COUPLING AND POISSON APPROXIMATION

Svante Janson

Abstract. We give an overview of the Stein–Chen method for establishing Poisson approximations of various random variables. Couplings of certain variables are used to gives explicit bounds for the total variation distance between the distribution of a random variable and a Poisson variable. Some applications are given. In some cases, explicit couplings may be used to obtain good estimates; in other applications it suffices to show the existence of couplings with certain monotonicity properties.

1. The Stein–Chen method

Consider a random variable W that can be written as a sum $\sum_{\alpha\in\Gamma} I_\alpha$ of 0–1 random variables, where Γ is a finite index set. The variables I_α may be dependent, but we will be interested in cases where the dependencies are rather weak and the individual probabilities $\mathrm{P}(I_\alpha = 1)$ rather small. It is then reasonable to try to approximate the distribution of W by a Poisson distribution.

The Stein–Chen method is a powerful tool to justify such approximations by establishing an upper bound to the error, measured as the total variation distance between $\mathcal{L}(W)$ and the Poisson distribution $\mathrm{Po}(\lambda)$ with a suitably chosen mean λ. In many cases, this method can be fruitfully combined with coupling methods to yield simple upper bounds. Furthermore, the bounds often turn out to be of the right order of magnitude.

The method was originally formulated for normal approximation by Stein in 1970, but the basic idea applies also to other cases. The Poisson approximation version was worked out by Chen (1975). See also Stein (1986). The method has successfully been applied to a number of different problems by various authors. The coupling approach described here is developed in detail in Barbour, Holst and Janson (1992), to which we refer for proofs, further details, additional references, and many more applications.

The central idea in the Stein–Chen method is that for any $\lambda > 0$ and $A \subset \mathbf{Z}^+$ there exists a function $g_{\lambda,A}$ on $\mathbf{Z}^+$ satisfying Stein's equation

$$\lambda g_{\lambda,A}(j+1) - j g_{\lambda,A}(j) = I(j \in A) - \mathrm{Po}(\lambda)\{A\}, \quad j \geq 0. \tag{1}$$

Consequently, substituting W, taking expectations and interchanging the sides, we get

$$\mathrm{P}(W \in A) - \mathrm{Po}(\lambda)\{A\} = \mathrm{E}\big(\lambda g_{\lambda,A}(W+1) - W g_{\lambda,A}(W)\big). \tag{2}$$

The function $g_{\lambda,A}$ can be given explicitly, but the method ignores the exact form of $g_{\lambda,A}$ and uses only some simple estimates. In particular, we will here only use the estimate

$$|g_{\lambda,A}(j+1) - g_{\lambda,A}(j)| \leq 1 \wedge \lambda^{-1}, \quad j \geq 0,\ A \subset \mathbf{Z}^+. \tag{3}$$

Supported by the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine.

(The proof is elementary but too long to be included here.) We define

$$\mathcal{G} = \{g : \mathbf{Z}^+ \to \mathbf{R} \mid |g(j+1) - g(j)| \le 1,\, j \ge 0\}$$

and obtain by (2) and (3)

$$\begin{aligned} d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(\lambda)\big) &= \sup_A \big(\mathrm{P}(W \in A) - \mathrm{Po}(\lambda)\{A\}\big) \\ &= \sup_A \mathrm{E}\big(\lambda g_{\lambda,A}(W+1) - W g_{\lambda,A}(W)\big) \\ &\le (1 \wedge \lambda^{-1}) \sup_{g \in \mathcal{G}} \mathrm{E}\big(\lambda g(W+1) - W g(W)\big). \end{aligned} \tag{4}$$

So far W could be any $\mathbf{Z}^+$-valued random variable and λ any positive number (assuming that the expectations above are finite), but we will now use our assumption $W = \sum_\alpha I_\alpha$. Furthermore, we will for simplicity consider only the case

$$\lambda = \mathrm{E}\, W = \sum_\alpha p_\alpha,$$

where

$$p_\alpha = \mathrm{E}\, I_\alpha = \mathrm{P}(I_\alpha = 1).$$

We then have

$$\begin{aligned} \mathrm{E}\big(\lambda g(W+1) - W g(W)\big) &= \lambda\, \mathrm{E}\, g(W+1) - \sum_\alpha \mathrm{E}\big(I_\alpha g(W)\big) \\ &= \sum_\alpha p_\alpha\, \mathrm{E}\, g(W+1) - \sum_\alpha p_\alpha\, \mathrm{E}\big(g(W) \mid I_\alpha = 1\big). \end{aligned} \tag{5}$$

We now introduce couplings. Assume that for each α we can construct a random variable W_α, on the same probability space as W, such that the distribution $\mathcal{L}(W_\alpha)$ equals the conditional distribution $\mathcal{L}(W - I_\alpha \mid I_\alpha = 1)$. The right hand side of (5) can then be further developed as

$$\sum_\alpha p_\alpha\, \mathrm{E}\big(g(W+1) - g(W_\alpha + 1)\big) \le \sum_\alpha p_\alpha\, \mathrm{E}\,|W - W_\alpha|, \tag{6}$$

for any $g \in \mathcal{G}$. Combining the estimates above, we arrive at our fundamental result.

Theorem 1. *If W and W_α are as above, then*

$$d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(\lambda)\big) \le (1 \wedge \lambda^{-1}) \sum_\alpha p_\alpha\, \mathrm{E}\,|W - W_\alpha|. \tag{7}$$

In order to justify the Poisson approximation we therefore only have to find couplings (W, W_α) such that the the right hand side of (7) is small. Before proceeding to examples of such couplings, let us note that the natural way to construct W_α is to find variables $J_{\beta\alpha}$, $\alpha, \beta \in \Gamma$, such that for each α

$$\mathcal{L}(J_{\beta\alpha}, \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma \mid I_\alpha = 1), \tag{8}$$

and then set $W_\alpha = \sum_{\beta \ne \alpha} J_{\beta\alpha}$. Since $W - W_\alpha = I_\alpha + \sum_{\beta \ne \alpha}(I_\beta - J_{\beta\alpha})$, we obtain the following corollary.

Theorem 2. *With* $(J_{\beta\alpha})$ *as above,*

$$d_{TV}\big(\mathcal{L}(W),\mathrm{Po}(\lambda)\big) \le (1\wedge\lambda^{-1})\Big(\sum_{\alpha\in\Gamma} p_\alpha^2 + \sum_{\alpha\in\Gamma}\sum_{\beta\neq\alpha} p_\alpha \,\mathrm{E}\,|I_\beta - J_{\beta\alpha}|\Big).$$

We will, somewhat informally, also speak of $(J_{\beta\alpha})$ as a coupling in this case. For later use we note that

$$p_\alpha \,\mathrm{E}\, J_{\beta\alpha} = \mathrm{P}(I_\alpha = 1)\,\mathrm{P}(I_\beta = 1 \mid I_\alpha = 1) = \mathrm{P}(I_\alpha = I_\beta = 1) = \mathrm{E}(I_\alpha I_\beta). \tag{9}$$

Remark. The method gives estimates for the total variation distance between a finite sum and a Poisson distribution; no asymptotics are involved. On the other hand, limit theorems yielding convergence in distribution often follow directly from these estimates. Indeed, suppose that the variables and the index set depend on a parameter n, indicated by a subscript, and that the estimate of $d_{TV}\big(\mathcal{L}(W_n),\mathrm{Po}(\lambda_n)\big)$ given by Theorem 1 or Theorem 2 tends to 0 as $n\to\infty$. If $\lambda_n\to\lambda$ for some $\lambda<\infty$, then $d_{TV}\big(\mathcal{L}(W_n),\mathrm{Po}(\lambda)\big)\to 0$ and thus $W_n \xrightarrow{d} \mathrm{Po}(\lambda)$. If $\lambda_n\to\infty$, then $\mathrm{Po}(\lambda_n)$ is asymptotically normal and it follows easily that $(W_n - \mathrm{E}\,W_n)/(\mathrm{E}\,W_n)^{1/2} \xrightarrow{d} \mathrm{N}(0,1)$, i.e., W_n is asymptotically normal. Note that the factor $1\wedge\lambda^{-1}$ in the estimates may be ignored in the first case, and more generally in any situation where $\mathrm{E}\,W$ is bounded and we are only interested in the order of the total variation distance; but the factor gives an important improvement when $\mathrm{E}\,W\to\infty$.

In the next three sections, we show how the general results Theorems 1 and 2 can be applied in some special situations.

2. Independent summands

The simplest application of the estimates above is to a sum of independent indicators I_α. We then simply take $W_\alpha = \sum_{\beta\neq\alpha} I_\beta$, or equivalently, $J_{\beta\alpha} = I_\beta$, $\beta\neq\alpha$, and obtain from Theorem 1 or 2 the following estimate, due to Barbour and Hall (1984).

Theorem 3. *If* $W=\sum_{\alpha\in\Gamma} I_\alpha$*, where* I_α*,* $\alpha\in\Gamma$*, are independent indicators, then*

$$d_{TV}\big(\mathcal{L}(W),\mathrm{Po}(\lambda)\big) \le (1\wedge\lambda^{-1})\sum_{\alpha\in\Gamma} p_\alpha^2. \tag{10}$$

This estimate is of the right order. In fact, Barbour and Hall (1984) proved also the opposite inequality $d_{TV}\big(\mathcal{L}(W),\mathrm{Po}(\lambda)\big) \ge \frac{1}{32}(1\wedge\lambda^{-1})\sum_{\alpha\in\Gamma} p_\alpha^2$. (The constant $\frac{1}{32}$ has later been improved to $\frac{1}{14}$.)

As a corollary to Theorem 3, we obtain the simpler bound

$$d_{TV}\big(\mathcal{L}(W),\mathrm{Po}(\lambda)\big) \le \lambda^{-1}\sum_{\alpha\in\Gamma} p_\alpha^2 \le \max_\alpha p_\alpha, \tag{11}$$

but this estimate is inferior to (10) when a few p_α are much larger than the others. In the special case of identically distributed I_α, W has a binomial distribution and we obtain, for all n and p, the simple estimate

$$d_{TV}\big(\mathrm{Bi}(n,p),\mathrm{Po}(np)\big) \le p.$$

(This estimate is of the right order, provided, say, $np\ge 1$.)

3. Dissociated summands

Suppose that for each α there is a subset $\Gamma_\alpha \subset \Gamma$ such that I_α is independent of $(I_\beta, \beta \notin \Gamma_\alpha)$. We may then choose $J_{\beta\alpha} = I_\beta$ whenever $\beta \notin \Gamma_\alpha$, and construct $J_{\beta\alpha}$ for $\beta \in \Gamma_\alpha$ such that (8) holds. (This is always possible, at least if we enlarge the probability space.) For $\beta \in \Gamma_\alpha$ we may use the trivial estimate $p_\alpha \operatorname{E}|I_\beta - J_{\beta\alpha}| \le p_\alpha \operatorname{E}(I_\beta + J_{\beta\alpha}) = p_\alpha p_\beta + \operatorname{E} I_\alpha I_\beta$. This means that we can ignore the details of the construction of $J_{\beta\alpha}$; it is only the *existence* of a coupling that is important. We obtain from Theorem 2 the following result by Barbour and Eagleson (1984).

Theorem 4. *If Γ_α are as above, then*

$$d_{TV}\big(\mathcal{L}(W), \operatorname{Po}(\lambda)\big) \le (1 \wedge \lambda^{-1})\Big(\sum_{\alpha\in\Gamma} p_\alpha^2 + \sum_{\alpha\in\Gamma}\sum_{\beta\in\Gamma_\alpha\setminus\{\alpha\}} (p_\alpha p_\beta + \operatorname{E} I_\alpha I_\beta)\Big).$$

The standard situation where this theorem is applicable is when the family (I_α) is *dissociated*, meaning that the index set Γ is a collection $\{\alpha\}$ of sets $\alpha = \{\alpha_1, \dots, \alpha_k\}$ of cardinality k, for some $k \ge 2$, and that two sets of random variables $(I_\alpha, \alpha \in A)$ and $(I_\alpha, \alpha \in B)$ are independent whenever $(\bigcup_{\alpha\in A} \alpha) \cap (\bigcup_{\alpha\in B} \alpha) = \emptyset$; we may thus apply Theorem 4 with $\Gamma_\alpha = \{\beta : \alpha \cap \beta \ne \emptyset\}$.

The most common case of dissociated variables is that in which $I_\alpha = \phi_\alpha(Y_{\alpha_1}, \dots, Y_{\alpha_k})$, where the ϕ_α are symmetric 0–1 valued functions and the Y_i are independent random elements of some space $\mathcal{Y}$. A typical example is when $Y_1, \dots, Y_n$ are independent elements of some separable metric space and W is the number of pairs (Y_i, Y_j) with distance less than some fixed number (here $k = 2$ and Γ is the set of all pairs $\{i, j\}$, $1 \le i \le j \le n$).

4. Monotone couplings

The estimate in Theorem 2 can be significantly simplified if there exists a coupling that is *monotone* in the following sense.

Definition. The indicator random variables $(I_\alpha, \alpha \in \Gamma)$ are said to be *positively related* if for each α there exist random variables $(J_{\beta\alpha}, \beta \in \Gamma)$ such that (8) holds and $J_{\beta\alpha} \ge I_\beta$ whenever $\alpha \ne \beta$. The variables are *negatively related* if there exist such $J_{\beta\alpha}$ with $J_{\beta\alpha} \le I_\beta$ whenever $\alpha \ne \beta$.

Remark. If the indicators are positively (negatively) related, then they are necessarily positively (negatively) correlated, but the converse does not hold. Correlation is only a pairwise property, whereas being positively or negatively related depends on the whole joint distribution.

If the variables (I_α) are positively related, then, using (9), we have for $\beta \ne \alpha$,

$$p_\alpha \operatorname{E}|I_\beta - J_{\beta\alpha}| = p_\alpha \operatorname{E}(J_{\beta\alpha} - I_\beta) = \operatorname{E} I_\alpha I_\beta - p_\alpha p_\beta = \operatorname{Cov}(I_\alpha, I_\beta)$$

and thus

$$\begin{aligned}\sum_\alpha \sum_{\beta\ne\alpha} p_\alpha \operatorname{E}|I_\beta - J_{\beta\alpha}| &= \sum_\alpha\sum_\beta \operatorname{Cov}(I_\alpha, I_\beta) - \sum_\alpha \operatorname{Var}(I_\alpha) \\ &= \operatorname{Var}\Big(\sum_\alpha I_\alpha\Big) - \sum_\alpha (p_\alpha - p_\alpha^2) = \operatorname{Var} W - \lambda + \sum_\alpha p_\alpha^2.\end{aligned}$$

For negatively related variables we obtain the same result with opposite sign. Theorem 2 now yields:

Theorem 5. *If the variables I_α are positively related, then*

$$d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(\lambda)\big) \le (1 \wedge \lambda^{-1})\Big(\operatorname{Var} W - \lambda + 2\sum_\alpha p_\alpha^2\Big).$$

Theorem 6. *If the variables I_α are negatively related, then*

$$d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(\lambda)\big) \le (1 \wedge \lambda^{-1})(\lambda - \operatorname{Var} W).$$

These results are very attractive indeed, since the mean and variance coincide for a Poisson distribution so that it is reasonable to require that $|\operatorname{Var} W - \mathrm{E}\, W|$ is small as a first condition for Poisson approximation in some application. (Of course, counter examples can be constructed, but they do rarely occur in natural examples.) The theorems say that, in fact, this condition is sufficient when there exists a monotone coupling, provided that in the positively related case also $(1 \wedge \lambda^{-1})\sum_\alpha p_\alpha^2$ is small (a condition needed even in the independent case). Moreover, in the negatively related case, it may be shown that the estimate in Theorem 6 is off by at most logarithmic factors, and thus it is also necessary for good Poisson approximation that $|\operatorname{Var} W - \mathrm{E}\, W|$ is small. Hence, for a sequence W_n of sums of negatively related variables such that $\mathrm{E}\, W_n \to \lambda$, we have $W_n \xrightarrow{d} \mathrm{Po}(\lambda)$ if and only if also $\operatorname{Var} W_n \to \lambda$. (Similar results hold for sums of positively related variables under additional conditions.)

Note that the estimates in Theorems 5 and 6 involve only first and second moments of the variables I_α, which makes them easy to compute in most applications. The couplings $J_{\beta\alpha}$ do not appear in the estimates, it is only the *existence* of a monotone coupling that matters.

The question of existence of a monotone coupling may be put in a more general context, namely as the question whether there exist random variables I and J, with values in a given partially ordered set S and with given distributions, such that $I \le J$ a.s.; in our case, S is the finite set $\{0,1\}^{\Gamma\setminus\alpha}$ with the natural ordering $(x_1, \dots, x_n) \le (y_1, \dots, y_n)$ if $x_i \le y_i$ for all i. In general, such a coupling of I and J exists if and only if

$$\mathrm{E}\,\phi(I) \le \mathrm{E}\,\phi(J)$$

for every increasing indicator function ϕ on S, see e.g. Liggett (1985), page 72. In our case, this leads to the following result.

Proposition 1. *The variables I_α are positively (negatively) related if and only if I_α and $\phi(I_\beta, \beta \ne \alpha)$ are positively (negatively) correlated for every α and every increasing indicator function ϕ.*

This yields, by the FKG inequality, the following important case where Theorem 5 is applicable.

Proposition 2. *Let $(I_\alpha, \alpha \in \Gamma)$ be increasing functions of independent random variables $X_1, \dots, X_n$. Then the random variables I_α are positively related.*

More generally, indicator variables that are *associated* in the sense of Esary, Proschan and Walkup (1967) are positively related. Similarly, indicator variables that are *negatively associated* in the sense of Joag-Dev and Proschan (1983) are negatively related. General results on (negatively) associated variables may thus be used to show that certain families of indicator variables I_α are positively or negatively related, which enables us to apply Theorem 5 or 6 in some applications.

5. Examples

We end by giving some examples of couplings that are useful in applications of the theorems above. For the resulting Poisson approximation error bounds, and conditions on the parameters that make the bounds small, we refer the interested reader to Barbour, Holst and Janson (1992).

Occupancy problems. Let r balls be thrown independently of each other into a number of boxes, with probability p_α of hitting box α. Let W be the number of empty boxes. Then $W = \sum_\alpha I_\alpha$, where $I_\alpha = I(\text{box } \alpha \text{ is empty})$. In order to define a coupling $(J_{\alpha\beta})$, we remove all balls from box β and redistribute them according to the original distribution, repeating until box β is empty; we then define $J_{\alpha\beta} = I(\text{box } \alpha \text{ is empty})$. Obviously, $J_{\alpha\beta} \le I_\alpha$, so the coupling is monotone and the variables I_α are negatively related. Thus Theorem 6 applies.

Similar arguments work for other occupancy problems, giving for example a Poisson approximation result for the hypergeometric distribution.

Random graphs. Consider the complete graph K_n with n vertices and $\binom{n}{2}$ edges. Keep each edge with probability p independently of the other edges, and delete it otherwise; this gives a random graph $K_{n,p}$. Let $G \subset K_n$ be a given graph with $e(G)$ edges and $v(G)$ vertices, and let Γ be the set of all copies of G in K_n, i.e. all subgraphs of K_n isomorphic to G. The number of subgraphs of $K_{n,p}$ isomorphic to G can be written $W = \sum_{\alpha\in\Gamma} I_\alpha$, where $I_\alpha = I(\alpha \subset K_{n,p})$.

The variables I_α are increasing functions of the edge indicators, which by definition are independent. Hence Proposition 2 shows that the variables I_α are positively related, and Theorem 5 applies. This can also be shown directly, using the explicit coupling $J_{\beta\alpha} = I\big(\beta \subset (K_{n,p} \cup \alpha)\big) \ge I_\beta$, where the union of two graphs combines their vertex and edge sets.

Matching problems. We consider two examples; several others may be treated similarly. We prefer to use the indices i, j instead of α, β.

Treize. The game 'Treize' was studied already by de Montmort (1713). In this game 13 cards are drawn at random one at a time from an ordinary deck of 52 cards having 4 sets of face values $1, \dots, 13$. A match occurs if a card with face value i appears in the ith drawing. The original problem is to find the probability that there is no match.

We can write the number of matches as $W = \sum_1^{13} I_i$, where $I_i = I(\text{match at drawing } i)$. It is easy to construct a monotone coupling with $J_{ij} \ge I_i$, namely by swapping the card obtained in the jth drawing with a card with face value j chosen at random. Hence $I_1, \dots, I_{13}$ are positively related and we can apply Theorem 5 to obtain an estimate for $d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(1)\big)$.

The ménage problem. The classical ménage problem asks for the number of seatings of n couples at a circular table, with men and women alternating, so that no one sits next to his or her partner. More generally, we can ask for the probability that a random seating gives exactly k couples sitting together.

We number the places around the table from 1 to $2n$, and let $I_i = 1$ if a couple occupies seats i and $i+1$ (counting modulo $2n$). Then W, the number of couples sitting next to each other, can be expressed as $\sum_1^{2n} I_i$. A coupling may be constructed by interchanging the person sitting on seat j with the partner of the person on seat $j+1$ and defining $J_{ij} = 1$ if a couple then occupies seats i and $i+1$. For this coupling, $J_{ij} = 0 \le I_i$ if $i = j \pm 1$, but $J_{ij} \ge I_i$ for all other i. Hence, the coupling is not monotone, but of a mixed kind where monotonicity holds for the individual J_{ij}, although in different directions. Theorem 2 gives a good estimate for $d_{TV}\big(\mathcal{L}(W), \mathrm{Po}(2)\big)$.

References

A.D. Barbour and G.K. Eagleson (1984). Poisson convergence for dissociated statistics. *Journal of the Royal Statistical Society*, **B 46**, 397–402.

A.D. Barbour and P. Hall (1984). On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, **95**, 473–80.

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

L.H.Y. Chen (1975). Poisson approximation for dependent trials. *Annals of Probability*, **3**, 534–45.

J.D. Esary, F. Proschan, and D.W. Walkup (1967). Association of random variables, with applications. *Annals of Mathematical Statistics*, **38**, 1466–74.

K. Joag-Dev and F. Proschan (1983). Negative association of random variables, with applications. *Annals of Statistics*, **11**, 286–95.

P.R. de Montmort (1713). *Essay d'analyse sur les jeux de hazard,* (2nd edn). Jacques Quillau, Paris.

T.M. Liggett (1985). *Interacting particle systems.* Springer, New York.

C. Stein (1986). *Approximate computation of expectations.* Institute of Mathematical Statistics Lecture Notes – Monograph Series, Vol. **7**. Hayward, California.

Svante Janson, Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden

E-mail address: `svante.janson@math.uu.se`