Notes on Poisson Approximation

A. D. Barbour*
Universität Zürich

Progress in Stein's Method, Singapore, January 2009

These notes are a supplement to the article 'Topics in Poisson Approximation', which appeared in *Stochastic Processes: Theory and Methods*, Handbook of Statistics, Eds D. N. Shanbhag and C. R. Rao, Elsevier Science, 79–115 (2001)

6. Translated Poisson approximation

In our basic example for the Stein-Chen method, we take $I_1, I_2, \ldots$ to be independent indicator random variables. A Poisson approximation is then plausible if the probabilities $p_i := \mathbf{P}(I_i = 1)$ are generally small. The Stein–Chen method, as in (2.9), easily implies that

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \;\le\; \lambda^{-1} \sum_{i=1}^{n} p_i^2 \;\le\; \max_{1 \le i \le n} p_i, \tag{6.1}$$

where $\lambda := \mathbf{E}W = \sum_{i=1}^n p_i$.

Clearly, if the p_i's are not all small, there may be little content in (6.1). This is to be expected, since then $\mathbf{E}W = \lambda$ and $\mathrm{Var}\, W = \lambda - \sum_{i=1}^n p_i^2$ need no longer be close to one another, whereas Poisson distributions have equal mean and variance. This makes it more natural to try to find a family of distributions for the approximation within which both mean and variance can be matched, as is possible using the normal family in the classical central limit theorem. One choice is to approximate with a member of the family

* Partially supported by Schweizer Nationalfonds Grant Nr. 20–117625/1

of translated Poisson distributions $\{\mathrm{TP}\,(\mu,\sigma^2),\ (\mu,\sigma^2) \in \mathbf{R}\times\mathbf{R}_+\}$, where

$$\begin{aligned} \mathrm{TP}\,(\mu,\sigma^2)\{j\} &:= \mathrm{Po}(\sigma^2+\delta)\{j-\lfloor\mu-\sigma^2\rfloor\} \\ &= \mathrm{Po}(\lambda')\{j-\gamma\}, \quad j\in\mathbf{Z}, \end{aligned} \tag{6.2}$$

where

$$\begin{aligned} &\gamma \ := \ \gamma(\mu,\sigma^2) \ := \ \lfloor\mu-\sigma^2\rfloor, \quad \delta \ := \ \delta(\mu,\sigma^2) \ := \ \mu-\sigma^2-\gamma \\ &\text{and} \quad \lambda' \ := \ \lambda'(\mu,\sigma^2) \ := \ \sigma^2+\delta. \end{aligned} \tag{6.3}$$

The $\mathrm{TP}\,(\mu,\sigma^2)$ distribution is just that of a Poisson with mean $\lambda' := \lambda'(\mu,\sigma^2) := \sigma^2+\delta$, then shifted along the lattice by an amount $\gamma := \gamma(\mu,\sigma^2) := \lfloor\mu-\sigma^2\rfloor$. In particular, it has mean $\lambda'+\gamma=\mu$ and variance λ' such that $\sigma^2 \le \lambda' < \sigma^2+1$; note that $\lambda' = \sigma^2$ only if $\mu-\sigma^2\in\mathbf{Z}$. For sums of independent, integer-valued random variables Y_i, this idea has been exploited by Vaitkus & Čekanavičius (1998), and also in Barbour & Xia (1999), Čekanavičius & Vaitkus (2001) and Barbour & Čekanavičius (2002), using Stein's method. Error rates are obtained that are of the same order as in the classical central limit theorem, but now with respect to the much stronger total variation norm, provided that some 'smoothness' of the distribution of W can be established.

Just as in the Poisson case, the introduction of Stein's method raises the possibility of making similar approximations for sums of dependent random variables as well. However, the 'smoothness' needed is typically a bound of order $O(1/\sqrt{n})$ for $d_{TV}(\mathcal{L}(W+1),\mathcal{L}(W))$, and this can lead to much more delicate arguments than are required for Poisson approximation. Nonetheless, since a sum of *even* integer valued random variables can never be close in total variation to any translated Poisson distribution, it is clear that some such condition is necessary.

In this chapter, we discuss some of the ways in which total variation approximation by translated Poisson distributions can be achieved. We give an explicit inequality in Lemma 6.1, from which the accuracy of translated Poisson approximation can be directly deduced. We then illustrate its use in the context of sums of independent random variables. Applying it for sums of dependent random variables is a lot more difficult; for an example in the case of Markov dependence, see Barbour and Lindvall (2006). Here, we discuss instead one of Röllin's methods for making such approximations (Röllin 2005), which is effective in a wider range of circumstances, including many local and combinatorial dependence structures. Here, the key feature is to find a sum of (conditionally) independent Bernoulli random variables imbedded within the problem. In such circumstances, the method also yields local limit approximations of appropriate accuracy.

6.1 The basic lemma

Since the $\mathrm{TP}\,(\mu,\sigma^2)$ distributions are just translates of Poisson distributions, the

Stein–Chen method itself can be used to establish total variation approximation. In particular, from (6.2), $W \sim \mathrm{TP}\,(\mu,\sigma^2)$ if and only if

$$\mathbf{E}\{\lambda' g(W+1) - (W-\gamma)g(W)\} \;=\; 0 \tag{6.4}$$

for all bounded functions $g : \mathbf{Z} \to \mathbf{R}$, where $\lambda' = \lambda'(\mu,\sigma^2)$ and $\gamma = \gamma(\mu,\sigma^2)$ are as defined in (6.2). Define g^*_C for $C \subset \mathbf{Z}_+$ by

$$\begin{aligned} g^*_C(k) \;&=\; 0, \quad k \le 0; \\ \lambda' g^*_C(k+1) - k g^*_C(k) \;&=\; \mathbf{1}_C(k) - \mathrm{Po}(\lambda')\{C\}, \quad k \ge 0, \end{aligned}$$

as in (2.1) for the Stein–Chen method. It then follows that

$$\|g^*_C\| \;\le\; (\lambda')^{-1/2} \quad \text{and} \quad \|\Delta g^*_C\| \;\le\; (\lambda')^{-1}$$

(Barbour, Holst and Janson 1992, Lemma I.1.1), where $\Delta g(j) := g(j+1) - g(j)$ and, for bounded functions $g : \mathbf{Z} \to \mathbf{R}$, we let $\|g\|$ denote the supremum norm. Correspondingly, for $B \subset \mathbf{Z}$ such that $B^* := B - \gamma \subset \mathbf{Z}_+$, the function g_B defined by

$$g_B(j) \;:=\; g^*_{B^*}(j-\gamma), \quad j \in \mathbf{Z}, \tag{6.5}$$

satisfies

$$\begin{aligned} &\lambda' g_B(w+1) - (w-\gamma) g_B(w) \\ &\quad=\; \lambda' g^*_{B^*}(w-\gamma+1) - (w-\gamma) g^*_{B^*}(w-\gamma) \\ &\quad=\; \mathbf{1}_{B^*}(w-\gamma) - \mathrm{Po}(\lambda')\{B^*\} \\ &\quad=\; \mathbf{1}_B(w) - \mathrm{TP}\,(\mu,\sigma^2)\{B\} \end{aligned} \tag{6.6}$$

if $w \ge \gamma$, and

$$\lambda' g_B(w+1) - (w-\gamma) g_B(w) \;=\; 0 \tag{6.7}$$

if $w < \gamma$; and clearly

$$\|g_B\| \;\le\; (\lambda')^{-1/2} \quad \text{and} \quad \|\Delta g_B\| \;\le\; (\lambda')^{-1}. \tag{6.8}$$

This can be exploited to prove the closeness in total variation of $\mathcal{L}(W)$ to $\mathrm{TP}\,(\mu,\sigma^2)$ for an arbitrary integer-valued random variable W, as illustrated in the following result. Note that we make no assumptions about non-negativity or about the dependence structure among the random variables $Y_1, \ldots, Y_n$.

Lemma 6.1. *Let $Y_1, Y_2, \ldots, Y_n$ be integer valued random variables with finite means, and define $W := \sum_{i=1}^n Y_i$. Let $(a_i)_{i=1}^n$ and $(b_i)_{i=1}^n$ be real numbers such that, for all bounded $g : \mathbf{Z} \to \mathbf{R}$,*

$$|\mathbf{E}[Y_i g(W)] - \mathbf{E}[Y_i]\mathbf{E}g(W) - a_i\mathbf{E}[\Delta g(W)]| \ \leq \ b_i\|\Delta g\|, \quad 1 \leq i \leq n. \tag{6.9}$$

Then

$$d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mathbf{E}W, \sigma^2)) \ \leq \ (\lambda')^{-1}\left(\delta + \sum_{i=1}^n b_i\right) + \mathbf{P}[W < \lfloor \mathbf{E}W - \sigma^2 \rfloor],$$

where $\sigma^2 := \sum_{i=1}^n a_i$, $\delta = \delta(\mathbf{E}W, \sigma^2)$ and $\lambda' = \sigma^2 + \delta$.

Proof. Adding (6.9) over i, and then adding and subtracting $c\mathbf{E}g(W)$ for $c \in \mathbf{R}$ to be chosen at will, we get

$$|\mathbf{E}[(W-c)g(W)] - (\mathbf{E}W - c - \sigma^2)\mathbf{E}g(W) - \sigma^2\mathbf{E}[g(W+1)]| \ \leq \ \left(\sum_{i=1}^n b_i\right)\|\Delta g\|,$$

where $\sigma^2 = \sum_{i=1}^n a_i$ as above. Taking $c = \gamma = \lfloor \mathbf{E}W - \sigma^2 \rfloor$, so that the middle term (almost) disappears, the expression can be rewritten as

$$|\mathbf{E}[(W-\gamma)g(W)] - \lambda'\mathbf{E}[g(W+1)]| \ \leq \ \left(\delta + \sum_{i=1}^n b_i\right)\|\Delta g\|, \tag{6.10}$$

where δ and λ' are as above.

Fixing any set $B \subset \mathbf{Z}_+ + \gamma$, take $g = g_B$ as in (6.5). It then follows from (6.6) that

$$\begin{aligned}
&|\mathbf{P}(W \in B) - \mathrm{TP}\,(\mathbf{E}W, \sigma^2)\{B\}| \\
&\quad = \ |\mathbf{E}\{(\mathbf{1}_B(W) - \mathrm{TP}\,(\mathbf{E}W, \sigma^2)\{B\})(I[W \geq \gamma] + I[W < \gamma])\}| \\
&\quad \leq \ |\mathbf{E}\{(\lambda' g_B(W+1) - (W-\gamma)g_B(W))\,I[W \geq \gamma]\}| + \mathbf{P}(W < \gamma) \\
&\quad = \ |\mathbf{E}\{\lambda' g_B(W+1) - (W-\gamma)g_B(W)\}| + \mathbf{P}(W < \gamma),
\end{aligned} \tag{6.11}$$

this last from (6.7). Hence (6.10) and (6.11) show that, for any $B \subset \mathbf{Z}_+ + \gamma$,

$$\begin{aligned}
|\mathbf{P}(W \in B) - \mathrm{TP}\,(\mathbf{E}W, \sigma^2)\{B\}| \ &\leq \ \left(\delta + \sum_{i=1}^n b_i\right)\|\Delta g_B\| + \mathbf{P}(W < \gamma) \\
&\leq \ (\lambda')^{-1}\left(\delta + \sum_{i=1}^n b_i\right) + \mathbf{P}(W < \gamma).
\end{aligned} \tag{6.12}$$

Now the largest value D of the differences $\{\mathrm{TP}\,(\mathbf{E}W,\sigma^2)\{C\} - \mathbf{P}(W \in C)\}$, $C \subset \mathbf{Z}$, is attained at a set $C_0 \subset \mathbf{Z}_+ + \gamma$, and is thus bounded as in (6.12); the minimum is attained at $\mathbf{Z} \setminus C_0$ with the value $-D$. Hence

$$|\mathbf{P}(W \in C) - \mathrm{TP}\,(\mathbf{E}W,\sigma^2)\{C\}| \;\le\; (\lambda')^{-1}\left(\delta + \sum_{i=1}^n b_i\right) + \mathbf{P}(W < \gamma)$$

for all $C \subset \mathbf{Z}$, and the lemma follows. □

If the random variables Y_i have finite variances, both λ' and $\mathrm{Var}\,W$ are typically of order $O(n)$, so that letting $\bar{b} := n^{-1}\sum_{i=1}^n b_i$ and applying Chebyshev's inequality to bound the final probability, we find that then $d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mathbf{E}W,\sigma^2))$ is of order $O(n^{-1}+\bar{b})$. Hence we are interested in choosing $a_1, a_2, \ldots$ so that $b_1, b_2, \ldots$ are small.

6.2 Independent indicator random variables

As a first example, consider the case where $W := \sum_{i=1}^n Y_i$ when $Y_1, Y_2, \ldots$ are independent indicators, with $\mathbf{E}Y_i = p_i$. (6.1) is good if the p_i are small: how do things go when they are not? We start by computing the elements in (6.9). First, we note that

$$\begin{aligned}\mathbf{E}\{Y_i g(W)\} &= p_i \mathbf{E}g(W_i+1);\\ \mathbf{E}Y_i\,\mathbf{E}g(W) &= p_i(1-p_i)\mathbf{E}g(W_i) + p_i^2\mathbf{E}g(W_i+1),\end{aligned} \tag{6.13}$$

where $W_i := \sum_{j\ne i} Y_j$. Hence

$$\mathbf{E}\{Y_i g(W)\} - \mathbf{E}Y_i\mathbf{E}g(W) \;=\; p_i(1-p_i)\mathbf{E}\Delta g(W_i). \tag{6.14}$$

Comparing this expression with (6.9) suggests taking $a_i = p_i(1-p_i)$. We then need to evaluate the difference $\mathbf{E}\Delta g(W) - \mathbf{E}\Delta g(W_i)$. Since

$$\begin{aligned}\mathbf{E}\Delta g(W) &= (1-p_i)\mathbf{E}\Delta g(W_i) + p_i\mathbf{E}\Delta g(W_i+1)\\ &= \mathbf{E}\Delta g(W_i) + p_i\mathbf{E}\Delta^2 g(W_i),\end{aligned}$$

we can substitute this into (6.14) to give

$$\mathbf{E}\{Y_i g(W)\} - \mathbf{E}Y_i\,\mathbf{E}g(W) - p_i(1-p_i)\mathbf{E}\Delta g(W) \;=\; -p_i^2(1-p_i)\mathbf{E}\Delta^2 g(W_i). \tag{6.15}$$

The right hand side of (6.15) is not quite in the form expected in (6.9), because of the appearance of the *second* difference of the function g. One might expect this to work to our advantage, since, for the solutions $g_{\lambda,A}$ of the Stein Equation for the Poisson distribution $\mathrm{Po}(\lambda)$ corresponding to $f = \mathbf{1}_A$, the bound (2.3) on $\|\Delta g_{\lambda,A}\|$ is of smaller order in λ than that on $\|g_{\lambda,A}\|$. However, for A a singleton $\{a\}$, suitably chosen, the value of $\|\Delta g_{\lambda,a}\|$ can be arbitrarily close to the upper bound in (2.3), and, for this choice of A, $\|\Delta^2 g_{\lambda,a}\| > \|\Delta g_{\lambda,a}\|$. What is more, if we bounded $|\mathbf{E}\Delta^2 g_A(W_i)|$ in (6.15) by (say) $2\|\Delta g_A\|$, the conclusion of Lemma 6.1 could yield nothing better than

$$d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mathbf{E}W, \sigma^2)) \;\le\; 2\sigma^{-2}\sum_{i=1}^n p_i^2(1-p_i)$$

with $\sigma^2 := \sum_{i=1}^n p_i(1-p_i)$, again of order $\max_{1\le i\le n} p_i$ if the p_i are all of comparable size, and hence no better than the bound in (6.1).

The trick is to observe that, for any bounded function g and any random variable V,

$$|\mathbf{E}\Delta^2 g(V)| \;\le\; 2\|\Delta g\|\; d_{TV}(\mathcal{L}(V), \mathcal{L}(V+1)), \tag{6.16}$$

in view of the definition of total variation distance. Note that the modulus must be *outside* the expectation. Now we can use an inequality, which can be proved using the ‘Mineka coupling’ (Lindvall 2002, II.14),

$$d_{TV}(\mathcal{L}(W_i), \mathcal{L}(W_i+1)) \;\le\; c_{MR}\Big\{\sum_{j\ne i}\{1 - d_{TV}(\mathcal{L}(Y_j), \mathcal{L}(Y_j+1))\}\Big\}^{-1/2}, \tag{6.17}$$

true for *arbitrary* independent integer valued random variables Y_i: the improved version here with $c_{MR} = \sqrt{2/\pi}$ was established by Mattner and Roos (2006). For our indicator random variables,

$$1 - d_{TV}(\mathcal{L}(Y_i), \mathcal{L}(Y_i+1)) \;=\; \min\{p_i, 1-p_i\}. \tag{6.18}$$

If $\sum_{i=1}^n p_i^2 \ge 1$, this leads to the following bound.

Theorem 6.2. *If $W := \sum_{i=1}^n Y_i$, where $Y_1,\ldots,Y_n$ are independent and $Y_i \sim \mathrm{Be}\,(p_i)$, then*

$$d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\lambda, \sigma^2)) \;\le\; 2\sigma^{-2}\left(1 + \sum_{i=1}^n p_i^2(1-p_i)\varphi\right) \;\le\; 2\sigma^{-2} + 2\varphi, \tag{6.19}$$

with $\lambda := \sum_{i=1}^n p_i$, $\sigma^2 := \sum_{i=1}^n p_i(1-p_i)$ and

$$\varphi \;:=\; \max_{1\le i\le n} c_{MR}\Big\{\sum_{j\ne i}\min\{p_j, 1-p_j\}\Big\}^{-1/2}. \tag{6.20}$$

If $\sigma^2 \ge 1$, the bound can be simply majorized by $4/\sigma$.

Note that, if $\sum_{i=1}^n p_i^2 < 1$, the approximation is by the Poisson distribution, and (6.1) should be used instead. If not, then the bound is at most $2\sigma^{-1}(2c_{MR}\max_i p_i + \sigma^{-1})$, since

$\sigma \geq 1$ when $\sum_{i=1}^n p_i^2 \geq 1$, and this is of order $O(p/\sqrt{np} + 1/np)$ if all the p_i are equal to p, $1/\sqrt{n} \leq p \leq 1/2$; smaller than the error in Poisson approximation if $p \gg 1/\sqrt{n}$, and smaller than that in the central limit theorem for $p \ll 1$. So it is a useful result.

Now there is no need to restrict attention to indicator random variables; any independent integer valued random variables with finite means will do. Of course, the expressions analogous to (6.13)–(6.15) become much more complicated, and it is advisable to assume that the Y_i have third moments, if an error rate of order $O(n^{-1/2})$ is to be achieved — exactly as in the Berry–Esseen theorem. But the basic method of proof remains the same: see, for example, Barbour and Čekanavičius (2002, Theorem 3.1). The big concern is that (6.17) need not yield something small, unless the distributions of the Y_i overlap with their unit translates. Something of this sort is also to be expected. For instance, for *even* integer valued random variables Y_i, their sum is also necessarily even, and approximation by a translated Poisson distribution in *total variation* is clearly poor. Improving the central limit theorem to approximation in total variation is a sensitive issue, and 'relatively small' errors cannot simply be discounted.

Note that, in (6.9), the choice

$$a_i \;=\; \mathbf{E}[Y_i W] - \mathbf{E}[Y_i]\mathbf{E}[W] \tag{6.21}$$

is rather natural, and always implies that $\sigma^2 = \mathrm{Var}\, W$.

The argument above hinges on showing that, in (6.15),

$$|\mathbf{E}\Delta^2 g_A(W_i)| \;=\; O(\sigma^{-3}), \tag{6.22}$$

and this was achieved using the observation (6.16). The end result is then, broadly speaking, to be able to approximate the probability of any set A to accuracy $O(\sigma^{-1})$, and not just the probabilities of intervals, as in the central limit theorem. However, for local limit approximations, this accuracy is too weak, since it is typically of the same order as the probabilities being approximated. For singletons $A = \{a\}$, one can, however, exploit the extra properties of the corresponding functions g_a. In particular, the form of the function g_a easily implies the ℓ_1 estimate

$$\sum_j |\Delta g_a(j)| \;\leq\; 2\sigma^{-2} \tag{6.23}$$

for the solutions to the Stein equation for $\mathrm{Po}(\sigma^2)$. This allows one to prove the bound

$$\begin{aligned} |\mathbf{E}\Delta^2 g_a(W')| \;&=\; \Bigl|\sum_j \{\Delta g_a(j+1) - \Delta g_a(j)\}\mathbf{P}[W' = j]\Bigr| \\ &=\; \Bigl|\sum_j \{\mathbf{P}[W' = j] - \mathbf{P}[W' = j-1]\}\Delta g_a(j)\Bigr| \end{aligned}$$

$$
\begin{aligned}
&\le \max_j \bigl|\mathbf{P}[W' = j] - \mathbf{P}[W' = j-1]\bigr| \sum_j |\Delta g_a(j)| \\
&\le 2\sigma^{-2} \max_j \bigl|\mathbf{P}[W' = j] - \mathbf{P}[W' = j-1]\bigr|. \qquad (6.24)
\end{aligned}
$$

The trick is now to show that

$$\max_j \bigl|\mathbf{P}[W' = j] - \mathbf{P}[W' = j-1]\bigr| \;=\; O(\sigma^{-2}). \qquad (6.25)$$

To do so, we consider once again a sum of independent random variables Y_i, and we now suppose that, by using the total variation approximation techniques above, we have been able to show that $W' = W_1' + W_2'$, where W_1' and W_2' are independent, and such that $d_{TV}(\mathcal{L}(W_l'), \mathcal{L}(W_l' + 1)) = O(\sigma^{-1})$, $l = 1, 2$, using (6.17). Then, by independence,

$$\mathbf{P}[W' = j] - \mathbf{P}[W' = j-1] \;=\; \mathbf{E}\{h_j(W_1') - h_j(W_1' + 1)\},$$

where

$$h_j(k) \;:=\; \mathbf{P}[W_2' = j - k].$$

Now it is immediate that

$$|h_j(k)| \;=\; |\mathbf{P}[W_2' \le j - k] - \mathbf{P}[W_2' + 1 \le j - k]| \;\le\; d_{TV}(\mathcal{L}(W_2'), \mathcal{L}(W_2' + 1)),$$

so that

$$\|h_j\| \;\le\; d_{TV}(\mathcal{L}(W_2'), \mathcal{L}(W_2' + 1)) \;=\; O(\sigma^{-1}), \qquad (6.26)$$

uniformly in j. Hence it follows that

$$|\mathbf{E}\{h_j(W_1') - h_j(W_1' + 1)\}| \;\le\; 2\|h_j\|\, d_{TV}(\mathcal{L}(W_1'), \mathcal{L}(W_1' + 1)) \;=\; O(\sigma^{-2}),$$

again uniformly in j. This establishes (6.25), from which and (6.24) we obtain

$$|\mathbf{E}\Delta^2 g_a(W')| \;=\; O(\sigma^{-4}). \qquad (6.27)$$

This strengthening of (6.22) for singleton sets $A = \{a\}$ enables one to establish local limit approximations for sums of independent random variables Y_i to an error of order $O(\sigma^{-2})$, which, by analogy with the central limit theorem, is just what one would hope for. Here, it is once again critical that the counterparts of (6.17) for W_1 and W_2 should yield bounds of order $O(\sigma^{-1})$: exclusively *even integer* valued random variables cannot work.

For the sum W of independent indicator random variables Y_i, the considerations above yield the following theorem. Let λ and σ^2 be the mean and variance of W, as before. Now split the Y_j into two sets, denoted by $(Y_{1j},\, 1 \le j \le n_1)$ and $(Y_{2j},\, 1 \le j \le n_2)$, with $n_1 + n_2 = n$, and set $p_{lj} = \mathbf{E}Y_{lj}$, $l = 1, 2$. We apply the argument above with each of the random variables $W_i := W - Y_i$ in turn as W'; thus, for the quantities appearing in the bounds, we define

$$\left.\begin{aligned}\sigma_l^2 &:= \sum_{j=1}^{n_l} p_{lj}(1-p_{lj}) - \max_{1\le i\le n_l} p_{li}(1-p_{li}), \qquad l=1,2;\\ \varphi_l &:= \max_{1\le i\le n} c_{MR}\Bigl\{\sum_{j\ne i}\min\{p_{lj}, 1-p_{lj}\}\Bigr\}^{-1/2},\end{aligned}\right. \qquad l=1,2.$$

The notation

$$d_{loc}(P,Q) \;:=\; \max_i |P\{i\} - Q\{i\}|$$

is used to denote the largest difference between point probabilities for probability distributions P and Q on the integers.

Theorem 6.3. *With the above definitions and assumptions, we have*

$$d_{loc}(\mathcal{L}(W), \mathrm{TP}\,(\lambda,\sigma^2)) \;\le\; 8\left\{\frac{1}{\sigma_1^2} + \varphi_1\right\}\left\{\frac{1}{\sigma_2^2} + \varphi_2\right\}\max_{1\le i\le n} p_i.$$

In particular, for $\sigma^2 \ge 2$, the bound can be majorized by $280\sigma^{-2}\max_{1\le i\le n} p_i$.

Of course, the choice of split can be made to minimize the bound in the theorem.

6.3 Dependent random variables: Röllin's theorem

There are many arguments for proving sharp results for sums of independent random variables Y_i, and the results above could certainly be improved upon by other means. Our interest lies primarily in showing that this technique can also be used for sums of *dependent* random variables. Because total variation is a very sharp measure of distance, the arguments are unfortunately correspondingly tricky. The method above can be carried through, provided that the error terms left over when computing the left hand side of (6.9) can be expressed (more or less) in terms of quantities of the form $|\mathbf{E}\Delta^2 g_A(W')|$, for random variables W' such that $d_{TV}(\mathcal{L}(W'), \mathcal{L}(W'+1))$ is small enough. To establish this latter property, the usual approach is to find a sum of (conditionally) independent integer valued random variables within W', to which the inequality (6.17) can be applied.

The following theorem of Röllin (2005) gives one way of doing this. It yields approximations both in total variation and locally. Note that the left hand side of (6.9), the difficult part, does not come into the conditions of the theorem. Instead, it is replaced with an inequality (6.28) which is typically much easier to verify. This inequality can be recognised as expressing (using Stein's method) that a suitably chosen random variable U is close to the standard normal, with respect to expectations of *smooth* test functions. The random variable U is, however, not the same as the original W, even when standardized, and it would often be more convenient if it were.

Theorem 6.4. *Let W be an integer-valued random variable with expectation μ and variance σ^2 and let X be a random element of a Polish space on the same probability space. Define $\mu_X := \mathbf{E}(W|X)$, $\sigma_X^2 := \mathrm{Var}\,(W|X)$ and $\rho^2 := \mathbf{E}(\sigma_X^2)$, and let $\tau^2 = \mathrm{Var}\,(\mu_X)$, $\nu^2 = \mathrm{Var}\,(\sigma_X^2)$. Assume that there exists $\varepsilon \geq 0$ such that $U := (\mu_X - \mu)/\tau$ satisfies*

$$\bigl|\mathbf{E}\{f'(U) - Uf(U)\}\bigr| \;\leq\; \varepsilon\,\|f''\|, \qquad \text{for all } f \in C^2. \tag{6.28}$$

Then we have

$$\begin{aligned} d_{TV}\bigl(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)\bigr) \;&\leq\; \mathbf{E}D_{TV}(X) + \frac{1}{\rho}\left\{14 + \frac{\nu}{\rho} + \frac{5\varepsilon\tau^3}{\sigma^2}\right\}, \\ d_{loc}\bigl(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)\bigr) \;&\leq\; \mathbf{E}D_{loc}(X) + \frac{1}{\rho^2}\left\{49 + \frac{6\nu^2}{\rho^2} + \frac{20\varepsilon\tau^3}{\sigma^2}\right\}, \end{aligned}$$

where

$$D_{TV}(X) := d_{TV}\bigl(\mathcal{L}(W|X), \mathrm{TP}\,(\mu_X,\sigma_X^2)\bigr); \quad D_{loc}(X) := d_{loc}\bigl(\mathcal{L}(W|X), \mathrm{TP}\,(\mu_X,\sigma_X^2)\bigr).$$

The elements in the bounds reflect the way in which Röllin's proof runs. First, one approximates the conditional distribution $\mathcal{L}(W \,|\, X)$ by the translated Poisson distribution which matches its mean μ_X and variance σ_X^2. This yields the first term in the error bounds, which describes how well the distribution of W is approximated by that of a *mixture* of translated Poisson distributions, where mean and variance are chosen according to $\mathcal{L}(\mu_X, \sigma_X^2)$. The remaining term governs errors arising in two ways. First, this translated Poisson distribution is approximated by the mixture with mean chosen according to $\mathcal{L}(\mu_X)$ but with fixed variance $\rho^2 = \mathbf{E}(\sigma_X^2)$, and the appearance of ν/ρ reflects the fact that σ_X^2 must not be too variable if the approximation is to be good. Secondly, this approximation is replaced by one involving a single translated Poisson distribution. Here, significant variation can be accommodated in $\mathcal{L}(\mu_X)$, with the variance parameter increasing from ρ^2 to σ^2, provided that the (normalized) distribution of μ_X is suitably close to the standard normal, whence the appearance of $\varepsilon\tau^3/\sigma^2$ in the bound. Note that the last two steps are approximations within the family of (mixtures of) translated Poisson distributions, and are thus technically relatively easy to handle.

In order to exploit the theorem, we need to be able to approximate $\mathcal{L}(W \,|\, X)$ by $\mathrm{TP}\,(\mu_X, \sigma_X^2)$ in the appropriate fashion. Here, one usually tries to choose X in such a way as to make $\mathcal{L}(W \,|\, X)$ that of a sum of independent integer valued random variables. Theorems 6.2 and 6.3 can be invoked provided these random variables have Bernoulli distributions. We start by showing that Röllin's theorem can be used to extend these theorems to more general independent summands.

Theorem 6.5. *Let $X_1, \ldots, X_n$ be independent integer valued random variables, and let $0 \le q_j(l) \le 1$ for any $1 \le j \le n$ and $l \in \mathbf{Z}$; write $Y_j := X_j + J_j$, where, conditional on $X = (X_1, \ldots, X_n)$, the $J_j \sim \mathrm{Be}\,(q_j(X_j))$ are independent. Write $W := \sum_{j=1}^n Y_j$, $\rho^2 := \mathbf{E}\left\{\sum_{j=1}^n q_j(X_j)(1-q_j(X_j))\right\}$. Then*

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mu, \sigma^2)) &\le \rho^{-1}(27 + 5\Lambda); \\ d_{loc}(\mathcal{L}(W), \mathrm{TP}\,(\mu, \sigma^2)) &\le \rho^{-2}(891 + 20\Lambda), \end{aligned}$$

where $\mu := \mathbf{E}W$, $\sigma^2 := \mathrm{Var}\, W$, and

$$\Lambda \ := \ \sigma^{-2} \sum_{j=1}^n \mathbf{E}|Y_j - \mathbf{E}Y_j|^3$$

is σ times the usual Lyapounov ratio for W.

Proof. We can suppose that $\rho \ge 1$, since otherwise the bounds are vacuous. Setting $T(X) := \sum_{j=1}^n X_j$, we first observe that $\mathcal{L}(W - T(X) \,|\, X)$ is that of a sum of independent indicators with means $q_j(X_j)$, so that, from Theorems 6.2 and 6.3,

$$\begin{aligned} d_{TV}(\mathcal{L}(W \,|\, X), \mathrm{TP}\,(\mu_X, \sigma_X^2)) &\le \min\{1, 4\sigma_X^{-1}\}; \\ d_{loc}(\mathcal{L}(W \,|\, X), \mathrm{TP}\,(\mu_X, \sigma_X^2)) &\le \min\{1, 280\sigma_X^{-2}\}, \end{aligned}$$

where

$$\mu_X \ := \ T(X) + \sum_{j=1}^n q_j(X_j) \quad \text{and} \quad \sigma_X^2 \ := \ \sum_{j=1}^n q_j(X_j)(1 - q_j(X_j)).$$

Note that

$$\nu^2 \ := \ \mathrm{Var}\,(\sigma_X^2) \ \le \ \sum_{j=1}^n \mathbf{E}\{[q_j(X_j)(1-q_j(X_j))]^2\} \ \le \ \tfrac{1}{4}\mathbf{E}(\sigma_X^2) \ = \ \tfrac{1}{4}\rho^2, \qquad (6.29)$$

so that $\nu/\rho \le 1/2$. Also, in view of (6.29), by Chebyshev's inequality,

$$\mathbf{E}\min\{1, \sigma_X^{-1}\} \ \le \ \rho^{-2} + \sqrt{2}\rho^{-1} \ \le \ 3\rho^{-1}; \qquad \mathbf{E}\min\{1, \sigma_X^{-2}\} \ \le \ 3\rho^{-2}, \qquad (6.30)$$

and hence, by Theorems 6.2 and 6.3,

$$\mathbf{E}D_{TV}(X) \ \le \ 12\rho^{-1}; \qquad \mathbf{E}D_{loc}(X) \ \le \ 840\rho^{-2}.$$

Now μ_X can also be written as a sum $\sum_{j=1}^n \mathbf{E}(Y_j \,|\, X_j)$ of independent random variables, with

$$\tau^2 \ := \ \mathrm{Var}\,(\mu_X) \ \le \ \mathrm{Var}\, W \quad \text{and} \quad \sum_{j=1}^n \mathbf{E}|\mathbf{E}(Y_j \,|\, X_j) - \mathbf{E}Y_j|^3 \ \le \ \sum_{j=1}^n \mathbf{E}|Y_j - \mathbf{E}Y_j|^3,$$

and hence, using a standard argument for the normal approximation of sums of independent random variables by Stein's method, it follows that

$$\tau^3\varepsilon \;\le\; \frac{3}{2}\sum_{j=1}^n \mathbf{E}|Y_j - \mathbf{E}Y_j|^3.$$

The theorem follows by substituting these quantities into the bounds given in Theorem 6.4. □

Corollary 6.6. *Let $Y_1,\ldots,Y_n$ be independent integer valued random variables, and for $1 \le j \le n$ and $l \in \mathbf{Z}$ let $p_{jl} := \mathbf{P}[Y_j = l]$. Write $W := \sum_{j=1}^n Y_j$, $\mu := \mathbf{E}W$, $\sigma^2 := \mathrm{Var}\,W$ and*

$$\Lambda \;:=\; \sigma^{-2}\sum_{j=1}^n \mathbf{E}|Y_j - \mathbf{E}Y_j|^3.$$

Then

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)) &\;\le\; \chi^{-1}(27 + 5\Lambda); \\ d_{loc}(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)) &\;\le\; \chi^{-2}(891 + 20\Lambda), \end{aligned}$$

with $\chi^2 := \frac{1}{4}\sum_{j=1}^n \sum_l (p_{jl} \wedge p_{j,l+1})$.

Proof. To recover the setting of the previous theorem, define independent random variables X_j with $\pi_{jl} := \mathbf{P}[X_j = l] = p_{jl} + \frac{1}{2}\{p_{jl} \wedge p_{j,l+1} - p_{j,l-1} \wedge p_{jl}\}$. Then, given $X_1,\ldots,X_n$, let $J_1,\ldots,J_n$ be independent, with $J_j \sim \mathrm{Be}\,(q_j(X_j))$, where

$$q_j(l) \;:=\; \frac{1}{2}(p_{jl} \wedge p_{j,l+1})/\pi_{jl} \;\le\; \frac{1}{2},$$

and check that $X_j + J_j$ and Y_j have the same distribution. Now

$$\begin{aligned} \rho^2 &\;=\; \mathbf{E}\left\{\sum_{j=1}^n q_j(X_j)(1 - q_j(X_j))\right\} \\ &\;=\; \sum_{j=1}^n \sum_l \frac{1}{2}(p_{jl} \wedge p_{j,l+1})(1 - q_j(l)) \;\ge\; \frac{1}{4}\sum_{j=1}^n \sum_l (p_{jl} \wedge p_{j,l+1}) \;=\; \chi^2, \end{aligned}$$

and the corollary follows. □

Now a simple example with dependence, the number of 2–runs in a sequence of Bernoulli trials. Let $I_1,\ldots,I_{2n}$ be independent $\mathrm{Be}\,(p)$ random variables, and let $W := \sum_{i=1}^{2n-1} Y_i$ with $Y_i = I_i I_{i+1}$. The random variables Y_i are one–dependent, and the central limit theorem is straightforward, with error rate $O(\{np^2(1-p^2)\}^{-1/2})$ in Kolmogorov distance. What can be said about translated Poisson approximation?

Theorem 6.7. *With the definitions above, and setting* $\mu = \mathbf{E}W$, $\sigma^2 = \mathrm{Var}\,(W)$, *we have*

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)) &= O(\{np^2(1-p)^2\}^{-1/2}); \\ d_{loc}(\mathcal{L}(W), \mathrm{TP}\,(\mu,\sigma^2)) &= O(\{np^2(1-p)^2\}^{-1}). \end{aligned}$$

Proof. Define $X := (I_2, I_4, \ldots, I_{2n})$. Then, conditional on $X = (x_2, x_4, \ldots, x_{2n}) \in \{0,1\}^n$, we have

$$W = x_2 I_1 + (x_2 + x_4) I_3 + \cdots + (x_{2n-2} + x_{2n}) I_{2n-1},$$

a sum of independent integer valued random variables, with overlap χ_x^2 satisfying

$$\chi_x^2 \geq \frac{1}{4} N_1(x)\{p \wedge (1-p)\}, \tag{6.31}$$

where $N_1(x) := x_2 + \sum_{j=2}^n \mathbf{1}_{\{1\}}(x_{2j-2} + x_{2j})$, and with $\Lambda_x \leq 2$ for all x. Now

$$\mathbf{E}N_1(X) = p + 2(n-1)p(1-p) \quad \text{and} \quad \mathrm{Var}\,(N_1(X)) \leq 5(n-1)p(1-p),$$

where the latter inequality requires some calculation, and we assume that $(n-1)p(1-p) \geq 1$, which is unimportant for the bound we are proving. Arguing as for (6.30), it thus follows that

$$\mathbf{E}\min\{1, N_1(X)^{-1/2}\} \leq \frac{6}{\sqrt{(n-1)p(1-p)}}; \quad \mathbf{E}\min\{1, N_1(X)^{-1}\} \leq \frac{6}{(n-1)p(1-p)},$$

and hence, from Corollary 6.6, that

$$\mathbf{E}D_{TV}(X) = O(1/p(1-p)\sqrt{(n-1)}); \quad \mathbf{E}D_{loc}(X) = O(1/\{p(1-p)\}^2(n-1)), \tag{6.32}$$

with $\mu_X := p\{X_{2n} + 2\sum_{j=1}^{n-1} X_{2j}\}$ and $\sigma_X^2 := p(1-p)\{X_2^2 + \sum_{j=2}^n (X_{2(j-1)} + X_{2j})^2\}$.

Some calculation now shows that

$$\begin{aligned} \nu^2 &:= \mathrm{Var}\,(\sigma_X^2) \leq 11(n-1)p^3(1-p)^3; \\ \rho^2 &:= \mathbf{E}(\sigma_X^2) = p^2(1-p)\{1 + 2(n-1)(1+p)\} \sim 2(n-1)p^2(1-p^2); \\ \tau^2 &:= \mathrm{Var}\,(\mu_X) = (4n-3)p^3(1-p), \end{aligned}$$

with $\sigma^2 = \rho^2 + \tau^2$, and, by Stein's method for the normal distribution,

$$\tau^3 \varepsilon \leq \frac{3}{2}(8n-7)p^4(1-p).$$

Putting these values into the bound given in Theorem 6.4, and recalling (6.32), the theorem follows. □

References

[] Barbour, A.D. and Čekanavičius, V. (2002). Total variation asymptotics for sums of independent integer random variables. *Ann. Probab.* **30**, 509–545.

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

[] Barbour, A.D. and Lindvall, T. (2006). Translated Poisson approximation for Markov chains. *J. Theor. Probab.* **19**, 609–630.

[] Barbour, A.D. and Xia, A. (1999). Poisson perturbations. *ESAIM: P&S* **3**, 131–150.

[] Čekanavičius, V. and Vaitkus, P. (2001). Centered Poisson approximation by the Stein's method. *Lithuanian Math. J.* **41**, 319–329.

[] Lindvall, T. (2002). *Lectures on the coupling method.* Dover Publications.

[] Mattner, L. and Roos, B. (2006) A shorter proof of Kanter's Bessel function concentration bound. *Prob. Theory Rel. Fields* **139**, 191–205.

[] Röllin, A. (2005). Approximation by the translated Poisson distribution. *Bernoulli* **12**, 1115–1128.

[] Vaitkus, P. and Čekanavičius, V. (1998). On a centered Poisson approximation. *Lithuanian Math. J.* **38**, 391–404.