principal interest and for this we refer the reader to [25] and the references therein. Here however the focus is on measure-valued approximations and, more specifically, distributions on the nonnegative integers $\mathbb{Z}^+$.

The approximation of the Poisson-binomial law by more easily calculable laws on $\mathbb{Z}^+$ has a long history. The earliest work focused on the Poisson approximation, of which a detailed account can be found in [6]. Subsequently, variations of the Poisson law were used: Poisson–Charlier signed measures (see [6] and the references therein), (signed) compound Poisson [5], [9], [22], shifted Poisson [4], [13]; as were the binomial law and its variations: binomial [13], [17], [30], 'almost' binomial [33], signed binomial-Krawtchouk [28], (signed) compound binomial [10], [11], [12], shifted binomial [24], [27]. Among more recent work, a class of signed measures was introduced in [7] that yielded notably impressive approximations in the special case of counting records in an independent and identically distributed sequence,

Received 8 September 2011; revision received 30 January 2012.

* Postal address: School of Mathematics and Statistics, The University of Sydney, Sydney, 2006, Australia.
Email address: m.skipper@maths.usyd.edu.au

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

a two-parameter polynomial birth–death distribution was proposed in [8], and an arbitrary Gibbs measure approximation was developed in [18].

In this paper we propose a Pólya approximation to the Poisson-binomial law, there being several motivating factors. For starters, it has already been noted that approximations by signed measures (usually perturbations of common distributions) may be undesirable because they can be time consuming to use in practice (cf. [24]) and lead to the approximation of positive quantities with negative quantities (cf. [8]). Moreover, the three parameters of the proposed Pólya approximation are easy to interpret, relating to the fact that its support, mean, and variance *exactly* match those of the Poisson-binomial distribution—a property not shared by any previously studied approximation. Also, the Pólya approximation is exact in the two special cases where the Poisson-binomial distribution is in fact either binomial or hypergeometric (see [34] for a proof that the hypergeometric distribution is a special case of the Poisson-binomial distribution).

Under a mild restriction on the dispersion of the success probabilities of the individual trials, our main result, Theorem 1, quantifies the error of the Pólya approximation via an upper bound on the total variation distance between the Pólya and Poisson-binomial laws; it also includes a nonuniform upper bound for the distance between corresponding point probabilities. A subsequent look at the orders of known total variation approximations by distributions (rather than signed measures) shows that the Pólya approximation is most accurate at least when the success probabilities are sufficiently tightly clustered. A further numerical study builds a more precise, yet less general picture. In particular, the evidence would suggest that the Pólya approximation is preferable for the application of Bayesian hierarchical models discussed in [24]. However, among the approximations considered, ultimately we see that none is uniformly most accurate.

Theorem 1 is proved via Stein's method which is now widely known and used; for a general introduction, we refer the reader to [2] or [31]. Moreover, with the potential for generalization we make a modest attempt to give a constructive proof highlighting the mathematical rationale behind, firstly, the Pólya distribution as a 'target' measure and, secondly, the specific choice of parameters; the essence being a deductive process used to ensure that the characterising operator of the approximating measure closely resembles that of the Poisson-binomial law.

There are several directions in which hopefully our result can be extended. In particular, previous total variation approximations to the sum of independent Bernoulli variables have been extended to the Wasserstein metric [4], [36], generalised Poisson-binomial distributions [10], [11], [12], sums of independent Bernoulli vectors [1], Bernoulli processes [35], and sums of dependent Bernoulli variables [6], [27], [30]. In addition, the Pólya distribution is a three-parameter, quadratic polynomial birth–death distribution as defined in [8] and also a generalized hypergeometric factorial moment distribution as defined in [20]; thus, in combining what is known about the apparently similar families of distributions, there is the potential to increase the accuracy of approximations through deducing similar results for higher-order polynomials (see [29] for a more detailed sketch).

2. Definitions and main result

We let $\Pi(n, p, \theta)$ denote a Pólya law with parameters $n \in \mathbb{N}$, $p \in (0, 1)$, and $\theta \in \mathbb{R}$. The probability mass function $\pi_k \equiv \Pi(n, p, \theta)\{k\}$ at an arbitrary point k is given by

$$\pi_k = \binom{n}{k} \frac{p(p+\theta)\cdots(p+(k-1)\theta)q(q+\theta)\cdots(q+(n-k-1)\theta)}{(1+\theta)(1+2\theta)\cdots(1+(n-1)\theta)}.$$

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

Here $\Pi(n, p, \theta)$ is a distribution with support equal to $\{0, 1, \dots, n\}$ provided that

$$\theta > -\frac{p \wedge q}{n-1} \tag{1}$$

(where we use $x \wedge y \equiv \min(x, y)$, $x \vee y \equiv \max(x, y)$, and $q = 1 - p$ here and subsequently); otherwise, it is a signed measure, generally with infinite support. If, for some positive integers a and b, $a + b > n$, $p = a/(a+b)$, and $\theta = -1/(a+b)$, then a hypergeometric distribution results; if $\theta = 0$, a binomial law is recovered. When (1) is satisfied, the mean and variance of the Pólya distribution are np and $npq((1+n\theta)/(1+\theta))$, respectively. For more background on the Pólya distribution, see [20].

Now, let $I_1, \dots, I_n$ be independent indicator random variables with distribution

$$\mathrm{P}(I_i = 1) = p_i, \qquad \mathrm{P}(I_i = 0) = q_i, \qquad p_i + q_i = 1,\ i = 1, \dots, n,$$

and set $W = \sum_{i=1}^n I_i$, $\lambda = \mathrm{E}\, W$, $\sigma^2 = \operatorname{var} W$. From now on, assume that

$$p = \frac{1}{n}\sum_{i=1}^n p_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^n (p_i - p)^2, \qquad \theta = -\frac{ns^2}{\kappa}, \qquad \kappa = \lambda(n-\lambda) - \sigma^2.$$

Denoting the total variation distance between two measures μ and ν on $\mathbb{Z}^+$ by

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \subset \mathbb{Z}^+} |\mu(A) - \nu(A)|,$$

we may now state the main result.

Theorem 1. *If (1) holds then*

$$d_{\mathrm{TV}}(\mathcal{L}W, \Pi(n, p, \theta)) \le \frac{sK\sqrt{n}(\sqrt{\lambda_{02}} + \sqrt{\lambda_{20}})}{\kappa\sigma^2(1+\theta)(2^{-1} \vee \sqrt{\sigma^2 - 1})}, \tag{2}$$

$$|\mathrm{P}(W = m) - \pi_m| \le 2((\lambda \vee m) - mp + \theta m(n-m))^{-1} \frac{sK\sqrt{n}(\sqrt{\lambda_{02}} + \sqrt{\lambda_{20}})}{\kappa(2^{-1} \vee (\sigma^2 - 2))},$$

where $\lambda_{jk} = \sum_{i=1}^n p_i^j q_i^k$, $j, k = 0, 1, 2$, *and*

$$K^2 = (n-1)\lambda_{22}(\lambda_{20}\lambda_{02} - \lambda_{22}) - (\lambda_{12}\lambda_{10} - \lambda_{22})^2 - (\lambda_{01}\lambda_{21} - \lambda_{22})^2 + (\lambda_{11}^2 - \lambda_{22})^2.$$

Remarks 1. 1. Condition (1) is a restriction on the size of the sample variance of the p_i relative to the sample mean. In particular, a slightly stronger substitute condition is

$$s^2 \le pq(p \wedge q). \tag{3}$$

Thinking of the p_i as independent samples from some parent distribution on $[0, 1]$ helps build a picture of when this condition should be satisfied and when it should fail. For example, if the p_i resemble a typical sample from any uniform distribution on a subinterval of $[0, 1]$, then (3) should be satisfied; if they resemble a typical sample from a $\mathrm{Beta}(\alpha, \beta)$ distribution, then (3) should be satisfied if and only if

$$\frac{\alpha \vee \beta}{\alpha \wedge \beta} \le \alpha + \beta;$$

if they form a harmonic sequence (as studied in [7]) then $s^2 = O(n^{-1})$ and $p = O(n^{-1} \ln n)$ as $n \to \infty$, so (3) cannot hold for large n. The reader may extrapolate as they wish.

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

2. While Theorem 1 is only proved under assumption (1), some consideration suggests the condition is not necessary for the approximation to be a good one, only that the approximating measure be a distribution with support equal to $\{0, \dots, n\}$ rather than a signed measure with (generally) infinite support.

3. The bounds for $d_{\mathrm{TV}}(\mathcal{L}W, \Pi(n, p, \theta))$ are invariant with respect to the substitutions $p_i \to q_i$, so in some sense there is a degree of uniform applicability with respect to the magnitude of the p_is.

4. It is evident that as the p_i tend towards $p, \theta \to 0$ and so $\mathcal{L}W \to \Pi(n, p, 0) = \mathrm{Bi}(n, p)$ as is appropriate.

3. Proof of the main result

Theorem 1 was the result of an explorative process aimed at deducing an approximate characterising operator for $\mathcal{L}W$. Recall that the defining property of a characterising operator $\mathcal{A}$ for a measure μ is that $X \sim \mu$ if and only if $\mathrm{E}\,\mathcal{A}f(X) = 0$ for all functions f in the domain of $\mathcal{A}$. Presently, we give a characterising operator for $\mathcal{L}W$ and state some associated properties; the proofs are deferred to Appendix A. We will use the forward difference notation $\Delta f(x) = f(x+1) - f(x)$ and $\Delta^2 f(x) = \Delta(\Delta f(x))$.

Proposition 1. *A characterising operator for $\mathcal{L}W$ is given by*

$$\mathcal{B}f(x) = (\lambda - g(x))f(x+1) - (x - g(x))f(x),$$

where

$$g(x) = \sum_{j=1}^{n} p_j \,\mathrm{E}\{I_j \mid W = x\}. \tag{4}$$

Proposition 2. *With $g(x)$ defined by (4), we have*

1. $g(0) = 0$, $g(n) = \lambda$,
2. $0 < \Delta g(x) < 1$, $x = 0, \dots, n-1$,
3. $\Delta(g(x)/x) \le 0$, $x = 1, \dots, n-1$, *with equality if and only if all the p_i are equal.*

Given the results above, it is intuitive to attempt to approximate $\mathcal{L}W$ with a distribution μ whose characterising operator $\mathcal{A}$ takes the form

$$\mathcal{A}f(x) = (\lambda - h(x))f(x+1) - (x - h(x))f(x),$$

where the function h is chosen to approximate g. Unfortunately, unless all the $p_j = p$ are equal, in which case $g(x) = \sum_{j=1}^{n} p\,\mathrm{E}\{I_j \mid W = x\} = px$, the approximation of $g(x)$ is not entirely trivial. Nevertheless, we can at least require that any candidate h satisfy the three basic properties of g listed above; as a point of reference, note that taking $h \equiv 0$ leads to a $\mathrm{Po}(\lambda)$ approximation, while with $h(x) = \lambda x/m$ we recover a $\mathrm{Bi}(m, \lambda/m)$ approximation—neither choice meeting the requirement.

Arguably, the simplest h that meets the stated requirement takes the form

$$h(x) = px + \psi x(n - x),$$

with $\psi < (p \wedge q)/(n-1)$ to ensure that property 2 of Proposition 2 is satisfied and $\psi \ge 0$ to ensure that property 3 is satisfied. It is a simple exercise to verify that this latest choice implies

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

that $\mu \equiv \Pi(n, p, -\psi)$; that the natural choice for ψ is $\psi = -\theta$ will become evident shortly. We now proceed with Stein's method.

For any $A \subset \mathbb{Z}_+$, let f_A be the solution to the Stein equation

$$\mathcal{A} f_A(x) = \mathbf{1}_A(x) - \mu(A),$$

where $\mathbf{1}_A(x)$ is the indicator function of the set A. It follows that

$$d_{\mathrm{TV}}(\mathcal{L}W, \mu) := \sup_{A \subset \mathbb{Z}_+} |\mathrm{P}(W \in A) - \mu(A)| = \sup_{f \in \mathcal{F}} |\mathrm{E}\mathcal{A} f(W)|,$$

where $\mathcal{F} = \{f_A \colon A \subset \mathbb{Z}_+\}$—we shall proceed to bound the right-hand side. Making use of the conventional abbreviation $W_{j_1 \cdots j_k} = W - \sum_{j \in \{j_1, \ldots, j_k\}} I_j$ and assuming that $\psi = -\theta$, we first show that

$$|\mathrm{E}\mathcal{A} f(W)| \le \frac{C_f s K \sqrt{n}(\sqrt{\lambda_{02}} + \sqrt{\lambda_{20}})}{\kappa},$$

where

$$C_f = \max_{m=1,2} \max_{i,j,k,l} |\mathrm{E}\Delta^2 f(W_{ijkl} + m)|.$$

Then, to complete the proof, we verify that

$$\sup_{f \in \mathcal{F}} C_f \le (2 \wedge (\sigma^2 - 1)^{-1/2})[(1+\theta)\sigma^2]^{-1}, \tag{5}$$

$$C_{f_{\{m\}}} \le 2(2 \wedge (\sigma^2 - 2)^{-1})((\lambda \vee m) - mp + \theta m(n-m))^{-1}, \tag{6}$$

provided (1) holds.

Using the fact that $\mathrm{E}\,\mathcal{B} f(W) = 0$, we obtain

$$\begin{aligned} n\,\mathrm{E}\,\mathcal{A} f(W) &= n\,\mathrm{E}\{g(W) - h(W)\}\Delta f(W) \\ &= \mathrm{E}\left\{n \sum p_i I_i - \lambda W - n\psi W(n - W)\right\}\Delta f(W) \\ &= \sum_{i,j} \mathrm{E}\{p_i I_i - p_i I_j - n\psi I_i(1 - I_j)\}\Delta f(W). \end{aligned}$$

Now, by interchanging i and j then adding, we obtain

$$\begin{aligned} 2n\,\mathrm{E}\,\mathcal{A} f(W) &= \sum_{i,j} \mathrm{E}\{(p_i - p_j)(I_i - I_j) - n\psi(I_i - 2I_iI_j + I_j)\}\Delta f(W), \\ &= \sum_{i,j} \mathrm{E}\{(p_i - p_j)(I_i - I_j) - n\psi(I_i - I_j)^2\}\Delta f(W), \end{aligned} \tag{7}$$

since $I_j^2 = I_j$. Furthermore, since $I_i + I_j = 1$ whenever $I_i \ne I_j$, it follows that

$$(I_i - I_j)\Delta f(W) = (I_i - I_j)\Delta f(W_{ij} + I_i + I_j) = (I_i - I_j)\Delta f(W_{ij} + 1),$$

even if $i = j$, and so (7) may be written as

$$n\,\mathrm{E}\,\mathcal{A} f(W) = \sum_{i<j} \mathrm{E}\{(p_i - p_j)(I_i - I_j) - n\psi(I_i - I_j)^2\}\Delta f(W_{ij} + 1).$$

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

The independence of I_i, I_j, and W_{ij} then implies that

$$n\,\mathrm{E}\,\mathcal{A}f(W)=\sum_{i<j}(\mathrm{E}^2(I_i-I_j)-n\psi\,\mathrm{E}(I_i-I_j)^2)\,\mathrm{E}\,\Delta f(W_{ij}+1). \tag{8}$$

Now, if we define

$$d^{(x)}_{ij,kl}=\mathrm{P}(I_i+I_j=2x)-\mathrm{P}(I_k+I_l=2x), \qquad x=0,1,$$

then a little algebra establishes the identity

$$\mathrm{E}^2(I_i-I_j)\,\mathrm{E}(I_k-I_l)^2-\mathrm{E}(I_i-I_j)^2\,\mathrm{E}^2(I_k-I_l)=d^{(0)}_{ik,jl}d^{(1)}_{jk,il}+d^{(1)}_{ik,jl}d^{(0)}_{jk,il}.$$

Thus, if we choose to let

$$\psi=\frac{\sum_{i<j}\mathrm{E}^2(I_i-I_j)}{n\sum_{i<j}\mathrm{E}(I_i-I_j)^2},$$

then substitute into (8), we arrive at

$$n\kappa\,\mathrm{E}\,\mathcal{A}f(W)=\sum_{i<j,k<l}\sum_{x=0,1}d^{(x)}_{ik,jl}d^{(1-x)}_{jk,il}\,\mathrm{E}\,\Delta f(W_{ij}+1),$$

having made use of the identity $\sum_{i<j}\mathrm{E}(I_i-I_j)^2=\kappa$. Note that our choice implies that $\psi=-\theta$. This time, using the symmetry of the pairs (i,j) and (k,l), we obtain

$$2n\kappa\,\mathrm{E}\,\mathcal{A}f(W)=\sum_{i<j,k<l}\sum_{x=0,1}d^{(x)}_{ik,jl}d^{(1-x)}_{jk,il}\,\mathrm{E}\,\Delta\{f(W_{ij}+1)-f(W_{kl}+1)\}.$$

Now, introduce $f^*_{ijkl}(x):=\mathrm{E}\,f(W_{ijkl}+x+1)$ and $x_{ijkl}\in\{0,1\}$, where $x_{ijkl}=1$ if i, j, k, and l are all mutually distinct, and $x_{ijkl}=0$ otherwise. Supposing that $x_{ijkl}=1$, the use of summation by parts yields

$$\begin{aligned}\mathrm{E}\,\Delta\{f(W_{ij}+1)-f(W_{kl}+1)\}&=\mathrm{E}\,\Delta\{f^*_{ijkl}(I_k+I_l)-f^*_{ijkl}(I_i+I_j)\}\\&=\sum_{x=0}^{2}\Delta^2 f^*_{ijkl}(x)[\mathrm{P}(I_k+I_l>x)-\mathrm{P}(I_i+I_j>x)]\\&=d^{(0)}_{ij,kl}\Delta^2 f^*_{ijkl}(0)+d^{(1)}_{kl,ij}\Delta^2 f^*_{ijkl}(1).\end{aligned}$$

A similar calculation accompanies the case $x_{ijkl}=0$ and so we obtain

$$2n\kappa\,\mathrm{E}\,\mathcal{A}f(W)=\sum_{i<j,k<l}\sum_{x=0,1}d^{(x)}_{ik,jl}d^{(1-x)}_{jk,il}(d^{(0)}_{ij,kl}\Delta^2 f^*_{ijkl}(0)+d^{(1)}_{kl,ij}\Delta^2 f^*_{ijkl}(x_{ijkl})).$$

In the interests of pursuing best-possible bounds, we note that, since $\Delta^2 f^*_{ijkl}(x)$ is constant with respect to permutations of i, j, k, and l, we could rewrite the above in the form

$$2n\kappa\,\mathrm{E}\,\mathcal{A}f(W)=\sum_{\alpha<\beta\le\gamma<\delta}\sum_{x=0,1}c^{(x)}_{\alpha\beta\gamma\delta}\Delta^2 f^*_{ijkl}(x)$$

for some coefficients $c^{(x)}_{\alpha\beta\gamma\delta}$ dependent only on $\{p_\alpha,p_\beta,p_\gamma,p_\delta\}$. Unfortunately, given an arbitrary set of probabilities $\{p_i\}_{i=1}^n$, it cannot be guaranteed that, for fixed x, the coefficients

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

$c^{(x)}_{\alpha\beta\gamma\delta}$ are all of the same sign for each choice of α, β, γ, and δ—this is in contrast to analogous binomial and Poisson approximations. Thus, we must trade some accuracy to obtain bounds computable in a single parse of the p_i. Using Cauchy's inequality, we follow one particular route among many:

$$\begin{aligned}
|\mathrm{E}\mathcal{A}f(W)| &\le \frac{1}{2n\kappa}C_f \sum_{i<j,k<l}\sum_{x=0,1}\sum_{y=0,1} |d^{(x)}_{ik,jl}d^{(1-x)}_{jk,il}\,d^{(y)}_{ij,kl}| \\
&= \frac{1}{4n\kappa}C_f \sum_{x=0,1}\sum_{i\neq j,\,k\neq l} |d^{(x)}_{ik,jl}d^{(1-x)}_{jk,il}\,d^{(1-x)}_{ij,kl}| \\
&\le \frac{1}{4n\kappa}C_f \bigg[\sum_{i\neq j,\,k\neq l}(d^{(0)}_{ik,jl}d^{(1)}_{jk,il})^2\bigg]^{1/2} \sum_{x=0,1}\bigg[\sum_{i,j,k,l}(d^{(x)}_{ij,kl})^2\bigg]^{1/2} \\
&\le \frac{C_f sK\sqrt{n}(\sqrt{\lambda_{02}}+\sqrt{\lambda_{20}})}{\kappa}.
\end{aligned}$$

Here we have used $\sum_{i\neq j,\,k\neq l}(d^{(0)}_{ik,jl}d^{(1)}_{jk,il})^2 = 4K^2$,

$$\sum_{i,j,k,l}(d^{(1)}_{ij,kl})^2 = 2(n^2\lambda_{20}^2 - \lambda_{10}^4) \le 4n^2\lambda_{20}\bigg(\lambda_{20} - \frac{\lambda_{10}^2}{n}\bigg) = 4n^3s^2\lambda_{20},$$

and, similarly, $\sum_{i,j,k,l}(d^{(0)}_{ij,kl})^2 \le 4n^3s^2\lambda_{02}$.

Now, to establish (5) and (6), we need only combine the following facts. We use the abbreviation $D(X) \equiv d_{\mathrm{TV}}(\mathcal{L}(X+1), \mathcal{L}X)$.

1. Using summation by parts,

$$C_f \le 2\max_x |\Delta f(x)| \max_{i,j,k,l} D(W_{ijkl}),$$

$$C_f \le 4\max_x |f(x)| \max_{i,j,k,l} D(X^{(1)}_{ijkl})D(X^{(2)}_{ijkl}),$$

   where the $X^{(m)}_{ijkl}$, $m = 1, 2$, are the Poisson-binomial random variables defined below.

2. An explicit formula for $f_{\{m\}}(x)$ is given in [8, Lemma 2.3], from which it follows that

$$\max_x |f_{\{m\}}(x)| \le \frac{1}{\lambda - h(m)} \wedge \frac{1}{m-h(m)} = ((\lambda \vee m) - mp + \theta m(n-m))^{-1}.$$

3. Also, since (1) implies that $\Delta(\lambda - h(x)) \le 0 \le \Delta(x - h(x))$ for $x = 0, \dots, n-1$, it follows from [8, Theorem 2.10] that

$$\begin{aligned}
\max_x \sup_{f\in\mathcal{F}} |\Delta f(x)| &\le \max_x((\lambda \vee x) - xp + \theta x(n-x))^{-1} \\
&\le (\lambda(1-p) + \theta\lambda(n-\lambda))^{-1} \\
&= \frac{1}{(1+\theta)\sigma^2}.
\end{aligned}$$

4. For any Poisson-binomial random variable X,

$$D(X) = \max_x \mathrm{P}(X = x) \le \tfrac{1}{2}(\operatorname{var} X)^{-1/2},$$

   where the equality holds since X is unimodal (Newton's inequality—cf. [19, p. 249]) and the inequality follows from [3, Lemma 1].

5. We may decompose $W_{ijkl} = X^{(1)}_{ijkl} + X^{(2)}_{ijkl}$ in such a way that
$$\operatorname{var} X^{(1)}_{ijkl} \geq \operatorname{var} X^{(2)}_{ijkl} \geq \operatorname{var} X^{(1)}_{ijkl} - \nu^*,$$
where $\nu^* = \max_k \operatorname{var} I_k \leq \frac{1}{4}$. Thus,
$$\min_{i,j,k,l} \operatorname{var} W_{ijkl} \geq \sigma^2 - 1, \qquad \min_{i,j,k,l} \operatorname{var} X^{(m)}_{ijkl} > \frac{\sigma^2 - 1}{2}, \quad m = 1, 2.$$

Note that we used $D(W_{ijkl}) \leq \frac{1}{2}(\operatorname{var} W_{ijkl})^{-1/2}$ since despite a larger asymptotically non-dominant term, it improves an analogous bound derived from [23, Corollary 1.6] (and used in [24]) by a factor of roughly 1.6 in the asymptotically dominant term.

4. Comparison of distributional approximations

It was proved in [16] that, under a mild restriction on λ, a binomial law with parameters n and p is a more accurate approximation in total variation distance to the Poisson-binomial law than a Poisson law with parameter λ. However, it remains a challenging open problem to find nontrivial subsets of the parameter domain for which an ordering on the accuracies of known approximations to the Poisson-binomial law can be proved. In the absence of such results, the best we can offer is an attempt (under space constraints) to give a representative numerical comparison of the various distributional approximations to the Poisson-binomial law. The distributions we consider, along with corresponding orders of approximation (which we do not claim are necessarily sharp), are listed in Table 1. Three numerical case studies and a summary of conclusions follow.

For each $i = 1, \ldots, 7$ in the first column of Table 1, let μ_i denote the corresponding distribution listed in the second column and let m_i be the number of moments of $\mathcal{L}W$ that, by design, are at least approximately matched to those of the approximating law μ_i. Furthermore, let $e_i = d_{\text{TV}}(\mathcal{L}W, \mu_i)$ and let b_i be the upper bound for e_i as given in the corresponding reference. The order estimates in the table are derived from the b_i and are intended to be understood in some asymptotic regime where $p \to 0$, $l := \max_k p_k - \min_k p_k \to 0$, but $\lambda \not\to 0$. We feel that the order estimates give a reasonable indication of the performance of each approximation but that they do not tell the whole story. The numerical examples are intended to advance a broader intuition.

Example 1. In this example we intend to convey the typical behaviour of the various approximations when applied to the Bayesian hierarchical modelling problem described in [24].

TABLE 1: Approximating distributions for the Poisson-binomial law. The references do not necessarily cite the original results, merely the equations we use for the purpose of comparison. PBD means the polynomial birth–death distribution defined in [8].

i	μ_i	m_i	$c_i : b_i = O(c_i)$	Reference
1	Poisson	1	$p + l$	[6, Equation (1.23)]
2	Binomial	1	l^2/p	[24, Equation (1.1)]
3	Shifted Poisson	2	$(p + l)/\sqrt{\lambda}$	[13, Equation (2.1)]
4	PBD	2	$(p + l)^2/\sqrt{\lambda}$	[8, Equation (3.4)]
5	Binomial	2	$l^2/\sqrt{\lambda}$	[24, Equation (1.2)]
6	Pólya	2	$l^3/(p\sqrt{\lambda})$	Equation (2)
7	Shifted binomial	3	l^2/λ	[24, Equation (2.3)]

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

TABLE 2: Approximation error when the p_ks are assumed to come from a beta distribution with parameters $\alpha = 2$ and β as indicated.

i	$\beta = 2$		$\beta = 5$		$\beta = 8$		$\beta = 11$	
	e_i	b_i	e_i	b_i	e_i	b_i	e_i	b_i
1	0.217 75	0.599 62	0.112 90	0.374 24	0.076 65	0.271 88	0.058 08	0.213 45
2	0.053 71	0.199 03	0.032 08	0.124 00	0.022 85	0.089 95	0.017 74	0.070 55
3	0.008 92	0.039 24	0.006 34	0.029 65	0.005 22	0.022 93	0.004 75	0.024 66
4	0.004 63	0.113 15	0.001 86	0.038 56	0.001 08	0.021 56	0.000 73	0.014 32
5	0.001 81	0.048 07	0.000 77	0.020 30	0.000 46	0.012 14	0.000 32	0.008 32
6	0.000 01	0.007 36	0.000 11	0.003 45	0.000 09	0.002 14	0.000 07	0.001 49
7	0.000 19	0.015 71	0.000 25	0.010 82	0.000 71	0.018 73	0.000 76	0.015 95

TABLE 3: The point tabulated is $-\log_{10} b_i$, where the p_ks are assumed to come from a scaled binomial distribution with parameters $(50, 0.5)$.

i	λ									
	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
1	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
2	1.80	1.72	1.70	1.70	1.70	1.70	1.70	1.70	1.70	1.70
3	0.27	1.05	1.61	2.13	2.64	3.15	3.65	4.15	4.65	5.15
4	0.28	0.82	1.32	1.82	2.32	2.82	3.32	3.82	4.32	4.82
5	0.45	1.99	2.52	3.02	3.52	4.02	4.53	5.03	5.53	6.03
6	3.00	3.42	3.89	4.38	4.88	5.38	5.88	6.38	6.88	7.38
7	0	1.22	2.23	3.22	4.23	5.24	6.24	7.25	8.26	9.25

In this context, the probability p_k represents the chance that patient k has an adverse event at a given hospital, with the number of patients at a typical hospital estimated to be around $n = 1000$. A reasonable comparison of the approximations of $\mathcal{L}W$ might assume that the p_ks are independent samples from some beta distribution. However, in the interests of reproducibility, we instead assume that $p_k = F^{-1}(k/(n + 1))$, $k = 1, \dots, n$, where F is the cumulative distribution function of a beta distribution, that is, $p_k = \mathrm{E}\, X_{(k:n)}$, where $X_{(k:n)}$ is the kth order statistic from a random sample of size n. A summary of the results is compiled in Table 2 for a small variety of beta distributions. The results suggest that the Pólya distribution is the preferred approximation.

Example 2. Recognizing that the number of moments (at least approximately) matched should have a growing influence on the e_is and b_is as $n \to \infty$, despite qualitative differences in 'shape' between the μ_i and $\mathcal{L}W$, a comparison in this regime is also given. Specifically, so that the empirical distribution of the p_ks maintains a 'bell' shape, we let $p_k = F^{-1}(k/(n + 1))/50$, $k = 1, \dots, n$, where this time F is the cumulative distribution function of the $\mathrm{Bi}(50, 0.5)$ law. We observe b_i for $\lambda = 10^m$, $m = 1, \dots, 10$ (due to memory overflow, the e_i could not be calculated for the larger values of λ). The results in Table 3 show that eventually the shifted binomial approximation of [24] becomes the clear stand out as the most accurate approximation.

Example 3. Finally, we tabulate results for the example cited in both [8] and [16]. For this, we assume that $W = B_1 + B_2 + B_3$, where $B_1 \sim \mathrm{Bi}(2, 1/\sqrt{2})$, $B_2 \sim \mathrm{Bi}(9, \frac{1}{3})$, and $B_3 \sim \mathrm{Bi}(70, 0.1)$. The results in Table 4 show that in this case e_4, the PBD approximation of [8], is

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

TABLE 4: Approximation error for the test case of [8] and [16].

i	e_i	b_i
1	0.065 91	0.236 54
2	0.029 38	0.109 96
3	0.022 74	0.245 60
4	0.000 40	0.074 69
5	0.004 80	0.235 99
6	0.003 43	0.020 14
7	0.013 10	0.669 64

smallest but b_4 is not as well performing. However, since $b_6 = \min_i b_i$, the Pólya approximation would be preferred if it was necessary to quantify the error.

Based on these examples and further unreported numerical exploration, we conclude that there is no 'one-size-fits-all' uniformly preferable approximation. Moreover, nor can we state categorically that any approximations have been made redundant in the face of others. Some more detailed conclusions now follow.

- Given the range of available approximations, it would be surprising if the single moment matching candidates were ever optimal in terms of accuracy. However, for their simplicity of use, they may still be preferred by some practitioners.
- If the empirical distribution of the p_ks is highly skewed and $n \ll 10^5$, the double moment matching PBD may be optimal, though the estimated approximation error for the PBD may not reflect well the true accuracy.
- If the empirical distribution of the p_ks is not highly skewed and $\lambda \ll 10^6$, the double moment matching Pólya approximation is likely to be the most accurate approximation and accompanied by relatively sharp error estimates.
- If $\lambda \gg 10^6$, the three moment matching shifted binomial approximation is very likely to be the most accurate approximation.

Appendix A

Proof of Proposition 1. We use Stein's method of exchangeable pairs. The initial part of the derivation appears in [14].

Construct an exchangeable pair (W, W') by choosing an index J uniformly from $\{1, \dots, n\}$ and recasting the random variable I_J. That is, we set $W' = W - I_J + I'$, where I' is independent of the $\{I_j\}$ and $\mathrm{P}(I' = 1 \mid J = j) = p_j = 1 - \mathrm{P}(I' = 0 \mid J = j)$. Now, since $\mathrm{E}\, F(W, W') \equiv 0$ for any antisymmetric F (see [31]), we see that if

$$\mathcal{B} f(x) = n\,\mathrm{E}\{\mathbf{1}_{\{W=W'-1\}}\, f(W') - \mathbf{1}_{\{W'=W-1\}}\, f(W) \mid W = x\}$$

then $\mathcal{B}$ is a characterising operator for W. Evaluating the expectation, we obtain

$$\begin{aligned}\mathcal{B} f(x) &= n\,\mathrm{P}(W' = W + 1 \mid W = x) f(x+1) - n\,\mathrm{P}(W' = W - 1 \mid W = x) f(x)\\ &= \sum_{j=1}^{n} \mathrm{E}\{(1 - I_j)p_j \mid W = x\} f(x+1) - \sum_{j=1}^{n} \mathrm{E}\{I_j(1 - p_j) \mid W = x\} f(x)\\ &= (\lambda - g(x)) f(x+1) - (x - g(x)) f(x).\end{aligned}$$

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

Proof of Proposition 2. The first statement of the proposition is immediate from the definition of g. For statement 2, note that if the result holds for arbitrary $\{p_i\}$ then $\Delta g(x) > 0$ implies that $\Delta g(x) < 1$ through making the substitutions $I'_k = 1 - I_k$, $p'_k = 1 - p_k = \mathrm{E}\, I'_k$, and $W' = n - W = \sum_k I'_k$. Thus, we need only show that $\mathrm{E}\{I_k \mid W = x\}$ is increasing in x for arbitrary k. Using the shorthand $a_x = \mathrm{P}(W_k = x)$ and the identity $\mathrm{P}(W = x) = p_k a_{x-1} + (1 - p_k) a_x$, simple algebra yields

$$\mathrm{E}\{I_k \mid W = x + 1\} - \mathrm{E}\{I_k \mid W = x\} = \frac{p_k(1 - p_k)}{\mathrm{P}(W = x + 1)\,\mathrm{P}(W = x)}(a_x^2 - a_{x+1}a_{x-1}),$$

which is positive on account of Newton's inequality (cf. [19, p. 249]).

Now for statement 3. If all the $p_i = p$ are equal then $g(m) = pm$ and, clearly, $\Delta(g(m)/m) = 0$ for each m. Thus, assuming that $p_1 \geq p_2 \geq \cdots \geq p_n$ with at least one p_i distinct from another, we shall now prove the strict inequality by establishing a strong stochastic ordering of the random variables $\{Y_m\}_{m=1}^n$ with distributions

$$\mathrm{P}(Y_m = k) = \frac{1}{m}\mathrm{E}\{I_k \mid W = m\}, \qquad k = 1, \dots, n.$$

For arbitrary but fixed $m \geq 1$ and each k, let

$$b_k := \frac{\mathrm{P}(Y_{m+1} = k)}{\mathrm{P}(Y_m = k)} = \frac{m}{m+1}\frac{\mathrm{P}(W_k = m)}{\mathrm{P}(W_k = m - 1)}\frac{\mathrm{P}(W = m)}{\mathrm{P}(W = m + 1)}.$$

Clearly, $b_k > b_j$ if and only if

$$\frac{\mathrm{P}(W_k = m)}{\mathrm{P}(W_k = m - 1)} - \frac{\mathrm{P}(W_j = m)}{\mathrm{P}(W_j = m - 1)} > 0. \tag{9}$$

Cross multiplying and using the identity

$$\mathrm{P}(W_k = m) = p_j\,\mathrm{P}(W_{jk} = m - 1) + (1 - p_j)\,\mathrm{P}(W_{jk} = m),$$

it is seen that (9) is equivalent to

$$(p_j - p_k)(\mathrm{P}(W_{jk} = m - 1)^2 - \mathrm{P}(W_{jk} = m)\,\mathrm{P}(W_{jk} = m - 2)) > 0,$$

and, hence, also $p_j > p_k$ (again by Newton's inequality). Combining this with the ordering of the p_ks, we conclude that $b_1 \leq b_2 \leq \cdots \leq b_n$ and at least one b_k differs from another. Now, by Lemma 1 below, this further implies that

$$\mathrm{P}(Y_{m+1} \geq k) > \mathrm{P}(Y_m \geq k), \qquad k = 2, \dots, n,$$

from which a simple summation by parts yields

$$\frac{g(m+1)}{m+1} = \mathrm{E}\, p_{Y_{m+1}} < \mathrm{E}\, p_{Y_m} = \frac{g(m)}{m}.$$

Lemma 1. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be two sequences of positive real numbers such that $b_1 \leq b_2 \leq \cdots \leq b_n$ with at least one b_i distinct from another. Then,*

$$\frac{\sum_{i=k}^n a_i b_i}{\sum_{i=1}^n a_i b_i} > \frac{\sum_{i=k}^n a_i}{\sum_{i=1}^n a_i}, \qquad k = 2, \dots, n. \tag{10}$$

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

Proof. By cross multiplying and using the decomposition $\sum_{i=1}^{n} = \sum_{i=1}^{k-1} + \sum_{i=k}^{n}$, we see that (10) is equivalent to

$$\sum_{i=k}^{n} a_i b_i \sum_{i=1}^{k-1} a_i > \sum_{i=k}^{n} a_i \sum_{i=1}^{k-1} a_i b_i. \tag{11}$$

Now, applying the ordering of the b_is, it follows that

$$\frac{\sum_{i=k}^{n} a_i b_i}{\sum_{i=k}^{n} a_i} \geq \frac{b_k \sum_{i=k}^{n} a_i}{\sum_{i=k}^{n} a_i} = \frac{b_k \sum_{i=1}^{k-1} a_i}{\sum_{i=1}^{k-1} a_i} \geq \frac{\sum_{i=1}^{k-1} a_i b_i}{\sum_{i=1}^{k-1} a_i},$$

where at least one of the inequalities is strict if not all the b_is are equal. Cross multiplying again, we recover (11) and, thus, the lemma is proved.

Acknowledgements

The author would like to thank Terry Lyons, Gesine Reinert, and an anonymous referee for helpful suggestions that led to improvements in the exposition. He also wishes to thank the Commonwealth Scholarship Commission, UK, for funding support while a DPhil student at the University of Oxford, and the School of Mathematics and Statistics, The University of Sydney, for providing subsequent support that helped in publishing this work.

References

[1] Barbour, A. D. (2005). Multivariate Poisson-binomial approximation using Stein's method. In *Stein's Method and Applications* (Lecture Notes Ser. Inst. Math. Sci. Natl. Univ. Singapore **5**), Singapore University Press, pp. 131–142.

[2] Barbour, A. D. and Chen, L. H. Y. (eds) (2005). *An Introduction to Stein's Method* (Lecture Notes Ser. Inst. Math. Sci. Natl. Univ. Singapore **4**). Singapore University Press.

[3] Barbour, A. D. and Jensen, J. L. (1989). Local and tail approximations near the Poisson limit. *Scand. J. Statist.* **16,** 75–87.

[4] Barbour, A. D. and Xia, A. (2006). On Stein's factors for Poisson approximation in Wasserstein distance. *Bernoulli* **12,** 943–954.

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

[6] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation* (Oxford Stud. Prob. **2**). Clarendon Press, Oxford University Press, New York.

[7] Borovkov, K. and Pfeifer, D. (1996). On improvements of the order of approximation in the Poisson limit theorem. *J. Appl. Prob.* **33,** 146–155.

[8] Brown, T. C. and Xia, A. (2001). Stein's method and birth–death processes. *Ann. Prob.* **29,** 1373–1403.

[9] Čekanavičius, V. (1997). Asymptotic expansions in the exponent: a compound Poisson approach. *Adv. Appl. Prob.* **29,** 374–387.

[10] Čekanavičius, V. and Roos, B. (2004). Two-parametric compound binomial approximations. *Liet. Mat. Rink.* **44,** 443–466. English translation: *Lithuanian Math. J.* **44**, 354–373.

[11] Čekanavičius, V. and Roos, B. (2006). Compound binomial approximations. *Ann. Inst. Statist. Math.* **58,** 187–210.

[12] Čekanavičius, V. and Roos, B. (2006). An expansion in the exponent for compound binomial approximations. *Liet. Mat. Rink.* **46,** 67–110. English translation: *Lithuanian Math. J.* **46,** 54–91.

[13] Čekanavičius, V. and Vaĭtkus, P. (2001). Centered Poisson approximation by the Stein method. *Liet. Mat. Rink.* **41,** 409–423 (in Russian). English translation: *Lithuanian Math. J.* **41,** 319–329.

[14] Chatterjee, S., Diaconis, P. and Meckes, E. (2005). Exchangeable pairs and Poisson approximation. *Prob. Surveys* **2,** 64–106.

[15] Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statist. Sinica* **7,** 875–892.

[16] Choi, K. P. and Xia, A. (2002). Approximating the number of successes in independent trials: binomial versus Poisson. *Ann. Appl. Prob.* **12,** 1139–1148.

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press

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

[18] Eichelsbacher, P. and Reinert, G. (2008). Stein's method for discrete Gibbs measures. *Ann. Appl. Prob.* **18,** 1588–1618.

[19] Hoggar, S. G. (1974). Chromatic polynomials and logarithmic concavity. *J. Combinatorial Theory B* **16,** 248–254.

[20] Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, 2nd edn. John Wiley, New York.

[21] Jurgelenaite, R., Lucas, P. and Heskes, T. (2005). Exploring the noisy threshold function in designing Bayesian networks. In *Research and Development in Intelligent Systems XXII* (Proc. SGAI Internat. Conf. Innovative Techniques and Applications of Artificial Intelligence), eds M. Bramer, F. Coenen and T. Allen, Springer, London, pp. 133–146.

[22] Kruopis, Y. (1986). The accuracy of approximation of the generalized binomial distribution by convolutions of Poisson measures. *Litovsk. Mat. Sb.* **26,** 53–69.

[23] Mattner, L. and Roos, B. (2007). A shorter proof of Kanter's Bessel function concentration bound. *Prob. Theory Relat. Fields* **139,** 191–205.

[24] Peköz, E. A., Röllin, A., Čekanavičius, V. and Shwartz, M. (2009). A three-parameter binomial approximation. *J. Appl. Prob.* **46,** 1073–1085.

[25] Percus, O. E. and Percus, J. K. (1985). Probability bounds on the sum of independent nonidentically distributed binomial random variables. *SIAM J. Appl. Math.* **45,** 621–640.

[26] Pitman, J. (1997). Probabilistic bounds on the coefficients of polynomials with only real zeros. *J. Combinatorial Theory A* **77,** 279–303.

[27] Röllin, A. (2008). Symmetric and centered binomial approximation of sums of locally dependent random variables. *Electron. J. Prob.* **13,** 756–776.

[28] Roos, B. (2000). Binomial approximation to the Poisson binomial distribution: the Krawtchouk expansion. *Teor. Veroyat. Primen.* **45,** 328–344. English translation: *Theory Prob. Appl.* **45** (2001), 258–272.

[29] Skipper, M. M. (2010). Some approximation theorems in discrete probability. Doctoral Thesis, University of Oxford.

[30] Soon, S. Y. T. (1996). Binomial approximation for dependent indicators. *Statist. Sinica* **6,** 703–714.

[31] Stein, C. (1986). *Approximate Computation of Expectations* (Inst. Math. Statist. Lecture Notes—Monogr. Ser. **7**). Institute of Mathematical Statistics, Hayward, CA.

[32] Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson binomial distribution. Tech. Rep. 354, Department of Statistics, Stanford University.

[33] Thompson, P. (2002). Almost-binomial random variables. *College Math. J.* **33,** 235–237.

[34] Vatutin, V. A. and Mikhailov, V. G. (1983). Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory Prob. Appl.* **27,** 734–743.

[35] Xia, A. and Zhang, F. (2008). A polynomial birth-death point process approximation to the Bernoulli process. *Stoch. Process. Appl.* **118,** 1254–1263.

[36] Xia, A. and Zhang, F. (2009). Polynomial birth-death distribution approximation in the Wasserstein distance. *J. Theoret. Prob.* **22,** 294–310.

https://doi.org/10.1239/jap/1346955331 Published online by Cambridge University Press