Topics in Poisson approximation

Barbour, A D
Topics in Poisson Approximation

A. D. Barbour*
Universität Zürich

Abstract.

Keywords:

AMS 1991 Subject Classifications:

* Partially supported by Schweizer Nationalfonds Grant 20-50686.97
1. Introduction

One of the many problems addressed by de Moivre in his fundamental treatise of (1712) was that of finding the value of \( m \) which makes the binomial probability \( \text{Bi}(n, p)\{[0, m]\} \) closest to \( \frac{1}{2} \), for given \( n \) and \( p \). In considering small values of \( p \), he gives the approximation

\[
\text{Bi}(n, p)\{[0, m]\} \approx \sum_{k=0}^{m} e^{-np}(np)^k/k!
\]  

(\text{Hald 1988, pp 214–5}). However, a formal statement of the Poisson approximation to the binomial distribution first appeared over a century later in Poisson (1837). Even then, the relevance of the Poisson ‘law of small numbers’ in statistical analyses, in stark contrast to that of the normal law, apparently remained completely neglected, until von Bortkewitsch (1898) demonstrated with a variety of sets of observational data that many random counts of rare and unrelated occurrences are well modelled as having Poisson distributions. This fact is all the more surprising because the Poisson family has only one parameter to be fitted, as opposed to the two parameters of the normal distribution, so that good fits to observational data are strongly indicative of some real underlying mechanism having produced them, rather than of the flexibility of the probability model. We shall see that the mechanism is very simple, and can indeed be summarized by the adjectives ‘rare’ and ‘unrelated’.

The natural proof of (1.1) is to observe that

\[
\text{Bi}(n, p)\{k\} = \binom{n}{k} p^k (1-p)^{n-k}
\]

\[
= \{e^{-np}(np)^k/k!\} \left\{ \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) \right\} e^n (p+\log(1-p))(1-p)^{-k}.
\]  

(1.2)

If \( n \) is large and \( p = p(n) \sim \lambda/n \) for fixed \( \lambda \), then

\[
n\{p + \log(1-p)\} \sim -\lambda^2/(2n) \to 0,
\]

and, for any fixed \( k \),

\[
(1-p)^{-k} \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) \to 1;
\]

hence

\[
\lim_{n \to \infty} \text{Bi}(n, p)\{k\} = e^{-\lambda} \lambda^k/k!,
\]  

(1.3)
and (1.1) follows. However, for any particular finite \( n \) and \( p \), limit asymptotics alone give no idea of the accuracy of the approximation in (1.1); for this, a finer analysis of (1.2) is needed. With a little more care, it follows from (1.2) that

\[
\text{Bi}(n, p) \{ k \} = \{ e^{-np}(np)^k/k! \} \left\{ 1 + O(np^2, k^2 n^{-1}) \right\},
\]

and this in turn can be used to show that there exists a constant \( c > 0 \) such that

\[
d_{TV}(\text{Bi}(n, p), \text{Po}(np)) \leq cp(1 + np),
\]

for all choices of \( n \) and \( p \), where, for probability distributions \( P \) and \( Q \) on \( \mathbb{Z}_+ \),

\[
d_{TV}(P, Q) := \sup_{A \subseteq \mathbb{Z}_+} |P(A) - Q(A)|.
\]

Hence de Moivre's approximation (1.1) is accurate to within an error of \( cp(1 + np) \), for a constant \( c \) which can be explicitly computed.

The estimate given in (1.5) suffices to show that \( \text{Bi}(n, p) \) and \( \text{Po}(np) \) are close, provided that both \( p \) and \( np^2 \) are small. If \( p = p(n) \sim \lambda / n \) for fixed \( \lambda \) as \( n \to \infty \), these two quantities are both of order \( n^{-1} \), and the Poisson approximation becomes ever more exact as \( n \to \infty \); the bound (1.5) shows that the same is true, provided only that \( p(n) = o(n^{-1/2}) \). However, the true accuracy of the approximation is actually better still. Prohorov (1953) proved that there is a universal constant \( c \) such that

\[
d_{TV}(\text{Bi}(n, p), \text{Po}(np)) \leq cp,
\]

for all \( n \) and \( p \), from which it follows that \( p(n) \to 0 \) is enough for asymptotic accuracy; he also showed that this order of accuracy is best possible, if \( np \) is bounded away from 0.

If Poisson approximation were only appropriate for the binomial distribution, it would be of limited importance, but von Bortkewitsch’s data suggest otherwise; the random mechanisms underlying his data are unlikely to correspond exactly to sums of \( n \) independent Bernoulli \( \text{Be}(p) \) random variables, all with the same \( p \). In fact, it turns out that the ‘law of small numbers’ is universally appropriate for sums of weakly dependent Bernoulli random variables with small, but possibly differing, values of \( p \).

The first step in checking such a claim is to retain independence, but to let the \( p \)'s vary. So take \( W = \sum_{i=1}^n I_i \), with \( I_1, \ldots, I_n \) independent and \( I_i \sim \text{Be}(p_i) \); how well is the distribution of \( W \) approximated by a Poisson distribution with the same mean \( \lambda = \sum_{i=1}^n p_i \)? A calculation in similar vein to that given in (1.2) can be undertaken, in order to compute \( \mathbb{P}[W = k] \), though now the combinatorics becomes rather more complicated, because each possible \( k \)-subset of indices at which the \( k \) values of 1 can occur yields a different probability in the sum. Nonetheless, with rather more effort, a formula

\[
\mathbb{P}[W = k] = \{ e^{-\lambda} \lambda^k / k! \} \exp \left\{ O\left( \lambda \max_{1 \leq i \leq n} p_i, k^2 \lambda^{-1} \max_{1 \leq i \leq n} p_i \right) \right\}
\]

(1.8)
can be derived, the order term being uniform in \( \max_{1 \leq i \leq n} p_i \leq x \), for any \( 0 < x < 1 \). In the case of equal \( p_i \)'s, \( \lambda = np \) and \( p = \max_{1 \leq i \leq n} p_i \), and expression (1.8) becomes essentially equivalent to (1.4); and, mirroring the situation with equal \( p_i \)'s, a bound

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq c \max_{1 \leq i \leq n} p_i (1 + \lambda)
\]

(1.9)

for the total variation distance between the distribution \( \mathcal{L}(W) \) of \( W \) and \( \text{Po}(\lambda) \) can be deduced from it.

Once again, as when the \( p_i \)'s were equal, the factor \( \lambda \) in (1.9) turns out to be unnecessary when \( \lambda \geq 1 \), Poisson approximation to \( \mathcal{L}(W) \) being accurate whenever \( \max_{1 \leq i \leq n} p_i \) is small. The first result in this direction was that of Hodges and Le Cam (1960), who gave the bound

\[
\max_{j \geq 0} \left| \mathbb{P}[W \leq j] - \sum_{k=0}^{j} e^{-\lambda} \lambda^k / k! \right| \leq 3 \left( \max_{1 \leq i \leq n} p_i \right)^{1/3}
\]

(1.10)

for the difference of the two distribution functions. This was immediately improved upon in Le Cam (1960), who showed that

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq 4.5 \max_{1 \leq i \leq n} p_i
\]

(1.11)

and

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq 8 \lambda^{-1} \sum_{i=1}^{n} p_i^2
\]

(1.12)

the latter under the restriction that \( \max_{1 \leq i \leq n} p_i \leq 1/4 \); then Kerstan (1974) improved upon (1.12) by reducing the factor 8 to 1.05. Both (1.11) and (1.12) give bounds of the form \( cp \) if all the \( p_i \)'s are equal, and thus provide analogues of Prohorov's bound (1.7) in the more general setting of unequally distributed summands.

The methods used by Kerstan and Le Cam, although quite different, both rely on the multiplicative properties of convolution. Kerstan compares the probability generating function of the sum of indicators with the Poisson probability generating function in an ingenious way, and then uses Cauchy's formula as a basic tool in obtaining the estimate of total variation distance from the difference of the probability generating functions. Le Cam also uses complex variables, in his case Fourier methods, for the detailed estimates, but his basic approach is to represent distributions over the non-negative integers \( \mathbb{Z} \) as operators on the space of bounded measurable functions over \( \mathbb{Z} \), with multiplication of operators corresponding to convolution. This method has since been developed by Deheuvels and Pfeifer, who have used it to obtain the sharpest results yet for the Poisson approximation of sums of independent indicators: in particular, in Deheuvels and Pfeifer (1988), the
complex variable techniques of Uspensky (1931) and Shorgin (1977) are combined with the operator method to yield asymptotic expansions of arbitrary order for set probabilities, together with explicit bounds on the error involved.

The next step is to relax the assumption of independence between the summands. Here, the above methods all immediately run into problems, because, without independence, the multiplicative structure associated with convolution can no longer be simply invoked—except in some combinatorial settings, where generating functions still arise naturally, as for instance in Hwang (1998). However, this difficulty has already been faced in connection with the central limit theorem for dependent random variables, and other techniques have been devised.

The oldest of these is probably the method of moments. In the Poisson context, it is most conveniently expressed in the following terms. A sequence $W_n$ of non-negative integer valued random variables such that

$$\lim_{n \to \infty} \mathbb{E}\{W_n(W_n - 1) \ldots (W_n - k + 1)\} = \lambda^k, \quad k = 1, 2, \ldots, \quad (1.13)$$

for some fixed $\lambda$, $0 \leq \lambda < \infty$, converges in distribution to the Poisson distribution $\text{Po}(\lambda)$: $W_n \xrightarrow{D} \text{Po}(\lambda)$ as $n \to \infty$ (c.f. Bollobas (1985, Theorem I.20)). If $W_n = \sum_{i=1}^{n} I_i$, where the $I_i$ are indicator random variables, then $(k!)^{-1}W_n(W_n - 1) \ldots (W_n - k + 1)$ counts the number of $k$-subsets $\{i_1, \ldots, i_k\}$ of $\{1, 2, \ldots, n\}$ for which $I_{i_1} = \cdots = I_{i_k} = 1$, so that (1.13) can be checked whenever it is possible to compute the expectations $\mathbb{E}\left( \prod_{j=1}^{k} I_{i_j} \right)$ for each $k$. This proves to be useful in many combinatorial contexts, including the study of random graphs. A probability approximation associated with (1.13) follows from the Bonferroni (1936) inequalities: in the representation

$$\mathbb{P}[W_n \geq k] = \sum_{r=k}^{n} (-1)^{k+r} \binom{r-1}{k-1} \mathbb{E}\{W_n(W_n - 1) \ldots (W_n - r + 1)\}/r!, \quad (1.14)$$

the sequence of partial sums successively bracket the total sum, and

$$\text{Po}(\lambda)[k, \infty) = \sum_{r=k}^{\infty} (-1)^{k+r} \binom{r-1}{k-1} \lambda^r/r!.$$  

However, an approach through (1.13) and (1.14) has two practical drawbacks. The first is that, with more complicated dependence structures, the computation of $\mathbb{E}\left( \prod_{j=1}^{k} I_{i_j} \right)$ can become rapidly more difficult with increasing $k$, rendering the method impracticable. The second is that (1.13) presupposes that $\lambda < \infty$ remains fixed, whereas, frequently, an asymptotic equivalence as $n \to \infty$ of the form

$$\mathbb{E}\{W_n(W_n - 1) \ldots (W_n - k + 1)\} \sim \lambda_n^k, \quad k = 1, 2, \ldots, \quad (1.15)$$
may hold for each $k$, but with $\lim_{n \to \infty} \lambda_n = \infty$. Unfortunately, (1.15) alone is not enough to prove that $\mathcal{L}(W_n)$ and $\text{Po}(\lambda_n)$ become close as $n \to \infty$ if $\lambda_n \to \infty$, principally because (1.14) then has many large terms of both signs, the probability emerging as the total sum only after many cancellations, and equivalence statements such as (1.15) are not precise enough for a passage to the limit. In practical terms, the approximation in (1.15) would need to be proved to an extremely high relative accuracy, and this greatly detracts from the usefulness of the moment method.

A particularly successful approach in the central limit theorem has been to identify a natural flow of time, and then to make use of martingales with respect to the associated filtration, since many of the salient properties of sequences of independent random variables are found to carry over to martingale difference sequences. For sums of indicators, this approach was taken up by Freedman (1974) and Serfling (1975). They suppose that there is a filtration $(\mathcal{F}_i, 0 \leq i \leq n)$ such that, for each $i$, $I_i$ is $\mathcal{F}_i$-measurable, and their bounds are then expressed in terms of properties of the distributions of the random variables

$$p_i = \mathbb{E}(I_i | \mathcal{F}_{i-1}).$$

(1.16)

For instance, Serfling (1975, Theorem 1) gives the bound

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \sum_{i=1}^{n} \{[\mathbb{E}(p_i)]^2 + \mathbb{E}|p_i - \mathbb{E}p_i|\},$$

(1.17)

for $\lambda = \sum_{i=1}^{n} \mathbb{E}p_i$.

However, if Serfling’s bound (1.17) is evaluated in the binomial context, with independent $I_i$’s and with $p_i = p$, fixed and non-random, it yields the value $np^2$, of the weaker order found in (1.5) and (1.9), rather than of the improved order $p$ given in (1.7) or (1.12). For large $\lambda$, this represents a substantial weakness. Indeed, in the binomial context, one way of achieving a bound of order $np^2$ for $d_{TV}(\text{Bi}(n,p), \text{Po}(np))$ is independently to couple each of the summands $I_i \sim \text{Be}(p)$ with $Z_i \sim \text{Po}(p)$ — when $Z_i = 0$, take $I_i = 0$ with probability $e^{p}(1-p) \leq 1$; take $I_i = 1$ otherwise — in such a way that $\mathbb{P}[Z_i \neq I_i] = p(1-e^{-p})$. Then, by independence and the triangle inequality,

$$d_{TV}(\text{Bi}(n,p), \text{Po}(np)) \leq \sum_{i=1}^{n} d_{TV}(\mathcal{L}(I_i), \mathcal{L}(Z_i)) = np(1-e^{-p}) \leq np^2;$$

(1.18)

this coupling idea is in fact the essence of Serfling’s proof of (1.17). Bounds like (1.7) and (1.12), by bringing in an extra factor of $1/(np)$, reflect a much more subtle matching of the distributions of the sums than can be achieved by matching each of the summands individually; roughly speaking, the same contrast as that between Skorohod imbedding and the Komlós, Major and Tusnády approach to strong approximation in the functional
central limit theorem. In order to reproduce bounds such as in (1.7) and (1.12) for sums of dependent indicators, a new idea is needed.

2. The Stein–Chen method

An entirely new method for approximating the distributions of random elements was introduced by Stein (1970). His original application was to the central limit theorem for sums of dependent random variables, and the version appropriate for Poisson approximation of sums of dependent indicator random variables was first developed by Chen (1975a); see also Stein (1986, 1992). The method hinges on the following observations. First, take $A$ to be any subset of $\mathbb{Z}_+$, and take $\lambda > 0$. Then it is possible to express the indicator function $1_A$ in the form

$$1_A(j) = \text{Po}(\lambda)\{A\} + \lambda g_{\lambda,A}(j + 1) - j g_{\lambda,A}(j), \quad j \geq 0,$$

the Stein Equation, where the function $g_{\lambda,A} : \mathbb{Z}_+ \to \mathbb{R}$ can be recursively determined on $\mathbb{Z}_+ \setminus \{0\}$; the value of $g_{\lambda,A}(0)$ is in any case irrelevant. Furthermore, if $M_0$ and $M_1$ are defined by

$$M_0(g) := \sup_{j \geq 1} |g(j)|; \quad M_1(g) := \sup_{j \geq 1} |g(j + 1) - g(j)|;$$

then

$$J_1 := \sup_{A \subseteq \mathbb{Z}_+} M_0(g_{\lambda,A}) \leq \min\{1, \lambda^{-1/2}\};$$

$$J_2 := \sup_{A \subseteq \mathbb{Z}_+} M_1(g_{\lambda,A}) \leq \lambda^{-1}(1 - e^{-\lambda}) \leq \min\{1, \lambda^{-1}\}.$$ (Barbour and Eagleson, 1983). Hence, for $W$ a random variable on $\mathbb{Z}_+$, it follows that

$$\mathbb{P}[W \in A] = \text{Po}(\lambda)\{A\} + \mathbb{E}\{\varepsilon_\lambda(g_{\lambda,A};W)\},$$

and thus

$$d_{TV}(\mathcal{L}(W),\text{Po}(\lambda)) \leq \sup_{A \subseteq \mathbb{Z}_+} |\mathbb{E}\{\varepsilon_\lambda(g_{\lambda,A};W)\}|,$$

where

$$\varepsilon_\lambda(g;w) = \lambda g(w + 1) - wg(w).$$

The second observation is that, if $W = \sum_{i=1}^n I_i$ is a sum of weakly dependent indicators and $\lambda = \mathbb{E}W = \sum_{i=1}^n p_i$, then $|\mathbb{E}\{\varepsilon_\lambda(g;W)\}|$ can frequently be shown to be small by relatively simple arguments.
For example, take the $I_i$ once again to be independent, with $I_i \sim \text{Be}(p_i)$ and $\lambda = \sum_{i=1}^n p_i$ as above. Then, writing $W_i = \sum_{i' \neq i} I_{i'} = W - I_i$, it follows for any bounded $g$
that
\begin{equation}
\mathbb{E}\{I_i g(W)\} = \mathbb{E}\{I_i g(W_i + 1)\} = p_i \mathbb{E}g(W_i + 1),
\end{equation}
this last because $I_i$ and $W_i$ are independent. Hence, from (2.5), and using the definitions
of $\lambda$ and $W$, we find that
\begin{equation}
\mathbb{E}\{\varepsilon_\lambda(g; W)\} = \left(\sum_{i=1}^n p_i\right) \mathbb{E}g(W + 1) - \sum_{i=1}^n p_i \mathbb{E}g(W_i + 1)
= \sum_{i=1}^n p_i \mathbb{E}\{g(W_i + 1 + I_i) - g(W_i + 1)\}.
\end{equation}
Now the quantity in braces is zero if $I_i = 0$, and is in modulus at most $M_1(g)$ if $I_i = 1$. Hence we have shown that
\begin{equation}
|\mathbb{E}\{\varepsilon_\lambda(g; W)\}| \leq \sum_{i=1}^n p_i^2 M_1(g).
\end{equation}
Thus, taking $g = g_{\lambda,A}$ for any $A \subset \mathbb{Z}_+$, and using (2.8) with (2.3) in (2.4), it follows that
\begin{equation}
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq J_2 \sum_{i=1}^n p_i^2 \leq \min\{1, \lambda^{-1}\} \sum_{i=1}^n p_i^2 \leq \max_{1 \leq i \leq n} p_i.
\end{equation}
These bounds are of the same orders as those (1.11), (1.12) of Le Cam, though with better constants even than Kerstan’s improvement, and without restriction on the value of $\max_{1 \leq i \leq n} p_i$; most importantly, the factor $\lambda^{-1}$ is present when $\lambda \geq 1$. Lower bounds of exactly the same order are proved in Barbour and Hall (1984), showing that no essential improvement is possible.

2.1 The local approach

Turning to dependent indicator random variables $I_i$, the main requirement is a substitute for
the last step in (2.6), the only place where independence was invoked. Clearly, $I_i$ and $\sum_{i' \neq i} I_{i'}$ cannot normally be expected to be independent, but it is often reasonable to
suppose that $I_i$ and $W_i = \sum_{i' \notin N_i} I_{i'}$ are almost independent, if, for each $i$, the subset $N_i$ of
indices is carefully chosen. For instance, if the $I_i$ are an $m$-dependent sequence, then the
choice $N_i = \{i' : |i' - i| \leq m\}$ is such that $W_i$ and $I_i$ are independent; if the $I_i$ form a mixing
sequence, the dependence between $W_i$ and $I_i$ is controlled by a mixing coefficient $\alpha(m)$.
In either case, one finds that
\begin{equation}
|\mathbb{E}\{I_i g(W_i + 1)\} - p_i \mathbb{E}g(W_i + 1)| \leq \chi_i M_0(g),
\end{equation}
uniformly for all bounded $g$, with $\chi_i$ either zero or suitably small; for instance, one can take
\[ \chi_i = \mathbb{E}[\mathbb{E}(I_i | W_i) - p_i], \tag{2.11} \]
zero if $I_i$ and $W_i$ are independent. It then remains only to reformulate the first equality in (2.6), which, because of the new definition of $W_i$, is now no longer valid.

So suppose that, for each $i$, $W$ can be written in the form
\[ W = W_i + Z_i + I_i, \tag{2.12} \]
where $Z_i$ and $W_i$ are any non-negative integer valued random variables. Then
\[
\mathbb{E}\{I_i g(W)\} = \mathbb{E}\{I_i g(W_i + Z_i + 1)\} \\
= \mathbb{E}\{I_i g(W_i + 1)\} + \mathbb{E}\{I_i (g(W_i + Z_i + 1) - g(W_i + 1))\},
\tag{2.13}
\]
whereas
\[
p_i \mathbb{E}g(W + 1) = p_i \mathbb{E}g(W_i + Z_i + I_i + 1) \\
= p_i \mathbb{E}g(W_i + 1) + p_i \mathbb{E}\{g(W_i + Z_i + I_i + 1) - g(W_i + 1)\}. \tag{2.14}
\]
Now the difference between (2.14) and (2.13) can be bounded by using (2.10), together with the direct estimates
\[
|\mathbb{E}\{I_i (g(W_i + Z_i + 1) - g(W_i + 1))\}| \leq \mathbb{E}(I_i Z_i) M_1(g)
\]
and
\[
|\mathbb{E}\{g(W_i + Z_i + I_i + 1) - g(W_i + 1)\}| \leq (p_i + \mathbb{E}Z_i) M_1(g).
\]
Adding over $i$ and substituting into (2.5) thus yields
\[
|\mathbb{E}\{\varepsilon_\lambda(g; W)\}| \leq \sum_{i=1}^n \left( \chi_i M_0(g) + \{\mathbb{E}(I_i Z_i) + p_i \mathbb{E}(Z_i + I_i)\} M_1(g) \right), \tag{2.15}
\]
where $\chi_i$ is such that (2.10) holds for all $g$, as with the choice (2.11). Now, again arguing as for (2.9), take $g = g_\lambda, A$ for any $A \subset \mathbb{Z}_+$, and use (2.3) to bound the right hand side of (2.15); then substitute the result into (2.4). This leads to the

Poisson Local Estimate (Chen, 1975a)
\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq J_1 \sum_{i=1}^n \chi_i + J_2 \sum_{i=1}^n \left[ p_i^2 + p_i \mathbb{E}Z_i + \mathbb{E}(I_i Z_i) \right], \tag{2.16}
\]
true for any collection of representations of $W$ as in (2.12), with $J_1 \leq \min(1, \lambda^{-1/2})$ and $J_2 \leq \min(1, \lambda^{-1})$. 

9
If \( Z_i = \sum_{\nu \in N_i} I_\nu \), where \( N_i \) consists of the ‘near’ neighbours of \( i \), the first term in the bound in (2.16) can be thought of as taking care of dependence at long range, since \( \chi_i \) measures the effect of the configuration far from \( i \) upon the probability of having \( I_i = 1 \). The second term then relates to the effects of local dependence. The quantity

\[
\lambda^{-1} \sum_{i=1}^{n} p_i (p_i + \mathbb{E}Z_i)
\]

(2.17)

reflects the expected number of 1’s, including that at \( i \) if there is one, in a typical neighbourhood \( N_i \), and represents a general penalty for choosing the \( N_i \) large in order to make the long range dependence weaker. Similarly, the quantity

\[
\lambda^{-1} \sum_{i=1}^{n} \mathbb{E}(I_iZ_i) = \lambda^{-1} \sum_{i=1}^{n} p_i \mathbb{E}(Z_i | I_i = 1)
\]

(2.18)

reflects the expected number of 1’s, excluding that at \( i \), in a typical neighbourhood \( N_i \) for which \( I_i = 1 \); it contrasts with (2.17) in that it measures the tendency for there to be local clumps of 1’s, and can be large even when the neighbourhoods \( N_i \) are small as measured by (2.17). Since \( J_2 \leq \lambda^{-1} \), the second term in the bound in (2.16) is no larger than the sum of (2.17) and (2.18). Note that local clumping can easily make Poisson approximation poor: for instance, if \( I_{2j} = I_{2j+1} = I_j' \) for each \( j \), where the \( I_j' \sim \text{Be}(p) \) are independent, then \( W = \sum_{i=1}^{n} I_i \) is not close to Poisson for large \( n \) when \( p \sim \lambda / n \), and this is reflected in (2.18) for choices of \( N_i \) which include the nearest neighbours of \( i \).

Note that the bound (2.16) does not involve the higher moments of \( W \). It is enough to be able to control \( \mathbb{E}Z_i \), \( \mathbb{E}(I_iZ_i) \) and the differences \( |\mathbb{E}(I_i | W_i) - p_i| \). Note also that any choices of \( W_i \) and \( Z_i \) satisfying (2.12) are in order, and that they do not have to be defined as sums of the original \( I_i \) over fixed neighbourhoods \( N_i \); the aim is merely to find choices \( W_i \) and \( Z_i \) which make the right hand side of (2.16) small.

**Example 2.1.** As an example of the application of the Poisson Local Estimate (2.16), suppose that \( Y_1, \ldots, Y_{n+m-1} \) are independent random variables with common distribution \( \mu \) over a finite alphabet \( \mathcal{A} \). Let \( \alpha = a_1a_2 \ldots a_m \) be a specific sequence of \( m \) letters of \( \mathcal{A} \) without self overlaps: that is, such that, for all \( r = 1, 2, \ldots, m - 1 \),

\[
a_1a_2 \ldots a_r \neq a_{m-r+1}a_{m-r+2} \ldots a_m.
\]

Define \( I_i = I[Y_i = a_1, \ldots, Y_{i+m-1} = a_m] \), \( 1 \leq i \leq n \), and set

\[
W = \sum_{i=1}^{n} I_i; \quad p = \mathbb{E}I_i = \prod_{l=1}^{m} \mu(a_l); \quad \lambda = np.
\]
Thus $W$ counts the number of times that the sequence $\alpha$ appears in the long random string $Y_1, \ldots, Y_{n+m-1}$. The $I_i$'s are dependent, but only locally, in the sense that $I_i$ is independent of $(I_j, j \notin N_i)$, where $N_i = \{ j : |i - j| < m \}$. So define

$$Z_i = \sum_{j \in N_i \setminus \{i\}} I_j; \quad W_i = \sum_{j \notin N_i} I_j,$$

satisfying (2.12), and observe that $\mathbb{E}(I_i | W_i) = p$, by the independence of $I_i$ and $W_i$, allowing the choice $\chi_i = 0$ for all $i$. Furthermore, $\mathbb{E}Z_i \leq 2(m - 1)p$ and, because self overlaps do not occur in $\alpha$, $I_iZ_i = 0$ a.s. for each $i$. Substituting these values into the Poisson Local Estimate (2.16) yields

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{1, \lambda^{-1}\} np^2(2m - 1) \leq (2m - 1)p. \quad (2.19)$$

Thus, if $\mu_+ = \max a \in \mathcal{A} \mu(a)$, then $p \leq \mu_+^m$, and it follows that

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq (2m - 1)\mu_+^m, \quad (2.20)$$

small whenever $m$ is large, except in the trivial case where $\mu$ makes one of the letters in $\mathcal{A}$ certain to occur every time.

This very elementary application of the Stein–Chen method has been widely generalized (Arratia, Gordon and Waterman 1990, Chryssaphinou and Papastavridis 1988, Geske et al. 1995, Neuhäuser 1994, Schbath 1995). One reason for the interest is that searching molecular sequences for interesting motifs and searching pairs of molecular sequences for local similarities are both basic tools in computational biology. In order to assess the significance of the results obtained by these search methods, a suitable ‘null hypothesis’ distribution is required for comparison. That derived from the independent letter model is frequently used for this purpose; moreover, the Stein–Chen method is flexible enough to show that a suitable Poisson or compound Poisson approximation may also be used under more general null hypotheses (Reinert and Schbath, 1998). Another field where Poisson approximation is a fundamental tool is extreme value theory and the study of rare events, beginning with the work of von Bortkiewitsch (1898). The decomposition of the Poisson Local Estimate into short range and long range parts is similar to that commonly found in extreme value theory: see Smith (1988) for the use of the Stein–Chen method in this context, and Dembo and Karlin (1992) for DNA–motivated applications. A related area is the analysis of coincidences, such as birthday problems and the space–time statistics of Knox (1964). Many further applications of the Poisson Local Estimate are given in Arratia, Goldstein and Gordon (1989,1990).

2.2 The coupling approach
An alternative approach to sums of dependent indicators is to modify (2.6) by using conditional expectation directly, writing

$$\mathbb{E}\{I_i g(W)\} = p_i \mathbb{E}\{g(W) \mid I_i = 1\}.$$  

This yields a general version of (2.7) in the form

$$\mathbb{E}\{\varepsilon(g; W)\} = \sum_{i=1}^{n} p_i \{\mathbb{E}\{g(W + 1) - \mathbb{E}(g(W) \mid I_i = 1)\}\}.$$  

(2.21)

In particular, if, for each $i$, $W_i^*$ is any random variable constructed on the same probability space as $W$ which has the conditional distribution of $\sum_{i' \neq i} I_{i'}$ given that $I_i = 1$, then (2.21) implies that

$$|\mathbb{E}\{\varepsilon(g; W)\}| = \left| \sum_{i=1}^{n} p_i \{\mathbb{E}\{g(W + 1) - g(W_i^* + 1)\}\} \right|$$

$$\leq \sum_{i=1}^{n} p_i \mathbb{E}|W - W_i^*|M_1(g).$$  

(2.22)

Hence, taking $g = g_{\lambda, A}$, (2.3) and (2.4) imply the

**Poisson Coupling Estimate**

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq J_2 \sum_{i=1}^{n} p_i \mathbb{E}|W - W_i^*|,$$  

(2.23)

with $J_2 \leq \min\{1, \lambda^{-1}\}$.

A variant of this coupling estimate can be used, when the construction of the $W_i^*$ is made easier by taking into account the value of some other random element $U_i$. Suppose that, for each possible value $u$ of a random element $U_i$, it is possible to construct $W_i^*(u)$ on the same probability space as $W$, so as to have the conditional distribution of $\sum_{i' \neq i} I_{i'}$ given that $I_i = 1$ and that $U_i = u$. Then

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq J_2 \sum_{i=1}^{n} p_i \int \mathbb{E}|W - W_i^*(u)|F_i(du),$$  

(2.24)

where $F_i$ is the conditional distribution of $U_i$, given that $I_i = 1$.

In both (2.23) and (2.24), the choice of the how to ‘couple’ a $W$ and a $W_i^*$ on the same probability space is entirely free, the aim being to make the right hand sides of (2.23) or (2.24) as small as possible. The estimates are particularly useful when the dependence between the $I_i$ is global rather than local, and there is no natural choice of subset $N_i \subseteq \{1, 2, \ldots, n\}$ for which $I_i$ and $W_i = \sum_{i' \in N_i} I_{i'}$ are almost independent.
Example 2.2. Let $G_{n, \theta}$ be the Bernoulli model for a random graph on $n$ vertices, in which each possible edge $e_{ij}$ is present with probability $\theta$, independently of all others: the edge indicators $E_{ij} \sim \text{Be}(\theta)$ are independent. Let $I_i = \prod_{j \neq i} (1 - E_{ij})$ be the indicator of the event that vertex $i$ is isolated, so that $W = \sum_{i=1}^{n} I_i$ is the number of isolated vertices in $G_{n, \theta}$. Then

$$p_i = \mathbb{P}[I_i = 1] = (1 - \theta)^{n-1} := p$$

is the same for each $i$, and $\lambda = \mathbb{E}W = np$. The indicators $I_1, \ldots, I_n$ are not independent, because $I_1 = 1$ implies that each of the vertices $I_2, \ldots, I_n$ is also more likely to be isolated, because the edges $e_{12}, e_{13}, \ldots, e_{1n}$ are not present. However, this dependence is rather weak, and a Po$(np)$ approximation appears plausible if $p$ is small, or, equivalently, if $n\theta$ is large.

To make this precise using (2.23), observe that, given any realization of $G_{n, \theta}$, an associated realization of $G_{n, \theta}$ conditional on $I_i = 1$ is obtained simply by deleting all the edges $e_{ij}$, $1 \leq j \leq n$, $j \neq i$. Hence we can take

$$W_i^* = W + \sum_{j = 1}^{n} U_j - I_i,$$

(2.25)

where $U_j = E_{ij} \prod_{j \neq i} (1 - E_{ij})$. It is then immediate that

$$\mathbb{E}|W - W_i^*| \leq (n - 1)\theta(1 - \theta)^{n-2} + (1 - \theta)^{n-1}.\tag{2.26}$$

Substituting (2.26) into (2.23), we find that

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{1, \lambda^{-1}\} \lambda \left\{ (n - 1)\theta(1 - \theta)^{n-2} + (1 - \theta)^{n-1} \right\},$$

(2.27)

$$\leq (1 + n\theta)e^{-n\theta}e^{2\theta},$$

which is indeed small whenever $n\theta$ is large.

2.3 Monotone couplings

The example above is interesting, because $W_i^* \geq W$ except when $I_i = 1$, so that, for each $i$,

$$|W - W_i^*| = W_i^* - W + 2I_i.$$

(2.28)

In any application in which (2.28) holds, it follows from (2.23) that

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{1, \lambda^{-1}\} \sum_{i=1}^{n} p_i \mathbb{E}(W_i^* - W + 2I_i)$$

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

13
since \( W_i^* \) has the distribution of \( W - I_i \), given that \( I_i = 1 \); now, using \( \sum_{i=1}^{n} p_i = \lambda \), we obtain the bound

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{1, \lambda^{-1}\} \left( \text{Var} W - \lambda + 2 \sum_{i=1}^{n} p_i^2 \right). \tag{2.29}
\]

Hence, so long as couplings satisfying (2.28) can be shown to exist, (2.29) gives a simple upper bound for the accuracy of Poisson approximation, which requires only that the mean and variance of \( W \) can be calculated.

In a similar way, if couplings exist such that

\[
|W - W_i^*| = W - W_i^* \tag{2.30}
\]

for all \( i \), then it follows that

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{1, \lambda^{-1}\}(\lambda - \text{Var} W), \tag{2.31}
\]

again requiring only the calculation of the mean and variance of \( W \).

**Example 2.3.** Sample \( n < N \) times without replacement from an urn containing \( M \) white and \( N - M \) black balls. Let \( I_i \) be the indicator of the event that the \( i \)’th ball drawn is white, so that \( W = \sum_{i=1}^{n} I_i \) is the number of white balls sampled; set \( p = \mathbb{E}I_i = M/N \).

For any realization of an \( n \)-sample from the urn, a realization of an \( n \)-sample conditional on \( I_i = 1 \) is obtained by swapping the \( i \)’th ball chosen, if black, for a randomly chosen white ball. Thus

\[
W_i^* = W - I_i - \sum_{j \neq i}^{n} I_j E_j \leq W, \tag{2.32}
\]

where \( E_j \) is the indicator of the event that the \( j \)’th ball in the sample is chosen to be swapped with the \( i \)’th, and (2.30) holds. Here, \( W \) has a hypergeometric distribution with mean \( \lambda = np = nM/N \) and variance \( \lambda(1 - \{p + (n - 1)(1 - p)/(N - 1)\}) \), and we have shown from (2.31) that

\[
d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \min\{\lambda, 1\}\{p + (n - 1)(1 - p)/(N - 1)\}. \tag{2.33}
\]

Conditions under which *monotone couplings* satisfying (2.28) or (2.30) exist, and further general results, including lower bounds on the accuracy of the approximations, can be found in Barbour, Holst and Janson (1992), Chapters 2 and 3.

### 3. Probabilities of small counts
The emphasis up to now has been on approximating the whole of the distribution of a sum \( W = \sum_{i=1}^{n} I_i \) of indicator random variables. However, it is often of greatest interest just to be able to approximate the probability \( \mathbb{P}[W = 0] \) accurately. For instance, in reliability theory, \( I_i \) may be the event that component \( i \) in a large system has failed, and the failure of any single component may entail the failure of the whole system, in which case \( \mathbb{P}[W = 0] \) is the probability that the system is still in operation. In such applications, \( \lambda = \mathbb{E}W \) is typically of order \( O(1) \) or smaller, and the small system failure probability \( \mathbb{P}[W > 0] \) is approximated by \( 1 - e^{-\lambda} \), of order \( O(\lambda) \). In contrast, the bound in (2.16) is typically of order \( \lambda m(p + q) \), where \( m \) is the size of the neighbourhood of dependence, \( p = n^{-1}\lambda \) and \( mq \sim \mathbb{E}(Z_i | I_i = 1) \). This is small in comparison to \( \lambda \) if \( mq \) is small, as is the case when failures do not tend to occur in clumps (Godbole, 1993). Note that, as observed in Kouniias (1995), the second term in the general bound (2.16) can be improved, if only the probability \( \mathbb{P}[W = 0] \) is of interest, since

\[
M_1(g_{A, \{0\}}) = \lambda^{-2}(\lambda - 1 + e^{-\lambda}),
\]  

(3.1)

better by a factor which approaches \( 1/2 \) as \( \lambda \to 0 \).

In other applications, it is of interest to approximate \( \mathbb{P}[W = 0] \) when \( \lambda \) is large. Here, the problem is more delicate, because the total variation error bounds are then at best of order \( p = n^{-1}\lambda \), whereas the probability to be approximated should be about \( e^{-\lambda} \), which can be much smaller; asymptotically, this will happen if \( \lambda = \lambda_n \to \infty \) faster than \( c \log n \), for some \( c > 1 \), and then the relative error in the Poisson approximation could be huge. Thus other methods are needed. We give two useful approximations, which require some preparation.

Let \( H \) be a dependency graph on the vertices \( \{1, 2, \ldots, n\} \) for the indicator random variables \( (I_i, \ 1 \leq i \leq n) \). This means that the collections of random variables \( (I_i, \ i \in A) \) and \( (I_i, \ i \in B) \) are independent of one another whenever \( A \) and \( B \) are such that no edge in \( H \) has one vertex in \( A \) and the other in \( B \). For example, if \( (Y_j, \ 1 \leq j \leq m) \) are independent random variables, and, for each \( i \),

\[
I_i = f((Y_j, \ j \in M_i)),
\]  

(3.2)

for some subset \( M_i \subset \{1, 2, \ldots, m\} \), then \( H \) can be taken to consist of all pairs \( \{i, i'\} \) such that \( M_i \cap M_{i'} \neq \emptyset \). If \( H \) is fairly sparse, as in the case of dissociated random variables (McGinley and Sibson, 1975), the dependence between the \( I_i \) can be expected to be relatively weak.

For the first approximation, we assume that the \( I_i \) are as in (3.2), but with

\[
f((Y_j, \ j \in M_i)) = \prod_{j \in M_i} Y_j,
\]  

(3.3)
for independent indicator random variables $Y_j$. We set $W = \sum_{i=1}^{n} I_i$, $p_i = \mathbb{E} I_i$ and $\lambda = \mathbb{E} W$ as usual, and we define

$$\delta = \lambda^{-1} \sum_{\{i, i'\} \in H} \mathbb{E}(I_i I_{i'}).$$

Then, for any $0 \leq \eta \leq 1$, we have

**Janson’s Inequality** (Janson 1990):

$$\mathbb{P}[W \leq \eta \lambda] \leq \exp \{-\lambda[(1 - \eta) + \eta \log \eta]/(1 + \delta)\}. \tag{3.4}$$

In the particular case $\eta = 0$, this gives

$$\prod_{i=1}^{n}(1 - p_i) \leq \mathbb{P}[W = 0] \leq \exp\{-\lambda/(1 + \delta)\}, \tag{3.5}$$

where the first inequality is a direct consequence of the FKG–inequality, because of the special form (3.3). Hence, by elementary analysis,

$$\exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left( \frac{p_i}{1 - p_i} \right)^2 \right\} \leq e^{\lambda} \mathbb{P}[W = 0] \leq e^{\lambda \delta}. \tag{3.6}$$

Thus, provided that $\delta$ and $\lambda^{-1} \sum_{i=1}^{n} \{p_i/(1 - p_i)\}^2$ are small, $\log \mathbb{P}[W = 0]$ is close to $-\lambda$; if $\lambda \delta$ and $\sum_{i=1}^{n} \{p_i/(1 - p_i)\}^2$ are small, the approximation of $\mathbb{P}[W = 0]$ by $e^{-\lambda}$ has small relative error. Note that $\delta$ appears as that part of the Poisson Local Estimate (2.16) which represents the effect of local clumping, if we take $Z_i = \sum_{i' : \{i, i'\} \in H} I_{i'}$. If $\delta$ is not small enough, because of clumping, for this Poisson approximation to be accurate, an extension based on compound Poisson approximation can give better results; see M. Roos (1996).

The particular form (3.3) may be too strong for many purposes. However, provided only that the dependency graph $H$ is sufficiently sparse, the following inequality can be used. Defining

$$u_i = \prod_{i' : \{i, i'\} \in H} (1 - p_i)^{-1}; \quad y_{i, i'} = 2\{\mathbb{E}(I_i I_{i'}) + p_i p_{i'}\} u_i u_{i'}; \quad \delta' = \lambda^{-1} \sum_{\{i, i'\} \in H} y_{i, i'} \geq 2\delta, \tag{3.7}$$

we have

**Suen’s Inequality** (Suen, 1990):

$$\left| \mathbb{P}[W = 0] - \prod_{i=1}^{n}(1 - p_i) \right| \leq \prod_{i=1}^{n}(1 - p_i) \left\{ e^{\lambda \delta'} - 1 \right\}. \tag{3.8}$$
Janson’s inequality (3.5) is sharper than Suen’s (3.8) whenever (3.3) is satisfied, though, if $\lambda\delta$ and $\lambda\delta'$ are of the same order of magnitude and both are small, they give comparable results; Suen’s inequality is more widely applicable. Janson (1998) has sharpened Suen’s inequality somewhat.

The Bernoulli random graph model has been the setting for a number of successful applications of Janson’s inequality. One illustration, given in Spencer (1995, Section 1.1), concerns estimating the probability that the Bernoulli random graph is triangle free. Here, we demonstrate the use of the Janson and Suen inequalities in an analogous context, slightly improving upon a result given in Alon, Erdős and Spencer (1992, pp 101–2).

**Example 3.1.** Let $G_{n,\theta}$ as before denote the Bernoulli random graph model, and set

$$J_{ijk} = E_{ij}E_{jk}E_{ik},$$  

(3.9)

the indicator of the event that the triangle with vertices $i$, $j$ and $k$ is present. Set

$$I_i = \prod_{\substack{j<k \ w.t.\ j,k \neq i}} (1 - J_{ijk}),$$

the indicator of the event that $i$ is contained in no triangle, and let $W = \sum_{i=1}^n I_i$. Then the event $\{W = 0\}$ is the event that every vertex is contained in at least one triangle, and we wish to approximate its probability.

The first step is to observe that each of the $I_i$ is a decreasing function of the independent edge indicators $E_{ij}$. It then immediately follows (Barbour, Holst and Janson, 1992, Theorem 2.G and Corollary 2.C.4) that a coupling satisfying (2.28) exists, and hence, from (2.29) and (3.1), we have

$$|\mathbb{P}[W = 0] - e^{-\lambda}| \leq \lambda^{-2}(\lambda - 1 + e^{-\lambda})(\text{Var} W - \lambda + 2np^2),$$  

(3.10)

where $\lambda = np = \mathbb{E}W$; note that, in this step, we use neither Suen’s nor Janson’s inequality, since no sparse dependency graph $H$ for the $I_i$ springs readily to mind.

It thus remains to evaluate $\lambda$ and $\text{Var} W$, for which it is enough to compute $p = \mathbb{P}[I_1 = 1]$ and $q = \mathbb{P}[I_1 = I_2 = 1]$. If (3.10) is to be useful, then $\lambda$ should be neither too large nor too small, suggesting in the first instance that $\theta = \theta(n)$ should be chosen such that $p = p(n) \approx n^{-1}\psi$ for some $0 < \psi < \infty$. Now $I_1 = 1$ exactly when $J_{ijk} = 0$ for all $2 \leq j < k \leq n$, so that we have

$$\mathbb{P}[I_1 = 1] = \mathbb{P}[W_1 = 0], \quad \text{where} \quad W_1 = \sum_{2 \leq j < k \leq n} J_{ijk}. $$  

(3.11)

Since the $(J_{ijk}, 2 \leq j < k \leq n)$ are a collection of $\text{Be}(\theta^3)$ indicators, which satisfy (3.3) with $(E_{rs}, 1 \leq r < s \leq n)$ in place of the $Y_j$ and with

$$M_{jk} = \{(j, l), (k, l); l \neq 1, j, k\},$$

(3.12)
we can apply Janson’s inequality. In fact, \( \lambda_1 = \mathbb{E}W_1 = \binom{n-1}{2}\theta^3 \),

\[
\{(j, k), (j', k')\} \in H \text{ if } |\{j, k\} \cap \{j', k'\}| = 1,
\]

and hence

\[
\lambda_1 \delta_1 = \binom{n-1}{2} 2(n-3)\theta^5.
\]

Thus, by (3.6),

\[
\exp \left\{ -\frac{1}{2} \binom{n-1}{2} \left( \frac{\theta^3}{1 - \theta^3} \right)^2 \right\} \leq e^{\lambda_1} \mathbb{P}[W_1 = 0] \leq \exp \{(n-1)(n-2)(n-3)\theta^5\}. \tag{3.15}
\]

Now to achieve \( p = \mathbb{P}[I_1 = 1] \sim n^{-1}\psi \) requires that \( e^{-\lambda_1} \sim n^{-1}\psi \), from (3.11) and (3.15): so define \( \theta = \theta(n, \psi) \) so that

\[
\left( \frac{n-1}{2} \right) \theta^3 = \log n - \log \psi,
\]

for some \( 0 < \psi < \infty \). Then

\[
\left( \frac{n-1}{2} \right) \left( \frac{\theta^3}{1 - \theta^3} \right)^2 \asymp (\log n)^2 n^{-2}
\]

and

\[
(n-1)(n-2)(n-3)\theta^5 \asymp n^3(n^{-2}\log n)^{5/3} = n^{-1/3}(\log n)^{5/3} = \varepsilon_n,
\]

both of which become small as \( n \to \infty \). Hence, from (3.15), if \( \theta \) is defined as in (3.16), then

\[
\lambda_1 = n\mathbb{P}[I_1 = 1] = \psi(1 + O(\varepsilon_n)). \tag{3.18}
\]

In similar style, we have

\[
q = \mathbb{P}[I_1 = I_2 = 1] = \mathbb{P}[W_2 = 0],
\]

where

\[
W_2 = \sum_{3 \leq j < k \leq n} (J_{1jk} + J_{2jk}) + \sum_{3 \leq j \leq n} J_{12j},
\]

again a sum of indicator random variables satisfying (3.3) with the \( E_{rs} \) in place of the \( Y_j \), and now with

\[
M_{12j} = \{(1, j, l), (2, j, l); l \neq 1, 2, j\} \cup \{(1, 2, l); l \neq 1, 2, j\}, \quad 3 \leq j \leq n;
\]

\[
M_{1jk} = \{(1, j, l), (1, k, l); l \neq 1, j, k\} \cup \{(2, j, k)\}, \quad 3 \leq j < k \leq n,
\]
and with $M_{2jk}$ defined analogously to $M_{1jk}$. Here,

$$
\lambda_2 = \mathbb{E} W_2 = \left\{ 2 \left( \frac{n-2}{2} \right) + (n-2) \right\} \theta^3 = (n-2)^2 \theta^3
$$

and

$$
\lambda_2 \delta_2 = \left\{ 2 \left( \frac{n-2}{2} \right) \{2(n-3) + 1\} + (n-2) 3(n-3) \right\} \theta^5
= 2(n-1)(n-2)(n-3)\theta^5,
$$

so that, by (3.6),

$$
\exp \left\{ -\frac{1}{2} (n-2)^2 \left( \frac{\theta^3}{1 - \theta^3} \right)^2 \right\} \leq e^{\lambda_2} \mathbb{P}[W_2 = 0] \leq \exp \{2(n-1)(n-2)(n-3)\theta^5\}.
$$

Thus, with the same choice of $\theta = \theta(n, \psi)$ and with $\varepsilon_n$ as before, we have

$$
q = \mathbb{P}[I_1 = I_2 = 1] = e^{-\lambda_2} (1 + O(\varepsilon_n))
= \exp \left\{ -\frac{2(n-2)}{n-1} [\log n - \log \psi] \right\} (1 + O(\varepsilon_n))
= (n^{-1} \psi)^2 (1 + O(\varepsilon_n)),
$$

yielding

$$
\text{Var } W = np(1-p) + n(n-1)(q - p^2) = \lambda + O(\varepsilon_n). \quad (3.19)
$$

In conclusion, we obtain our approximation for the probability of the event that every vertex in $G_n, \theta(n, \psi)$ is contained in at least one triangle, here re-expressed as $\mathbb{P}[W = 0]$ for $W$ defined as above, by substituting from (3.18) and (3.19) into (3.10), giving

$$
|\mathbb{P}[W = 0] - e^{-\psi}| = O(\varepsilon_n), \quad (3.20)
$$

where $\theta(n, \psi)$ is defined in (3.16) in terms of $\psi$, for any $0 < \psi < \infty$, and $\varepsilon_n$ is as in (3.17); note that, for fixed $n$, $\theta(n, \psi)$ is a decreasing function of $\psi$. This narrow range of values of $\theta$ covers the interesting region where the probability of having vertices not contained in any triangle changes from 1 to 0.

In fact, the range of values of $\theta$ can be extended slightly, by allowing $\psi$ also to vary with $n$, while still maintaining small relative error in the approximation of the smaller of $\mathbb{P}[W = 0]$ and $\mathbb{P}[W > 0]$. The calculation above shows that $\text{Var } W = \lambda + O(\psi^2 \varepsilon_n)$, and (3.10) remains valid; it then follows that

$$
|e^\psi \mathbb{P}[W = 0] - 1| = O(\psi e^\psi \varepsilon_n) \quad \text{in } 1 \leq \psi \leq \frac{1}{3} \log n,
$$

and that

$$
|\psi^{-1} \mathbb{P}[W > 0] - 1| = O(\psi \varepsilon_n) \quad \text{in } n^{-1} \leq \psi \leq 1.
$$
This covers a range of probabilities for the event $P[W = 0]$, running from a little more than $n^{-1/3}$ to $1 - n^{-1}$.

4. Poisson point process approximation

Under circumstances in which a sum $W = \sum_{i=1}^{n} I_i$ of weakly dependent indicator random variables has a distribution close to Poisson, it is usually also the case that $W_A = \sum_{i \in A} I_i$ is almost Poisson distributed, for any fixed subset $A \subset \{1, 2, \ldots, n\}$. This suggests that, if each $i$ is associated with a point $y_i$ in some carrier space $\mathcal{Y}$, then the random configuration of points described by the measure $\Xi = \sum_{i=1}^{n} I_i \delta_{y_i}$, where $\delta_y$ denotes the point mass at $y$, might be distributed almost as a Poisson process on the set $\mathcal{Y}_n = \{y_i, 1 \leq i \leq n\} \subset \mathcal{Y}$.

For instance, if $y_i = y \in \mathcal{Y}$ were the same for all $i$, then $\mathcal{Y}_n$ would consist of the singleton $\{y\}$, and $\Xi = W \delta_y$ would represent $W$ points at location $y$; in this case, Poisson process approximation for $\Xi$ is exactly the same as Poisson approximation for $W$, since there is no spatial distribution at all. A little more generally, the $y_i$ could be chosen as realizations $Y_i(\omega)$ of independent random variables $Y_i$ with common distribution $\mu$, which are independent also of the $I_i$. In this case, the process $\Xi = \sum_{i=1}^{n} W \delta_{y_i}$ has a non-degenerate distribution over $\mathcal{Y}$, being distributed as $\sum_{i=1}^{W} \delta_{Z_i}$, where the $Z_j$ are independent, with common distribution $\mu$, and are independent of $W$. Then, if $W$ is close in distribution to $\text{Po}(\lambda)$, the process $\Xi$ is close to the Poisson process $\text{PP}(\lambda \mu)$ on $\mathcal{Y}$ with mean measure $\lambda \mu$, which can be realized in similar fashion as $\sum_{j=1}^{W^*} \delta_{Z_j}$, where $W^* \sim \text{Po}(\lambda)$ is independent of the $Z_j$. Indeed, the latter representation easily shows that

$$d_{TV}(\mathcal{L}(\Xi), \text{PP}(\lambda \mu)) \leq d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)), \quad (4.1)$$

where $d_{TV}$ for probability measures $P$ and $Q$ over the space $\mathcal{H}$ of integer valued Radon measures on $\mathcal{Y}$ is defined analogously to (1.6) by

$$d_{TV}(P, Q) = \sup_{A \subset \mathcal{H}} |P(A) - Q(A)|. \quad (4.2)$$

What can be said for other choices of $y_i$?

An obvious possibility is to take $y_i = i$, and to look at the measure $\Xi = \sum_{i=1}^{n} I_i \delta_i$ as a point process on $\mathbb{R}_+$. It is then natural to introduce the filtration $(\mathcal{F}_i, 0 \leq i \leq n)$, and to suppose for each $i$ that $I_i$ is $\mathcal{F}_i$-measurable, as in Serfling (1975); once again, the martingale characteristics of the process $(\sum_{i=1}^{m} I_i \delta_i, 0 \leq m \leq n)$ can be expected to play a part in the approximation bounds. Indeed, if $p_i$ denotes $\mathbb{E}(I_i | \mathcal{F}_{i-1})$ and $\lambda$ is the measure which puts mass $\lambda_i = \mathbb{E}p_i$ on $i$, then

$$d_{TV}(\mathcal{L}(\Xi), \text{PP}(\lambda)) \leq \sum_{i=1}^{n} \{\mathbb{E}[p_i^2] + \mathbb{E}|p_i - \mathbb{E}p_i|\}; \quad (4.3)$$

20
the expression on the right hand side is almost the same as that in (1.17), yet the distance being bounded is now between the distributions of whole processes, and not just of total numbers of points. This inequality is a special case of a theorem of Brown (1983), which compares the distribution of a simple point process \( M \) on \( \mathbb{R}_+ \) to that of any Poisson process \( \text{PP}(\mu) \), and shows that

\[
d_{TV}(\mathcal{L}(M^t), \text{PP}(\mu^t)) \leq \mathbb{E}|A - \mu|_t + \mathbb{E}\left\{ \sum_{s \leq t} \Delta A^2(s) \right\}
\]

(4.4)

for any fixed \( t > 0 \), where \( M^t \) and \( \mu^t \) denote the restrictions of the corresponding measures to \([0, t]\), \( A \) is the compensator of \( M \), and \(|A - \mu|_t\) is the pathwise absolute variation of the signed measure \( A - \mu \) on \([0, t]\).

These inequalities provide a rather satisfactory basis for Poisson process approximation on the line. However, two obvious questions remain. The first is clearly how to proceed when the carrier space is not the line, and there is no preferred direction for time. The second concerns the bound (4.3): is it best possible? This latter question is illustrated by two examples.

**Example 4.1.** In the simple case of independent \( I_i \sim \text{Be}(p) \), it is immediate that (4.3), like (1.17), is of order \( np^2 \), and not of the better order \( p \) found in (1.7) and (1.11).

**Example 4.2.** Take a Cox process on \([0, 1]\) in which, conditional on \( \Lambda, \Xi \sim \text{PP}(\Lambda, \nu) \), where \( \nu \) denotes Lebesgue measure. Suppose that \( \mathbb{P}[\Lambda = 1 + \eta] = \mathbb{P}[\Lambda = 1 - \eta] = \frac{1}{2} \). Then, taking \( \mu = \nu \) in (4.4), and including the random element \( \Lambda \) in \( \mathcal{F}_0 \), it follows that \(|A - \nu|_1 = |\Lambda - 1| \) and that \( \Delta A(s) = 0 \) a.s. for all \( s \), so that (4.4) yields

\[
d_{TV}(\mathcal{L}(\Xi), \text{PP}(\nu)) \leq \mathbb{E}|\Lambda - 1| = \eta.
\]

(4.5)

On the other hand, as for (4.1), it is immediate that

\[
d_{TV}(\mathcal{L}(\Xi), \text{PP}(\nu)) \leq d_{TV}(\text{Po}(\Lambda), \text{Po}(1)) \leq \eta^2;
\]

(4.6)

the final estimate following by the Stein–Chen method (Barbour, Holst and Janson 1992, Theorem 1.C(ii)); this is of better order for small \( \eta \).

The process bound given in (4.3) is an analogue of Serfling’s bound (1.17). There are also process analogues of the Stein–Chen Poisson Local and Coupling Estimates. The link to the random variable context (Barbour, 1988) is afforded by rewriting (2.1) in the form

\[
\lambda[h_{\lambda,A}(j + 1) - h_{\lambda,A}(j)] + j[h_{\lambda,A}(j - 1) - h_{\lambda,A}(j)] = 1_A(j) - \text{Po}(\lambda\{A\}),
\]

(4.6)
where $g_{\lambda, A} = \Delta h_{\lambda, A}$, and by observing that the left hand side takes the form $(A h_{\lambda, A})(j)$, where $A$ is the infinitesimal generator of a simple immigration-death process with immigration at rate $\lambda$ and unit per-capita death rate. For process approximation, define a generator $A$ on $H$ analogously as follows:

$$A h(\xi) = \int_0^S [h(\xi + \delta_s) - h(\xi)] \lambda(ds) + \int_0^S [h(\xi - \delta_s) - h(\xi)] \xi(ds),$$  

for all $\xi \in H$, for a suitable class of functions $h$ on $H$. With this definition, $A$ is the generator of an immigration-death process $Z$ with state space $H$, immigration intensity $\lambda$ and unit per capita death rate. The process $Z$ has equilibrium distribution $\text{Po}(\lambda)$, the distribution of a Poisson process with mean measure $\lambda$. Now, by analogy with (4.6), given any $f$ belonging to a suitable family $F$ of bounded test functions, we wish to find an $h = h_f$ on $H$ such that

$$(A h)(\xi) = f(\xi) - \text{Po}(\lambda)(f),$$

where $\mu(f)$ denotes $\int f d\mu$. For total variation distance, $F = F_{TV}$ consists of all indicators of measurable subsets $A \subset H$, mirroring the approach of Section 2, but other choices of $F$ may also be appropriate. It can now be shown that Equation (4.8), the Stein Equation for the generator $A$, has a solution given by

$$h(\xi) = -\int_0^\infty [E^f f(Z(t)) - \text{Po}(\lambda)(f)] dt,$$

where $E^f$ is the conditional expectation given $Z(0) = \xi$. Then, for any point process $\Xi$ defined on $[0, S]$,

$$d_F(L(\Xi), \text{Po}(\lambda)) := \sup_{f \in F} |E^f f(\Xi) - \text{Po}(\lambda)(f)| = \sup_{f \in F} |E A h_f (\Xi)|,$$

where

$$E A h(\Xi) = E \int_0^S [h(\Xi + \delta_s) - h(\Xi)] \lambda(ds) + E \int_0^S [h(\Xi - \delta_s) - h(\Xi)] \Xi(ds).$$  

In particular, take $\Xi = \sum_{i=1}^n I_i \delta Y_i$, and consider approximating $L(\Xi)$ as a point process by Poisson $(\lambda)$, where $\lambda = \sum_{i=1}^n p_i F_i$ and $F_i$ is the distribution of $Y_i$ given $I_i = 1$; in this setting, (4.11) becomes

$$E A h(\Xi) = E \left\{ \sum_{i=1}^n p_i \int [h(\Xi + \delta_y) - h(\Xi)] F_i(dy) + \sum_{i=1}^n I_i [h(\Xi - \delta Y_i) - h(\Xi)] \right\}.$$
Just as in the Stein–Chen situation, estimation of the final expression in (4.12) is often easier than trying to work directly with $\mathbb{E} f(\Xi)$, and leads to tractable expressions; these involve the quantities

$$J_1(\mathcal{F}) := \sup_{f \in \mathcal{F}} M_1(h_f) \quad \text{and} \quad J_2(\mathcal{F}) := \sup_{f \in \mathcal{F}} M_2(h_f)$$

(4.13)

in place of the $J_1$ and $J_2$ of (2.3), where

$$M_1(h) := \sup_{\xi \in \mathbb{R}, \delta_i \leq i \leq n} |h(\xi + \delta_i) - h(\xi)|;$$

$$M_2(h) := \sup_{\xi \in \mathbb{R}, \delta_i \leq i \leq n} |h(\xi + \delta_i + \delta_j) - h(\xi + \delta_i) - h(\xi + \delta_j) + h(\xi)|.$$  

(4.14)

In the case of total variation distance, the bounds

$$J_1(\mathcal{F}_{TV}) \leq 1; \quad J_2(\mathcal{F}_{TV}) \leq 1,$$

(4.15)

are easily derived using the representation (4.9).

To show how (4.10) and (4.12) can be applied, take first the setting of Example 4.1, where the $I_i \sim \text{Be}(p_i)$ are independent. Then, setting $\Xi_i = \sum_{j \neq i} I_j \delta_j$, and retracing the argument used in the case of random variables from (2.6) to (2.9), it follows that

$$\mathbb{E}\{I_i[h(\Xi - \delta_i) - h(\Xi)]\} = \mathbb{E}\{I_i[h(\Xi_i) - h(\Xi_i + \delta_i)]\}$$

$$= p_i \mathbb{E}\{h(\Xi_i) - h(\Xi_i + \delta_i)\},$$

(4.16)

since $I_i$ and $\Xi_i$ are independent. Thus, from (4.12),

$$\mathbb{E}\mathcal{A}h(\Xi) = \sum_{i=1}^{n} p_i \mathbb{E}\{h(\Xi + \delta_i) - h(\Xi) - h(\Xi_i + \delta_i) + h(\Xi_i)\}$$

$$= \sum_{i=1}^{n} p_i \mathbb{E}\{h(\Xi_i + \delta_i + I_i \delta_i) - h(\Xi_i + I_i \delta_i) - h(\Xi_i + \delta_i) + h(\Xi_i)\},$$

(4.17)

analogously to (2.7); and the quantity in braces is zero if $I_i = 0$, and is in modulus at most $M_2(h)$ if $I_i = 1$. As a result, we obtain the estimate

$$|\mathbb{E}\mathcal{A}h(\Xi)| \leq \sum_{i=1}^{n} p_i^2 M_2(h),$$

(4.18)

and conclude from (4.10) that

$$d_{\mathcal{F}}(\mathcal{L}(\Xi), \mathbb{P}(\mathcal{A})) \leq \sum_{i=1}^{n} p_i^2 J_2(\mathcal{F}).$$

(4.19)
The argument was exactly as for the sum \( W = \sum_{i=1}^{n} I_i \), apart from the slightly different notation. For total variation distance, as originally considered in Example 4.1, it follows from (4.15) and (4.19) that 
\[ d_{TV}(\mathcal{L}(\Xi), \text{PP}(\lambda)) \leq \sum_{i=1}^{n} p_i^2, \]
the same bound of order \( O(np^2) \) as followed also from (4.3), so that the random variable order \( O(p) \) is still not recovered for the process bound. In fact, consideration of the event \( \bigcup_{i=1}^{n} \{ \Xi_i \geq 2 \} \) shows that the total variation distance between \( \mathcal{L}(\Xi) \) and \( \text{PP}(\lambda) \) is at least
\[
\sum_{i=1}^{n} (1 - e^{-p_i}(1 + p_i)) \geq \sum_{i=1}^{n} \frac{1}{2} e^{-1} p_i^2,
\]
and thus that the order of approximation cannot be improved in this setting.

For dependent indicators \( I_i \), process approximation can be derived from the approaches used for random variables in Sections 2.1 and 2.2, in a similarly direct way. For the local approach, suppose that, for each \( i \), \( \Xi \) can be written in the form
\[
\Xi = \Xi_i + H_i + I_i \delta_{Y_i}
\]
for random elements \( \Xi_i \) and \( H_i \) of \( \mathcal{H} \). Modifying (4.16) in the same way that (2.6) was modified to obtain (2.13), we find that
\[
\mathbb{E}\{I_i[h(\Xi - \delta_{Y_i}) - h(\Xi)]\} = \mathbb{E}\{I_i[h(\Xi_i + H_i) - h(\Xi_i + H_i + \delta_{Y_i})]\}
= \mathbb{E}\{I_i[h(\Xi_i) - h(\Xi_i + \delta_{Y_i})]\} + \theta_1(h),
\]
where
\[
|\theta_1(h)| \leq \mathbb{E}(I_i Z_i) M_2(h)
\]
and \( Z_i = \|H_i\| \) is the total variation norm of \( H_i \). In a similar fashion, we obtain
\[
\left| p_i \int \mathbb{E}[h(\Xi + \delta_y) - h(\Xi)] F_i(dy) - p_i \int \mathbb{E}[h(\Xi_i + \delta_y) - h(\Xi_i)] F_i(dy) \right|
\leq p_i \mathbb{E}(Z_i + I_i) M_2(h),
\]
and
\[
\left| p_i \int \mathbb{E}[h(\Xi_i + \delta_y) - h(\Xi_i)] F_i(dy) + \mathbb{E}\{I_i[h(\Xi_i) - h(\Xi_i + \delta_{Y_i})]\} \right|
\leq p_i \mathbb{E} \left| \int [h(\Xi_i + \delta_y) - h(\Xi_i)] F_i(dy) - \int [h(\Xi_i + \delta_y) - h(\Xi_i)] F_i(dy | \Xi_i) \right|
\leq 2p_i \mathbb{E}\{d_{TV}(F_i, F_i(\cdot | \Xi_i))\} M_1(h) := \chi_i M_1(h),
\]
in direct analogy to (2.10) and (2.14), where \( F_i(\cdot | \Xi_i) \) denotes \( \mathcal{L}(Y_i | \Xi_i) \). This leads to the
\[ d_\mathcal{F}(\mathcal{L}(\Xi), \text{PP}(\lambda)) \leq J_1(\mathcal{F}) \sum_{i=1}^{n} \chi_i + J_2(\mathcal{F}) \sum_{i=1}^{n} \left\{ p_i^2 + p_i \mathbb{E}Z_i + \mathbb{E}(I_iZ_i) \right\}. \] (4.25)

Note that there is typically nothing more to be calculated here than for the Poisson Local Estimate (2.16). The main difference between the bound in (4.25) and that in (2.16) is that \( J_1(\mathcal{F}) \) and \( J_2(\mathcal{F}) \) replace \( J_1 \) and \( J_2 \) of (2.3), respectively. Note also that if \( Y_i = i \) a.s. for all \( i \), then one can take \( \chi_i = \mathbb{E}[\mathbb{E}(I_i|\Xi_i) - p_i] \), in direct parallel with the random variable case.

Alternatively, for a coupling version, (4.21) can be replaced by

\[ \mathbb{E}\{I_i[h(\Xi - \delta Y_i) - h(\Xi)]\} = p_i \mathbb{E}\int [h(\Xi_{i_y}^*) - h(\Xi_{i_y} + \delta_y)] F_i(dy), \]

where \( \Xi_{i_y}^* \) is any random element of \( \mathcal{H} \) on the same probability space as \( \Xi \), having the conditional distribution of \( \sum_{i' \neq i} I_{i'} \delta Y_{i'} \), given that \( I_i = 1 \) and \( Y_i = y \). This directly implies that

\[ |\mathbb{E}(\mathcal{A}h)(\Xi)| = \left| \mathbb{E} \sum_{i=1}^{n} p_i \int [h(\Xi + \delta_y) - h(\Xi) + h(\Xi_{i_y}^*) - h(\Xi_{i_y}^* + \delta_y)] F_i(dy) \right| \]

\[ \leq \sum_{i=1}^{n} p_i \int \mathbb{E}\|\Xi - \Xi_{i_y}^*\| F_i(dy) M_2(h), \] (4.26)

in analogy to (2.22), and taking \( h = h_f \) for \( f \in \mathcal{F} \) and using (4.10) gives the

**Poisson Process Coupling Estimate**

\[ d_\mathcal{F}(\mathcal{L}(\Xi), \text{PP}(\lambda)) \leq J_2(\mathcal{F}) \sum_{i=1}^{n} p_i \int \mathbb{E}\|\Xi - \Xi_{i_y}^*\| F_i(dy). \] (4.27)

The corresponding bounds for point processes not concentrated on a discrete carrier space are given in Barbour and Brown (1992b).

Both (4.25) and (4.27) can be used for a wide variety of metrics \( d_\mathcal{F} \), the differences between the estimates obtained lying in the values of the factors \( J_i(\mathcal{F}) \). For total variation, the bounds (4.15) are of best possible order, and do not decrease with increasing \( \lambda = |\lambda| \), as do the \( J_i \) in the random variable case. Certain Wasserstein metrics have been considered in place of total variation distance (see Barbour, Brown and Xia 1998, for example), for which \( \lambda \)-dependence similar to that in (2.3) can be established; these ideas have been applied in the context of networks of queues in Barbour and Brown (1996).
Multivariate Poisson approximation can also be undertaken, by applying the considerations above, but with the set $\mathcal{Y}$ consisting of only finitely many points. Some particular results in this direction are given in Barbour, Holst and Janson (1992, Theorems 10.J and 10.K). The best results for sums of independent random vectors are those of B. Roos (1998), who exploits an extension of Kerstan’s method.

**Example 4.3.** This example is a discrete version of the Cox process Example 4.2. Suppose that, conditional on $\Lambda$, the $I_i \sim \text{Be}(\Lambda)$ are independent, where $\mathbb{P}[\Lambda = 1 + \eta] = \mathbb{P}[\Lambda = 1 - \eta] = \frac{1}{2}$, and that $Y_i = i$ a.s. Then an easy calculation shows that, if we take $\Xi_i = \sum_{\nu \neq i} I_i \delta_{\nu}$ and $H_i = 0$ a.s. in (4.20), then $\chi_i = O(n^{-1} \eta^2)$; for $\lambda = \sum_{i=1}^n n^{-1} \delta_i$, the Poisson Process Local Estimate (4.25) yields the bound

$$d_{TV}(\mathcal{L}(\Xi), \mathbb{P}(\lambda)) = O(\eta^2 + n^{-1}).$$

This estimate is the discrete analogue of inequality (4.6), showing that a better bound than that obtained from (4.3) in (4.5) can indeed be derived.

**Example 4.4.** An old Poisson approximation problem is to find the number of close pairs, when $m$ points are uniformly distributed on a sphere; an early treatment of Perelson and Wiegal (1979) was motivated by the immunological phenomenon of the IgM/IgG switch. Here, we use Poisson process approximation to study a related problem, taken from Månsson (1997); see also Aldous (1988a), Alm (1983), Silverman and Brown (1978).

Suppose that $m$ independent points $\xi_1, \ldots, \xi_m$ are distributed uniformly at random in the square $A = [0, 1]^2$. Let the $k$–subsets $K_i$ of $\{1, 2, \ldots, m\}$ be indexed in some way by $i \in \{1, 2, \ldots, n\}$, where $n = \binom{m}{k}$. Let $I_i$ denote the indicator of the event $\bigcup_{a \in A} \bigcap_{j \in K_i} \{\xi_j \in C + a\}$, where $C + a$ denotes the $a$–translate of the square $C = [0, c]^2$ for some $c < \frac{1}{2}$, and where the torus convention is used throughout to avoid edge effects: $I_i = 1$ when the $k$–subset $K_i$ is covered by some square of side $c$. When $I_i = 1$, let $Y_i$ be the (torus) centre of gravity of $\{\xi_j, j \in K_i\}$, and set $Y_i = (0, 0)$ otherwise. How close is the distribution of the point process $\Xi = \sum_{i=1}^n I_i \delta_{Y_i}$ to that of a Poisson process on $A$ with mean measure $n \mathbb{E}[I_1] \nu$, where $\nu$ denotes Lebesgue measure? Clearly, if $|K_j \cap K_i| \geq 2$, then $\mathbb{P}[I_j = 1 | I_i = 1] > \mathbb{P}[I_j = 1]$, so that there is positive dependence between $I_i$ and $I_j$ when $K_i$ and $K_j$ overlap, and, in this case, $Y_i$ and $Y_j$ are also dependent. However, if $k$ is fixed and $c$ is small, the effect of this dependence is not too great.

To see that this is the case, we use the Poisson Process Local Estimate (4.25), setting $H_i = \sum_{j \in N_i} I_j \delta_{Y_j}$ and $\Xi_i = \Xi - H_i - I_i \delta_{Y_i}$, where $N_i = \{j : j \neq i, K_j \cap K_i \neq \emptyset\}$. Note that $\Xi_i$ depends only on the points $(\xi_l, l \in K_i)$, and is thus independent of $(I_i, Y_i)$, so that $\chi_i = d_{TV}(F_i, F_i(\cdot | \Xi_i)) = 0$ for each $i$; note also that $p_i = \mathbb{E}I_i = k^2 c^{2(k-1)} = p$ and that $F_i$ is uniform on $A$, so that $\sum_{i=1}^n p_i F_i = \lambda \nu$, with $\lambda = np$ as usual. Thus

$$\mathbb{E}Z_i \leq k \left( \frac{m}{k-1} \right) p \leq \beta_k (mc^2)^{k-1}$$

26
and
\[
E(I_i Z_i) \leq p \sum_{l=1}^{k-1} (4c^2)^{k-l} \binom{k}{l} \binom{m-k}{k-l} \leq p \sum_{s=1}^{k-1} \alpha_{ks} (mc^2)^s,
\]
for constants \( \alpha_{ks} \) and \( \beta_k \); hence, from (4.25),
\[
d_F(\mathcal{L}(\Xi), \text{PP}(\lambda)) \leq J_2(\mathcal{F}) \lambda O(mc^2),
\]
uniformly in \( mc^2 \leq 1 \). Thus if \( m \to \infty \) and \( c_m \to 0 \) in such a way that \( \lambda_m \sim k^2 m^k c_m^{2(k-1)}/k! \to \lambda \), it follows that \( d_{TV}(\mathcal{L}(\Xi_m), \text{PP}(\lambda_m)) \) is of order \( O(m^{-1/(k-1)}) \), and becomes small as \( m \to \infty \), so that Poisson process approximation is asymptotically accurate. However, if \( c_m^2 = m^{-k/(k-1) + \alpha} \) for any \( \alpha > (k-1)^{-2} \), then the bound in (4.28) becomes large with \( m \).

5. Compound Poisson approximation

As observed in Section 2, the Poisson Local Estimate for the total variation distance between the distribution of a sum \( W \) of indicators \( I_i \) and the Poisson \( \text{Po}(E W) \) distribution becomes large if the \( I_i \) have a tendency towards local clumping, because it includes the component (2.18), and this usually reflects a real departure from the Poisson. However, there are many probability models (Aldous, 1988b) in which rare, isolated and weakly dependent clumps of 1’s appear, and a compound Poisson approximation may be appropriate instead. The compound Poisson distribution \( \text{CP}(\lambda, \mu) \), where \( \mu \) is a probability measure on \( \mathbb{N} \), is defined by
\[
\text{CP}(\lambda, \mu) = \mathcal{L} \left( \sum_{j=1}^{M} Y_j \right) = \mathcal{L} \left( \sum_{i \geq 1} i M_i \right),
\]
where \( (Y_j, j \geq 1) \) are independent, have distribution \( \mu \) and are independent also of \( M \sim \text{Po}(\lambda) \); and where \( (M_i, i \geq 1) \) are independent, with \( M_i \sim \text{Po}(\lambda \mu_i) \). In the former representation, one thinks of \( M \sim \text{Po}(\lambda) \) clumps, whose sizes \( Y_j \) are independently sampled from the distribution \( \mu \).

As in the case of Poisson approximation, there have been a number of approaches to compound Poisson approximation using transform methods. Of particular note is that using signed compound Poisson measures, as illustrated, for instance, in Čekanavičius (1997), where very accurate approximations are derived. Unfortunately, the use of such methods for sums of dependent random variables seems at present to be limited to only the simplest of cases.

One way of approaching compound Poisson approximation to sums of dependent indicators is to proceed by way of Poisson point process approximation. The typical strategy is to mark exactly one of the 1’s in each clump as its representative, and to replace
$W = \sum_{i=1}^n I_i$ by a sum $\sum_{i=1}^n \sum_{t \geq 1} I_{it}$, where $I_{it}$ now denotes the indicator of the event that $i$ is the index of the representative of a clump of $1$'s of size $l$; thus, for each clump, exactly one of the $I_{it}$ takes the value $1$. Poisson process approximation in total variation to the point process $\Xi = \sum_{i=1}^n \sum_{t \geq 1} I_{it} \delta_t$ is then accomplished by the Poisson Process Local or Coupling Estimates (4.25) and (4.27), and compound Poisson approximation in total variation to the random variable $W = \sum_{i=1}^n \sum_{t \geq 1} I_{it}$, with exactly the same error bound, follows as a consequence. There have been many successful applications of this approach, a number of which are given in Arratia, Goldstein and Gordon (1989,1990). The results obtained from it are very good, provided that $\mathbb{E} W$ is not too large.

There are two drawbacks to the point process approach. First, the identification of a unique representative for each clump (‘declumping’) is rarely natural, and can pose difficulties. Secondly, if $\lambda$ is large, the error bounds derived in this way are frequently far from accurate, because they are of the weaker $O(np^2)$ variety, rather than of order $O(p)$. For point process approximation in total variation, this often represented the true order of approximation; in compound Poisson approximation, it is rather seldom the case, as is illustrated in the examples which follow. However, the problems associated with improving the approximation for large $\lambda$ have not been fully overcome.

**Example 5.1.** Let $V_{ij} = I_i I[Y_i \geq j]$, $1 \leq i \leq n$, $j \geq 1$, be a double array of indicators, in which the $I_i \sim \text{Be}(p_i)$ are independent, and the $Y_i \sim \mu^{(i)}$ are independent of each other and of the $I_i$. ‘Declumping’ is easily achieved by defining $I_{il} = I_i I[Y_i = l]$ for each $i,l$; then set $N_{it} = \{(i,j), j \neq l\}$ and write $\Xi = I_{it} \delta_t + H_{it} + \Xi_{it}$ as in (4.20), where

$$H_{it} = \sum_{j \in N_{it}} I_{ij} \delta_j; \quad \Xi_{it} = \sum_{j \neq i} \sum_{t \geq 1} I_{jt} \delta_t,$$

here, for all $i,l$, $\Xi_{it}$ is independent of $I_{it}$, so that $\chi_{it} = 0$, and $I_{it} Z_{it} = 0$ a.s., where $Z_{it} = \|H_{it}\| = \sum_{j \in N_{it}} I_{ij}$. Hence, with $\lambda \mu = \sum_{i=1}^n \sum_{t \geq 1} \mathbb{E} I_{it} \delta_t$, the Poisson Process Local Estimate (4.25) reduces to

$$d_{TV}(\mathcal{L}(\Xi), \mathcal{P}(\lambda)) \leq J_2(F_{TV}) \sum_{i=1}^n \sum_{t \geq 1} \mathbb{E} I_{it} \mathbb{E} I_i \leq \sum_{i=1}^n \mathbb{E} I_i^2. \quad (5.2)$$

To illustrate the implications of (5.2), let $p(n)$ be such that $p(n) \to 0$ and $np(n) \to \infty$ as $n \to \infty$, and consider three choices of the $p_i = p_i^{(n)}$ and $\mu^{(i)} = \mu^{(in)}$.

(a) Suppose that $p_i^{(n)} = p(n)$ and $\mu^{(in)} = \mu^{(i)}$ for all $i$. Then (5.2) gives a total variation bound of $np(n)^2$ for approximation by CP $(np(n), \mu)$; however, combining (4.1) and (2.9), the true error is much less, being at most $p(n)$.

(b) Suppose that the $p_i^{(n)}$ and $\mu^{(in)}$ are as above for $2 \leq i \leq n$, and that $\mu$ is such that $\mu_1 > 0$; suppose also that $p_1^{(n)} = \frac{1}{2}$ and that $\mu^{(1n)} = \delta_1$ for all $n$. Then (5.2) gives a
bound of order $O(1)$ for approximation by $\text{CP} \left( \lambda_n, \mu_n \right)$, where

$$
\lambda_n = (n - 1)p(n) + \frac{1}{2} \quad \text{and} \quad \mu_n = \lambda_n^{-1}\{(n - 1)p(n) + \frac{1}{2} \delta_1\};
$$

here, the true error is again smaller, and in fact tends to 0 with $n$, being of order $O(p(n) + [np(n)]^{-1})$.

(c) Suppose that everything is as in (b), except that now $\mu \{2 \mathbb{Z}_+ \} = 1$, so that, in particular, $\mu_1 = 0$. In this case, the bound of order $O(1)$ furnished by (5.2) is of the correct order.

The contrast between cases (b) and (c) indicates that improving the error bounds for compound Poisson approximation that are derived using the point process approach is likely to be a delicate matter.

An alternative to the route by way of Poisson process approximation is to aim for compound Poisson approximation directly, using the second representation in (5.1) to derive an analogue of (2.1) for compound Poisson distributions, the Stein Equation

$$
1_A(\lambda) = \text{CP} \left( \lambda, \mu \right) \{A\} + \sum_{i \geq 1} i \lambda \mu_i g_{\lambda, \mu, A}(j + i) - j g_{\lambda, \mu, A}(j),
$$

(5.3)
to be solved for the function $g_{\lambda, \mu, A} : \mathbb{N} \to \mathbb{R}$ for any given $A \subset \mathbb{Z}_+$. This accomplished, it follows just as for (2.4) and (2.5) that

$$
d_{TV}(\mathcal{L}(W), \text{CP} \left( \lambda, \mu \right)) \leq \sup_{A \subset \mathbb{Z}_+} \left| \mathbb{E}\{\varepsilon_{\lambda, \mu}(g_{\lambda, \mu, A}; W)\} \right|,
$$

(5.4)
where

$$
\varepsilon_{\lambda, \mu}(g; w) = \sum_{i \geq 1} i \lambda \mu_i g(w + i) - wg(w);
$$

(5.5)
once again, $|\mathbb{E}\{\varepsilon_{\lambda, \mu}(g; W)\}|$ can often be successfully bounded.

For instance, let $W = \sum_{i=1}^n X_i$, where the $X_i$ are nonnegative integer valued random variables with finite means $m_i$. For each $i$, analogously to (2.12), decompose $W$ in the form

$$
W = W_i + Z_i + U_i + X_i,
$$

(5.6)
where, for the representation to be useful, $W_i$ should be almost independent of $(X_i, U_i)$, and $U_i$ and $Z_i$ should not be too large; the sense in which these requirements are to be interpreted becomes clear shortly. Such a decomposition is often realized by partitioning the indices $\{1, 2, \ldots, n\}$ into subsets $\{i\}, S_i, N_i$ and $T_i$, and setting

$$
U_i = \sum_{j \in S_i} X_j \quad \text{and} \quad Z_i = \sum_{j \in N_i} X_j;
$$

(5.7)
$S_i$ contains those $X_j$ which strongly influence $X_i$, and $T_i$ those $X_j$ whose cumulative effect on $(X_i, U_i)$ is negligible. Define the parameters of the canonical approximating compound Poisson distribution by

$$
\lambda = \sum_{i=1}^{n} \mathbb{E} \left\{ \left( \frac{X_i}{X_i + U_i} \right)^{1\{X_i + U_i \geq 1\}} \right\},
$$

$$
\mu_l = \frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E} \{ X_i I[X_i + U_i = l] \}, \quad l \geq 1.
$$

(5.7)

Setting $\pi_{j,k}^{(i)} = j P[X_i = j, U_i = k]/m_{i1}$, $j \geq 1$, $k \geq 0$, define the four following quantities which appear in the error bounds, and which should be small for the bounds to be good:

$$
\delta_1 = \sum_{i=1}^{n} \sum_{j \geq 1} m_{i1} \sum_{k \geq 0} \pi_{j,k}^{(i)} \mathbb{E} \left[ \frac{P[X_i = j, U_i = k | W_i]}{P[X_i = j, U_i = k]} - 1 \right];
$$

$$
\delta_2 = 2 \sum_{i=1}^{n} \mathbb{E} \{ X_i d_{TV}(\mathcal{L}(W_i | X_i, U_i), \mathcal{L}(W_i)) \};
$$

$$
\delta_3 = \sum_{i=1}^{n} \mathbb{E} \{ X_i d_W(\mathcal{L}(W_i | X_i, U_i), \mathcal{L}(W_i)) \};
$$

$$
\delta_4 = \sum_{i=1}^{n} \{ \mathbb{E}(X_i; Z_i) + \mathbb{E}X_i \mathbb{E}\{X_i + U_i + Z_i\} \}.
$$

(5.8) (5.9) (5.10) (5.11)

In (5.10), the distance $d_W$ is the Wasserstein $L_1$ metric on probability measures over $\mathbb{Z}_+$:

$$
d_W(P, Q) = \sup_{\{f: M_1(f) \leq 1\}} \left| \int f \, dP - \int f \, dQ \right| \geq d_{TV}(P, Q).
$$

For $1 \leq l \leq 3$, $\delta_l / \mathbb{E} W$ is a measure of the average dependence between $W_i$ and $(X_i, U_i)$; $\delta_4$ is small in comparison with $\mathbb{E} W$ if $U_i$ and $Z_i$ are small in expectation, provided that $Z_i$ is not too strongly dependent on $X_i$. Note that this makes no restriction on the dependence between $X_i$ and $U_i$. If Poisson approximation is to be good, nothing can be too strongly dependent on any single $X_i$, and so no element $U_i$ is allowed to accommodate strong local dependence in the Poisson decomposition (2.12). In compound Poisson approximation, such local dependencies can be taken into account. The expectation $\mathbb{E}\{X_i U_i\}$ does not appear in $\delta_4$; instead, the $U_i$ are prominent in the definition (5.7) of the approximating compound Poisson distribution.

Analogously to (2.15) and (2.22), and by rather similar arguments (see also M. Roos (1994a,b)), it can be deduced that for the canonical $\lambda$ and $\mu$

$$
|\mathbb{E}\{\varepsilon \lambda, \mu(g; W)\}| \leq \varepsilon_0 M_0(g) + \varepsilon_1 M_1(g),
$$

(5.12)

(i) with $\varepsilon_0 = \min(\delta_1, \delta_2)$ and $\varepsilon_1 = \delta_4$, and (ii) with $\varepsilon_0 = 0$ and $\varepsilon_1 = \delta_3 + \delta_4$; these bounds are then to be combined with (5.4) to give total variation bounds. If, instead,
approximation by another compound Poisson distribution with parameters $\lambda'$ and $\mu'$ is preferred, note that

$$|\varepsilon_{N, \mu}(g; w) - \varepsilon_{N, \mu}(g; w)| = \left| \sum_{i \geq 1} i(\lambda'\mu'_i - \lambda\mu_i)g(w + i) \right|$$

$$\leq \sum_{i \geq 1} i\mu'_i|\lambda'' - \lambda'|g(w + i)| + \sum_{i \geq 1} i(\lambda''\mu'_i - \lambda\mu_i)g(w + i)$$

$$\leq m'_1|\lambda'' - \lambda'|M_0(g) + \lambda m_1 d_W(Q', Q)M_1(g),$$

where $m_1 = \sum_{i \geq 1} i\mu_i$, $m'_1 = \sum_{i \geq 1} i\mu'_i$, $\lambda'' = \lambda m_1/m'_1$, and the probability measures $Q$ and $Q'$ on $\mathbb{N}$ are such that $Q\{i\} = i\mu_i/m_1$ and $Q'\{i\} = i\mu'_i/m'_1$. This leads to the more general bound

$$|\mathbb{E}\varepsilon_{N, \mu}(g; W)| \leq \varepsilon'_0 M_0(g) + \varepsilon'_1 M_1(g),$$

with

$$\varepsilon'_0 = \varepsilon_0 + |\lambda m_1 - \lambda' m'_1| \quad \text{and} \quad \varepsilon'_1 = \varepsilon_1 + \lambda m_1 d_W(Q', Q).$$

In particular, if $\lambda'$ is chosen equal to $\lambda m_1/m'_1$, then $\varepsilon'_0 = \varepsilon_0$. The advantage of allowing distributions other than the canonical compound Poisson distribution as approximations is that the canonical distribution may be very complicated, whereas an approximation of the same order may be obtainable with a very simple compound Poisson distribution: see Example 5.2.

For independent $X_i$, one can take $Z_i = U_i = 0$, for which choice $\delta_1 = \delta_2 = \delta_3 = 0$, and $\delta_4$ reduces to $\sum_{i=1}^n (\mathbb{E}X_i)^2$. This observation can be used in the setting of Example 5.1 also, by setting $X_i = \sum_{j \geq 1} V_{ij}$. Alternatively, one could use the decomposition (5.6) with $U_{ij} = \sum_{t \neq j} V_{it}$ and $Z_{ij} = 0$, obtaining the same result. In general, when evaluating $\delta_2$ and $\delta_3$, it is often possible to compute the distances between distributions by means of couplings. Variant (ii), when applied with $Z_i = 0$, gives the analogue of the Poisson Coupling Estimate; variant (i) leads to the analogue of the Poisson Local Estimate.

In order to exploit (5.4) and (5.12), it thus remains to find suitable bounds for

$$J_1^{CP} = \sup_{A \subseteq \mathbb{Z}_+} M_0(g, \mu, A) \quad \text{and} \quad J_2^{CP} = \sup_{A \subseteq \mathbb{Z}_+} M_1(g, \mu, A).$$

However, this is not as easy as it might seem. The only known analogue of (2.3) is the bound

$$J_1^{CP}, J_2^{CP} \leq \min\{1, (\lambda\mu_1)^{-1}\}e^\lambda,$$

proved in Barbour, Chen and Loh (1992), and to get bounds which decrease with increasing $\lambda$ they needed to assume that

$$i\mu_i \geq (i + 1)\mu_{i+1}$$

(5.17)
for all $i \geq 1$, in which case

\[
J_1^{CP} \leq \min \left\{ 1, \frac{1}{\sqrt{\lambda(\mu_1 - 2\mu_2)}} \left( 2 - \frac{1}{\sqrt{\lambda(\mu_1 - 2\mu_2)}} \right) \right\};
\]

\[
J_2^{CP} \leq \min \left\{ 1, \frac{1}{\lambda(\mu_1 - 2\mu_2)} \left( \frac{1}{4\lambda(\mu_1 - 2\mu_2)} + \log^+ \{2\lambda(\mu_1 - 2\mu_2)\} \right) \right\}.
\]

(5.18)

These bounds, together with (5.4) and (5.12), yield the

**Compound Poisson Estimate 1:** If $W$ is decomposed as in (5.6), and $\lambda$, $\mu$ and $\delta_l$, $1 \leq l \leq 4$, are as in (5.7)–(5.11), then, for any $\lambda'$, $\mu'$, we have

\[
d_{TV}(\mathcal{L}(W), \text{CP} (\lambda', \mu')) \leq J_1^{CP} \varepsilon_0' + J_2^{CP} \varepsilon_1',
\]

(5.19)

where $\varepsilon_0'$ and $\varepsilon_1'$ are as in (5.14); $J_1^{CP}$ and $J_2^{CP}$ are bounded as in (5.16), or, if condition (5.17) holds, as in (5.18), in either case with $\lambda'$, $\mu'$ for $\lambda$, $\mu$.

Thus, in Example 5.1(a) and (b), if $m_1 < \infty$ and $\mu_1 > 2\mu_2$ and (5.17) holds, then bounds of order

\[
O(p(n) \log \{np(n)\}) \quad \text{and} \quad O((p(n) + [np(n)]^{-1}) \log \{np(n)\})
\]

are obtained; not quite of the correct order, because of the factor $\log \{np(n)\}$, but much sharper than those implied by (5.2). Condition (5.17) cannot hold in case (c), and if (5.17) fails to hold in cases (a) and (b), the bound is of order $p(n) \exp \{np(n)\}m_2^2$, rapidly becoming worse than those implied by (5.2) as $n$ increases, and giving no information at all if $m_1 = \infty$. What is more, Condition (5.17) is no artefact, since it is shown in Barbour and Utev (1998) that $J_1^{CP} \geq C e^{\alpha \lambda}$ for some $C, \alpha > 0$, whenever $\mu_1 + \mu_2 = 1$ and $\mu_1 < 2\mu_2$.

In order to try to avoid these difficulties, Barbour and Utev (1999) exploit the Stein Equation in a different way, by varying the usual argument so as to involve only $J_1^{(a)} = \sup_{A \in \mathcal{Z}_+} M_{l-1}^{(a)}(g, \mu, \lambda)$, $l = 1, 2$, where

\[
M_0^{(a)}(g) = M_0(g(\cdot + a)); \quad M_1^{(a)}(g) = M_1(g(\cdot + a)),
\]

for suitably chosen $a > 1$. This leads to the

**Compound Poisson Estimate 2:** If $W$ is decomposed as in (5.6), and $\lambda \geq 2$, $\mu$ and $\delta_l$, $1 \leq l \leq 4$, are as in (5.7)–(5.11), then, for any $\lambda'$, $\mu'$ satisfying $\lambda' m_1' = \lambda m_1$, and such that $\sum_{j \geq 1} \mu'_j r^j < \infty$ for some $r > 1$ and that $\mu'$ is aperiodic ($\mu'[\mathbb{Z}_+] < 1$ for all $l \geq 2$), we have

\[
d_{TV}(\mathcal{L}(W), \text{CP} (\lambda', \mu')) \leq \varepsilon_0 J_1^{(a)} + \varepsilon_1 K_1^{(a)} + \mathbb{P}[W \leq \frac{1}{2}(\lambda m_1 + a)] K_2^{(a)},
\]

(5.20)
for $a = \lambda v m_1$ and for $0 < v < 1$ suitably chosen, where
\[ K_1^{(a)} = J_1^{(a)} + 2 J_0^{(a)} / (\lambda m_1 (1-v)); \quad K_2^{(a)} = 1 + 2 m_2 J_0^{(a)} / (m_1 (1-v)); \]
and
\[ J_0^{(a)} \leq (\lambda')^{-1/2} C_0(\mu'), \quad J_1^{(a)} \leq (\lambda')^{-1} C_1(\mu'), \]  \hspace{1cm} (5.21)
with $C_0(\mu'), C_1(\mu') < \infty$. The detailed way in which $v$ is to be chosen and in which the $C_l(\mu')$ depend on the radius of convergence of the power series $\sum_{j \geq 1} \mu_j' z^j$ and on the nearness of $\mu'$ to being periodic are explicitly specified in Barbour and Utev (1999). The third term in (5.20) is a penalty incurred in exchanging the $J_l^{(a)}$ for the $J_l^{CP}$.

Applying (5.20) to Example 5.1, assume that $\mu$ has radius of convergence greater than 1 and is aperiodic. Take $\lambda'$ and $\mu'$ to be the canonical parameters. Then $v$ can be chosen fixed for all $n$, and $K_1^{(a)}$ and $K_2^{(a)}$ are uniformly of order $J_1^{(a)}$ and $1 + J_0^{(a)}$ respectively. In cases (a) and (b), both $C_0(\mu_n)$ and $C_1(\mu_n)$ are bounded uniformly in $n$, and hence, from (5.21), bounds of order $O(p(n) + \exp\{-np(n)\alpha\})$ for some $\alpha > 0$ are obtained, with Bernstein’s inequality being used to derive the exponential estimate of the large deviation probability in the third term in (5.20). This is of the ideal order $O(p(n))$, except when $np(n) \to \infty$ very slowly with $n$. At first sight, it appears that the same bound should also follow in case (c), contradicting the fact that the true distance in total variation is of order $O(1)$. The reason why this bound is not obtained in case (c) is that the distribution $\mu_n$ approaches the periodic limit $\mu$ as $n \to \infty$, with the result that $\lim_{n \to \infty} C_l(\mu_n) = \infty$, $l = 1, 2$; in fact, $C_l(\mu_n) \asymp [np(n)]^{-l}$ for $l = 1, 2$, and the bound obtained from (5.20) and (5.21) in case (c) is once again of order $O(1)$.

**Example 5.2.** We conclude with a second example of compound Poisson approximation. Let $\{V_{ij}, 1 \leq i, j \leq t\}$ be independent $\text{Be}(q)$ random variables, and define the indicators $I_{ij} = V_{ij} V_{i+1,j+1} V_{i+1,j+1} V_{i,j+1} V_{i+1,j}$ of the events that $V_{kl} = 1$ at each of the $2 \times 2$ square of lattice points $(k, l)$ with bottom left hand corner $(i, j)$; we adopt the torus convention throughout, in order to avoid edge complications. Define $n = t^2, p = \mathbb{E} I_{ij} = q^4$ and $\lambda = np$. What is the approximate distribution of $W = \sum_{1 \leq i, j \leq t} I_{ij}$ as $t \to \infty$, if $q = q(t) \to 0$ in such a way that $np(t) \to \infty$?

A first approximation is obtained from the Poisson Local Estimate (2.16), using a decomposition $W = W_{ij} + Z_{ij} + I_{ij}$ as in (2.12), defined by taking
\[ Z_{ij} = \sum_{(k, l) \in \mathcal{N}_{ij} \setminus \{(i, j)\}} I_{kl}, \]
where $\mathcal{N}_{ij}$ is $\{(k, l) : |k-i| \leq r, |l-j| \leq r\}$. Then $W_{ij}$ and $I_{ij}$ are independent, so that $\chi_{ij} = 0$ in (2.16) for all $i, j$, $\mathbb{E} I_{ij} \mathbb{E}(I_{ij} + Z_{ij}) = 9p^2$ and $\mathbb{E} (I_{ij} Z_{ij}) = 4(q^2 + q^3)p$, which from (2.16) gives the bound
\[ d_{TV}(\mathcal{L}(W), \text{Po}(np)) \leq 9p + 4(q^2 + q^3) = O(p^{1/2}). \]  \hspace{1cm} (5.22)
The leading term in this error bound comes from $E(I_{ij}Z_{ij})$. This indicates that local clustering is responsible for the main departure from the Poisson, suggesting that a compound Poisson approximation may be better.

For compound Poisson approximation by way of Poisson point process approximation, it is first necessary to identify clumps. One way of doing this is to construct the (random) subgraph $H$ of the $t \times t$ rectangular lattice which has edges only between neighbouring lattice points $(k,l)$ and $(k',l')$ at which $I_{kl} = I_{k'l'} = 1$. Clumps can then be identified with the connected components of $H$ which contain at least one $2 \times 2$ square; let $\sigma \in S$ index all possible components of this type. Then a point process $\Xi$ can be defined by $\sum_{\sigma \in S} J_\sigma \delta_\sigma$, where $J_\sigma = 1$ exactly when $\sigma$ appears as a component of $H$. The point process $\Xi$ can be decomposed (4.20) as $\Xi = J_\sigma \delta_\sigma + H_\sigma + \Xi_\sigma$, with $H_\sigma = \sum_{\sigma' \in S_\sigma} J_{\sigma'} \delta_{\sigma'}$, and $S_\sigma = \{ \sigma' \in S : \sigma' \neq \sigma, \sigma' \cap \sigma^+ \neq \emptyset \}$, where $\sigma^+$ consists of $\sigma$ and all its lattice neighbours. In this decomposition, $J_\sigma$ and $\Xi_\sigma$ are independent and $J_\sigma Z_\sigma = 0$ a.s., where $Z_\sigma = \| H_\sigma \|$, leading to a Poisson Process Local Estimate (4.25) of

$$d_{TV}(\mathcal{L}(\Xi), PP(\lambda)) \leq \sum_{\sigma \in S} p_\sigma \left( p_\sigma + \sum_{\sigma' \in S_\sigma} p_{\sigma'} \right), \quad (5.23)$$

where $p_\sigma = E J_\sigma = q^{|\sigma|}(1-q)^{|\sigma^+|-|\sigma|}$, and the same bound then holds also for the compound Poisson approximation error $d_{TV}(\mathcal{L}(W), CP(\lambda, \mu))$, with $\lambda$ and $\mu$ defined by

$$\lambda \mu = \sum_{\sigma \in S} p_\sigma \delta_{l(\sigma)}, \quad (5.24)$$

where $l(\sigma)$ denotes the number of $2 \times 2$ squares contained in $\sigma$.

To evaluate the bound in (5.23), note that $np = \sum_{\sigma \in S} p_\sigma$, and that $p_\sigma = p$ for each of the $n$ components $\sigma$ consisting of just a $2 \times 2$ square. Thus the bound is at best of order $O(np^2)$; this improves upon the order $O(p^{1/2})$ of the Poisson approximation (5.22) when $np$ is not too large, but is worse if $n \gg p^{-3/2}$. Note also that the compound Poisson distribution in (5.24) is complicated to compute, as is the bound (5.23), since both contain elements from almost all possible components of subgraphs of the (large) $t \times t$ lattice. In practice, one would probably prefer to incur an extra penalty of $\sum_{\sigma \in S_8} p_\sigma$ in exchange for neglecting $\sum_{\sigma \in S_8} J_\sigma \delta_\sigma$, where $S_8 = \{ \sigma \in S : |\sigma| \geq 8 \}$ contains all elements of $S$ with 8 or more vertices. This reduces $l(\sigma)$ to taking only two possible values, 1 and 2, in the compound Poisson approximation $CP(\lambda, \mu^{(8)})$, where

$$\lambda \mu^{(8)} = \sum_{\sigma \in S \setminus S_8} p_\sigma \delta_{l(\sigma)}, \quad (5.25)$$

but the bound remains awkward to compute, and is still of order only $O(np^2)$.
For direct compound Poisson approximation by Stein’s method, one can take the decomposition $W = W_{ij} + Z_{ij} + U_{ij} + I_{ij}$ in (5.6), where now $U_{ij} = \sum_{(k, l) \in N_i \setminus \{(i, j)\}} I_{kl}$ (the previous $Z_{ij}$) and $Z_{ij} = \sum_{(k, l) \in N_i \setminus N_j} I_{kl}$, so that $W_{ij} = \sum_{(k, l) \notin N_i} I_{kl}$ is smaller than in the Poisson case. Then $W_{ij}$ and $(I_{ij}, U_{ij})$ are independent, making $\delta_1 = \delta_2 = \delta_3 = 0$ in (5.8)-(5.10), and
\[ \mathbf{E} I_{ij} \mathbf{E} (I_{ij} + U_{ij} + Z_{ij}) + \mathbf{E} (I_{ij} Z_{ij}) = 41 p^2. \]

Now the canonical approximating compound Poisson distribution $\text{CP} (\lambda, \mu)$ has
\[ \lambda \mu_l = l^{-1} n \mathbf{E} \{ I_{11} I [U_{11} = l - 1] \}, \quad l \geq 1. \tag{5.26} \]

Since $U_{11}$ is a function of the configuration of 1’s on a $4 \times 4$ square, its distribution is easily calculable algebraically as a function of $q$. In particular, $0 \leq U_{11} \leq 8$ a.s., and $(l + 1) \mu_{t + 1} / (l \mu_l) = O(q)$ for each $1 \leq l \leq 8$, so that condition (5.17) is satisfied for all $t$ large enough; furthermore,
\[ \lambda \mu_1 = np \mathbb{P}[U_{11} = 0 | I_{11} = 1] \geq np(1 - 4(q^2 + q^3)); \]
\[ 2 \lambda \mu_2 = np \mathbb{P}[U_{11} = 1 | I_{11} = 1] \leq 4 np(q^2 + q^3). \]

Hence, from the Compound Poisson Estimate 1 (5.19), it follows that, for all $t$ sufficiently large,
\[ d_{TV}(\mathcal{L}(W), \text{CP} (\lambda, \mu)) \leq \frac{41 p}{1 - 8(q^2 + q^3)} \left\{ \frac{1}{4np(1 - 8(q^2 + q^3))} + \log[2np(1 - 8(q^2 + q^3))] \right\} \]
\[ = O(p(t) \log\{np(t)\}). \tag{5.27} \]

This is very much better than (5.23), and is only inferior to (5.22) for very large values of $n$: for instance, if $p(t) = t^{-\beta}$ for any $0 < \beta < 2$, then the order obtained from (5.27) is $O(t^{-\beta} \log t)$, almost the optimal $O(t^{-\beta})$, and much better than the order $O(p(t)^{1/2}) = O(t^{-\beta/2})$ obtained for Poisson approximation. Squares of any size $r \times r$ are treated in this way in Barbour, Chryssaphinou and Roos (1996).

If the Compound Poisson Estimate 2 is used instead, the result is less explicit than that of (5.27), and it is necessary to bound the probability $\mathbb{P}[W \leq y \mathbb{E} W]$ for $y < 1$. However, we can use Janson’s inequality (3.4) to do this, with $\delta = 4(q^2 + q^3)$, yielding a bound of $\exp\{-\frac{1}{2} np[1 - y + y \log y]\}$ for all $t$ large enough that $4(q^2(t) + q^3(t)) \leq \frac{1}{2}$. Hence the Compound Poisson Estimate 2 gives a rate
\[ d_{TV}(\mathcal{L}(W), \text{CP} (\lambda, \mu)) = O(p + e^{-\alpha np}) \tag{5.28} \]
for some $\alpha > 0$, which is of the ideal order $O(p)$ as soon as $np$ is at all large, and which, combined with (5.27), shows that
\[ d_{TV}(\mathcal{L}(W), \text{CP} (\lambda, \mu)) = O(p \log(1/p)) \tag{5.29} \]

35
as \( t \to \infty \), under all circumstances.

Note that, in (5.27)–(5.29), the compound Poisson distribution \( \mathbb{CP}(\lambda, \mu) \) can be replaced by \( \mathbb{CP}(\lambda', \mu') \), where \( \mu'_1 = 1 - 4(q^2 + q^3) \) and \( \mu'_2 = 4(q^2 + q^3) \), and \( \lambda' = n \rho (1 + 4(q^2 + q^3))^{-1} \) is chosen to satisfy \( \lambda' m'_1 = \lambda m_1 \). This is because the difference involved in replacing \( \varepsilon_0 \) and \( \varepsilon_1 \) in the Compound Poisson Estimates 1 and 2 by \( \varepsilon'_0 \) and \( \varepsilon'_1 \) as in (5.14) is only of order \( O(p) \), since, by the Bonferroni inequalities,

\[
|\mathbb{P}[U_{11} = 0 | I_{11} = 1] - \{1 - 4(q^2 + q^3)\}| \leq 28p
\]

and

\[
|\mathbb{P}[U_{11} = 1 | I_{11} = 1] - 4(q^2 + q^3)| \leq 56p,
\]

from which it follows that \( d_W(Q', Q) = O(p) \). The distribution \( \mu' \) satisfies (5.17) for all \( q \) such that \( q^2 + q^3 \leq 1/12 \).

The setting of Example 5.2 can be seen as a simplification of that discussed in the Poisson process context in Example 4.4. Compound Poisson approximation for the number of \( k \)-subsets covered by a square \( C \) of side \( c \), when \( m \) points are distributed uniformly at random on a square of side 1, is shown in Barbour and Månsson (1998) to be accurate to order \( O(m^{-1} \lambda + e^{-\alpha \lambda}) \), with \( \lambda = m(mc^2)^{k-1} \). The analogy is obtained by taking \( k = 4 \), and replacing \( m \) by \( t^2q \) and \( c \) by \( 2t^{-2} \), in which notation this error is \( O(q^{k-1} + e^{-\alpha \rho}) \), for some \( \alpha > 0 \), not quite as small as in (5.28). In this case, the bounds (5.18) cannot be used in general, since (5.17) is not satisfied. Further examples of Stein’s method used in compound Poisson approximation are to be found in Eichelsbacher and Roos (1998) and in Erhardsson (1998); in the latter paper, Stein’s method is combined with regenerative theory to give a very general treatment of compound Poisson approximation to the number of hits on a rare set in a Markov chain.

References


37


38


