Offprints from: **New Perspectives in Theoretical and Applied Statistics**
Edited by Dr. Madan L. Puri, Dr. Jose Perez Vilaplana and Dr. Wolfgang Wertz

Semigroups and Poisson Approximation

Paul Deheuvels
Institute de Statisique de Paris
Université de Paris VI
Paris, France

D. Pfeifer
Institut für Statistik
RWTH
Aachen, Federal Republic of Germany

SUMMARY

In this Chapter we extend previous work of the authors on *Poisson approximation* for (general) independent Bernoulli summands with respect to the total variation distance, without imposing any conditions on the underlying parameters. This enables one to study also the case of unbounded means, without asymptotic uniform "smallness" of the individual summands, provided that the variance increases with the same rate as the mean. An important practical situation in which such an asymptotic behavior occurs is described by Ross's Markov chain model for the *simplex algorithm in linear programming*, which will be discussed as an example of possible application.

The main tool for the derivation of the results is the same Poisson *convolution semigroup* approach as used formerly, allowing again for a

D. Pfeifer's research was supported in part by AFOSR Contract No. F49620 85 C 0144 while visiting the Center for Stochastic Processes, U.N.C. at Chapel Hill, U.S.A.

treatment of optimal choice problems for the Poisson parameter, in different asymptotic settings.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with means $p_i = P(X_i = 1) = 1 - P(X_i = 0)$, $0 \le p_i \le 1$, $i = 1, 2, \ldots, n$, and $T(\mu)$ a Poisson random variable with mean $\mu > 0$. We are concerned with the *approximation of the partial sums* $S_n = \sum_{k=1}^n X_k$ by $T(\mu)$, where the goodness of fit is measured by the *total variation distance*

$$d(S_n, T(\mu)) = \sup_{M \subseteq \mathbf{Z}^+} |P(S_n \in M) - P(T(\mu) \in M)|$$

$$= \frac{1}{2} \sum_{k=0}^{\infty} |P(S_n = k) - P(T(\mu) = k)| \tag{1.1}$$

Estimations and asymptotic expansions for the distance d have been given by different authors, for instance, LeCam (1960), Kerstan (1964), Chen (1974, 1975), Serfling (1975, 1978), Barbour and Hall (1984), and most recently by Deheuvels and Pfeifer (1986). Whereas in earlier approaches, mainly the choices $\mu = \sum_{k=1}^n p_k$ (LeCam, Kerstan, Chen, and Barbour and Hall) or $\mu = -\sum_{k=1}^n \log(1 - p_k)$ (Serfling) have been considered, which possess certain optimality characteristics in that they asymptotically minimize the distance $d(S_n, T(\mu))$ if $\sum_{k=1}^n p_k \gg 1$ and $\sum_{k=1}^n p_k^2 \ll 1$ (Deheuvels and Pfeifer, 1986) or $\sum_{k=1}^n p_k \ll 1$, resp. (Serfling, 1978; Deheuvels and Pfeifer, 1986), it is more generally of importance to consider also the cases where

$$\mu = \sum_{k=1}^n p_k + \gamma \sum_{k=1}^n p_k^2 \equiv \mu(\gamma)$$

with $0 \le \gamma \le \frac{1}{2}$, where $\gamma = \frac{1}{2}$ (asymptotically)

corresponds to Serfling's approach. This is because for large n, if $\sum_{k=1}^n p_k$ is close to some positive value a and $\sum_{k=1}^n p_k^2 \ll 1$, then there always exists some $\gamma = \gamma(a) \in [0, \frac{1}{2}]$ such that the choice $\mu = \mu(\gamma(a))$ asymptotically minimizes $d(S_n, T(\mu))$ (Deheuvels and Pfeifer, 1986). For instance, in the range $0 < a < 2$, we have

$$\text{if } 0 < a \le 1: \qquad \gamma(a) = \tfrac{1}{2},\ d(S_n, T(\mu(\gamma(a)))) \approx \tfrac{1}{2} e^{-a} \sum_{k=1}^{n} p_k^2$$

$$\text{if } 1 < a \le \sqrt{2}: \qquad \gamma(a) = \tfrac{1}{2},\ d(S_n, T(\mu(\gamma(a)))) \approx \tfrac{1}{2} a e^{-a} \sum_{k=1}^{n} p_k^2$$

$$\text{if } \sqrt{2} < a \le \sqrt[3]{6}: \qquad \gamma(a) = \frac{1}{2} - \frac{3}{2a} \frac{2-a}{3-a},$$

$$d(S_n, T(\mu(\gamma(a)))) \approx \frac{1}{2}\left[a + \left(1 - \frac{a^2}{2}\right)\frac{3}{a}\frac{2-a}{3-a}\right] e^{-a} \sum_{k=1}^{n} p_k^2$$

$$= \frac{1}{2}\left[\frac{a^2}{2} + \left(1 - \frac{a^3}{6}\right)\frac{3}{a}\frac{2-a}{3-a}\right] e^{-a} \sum_{k=1}^{n} p_k^2$$

$$\text{if } \sqrt[3]{6} < a < 2: \qquad \gamma(a) = 0,$$

$$d(S_n, T(\mu(\gamma(a)))) \approx \frac{1}{2}\left[\frac{a^2}{2} + \left(1 - \frac{a^3}{6}\right)\right] e^{-a} \sum_{k=1}^{n} p_k^2$$

This follows from Theorem 1.3 and Section 3 in Deheuvels and Pfeifer (1986) [note that for $0 < a \le \sqrt{2}$ in (1.2), we could as well have used Serfling's choice for μ].

Whereas in all former investigations the condition $\sum_{k=1}^{n} p_k^2 \ll 1$ is (more or less) necessary to obtain asymptotically sharp results, we should also like to deal with more general cases in which $\sum_{k=1}^{n} p_k^2$ may be arbitrary. It turns out again that the semigroup technique as developed in Deheuvels and Pfeifer (1986) can be fruitfully applied, as will be worked out in Section 2.

2. THE SEMIGROUP TECHNIQUE

In order to facilitate the understanding for readers who are not familiar with the language of functional analysis and operator theory, we shall first state the main results, without appealing to the concept of semigroups.

Theorem 2.1. Suppose that $X_1, \dots, X_n$ are independent Bernoulli random variables with means $p_i = P(X_i = 1) \in [0,1]$, $i = 1,2,\dots,n$ and $T(\mu)$ is a Poisson random variable with mean $\mu > 0$. For abbreviation, let

$$\lambda_j = \sum_{k=1}^{n} p_k^j, \quad j = 1,2,3 \quad \text{and} \quad \mu(\gamma) = \lambda_1 + \gamma\lambda_2 \quad \text{for } 0 \le \gamma \le \tfrac{1}{2}$$

Then

$$d(S_n, T(\mu(\gamma))) = \tfrac{1}{2}\lambda_2 K(\gamma, \lambda_1) + r_n \tag{2.1}$$

where

$$K(\gamma, \lambda) = \left\{ \frac{\lambda^{\delta-1}(\delta - (1-2\gamma)\lambda)}{\delta!} + \frac{\lambda^{\eta-1}((1-2\gamma)\lambda - \eta)}{\eta!} \right\} e^{-\lambda} \qquad (\lambda > 0) \tag{2.2}$$

with

$$\delta = \left[\!\left[(1-\gamma)\lambda + \tfrac{1}{2} + \sqrt{\gamma^2\lambda^2 + (1-\gamma)\lambda + \tfrac{1}{4}} \right]\!\right]$$

$$\eta = \left[\!\left[(1-\gamma)\lambda + \tfrac{1}{2} - \sqrt{\gamma^2\lambda^2 + (1-\gamma)\lambda + \tfrac{1}{4}} \right]\!\right]$$

where $[\![\ ]\!]$ means integer part and

$$|r_n| \le e^{22\lambda_2}\left[\frac{18\lambda_3}{\sqrt{\lambda_1^3}} + 3168\left(\frac{\lambda_2}{\lambda_1}\right)^2 \right] + \gamma^2\frac{\lambda_2^2}{\lambda_1} \tag{2.3}$$

Especially,

$$K(0, \lambda) = \left\{ \frac{\lambda^{\delta-1}(\delta - \lambda)}{\delta!} + \frac{\lambda^{\eta-1}(\lambda - \eta)}{\eta!} \right\} e^{-\lambda} \approx \frac{2}{\sqrt{2\pi e}}\frac{1}{\lambda} \quad \text{for } \lambda \to \infty$$

$$K(\tfrac{1}{2}, \lambda) = \frac{\lambda^{[\![\lambda]\!]}}{[\![\lambda]\!]!} e^{-\lambda} \approx \frac{1}{\sqrt{2\pi\lambda}} \quad \text{for } \lambda \to \infty \tag{2.4}$$

Theorem 2.2. Under the conditions of Theorem 2.1, we have

1. If $\lambda_2 \ll 1$, then also $|r_n/\lambda_2| \ll 1$, and

$$d(S_n, T(\mu(\gamma))) \approx \tfrac{1}{2}\lambda_2 K(\gamma, \lambda_1) \tag{2.5}$$

   which can be minimized (asymptotically) by choosing the corresponding value for γ that minimizes (2.2) for $\lambda = \lambda_1$. Here, the (asymptotically) optimal value for γ for $\lambda_1 \ll 1$ is $\gamma = \frac{1}{2}$, whereas the (asymptotically) optimal value for $\lambda_1 \gg 1$ is $\gamma = 0$.

2. If λ_2 remains bounded and $\lambda_1 \gg 1$, then $\gamma = 0$ is asymptotically optimal, and

$$d(S_n, T(\mu(\gamma))) = d(S_n, T(\lambda_1)) \approx \frac{1}{\sqrt{2\pi e}} \frac{\lambda_2}{\lambda_1} \tag{2.6}$$

Note that the condition under 2 implies that $\mathrm{Var}(S_n) = \lambda_1 - \lambda_2 \approx \lambda_1 = E(S_n)$ for $\lambda_1 \gg 1$, and that by Theorem 2.1, no Poisson convergence takes place unless $\lambda_2/\lambda_1 \to 0$, that is, for λ_2 bounded (and bounded away from 0), $\lambda_1 \gg 1$ is a necessary condition for Poisson convergence.

Theorems 2.1 and 2.2, which extend results of Deheuvels and Pfeifer (1985), will be proved by the following operator semigroup technique, paralleling (but improving) LeCam's (1960) approach.

Consider the Banach space l^1 of all absolutely summable sequences $f = (f(0), f(1), \dots)$ with norm

$$f = \sum_{k=0}^{\infty} |f(k)| \tag{2.7}$$

The convolution $f * g$ for $f, g \in l^1$ is defined by

$$f * g(n) = \sum_{k=1}^{n} f(k) g(n-k), \qquad n \geq 0 \tag{2.8}$$

being again an element of l^1 with

$$\|f * g\| \leq \|f\| \, \|g\| \tag{2.9}$$

Henceforth we shall think of discrete probability measures ν over $\mathbf{Z}^+$ as elements $(\nu(\{0\}), \nu(\{1\}), \dots) \in l^1$. Specifically, ε_k will denote the Dirac measure concentrated in $k \in \mathbf{Z}^+$. The *infinitesimal generator* of the Poisson convolution semigroup is defined by

$$Ag(n) = \begin{cases} g(n-1) - g(n), & n \geq 1 \\ -g(0), & n = 0 \end{cases} \quad g \in l^1 \tag{2.10}$$

such that

$$e^{tA}g = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k g = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} \varepsilon_k * g = \Pi(t) * g, \qquad t \geq 0, \quad g \in l^1 \tag{2.11}$$

where $\Pi(t)$ denotes the Poisson distribution with mean t. Also,

$$(I + p_k A)g = ((1 - p_k)\varepsilon_0 + p_k \varepsilon_1) * g = B(p_k) * g, \qquad g \in l^1 \tag{2.12}$$

where $B(p_k)$ denotes the binomial distribution with mean p_k over $\{0, 1\}$ and I denotes the identity operator. Because of (1.1), we can now rewrite

$$d(S_n, T(\mu)) = \frac{1}{2} \left\| \prod_{k=1}^{n} (I + p_k A) g_0 - e^{\mu A} g_0 \right\| \tag{2.13}$$

where $g_0 = (1, 0, 0, \dots) \in l^1$. Since $I + p_k A$ is the first-order Taylor expansion of the semigroup e^{tA} for $t = p_k$ [see, e.g., Butzer and Berens (1967)], it seems natural that the right-hand side of (2.13) could be attacked by general operator semigroup theory, in combination with the well-known fact that for commuting operators $U_1, \dots, U_n, V_1, \dots, V_n$ we always have

$$\prod_{k=1}^{n} U_k - \prod_{k=1}^{n} V_k = \sum_{k=1}^{n} U_{k+1} \cdots U_n (U_k - V_k) V_1 \cdots V_{k-1} \tag{2.14}$$

as can be easily proved, for example, by induction.

Consider first the case $\mu = \lambda_1$. From what has been said above, we should have, in some sense,

$$e^{\lambda_1 A} - \prod_{k=1}^{n} (I + p_k A) \simeq \frac{1}{2} \sum_{k=1}^{n} p_k^2 e^{\lambda_1 A} A^2 \tag{2.15}$$

A more precise statement is given below.

Theorem 2.3. We have, for $\lambda_1^{(k)} = \sum_{i \neq k} p_i$,

$$\left\| e^{\lambda_1 A} - \prod_{k=1}^{n} (I + p_k A) - \frac{1}{2} \sum_{k=1}^{n} p_k^2 e^{\lambda_1^{(k)} A} A^2 \right\|$$

$$\leq e^{22\lambda_2} \left[\tfrac{3}{2}\lambda_3 \| e^{\lambda_1 A} A^3 \| + 99\lambda_2^2 \| e^{\lambda_1 A} A^4 \| \right] \tag{2.16}$$

Proof. From (2.14) we obtain, as in the proof of Theorem 2.1 in Deheuvels and Pfeifer (1986),

$$\left\| e^{\lambda_1 A} - \prod_{k=1}^{n} (I + p_k A) - \frac{1}{2} \sum_{k=1}^{n} p_k^2 e^{\lambda_1^{(k)} A} A^2 \right\|$$

$$\leq \sum_{k=1}^{n} \left[\prod_{i=1}^{k-1} \| e^{-p_i A}(I + p_i A) \| \left\| e^{\lambda_1^{(k)} A} \left(e^{p_k A} - \left(I + p_k A + \frac{p_k^2}{2} A^2 \right) \right) \right\| \right.$$

$$\left. + \frac{p_k^2}{2} \left\| e^{\lambda_1^{(k)} A} A^2 \left\{ \prod_{i=1}^{k-1} e^{-p_i A}(I + p_i A) - I \right\} \right\| \right]$$

But

$$e^{-p_i A}(I + p_i A) = (I + p_i A)\left\{ I - p_i A + \int_0^{p_i} (p_i - u)\, e^{-uA} A^2\, du \right\}$$

$$= I - p_i^2 A^2 + (I + p_i A) \int_0^{p_i} (p_i - u)\, e^{-uA} A^2\, du \tag{2.17}$$

hence

$$\| e^{-p_i A}(I + p_i A) \| \leq 1 + p_i^2 \|A\|^2 + \tfrac{1}{2} p_i^2\, e^{p_i \|A\|} \|A\|^2 \tag{2.18}$$

where $\|A\| = 2$, such that, since $p_i \leq 1$, the left-hand side of (2.16) can be estimated by

$$e^{22\lambda_2} \sum_{k=1}^{n} \left[\frac{p_k^3}{6} \| e^{\lambda_1^{(k)} A} A^3 \| + 11 p_k^2 \sum_{i=1}^{n} \| e^{\lambda_1^{(k)} A} A^4 \| p_i^2 \right] \tag{2.19}$$

But

$$\| e^{\lambda_1^{(k)} A} A^j \| \leq e^{-p_k A} \| \, \| e^{\lambda_1 A} A^j \| \leq 9 \| e^{\lambda_1 A} A^j \|, \qquad j \geq 1$$

Hence the result. ∎

In a similar manner, the following result can be obtained.

Theorem 2.4. We have

$$\left\| e^{\mu(\gamma)A} - \prod_{k=1}^{n} (I + p_k A) \right\| = \tfrac{1}{2}\lambda_2 \|(2\gamma A + A^2)\, e^{\lambda_1 A}\| + R_n \quad (2.20)$$

with

$$|R_n| \le e^{22\lambda_2}\left[6\lambda_3 \|e^{\lambda_1 A}A^3\| + 99\lambda_2^2 \|e^{\lambda_1 A}A^4\|\right] + 2\gamma^2 \frac{\lambda_2^2}{\lambda_1} \quad (2.21)$$

Proof. The proof is analogous to the proof of Theorem 2.1 in Deheuvels and Pfeifer (1986).

From relation (3.19) in Pfeifer (1985a) we see that

$$\|e^{\lambda_1 A}A^3\| \le \min\left(8, 6/\sqrt{\lambda_1^3}\right) \quad (2.22)$$

but also

$$\|e^{\lambda_1 A}A^4\| \le \|e^{(1/2)\lambda_1 A}A^2\|^2 \le 16 \min\left(1, 4/\lambda_1^2\right) \quad (2.23)$$

Theorem 2.1 now follows from Theorem 2.4 and Theorem 1.2 in Deheuvels and Pfeifer (1986). Theorem 2.2 follows from Theorem 2.1 and the fact that for $\gamma > 0$, we have $\|(2\gamma A + A^2)\, e^{\lambda_1 A}\| = O(1/\sqrt{\lambda_1})$, but $\|A^2 e^{\lambda_1 A}\| = O(1/\lambda_1)$ for $\lambda_1 \gg 1$.

REMARK 2.1. An application of the above technique directly to the norm term

$$\left\| e^{\lambda_1 A} - \prod_{k=1}^{n} (I + p_k A) \right\| \text{ yields}$$

$$d(S_n, T(\lambda_1)) \le e^{22\lambda_2}\left\{ \tfrac{1}{2}\lambda_2 K(0, \lambda_1) + 27 \frac{\lambda_3}{\sqrt{\lambda_1^3}} \right\} \quad (2.24)$$

which is considerably sharper than the corresponding bound in Barbour and Hall (1984) for moderate and large values of λ_1 when $\lambda_2 \ll 1$.

3. APPLICATION

Here we shall consider Ross's (1983) Markov chain model for the simplex algorithm in linear programming and the *average-case analysis* corresponding to it. Characteristically, the simplex algorithm moves through the n

(say) extreme points of the feasible region defined by the linear constraints in such a way that at each step of the algorithm, all points visited and those that cannot increase the value of the goal function are excluded from further consideration. If we assume that the extreme points are ranked such that 1 corresponds to the optimal and n to the worst point, then the ranks visited can be considered as a homogeneous Markov chain with uniform initial distribution, absorbing state 1 and uniform transition probabilities for the remaining higher ranks [see Ross (1983, Chapter 4)]. If we assume that the random variable X_k takes the value 1 if rank k was visited and 0 otherwise ($k = 1, \dots, n$), then it is possible to show that by the structure of the underlying Markov chain, $X_1, \dots, X_n$ are independent with $p_k = 1/k$ (note that 1 is absorbing, hence $p_1 = 1$, and the first selected extreme point is the worst with probability $1/n$). Then S_n denotes the number of steps required for termination. Here,

$$\lambda_1 \approx \log n + C, \qquad \lambda_2 \sim \frac{\pi^2}{6} \quad \text{for } n \to \infty \tag{3.1}$$

(where $C = 0.5772$ denotes Euler's constant). Hence by Theorem 2.2, S_n is asymptotically Poisson with mean $\lambda_1 \approx \log n$ and

$$d(S_n, T(\lambda_1)) = \frac{\pi^2}{6\sqrt{2\pi e}} \frac{1}{\log n} + O\left((\log n)^{-3/2}\right) \quad \text{for } n \to \infty \tag{3.2}$$

Note that by methods similar to those above, we obtain only

$$d(S_n, T(\log n)) = \frac{C}{\sqrt{2\pi \log n}} + O\left((\log n)^{-1}\right) \quad \text{for } n \to \infty \tag{3.3}$$

which means that S_n is indeed (asymptotically) closer to a Poisson random variable with mean λ_1 than to a Poisson random variable with mean $\log n$.

4. FURTHER COMMENTS

It seems reasonable that the generalizations presented in this chapter should also apply to distance measures other than total variation, for instance, the Kolmogorov distance or Wasserstein distances [see, e.g., Deheuvels and Pfeifer (1985)], since these can also be treated in the operator semigroup framework, and Theorems 2.3 and 2.4 do not heavily depend on the underlying semigroup or the underlying Banach space.

A generalization to Poisson point process approximation by the semigroup approach seems also to be tractable quite well (Pfeifer, 1985b), although in this area we are still far from sufficient results.

Note that the constants which appear in our theorems can be precised. For instance the constant 22 in (2.24) originates from (2.18) and the crude inequalitites

$$1 + p^2\|A\|^2 + \tfrac{1}{2}p^2 e^{p\|A\|}\|A\|^2 \leq 1 + (4 + 2e^2)p^2 \leq 1 + 22p^2 \leq e^{22p^2}.$$

If $p \to 0$, it follows that 22 can be replaced by any constant $c > 6$, for asymptotic validity.

REFERENCES

Barbour, A. D. and Hall, P. (1984). On the rate of Poisson convergence. *Math. Prob. Camb. Phil. Soc.* **95**, 473–480.

Butzer, H. L. and Berens, H. (1967). *Semi-Groups of Operators and Approximation*, New York: Springer.

Chen, L. H. Y. (1974). On the convergence of Poisson binomial to Poisson distributions. *Ann. Prob.* **2**, 178–180.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3**, 534–545.

Deheuvels, P. and Pfeifer, D. (1985). A semigroup approach to Poisson approximation. II Tech. Report, L.S.T.A., Université Paris VI.

Deheuvels, P. and Pfeifer, D. (1986). A semigroup approach to Poisson approximation. *Ann. Prob.* **14**, 663–676.

Kerstan, J. (1964). Verallgemeinerung eines Satzes von Prochorow und LeCam. *Z. Wahrscheinlichkeitsth. verw. Geb.* **2**, 173–179.

LeCam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* **10**, 1181–1197.

Pfeifer, D. (1985a). On the distance between mixed Poisson and Poisson distributions. Center for Stochastic Processes, University of North Carolina at Chapel Hill, Technical Report No. 115.

Pfeifer, D. (1985b). A semigroup approach to Poisson point process approximation. Personal communication.

Ross, S. M. (1983). *Stochastic Processes*. New York: Wiley.

Serfling, R. J. (1975). A general Poisson approximation theorem. *Ann. Prob.* **3**, 726–731.

Serfling, R. J. (1978). Some elementary results on Poisson approximation in a sequence of Bernoulli trials. *SIAM Rev.* **20**, 567–679.