The Annals of Probability
1992, Vol. 20, No. 3, 1504–1527

THE STEIN–CHEN METHOD, POINT PROCESSES AND COMPENSATORS[1]

By A. D. Barbour and Timothy C. Brown

Universität Zürich and University of Western Australia

The paper gives bounds for the accuracy of Poisson approximation to the distribution of the number of points in a point process. There are two principal bounds, one in terms of reduced Palm probabilities for general point processes, and one involving compensators for point processes on the line. The latter bound is frequently sharper than the previously used compensator bounds when the expected number of points is large, and examples show that little improvement is possible without changing the form of the bound.

1. Introduction. There are two methods commonly used for obtaining bounds on the total variation accuracy of approximation of the distribution of a sum N_n of dependent indicators $I_1, \dots, I_n$ by a Poisson distribution. The first, initiated by Freedman (1974) and Serfling (1975), is to suppose that there is an associated filtration $(\mathscr{F}_i)_{0 \le i \le n}$ and that the conditional probabilities $p_i = P[I_i = 1 | \mathscr{F}_{i-1}]$ are known for each i. This structure leads naturally to bounds expressed in terms of martingale characteristics and, especially, of the compensator A of the associated point process: $A(i) = \sum_{j=1}^{i} p_j$.

A typical result is that of Brown (1983) [but see also Brown (1982), Valkeila (1982, 1984), Kabanov and Liptser (1983) and Kabanov, Liptser and Shiryayev 1986)]:

$$d_{\mathrm{TV}}(\mathscr{L}N_n, \mathrm{Poisson}_\mu) \le E|A(n) - \mu| + E\left\{ \sum_{j \le n} p_j^2 \right\}, \tag{1.1}$$

where d_{TV} denotes total variation distance, and the result carries over to counts of numbers of points in a point process on the line. In particular, the compensator bounds may be considered as perturbations of the celebrated result of Watanabe, that a point process is a Poisson process if it has a deterministic continuous compensator.

The second method, the Stein–Chen method, was developed by Chen (1975). Here, the dependence structures most easily exploited are those in which any given indicator is only strongly dependent on a few others or in which the dependence between indicators is essentially symmetrical. With dependence of either form, it may be possible to construct an associated filtration, but it is rarely natural to do so, and the bound thus derived using (1.1) may suffer in

Received June 1990; revised March 1991.
[1]Work partially supported by Schweiz. Nationalfonds Grant No. 21-25579.88.
AMS 1980 *subject classifications*. Primary 60G55; secondary 60E15, 60G44, 60J75.
Key words and phrases. Stein–Chen method, point process, compensator, Palm probability, martingale, Poisson approximation.

consequence. A good example is given by sums of dissociated indicators, where the estimate of Barbour and Eagleson (1983) using the Stein–Chen method improves significantly upon the bound obtained using compensator methods in Brown and Silverman (1979). However, even for independent indicators, taking $\mu = \sum_{j=1}^{n} p_j$ in (1.1) only gives a bound of $\sum_{j=1}^{n} p_j^2$, whereas the Stein–Chen method gives $(1 \wedge \mu^{-1})\sum_{j=1}^{n} p_j^2$ [Barbour and Hall (1984)], which is essentially sharp. The crucial point is that for large means the bound remains small if the maximum probability is small.

The aim of the paper is to combine the two methods in such a way that the extra precision of the Stein–Chen approach can be achieved for point processes, especially taking account of the fact that point processes are often defined by their compensators. There are two main results. The first, Theorem 3.1, is in essence a generalisation of Theorem 2.1 of Barbour and Holst (1989) to a point process N over a general carrier space; it expresses the total variation distance between $\mathscr{L}N(B)$ and $\text{Poisson}_{\nu(B)}$ as an average, weighted according to the mean measure ν over B, of the Wasserstein distance between $\mathscr{L}N(B)$ and the reduced Palm distribution of $N(B)$ given a point at $x \in B$. Note that, when N is not simple, the result goes beyond that of Barbour and Holst and is better than just admitting failure for all realisations of N which contain multiple points in B. The theorem requires no filtration, but computation of the Wasserstein distance may prove difficult, if no inspired coupling can be found.

The second result, Theorem 3.7, presupposes a point process N on the line, and the bound consists of three terms. Two of these are those of (1.1), multiplied by factors which become small if μ, which is usually close to EN_t, is large. The cost of this improvement lies in the third term, which again requires the computation of an average of Wasserstein distances, but now only involving the *future* development of the process beyond time s. As a result, the third term is more easily estimated than the bound in Theorem 3.1.

The remainder of the paper consists of examples illustrating the significance of the various terms in the bounds, as well as techniques useful for estimating the Wasserstein distances. The bounds are computed for point processes with conditionally independent increments, and in particular for Cox processes, for the Markov point process and for a self-compensating Bernoulli process. The examples illustrate that the improvement in the first two terms of Theorem 3.7 as compared with (1.1) is the best possible, in the sense that the new factors introduced have the right order of magnitude for large μ and that the third term also cannot be improved by the introduction of a better μ-dependent factor.

2. Notation and definitions. The most general result (Theorem 3.1) concerns a locally finite point process N whose carrier space $\mathbb{X}$ is assumed to be locally compact, second countable and Hausdorff. Informally this process is just a random countable collection of points on $\mathbb{X}$ and, for a Borel set B of $\mathbb{X}$, $N(B)$ counts the number of points in B. For general definitions and results about such processes, see Kallenberg (1976). The point process is *simple*, if, for ω outside a P-null set E, for all x, we have $N(\{x\}, \omega) = 0$ or 1. We always assume that the first moment measure ν of N is also locally finite, that is, finite on bounded Borel sets. Of particular importance here is the Palm point process N^x whose distribution, when N is simple, is that of N conditional on there being a point at $x \in \mathbb{X}$. Formally, let $\mathfrak{N}(\mathbb{X})$ denote the space of point process measures, let $f\colon \mathbb{X} \times \mathfrak{N}(\mathbb{X}) \to [0, \infty)$ be measurable and let B be a Borel set of $\mathbb{X}$. Then, by Lemma 10.1 of Kallenberg (1976),

$$E\left\{\int_B f(x, N)N(dx)\right\} = \int_B E\{f(x, N^x)\}\nu(dx) \tag{2.1}$$

(a special case of this identity actually defines the distribution of N^x as a Radon–Nikodym derivative). We depart from the notation of Kallenberg in using N^x, rather than N_x, for the Palm process because for a point process on the half-line N_x is commonly used to mean $N(0, x]$, a convention that we follow. It is sometimes convenient to remove an atom at x ($\in \mathbb{X}$) from the realisation of N^x by considering the process $N^x - \delta_x$, called the *reduced Palm process*, where δ_x is the measure which attributes 1 to the singleton $\{x\}$ and 0 to all sets not containing x. Lemma 10.2 of Kallenberg (1976) then gives

$$E\left\{\int_B f(x, N - \delta_x)N(dx)\right\} = \int_B E\{f(x, N^x - \delta_x)\nu(dx)\}. \tag{2.2}$$

Equation (2.2) is particularly useful when N is simple, for the reduced Palm process then describes the behaviour of the *rest* of the process conditional on there being a point at x.

The second main result concerns simple point processes whose carrier space is $(0, \infty)$. For such processes, there is a natural ordering of the points and this ordering is often important in specifying the probability law. Hence, we frequently use results from the French structural theory of stochastic processes indexed by $(0, \infty)$, referring to the two volumes by Dellacherie and Meyer (1978) and (1982). In particular, we always have a history, that is, a right-continuous, increasing set of σ-fields $\{\mathscr{F}_t\}_{t \ge 0}$, in the background. For the same reasons as in Jacod (1975), it is neither convenient nor necessary to assume that the history is complete with null sets. A stochastic process $\{X_t\}$ is normally assumed to be adapted (X_t is $\mathscr{F}_t$-measurable) unless it is stated that it is *raw*. We adopt the usual convention that random variables which are *a.s.* equal are declared equal and also that processes are equal if the projection onto Ω of the set on which they are unequal is contained in a set of probability zero, that is, that the set of inequality of the processes is *evanescent*. Of considerable importance to us are the *optional* and *previsible* projections of a raw process X, whose definitions are given in Dellacherie and Meyer [(1982), page 103]. The optional projection is denoted ${}^oX$ and the previsible projection is denoted ${}^pX$.

A point process whose carrier space is $(0, \infty)$ can be considered as a stochastic process $\{N_t\}_{t \ge 0}$ by setting $N_0 = 0$ and $N_t = N(0, t]$. An equivalent way to define such a process N is to suppose that $0 < T_1 < T_2 < \cdots$ is an unbounded sequence of stopping times. If the stochastic process N is defined for

$t \geq 0$ by

$$N_t = \sum_{n=0}^{\infty} I[T_n \leq t],$$

then N is a simple point process. If $Z_1, Z_2, \ldots$ is a sequence of random variables such that $Z_n \in \mathscr{F}(T_n)$ for all n and

$$N_t = \sum_{n=0}^{\infty} Z_n I[T_n \leq t],$$

then we say that N is a *jump process*, and if the Z's take nonnegative integer values, then N is a general point process. In this case, assuming as above that $\nu_t = \nu(0, t]$ is finite, the *compensator* of N is its dual previsible projection, the unique process A with increasing right-continuous paths and $A_0 = 0$ such that $N - A$ is a martingale. We frequently use the operator Δ on processes: $\Delta A_s = A_s - A_{s-}$.

The bounds involve metrics on the set $\mathscr{P}(\mathbb{Z}^+)$ of probability measures on $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$. The *total variation* metric d_{TV} is defined for $\mathscr{P}$ and $\mathscr{Q}$ in $\mathscr{P}(\mathbb{Z}^+)$ by

$$d_{\mathrm{TV}}(\mathscr{P}, \mathscr{Q}) = \sup|\mathscr{P}h - \mathscr{Q}h|, \tag{2.3}$$

where the supremum is taken over $h\colon \mathbb{Z}^+ \to [0, 1]$ and a probability applied to such a function is just the expectation of the function as a random variable on the nonnegative integers. We also have

$$d_{\mathrm{TV}}(\mathscr{P}, \mathscr{Q}) = \inf P(X \neq Y), \tag{2.4}$$

where the infimum is taken over all possible joint distributions for random variables (X, Y) such that X has distribution $\mathscr{P}$ and Y has distribution $\mathscr{Q}$, and

$$d_{\mathrm{TV}}(\mathscr{P}, \mathscr{Q}) = \tfrac{1}{2} \sum_{n=0}^{\infty} |\mathscr{P}(n) - \mathscr{Q}(n)| \tag{2.5}$$

[see Barbour, Holst and Janson (1992), Appendix]. For $h\colon \mathbb{Z}^+ \to \mathbb{R}$ define dh to be the function defined on $\{1, 2, \ldots\}$ with $dh(n) = h(n) - h(n-1)$. The *Wasserstein* metric d_W is defined by

$$d_W(\mathscr{P}, \mathscr{Q}) = \sup|\mathscr{P}h - \mathscr{Q}h|, \tag{2.6}$$

where the supremum is over bounded functions $h\colon \mathbb{Z}^+ \to \mathbb{R}$ such that dh has codomain $[-1, 1]$. Just as for the total variation metric, there are two equivalent ways of writing the Wasserstein metric,

$$d_W(\mathscr{P}, \mathscr{Q}) = \inf E|X - Y|, \tag{2.7}$$

where the infimum is as in (2.4), and

$$d_W(\mathscr{P}, \mathscr{Q}) = \sum_{n=0}^{\infty} |\mathscr{P}[n, \infty) - \mathscr{Q}[n, \infty)| \tag{2.8}$$

[see Barbour, Holst and Janson (1992), Appendix]. Equations (2.4) and (2.7) permit the estimation of distances by the construction of pairs of random variables and this technique is often referred to as *coupling*.

The Stein–Chen method for computing the total variation distance between the distribution of a random variable X with values in $\mathbb{Z}^+$ and a Poisson distribution with parameter μ was introduced in Chen (1975). The main plank [see, e.g., Barbour and Eagleson (1983)] is that for any h: $\mathbb{Z}^+ \to [0,1]$ there exists a function

$$g: \mathbb{Z}^+ \to [-K_1(\mu), K_1(\mu)] \tag{2.9}$$

such that

$$E(h(X)) - \text{Poisson}_\mu(h) = E\{\mu g(X+1) - Xg(X)\}, \tag{2.10}$$

where $K_1(\mu) = 1 \wedge 1.4\mu^{-1/2}$. It can also be shown that for each h and μ

$$\sup_{n=1,2,\dots} |dg(n)| \le K_2(\mu), \tag{2.11}$$

where $K_2(\mu) = (1 - \exp(-\mu))/\mu$. Thus, to bound the total variation distance we need to estimate the absolute value of the right-hand side of (2.10) for g as in (2.9) and (2.11). It is worth noting that K_1 and K_2 are both bounded by 1, both tend to 0 at ∞ and $K_1 \ge K_2$. Often, terms in the bounds given will consist of $K_2(\mu)$ multiplied by an integral with respect to a measure whose total mass is μ. Apart from the factor $(1 - \exp(-\mu))$, such terms are just averages of the integrand.

For a random element X of some space, $\mathscr{L}X$ denotes the distribution or probability law of X, that is, if F is a measurable set of the space, $\mathscr{L}X(F) = P(X \in F)$.

3. Main results. The first result concerns the distribution of the number of points in a set for general carrier spaces. It provides an explicit bound for the departure of the distribution from Poisson in terms of the average Wasserstein distance between the distribution and the reduced Palm distribution for the number of points in the set. The theorem is natural in that the reduced Palm distribution and the ordinary distribution for the number of points in a set coincide in the case of a Poisson process, as may easily be checked from (2.2) with $f(x, \xi)$ of the form $I[\xi(A) = m]$ for bounded A and $m = 0, 1, \dots$. In the case when $\mathbb{X}$ is a finite set, it coincides with that of Barbour and Holst [(1989), Theorem 2.1], but it does not seem that the connection with reduced Palm probabilities for the general case has been noted before.

THEOREM 3.1. *Suppose that N is a point process on $\mathbb{X}$ and that N has a locally finite mean measure ν. Then, for any bounded Borel set B of $\mathbb{X}$ and $\mu \ge 0$,*

$$\begin{aligned} &d_{\text{TV}}(\mathscr{L}N(B), \text{Poisson}_\mu) \\ &\quad \le K_1(\mu)|\nu(B) - \mu| \\ &\qquad + K_2(\mu)\int_B d_W(\mathscr{L}N(B), \mathscr{L}(N^x - \delta_x)(B))\nu(dx), \end{aligned} \tag{3.2}$$

where $K_1(\mu) = 1 \wedge (1.4\mu^{-1/2})$ *and* $K_2(\mu) = (1 - \exp(-\mu))/\mu$. *Usually,* μ *is taken equal to* $\nu(B)$, *in which case* (3.2) *reduces to*

$$d_{TV}(\mathscr{L}N(B), \text{Poisson}_{\nu(B)})$$
$$\le K_2(\nu(B)) \int_B d_W(\mathscr{L}N(B), \mathscr{L}\{N^x - \delta_x\}(B))\nu(dx).$$

PROOF. Let g be any function satisfying (2.9) and (2.11). The first term arises as an estimate of $|E(\{\mu - \nu(B)\}g\{N(B)+1\})|$. We have

$$E(N(B)g\{N(B)\}) = E\left(\int_B g\{N(B)\}N(dx)\right)$$
$$= E\left(\int_B g(\{(N - \delta_x)(B)\} + 1)N(dx)\right),$$

because the integral is a sum over the values at the atoms of N times the number of points at each atom. Applying (2.2), the left-hand side equals

$$\int_B E\{g((N^x - \delta_x)(B) + 1)\}\nu(dx)$$

and thus the absolute difference of $E\{N(B)g\{N(B)\}\}$ from

$$E\{\nu(B)g\{N(B)+1\}\} = \int_B E\{g\{N(B)+1\}\}\nu(dx)$$

is bounded by

$$\int_B \left| E\{g\{(N^x - \delta_x)(B) + 1\}\} - E\{g\{N(B) + 1\}\} \right| \nu(dx).$$

This gives the theorem, upon applying (2.6), (2.9), (2.10) and (2.11). □

We now turn to the problem of finding bounds suited to the special structure of point processes on the line. The following lemma plays an important role in the formulation of Theorem 3.7 in terms of compensators.

LEMMA 3.3. *Suppose* $\{X_s\}$ *is a raw nonnegative integer-valued process. For each* ω *and each* $s \ge 0$, *there exist probability distributions* $\mathscr{P}_s(\cdot, \omega)$ *and* $\mathscr{P}_s^-(\cdot, \omega)$ *on the nonnegative integers such that if, for each* $m \ge 0$, $\{g_s(m)\}$ *is a bounded previsible process,*

$$ {}^o g_s(X_s) = \sum_{m=0}^{\infty} g_s(m)\mathscr{P}_s(m) \tag{3.4}$$

and

$$ {}^p g_s(X_s) = \sum_{m=0}^{\infty} g_s(m)\mathscr{P}_s^-(m), \tag{3.5}$$

where $\mathscr{P}_s(m)$ *is the process* $(s, \omega) \mapsto \mathscr{P}_s(\{m\}, \omega)$ *and similarly for* $\mathscr{P}_s^-$. *Furthermore,* $\mathscr{P}_s(m)$ *is* ${}^oI[X_s = m]$ *and* $\mathscr{P}_s^-(m)$ *is* ${}^pI[X_s = m]$.

REMARK 3.6. The optional projection of a right-continuous left limits process X is the unique right-continuous left limits process ${}^oX$ such that ${}^oX_s$ is a version of $E_s(X_s)$: see, for example, Rogers and Williams [(1987), page 320, Theorem 7.10]. This makes it easy to identify the optional projection later.

PROOF OF LEMMA 3.3. We define $\mathscr{P}_s(m)$ and $\mathscr{P}_s^-(m)$ so that the last sentence of the lemma is true. From the construction of the optional and previsible projections, a nonnegative raw process has a nonnegative optional and previsible projection. Furthermore, the optional and previsible projections are linear operators. Thus, we may modify $\mathscr{P}_s(m)$ and $\mathscr{P}_s^-(m)$ on a single evanescent set so that, for all ω and s,

$$0 \leq \mathscr{P}_s(m, \omega), \qquad \mathscr{P}_s^-(m, \omega) \leq 1.$$

Furthermore,

$$\sum_{m=0}^{\infty} I[X_s = m] = 1.$$

because X_s takes values in $\{0, 1, 2, \dots\}$. Now the optional projection of the sum on the left is the sum of the optional projections as is readily checked from the definition and the monotone convergence theorem for conditional expectations. Likewise, the previsible projection of the sum on the left is the sum of the previsible projections. Hence the sum of $\mathscr{P}_s(m)$ and $\mathscr{P}_s^-(m)$ over m in $\{0, 1, 2, \dots\}$ is the optional projection of the constant 1. Making further modifications on a single evanescent set we can ensure these sums are globally 1 in s and ω. We may therefore define, for a set of nonnegative integers I,

$$\mathscr{P}_s(I) = \sum_{m \in I} \mathscr{P}_s(m) \quad \text{and} \quad \mathscr{P}_s^-(I) = \sum_{m \in I} \mathscr{P}_s^-(m),$$

and these have the property that, pointwise in ω, $\mathscr{P}_s$ and $\mathscr{P}_s^-$ are probability measures on $\{0, 1, 2, \dots\}$.

Suppose T is a bounded stopping time. For each ω and s,

$$g_s(X_s) = \sum_{m=0}^{\infty} g_s(m) I[X_s = m].$$

Because g is adapted, ${}^o(g_s(m)I[X_s = m]) = g_s(m)\mathscr{P}_s(m)$ and thus $E_T\{g_T(m)I[X_T = m]\} = g_T(m)\mathscr{P}_T(m)$. Hence, by bounded convergence and linearity for conditional expectations,

$$E_T(g_T(X_T)) = \sum_{m=0}^{\infty} g_T(m)\mathscr{P}_T(m).$$

Now the right-hand side of (3.4) is the limit of optional processes and so is optional. This and the last equation demonstrate (3.4), by the definition of the optional projection. Similar considerations with T being previsible give (3.5). □

THEOREM 3.7. *Suppose that N is a simple point process with compensator A. For each $t \geq 0$,*

$$d_{\mathrm{TV}}(\mathscr{L}N_t, \mathrm{Poisson}(\mu)) \leq K_1(\mu)E|A_t - \mu| + K_2(\mu)E\left\{\sum_{s \leq t} \Delta A^2(s)\right\} + K_2(\mu)E\left\{\int_0^t d_W(\mathscr{P}_s, \mathscr{P}_s^-)\, dN_s\right\}, \tag{3.8}$$

where $\mathscr{P}_s$, $\mathscr{P}_s^-$ are the probability distributions defined in Lemma 3.3 for the process $X_s = N_t - N_s$, $K_1(\mu) = 1 \wedge (1.4\mu^{-1/2})$ and $K_2(\mu) = (1 - e^{-\mu})/\mu$.

REMARK 3.9. The first two terms of the above bound without the μ-dependent constants constitute the bound that appeared in Brown (1983). In many cases (see the examples in Sections 4 and 5) it turns out that the third term of the above bound can be shown to be of no greater order than the first two. In these cases, the above bound will therefore be of lower order of magnitude for large μ than the bound of Brown (1983). This partly explains why previous compensator bounds have not produced the correct orders of magnitude. All the terms of the bound are zero if N is a simple Poisson process with history generated by the process.

REMARK 3.10. The Wasserstein distance in the above integral can, by (2.8) and the last sentence of Lemma 3.3, be expressed as an infinite sum of absolute differences of previsible and optional processes and is therefore optional. The integral is thus well defined.

REMARK 3.11. The third term has a nice intuitive explanation. The distance is a measure of the instantaneous change in our best estimation of the distribution of the remaining number of points before t, and by integrating with respect to dN_s/μ we are averaging this instantaneous change over the points of N in $[0, t]$. For a good Poisson approximation, we should not be too surprised about the number of points that still remain before t at each point of the process.

REMARK 3.12. The argument for Theorem 3.7 is simple relative to the arguments used, for example, in Brown (1983), which relied heavily on coupling. However coupling is used in several of the examples below, to estimate the third term in the bound.

PROOF OF THEOREM 3.7. We need to bound the right-hand side of (2.10) for g satisfying (2.9) and (2.11). The first term of (3.8) is an upper bound for $|E\{\mu g(N_t + 1)\} - E\{A_t g(N_t + 1)\}|$, and it is therefore enough to show that the second and third terms are an upper bound for $|E\{A_t g(N_t + 1)\} - E(N_t g(N_t))\}|$.

We can write

$$E\{A_t g(N_t+1)\} = E\left\{\int_0^t g(N_t+1)\,dA_s\right\}$$

$$= E\left\{\int_0^t {}^p g(N_t+1-\Delta N_s)\,dN_s\right\}$$

$$+ E\left\{\int_0^t {}^p(dg(N_t+1)\Delta N_s)\,dA_s\right\},$$

in view of the identity $g(N_t+1) = g(N_t+1-\Delta N_s) + dg(N_t+1)\,\Delta N_s$ and elementary facts about optional, previsible and dual previsible projections [see, e.g., Dellacherie and Meyer (1982), pages 122 and 135]. If the previsible projection in the first term were an optional projection, the projection could be removed, and that term would be $E\{\int_0^t g(N_t)\,dN_s\}$ because N only increases at jumps and at each jump $\Delta N = 1$. We introduce $g_s(n) = g(N_{s-}+1+n)$ so that $g_s(N_t - N_s) = g(N_t + 1 - \Delta N_s)$ and $g_s(n)$, being left-continuous and adapted, is previsible. We then have

$$E\{A_t g(N_t+1)\} - E\{N_t g(N_t)\} = E\left\{\int_0^t [{}^p g_s(N_t - N_s) - {}^o g_s(N_t - N_s)]\,dN_s\right\}$$

$$+ E\left\{\int_0^t {}^p(dg(N_t+1)\,\Delta N_s)\,dA_s\right\},$$

and it remains to bound the absolute value of the right-hand side. The absolute value of the first term is bounded by the expected integral of the absolute value of the integrand. Using Lemma 3.3, (2.6), (2.9) and (2.11), the absolute value of the integrand is bounded above by $K_2(\mu)d_W(\mathscr{P}_s, \mathscr{P}_s^-)$ and thus the absolute value of the first term is bounded by the third term of (3.8). The absolute value of the second term is bounded by the expected integral of the previsible projection of the absolute value of $dg(N_t+1)\,\Delta N_s$, since for any process $|{}^pX| = |{}^pX^+ - {}^pX^-| \le {}^p|X|$. Since $|dg(N_t+1)\,\Delta N_s| \le K_2(\mu)\,\Delta N_s$ and the previsible projection of ΔN is ΔA [Dellacherie and Meyer (1982), page 136], the second term is bounded by the second term of (3.8) and the proof is complete. □

It is possible to generalise the above bound to general point processes.

THEOREM 3.13. *Suppose that N is a point process and that A is the compensator of N. Then the bound of Theorem* 3.7 *remains true provided the following extra term is added*:

$$K_2(\mu)E\left\{\sum_{s\le t}(\Delta N_s - 1)\,\Delta N_s\right\}.$$

PROOF. The proof is similar to the proof of Theorem 3.7 with the following changes. The identity in the point process case

$$g(N_t + 1) = g(N_t + 1 - \Delta N_s) + dg(N_t + 1)\,\Delta N_s$$

has the second term replaced by $g(N_t + 1) - g(N_t + 1 - \Delta N_s)$ but $K_2(\mu)\,\Delta N_s$ is still an upper bound for the absolute value of this term. The other change is that $E\{g(N_t)N_t\}$ is now

$$E\left\{\int_0^t g(N_t + 1 - 1)\,dN_s\right\}$$

and it is the absolute difference of this expression from

$$E\left\{\int_0^t g(N_t + 1 - \Delta N_s)\,dN_s\right\}$$

which is bounded by the additional term. □

We can also consider the number of points at a stopping time.

THEOREM 3.14. *Suppose τ is a stopping time. Then Theorems* 3.7 *and* 3.13 *remain true with t everywhere replaced by τ, provided the process X which defines $\mathscr{P}_s$ and $\mathscr{P}_{s-}$ becomes $X_s = N_\tau - N_{\tau \wedge s}$.*

PROOF. We can take $t = \infty$ in Theorems 3.7 and 3.13 because either $EN_\infty = EA_\infty = \infty$, in which case the right-hand side of the bound is also ∞, or $EN_\infty = EA_\infty < \infty$, in which case the proofs work. Thus Theorems 3.7 and 3.13 may be applied to the point process N stopped at τ evaluated at ∞, and this gives the present theorem. □

It might be thought that the Wasserstein distance would always be difficult to measure. However, it can be calculated relatively easily in a number of special cases, as is seen in the next section.

4. Processes with conditionally independent increments. We say that a point process N (on a general carrier space) has *conditionally independent increments* if, for some locally finite random measure A, for all $n = 1, 2, 3, \dots$ and for all bounded, disjoint, Borel sets $B_1, B_2, \dots, B_n$, we have $N(B_1), N(B_2), \dots, N(B_n)$ independent with means $A(B_1), A(B_2), \dots, A(B_n)$ conditional on A. For any $h\colon \mathbb{Z}^+ \to [0,1]$,

$$\begin{aligned} |Eh\{N(B)\} - \text{Poisson}_\mu\, h| &\le E\big|E(h\{N(B)\}|A) - \text{Poisson}_\mu\, h\big| \\ &\le E\big(d_{\text{TV}}(\mathscr{L}N_A(B), \text{Poisson}_\mu)\big), \end{aligned} \tag{4.1}$$

where N_A has distribution that of N conditional on A. Thus the total variation distance of $N(B)$ from Poisson_μ is dominated by the right-hand side of (4.1). Two cases are of special interest.

4.2. N IS A COX PROCESS. We say N is a *Cox process* if conditional on A it is a Poisson process with mean measure A. The law of $N_A(B)$ and $(N_A^x - \delta_x)(B)$ coincide (as remarked before Theorem 3.1) and thus from (4.1) and (3.2) we have

$$(4.3) \qquad d_{\mathrm{TV}}\big(\mathscr{L}N(B), \mathrm{Poisson}_\mu\big) \le K_1(\mu) E|A(B) - \mu|.$$

This estimate is Theorem I.C(i) of Barbour, Holst and Janson (1991).

If N has carrier space $(0, \infty)$, $B = (0, t]$, and A is continuous and $\mathscr{F}_0$-measurable, we can derive the same bound as (4.3) from Theorem 3.7. In this case, N is a Cox process [see, e.g., Brown (1981), Theorem 2], but it is instructive nonetheless to see how the compensator bound works. The first term of (3.8) coincides with the right-hand side of (4.3) and the second term is zero by the assumption of continuity of A. Select a regular conditional distribution $\mathscr{P}_0$ for N given $\mathscr{F}_0$. If $Y_s = I[N_t - N_s = m]$, for $0 \le s \le t$, $m \in \mathbb{Z}^+$, then, by dominated convergence using $\mathscr{P}_0$, Z: $s \mapsto \mathscr{P}_0[N_t - N_s = m]$ is right-continuous, since Y is right-continuous and Y_s is dominated by 1. Moreover, because N is Cox,

$$E(Y_s|\mathscr{F}_s) = \mathscr{P}_0(N_t - N_s = m)$$

and thus Z is the optional projection of Y [Rogers and Williams (1987), Theorem VI.7.10]. It is also the previsible projection of Y because Z, being $\mathscr{F}_0$-measurable, is also previsible. Thus, in this case, $d_W(\mathscr{P}_s, \mathscr{P}_s^-) = 0$ identically. It is possible to extend this argument to the general Cox process on $(0, \infty)$ by approximating such a process by one having the atoms of A replaced by diffuse components of the same size concentrated near to the atoms of A.

4.4. N IS SIMPLE. Suppose that N is a simple point process with independent increments on a general carrier space. Then, we may write [see Kallenberg (1976), Theorem 7.1 and Corollary 7.4],

$$(4.5) \qquad N = N_1 + N_2,$$

where N_1 is a Poisson process and N_2 is an independent simple point process whose points are concentrated on the atoms of ν. Now the reduced Palm process of N_1 is N_1, while, for an atom x of ν, the reduced Palm process of N_2 is easily seen to be $N_2 - N_2(\{x\})\delta_x$ (checking the required identity on a suitable generating class as in the case of a Poisson process). Using Kallenberg [(1976), Problem 10.3], we have the reduced Palm process for N as N except at atoms x of ν, where it is $N - N(\{x\})\delta_x$. Thus the Wasserstein distance between the law of $N(B)$ and that of $(N^x - \delta_x)(B)$ is 0 except at the atoms x of ν in B where it is $\nu(\{x\})$ [using (2.7)]. Thus, from (3.2), for $\mu > 0$.

$$(4.6) \qquad \begin{aligned} d_{\mathrm{TV}}(\mathscr{L}N(B), \mathrm{Poisson}(\mu)) &\le K_1(\mu)|\nu(B) - \mu| \\ &\quad + K_2(\mu)\int_B \nu(\{x\})\nu(dx). \end{aligned}$$

It is worth remarking that the second term of the right-hand side of (4.6) is a sum of squares of jumps of the atoms of ν, just as in the second term of the linear case.

Now suppose that N is a simple point process with conditionally independent increments. The reasoning leading to (4.6) may be applied to compute the total variation distance of $\mathscr{L}N_A(B)$ from Poisson(μ). Equation (4.1) then gives

$$d_{\mathrm{TV}}\big(\mathscr{L}N(B), \mathrm{Poisson}_\mu\big) \le K_1(\mu)E|A(B) - \mu| \\ + K_2(\mu)E\int_B A(\{x\})A(dx). \tag{4.7}$$

It is interesting that this is exactly the bound obtained by applying Theorem 3.7 for point processes on the line, the proof that $d_W(\mathscr{P}_s, \mathscr{P}_s^-) = 0$ being exactly as in example 4.2.

Two special cases are now included to show that the first two terms of (3.8) cannot be essentially improved.

4.8. Bernoulli trials. Suppose that $I_1, \dots, I_n$ are *independent* Be(p_i) random variables. This is the special case of (4.7) in which the carrier space is $\{1, \dots, n\}$ ($n = 1, 2, \dots$) and the process has independent increments. Taking $\mu = \Sigma_{j=1}^n p_j$, (4.7) gives the bound $K_2(\mu)\Sigma_{j=1}^n p_j^2$. The bound is the same as was established in Barbour and Hall (1984), who also showed that the bound is of best μ-order. Hence the μ-order of the multiplier of the second term in (3.8) is also best possible.

4.9. Two values of A. Suppose that A can take only two values, either having constant intensity $\mu + 10\mu^{1/2}$, with probability $\mu^{-1/2}$, or having constant intensity $\mu - 10/(1 - \mu^{-1/2})$, with probability $1 - \mu^{-1/2}$, where μ is thought of as being large. Then only the first of the three terms in (3.8) is nonzero, and at $t = 1$ it takes the value

$$K_1(\mu)E|A_1 - \mu| = 20K_1(\mu) \asymp \mu^{-1/2}.$$

On the other hand, it is not difficult to see directly that $d_{\mathrm{TV}}(\mathscr{L}N_1, \mathrm{Poisson}_\mu)$ is of order $\mu^{-1/2}$, as follows also from Theorem III.F of Barbour, Holst and Janson (1991). Hence this example shows that the μ-order of the multiplier of $E|A_t - \mu|$ in (3.8) cannot in general be improved, even when μ is chosen to be EA_t.

5. Processes with internal history. Here we consider only point processes on the half-line and insist that the history used is internal, in that it is generated by the point process and some extra information at time 0. It happens that in this case we can derive an explicit formula for the Wasserstein distance in (3.8) in terms of the conditional distributions for points given the previous point locations and the initial information. These conditional distributions are also those which are required for the computation of the compensator [see, e.g., Jacod (1975)]. Hence, Theorem 5.2 only concerns the Wasserstein distance term of the bound (3.8).

Suppose therefore that Y is a random element of some measurable space $(S, \mathscr{S})$ and that $\mathscr{F}_s = \sigma(Y) \vee \sigma(N_z,\ z \le s)$. Thus the initial information is contained in Y: It will usually be possible to encapsulate such information in some random element Y, since no restrictions are made on the space in which Y takes its values. The purpose of requiring this is to avoid the notational complexities that are inherent in using regular conditional distributions given an arbitrary σ-algebra at time 0—these notational complexities are, however, the only problem encountered in the generalization.

The quantity that has to be calculated in order to compute the third term of the bound is

$$P(N_t - N_s = m \mid Y = y, T_1 = t_1, \dots, T_n = t_n, T_{n+1} > s) \tag{5.1}$$

for y in S and $0 < t_1 < t_2 < \cdots < t_n \le s$ and $m \in \{0, 1, 2, \dots\}$. Because we require various properties of the quantity (5.1), the following theorem defines it in a precise way. Let F_n be the joint probability distribution of $(Y, T_1, \dots, T_n)$.

THEOREM 5.2. *Suppose, for each* $n = 0, 1, 2, \dots,$ *that* $f_n\colon S \times (\mathbb{R}^+)^{n+1} \to \mathscr{P}(\mathbb{Z}^+)$ *so that the following hold*:

(a) *for each* $m \in \{0, 1, 2, \dots\}$, *the mapping*

$$(y, s, t_1\, t_2, \dots, t_n) \mapsto f_n(y, s, t_1, \dots, t_n)(m)$$

is product-measurable;

(b) *for each set* F *in* $\mathscr{S}$, *Borel set* B *of* $\mathbb{R}^n$, m *in* $\mathbb{Z}^+$ *and* $s \ge 0$,

$$\int_{F \times B \cap [0, s]^n \times (s, \infty)} f_n(y, s, t_1, \dots, t_n)(m)\, dF_{n+1}(y, t_1, \dots, t_n, t_{n+1})$$
$$= P\big(N_t - N_s = m, (T_1, \dots, T_n) \in B \cap [0, s]^n, T_{n+1} > s, Y \in F\big);$$

(c) *for each* y *in* S, $t_1, \dots, t_n$ *in* $\mathbb{R}^+$ *and* m *in* $\mathbb{Z}^+$,

$$s \mapsto f_n(y, s, t_1, \dots, t_n)(m)$$

is right-continuous.

Then

$$\mathscr{P}_s = \sum_{n=0}^{\infty} f_n(Y, s, T_1, \dots, T_n) I[T_n \le s < T_{n+1}] \tag{5.3}$$

and

$$\begin{aligned} \mathscr{P}_s^- = \sum_{n=0}^{\infty} \{ & f_{n+1}(Y, s, T_1, \dots, T_n, s)(\Delta A(s)) \\ & + f_n(Y, s, T_1, \dots, T_n)(1 - \Delta A(s))\} I[T_n < s \le T_{n+1}]. \end{aligned} \tag{5.4}$$

Furthermore,

$$(5.5)\qquad \begin{aligned} E\left\{\int_0^t d_W(\mathscr{P}_s,\mathscr{P}_s^-)\,dN_s\right\} &= E\left\{\int_0^t d_s(1-\Delta A(s))\,dN_s\right\} \\ &= E\left\{\int_0^t d_s(1-\Delta A(s))\,dA_s\right\}, \end{aligned}$$

where $d_s = d_W(f_{n+1}(Y,s,T_1,\dots,T_n,s), f_n(Y,s,T_1,\dots,T_n))$ *on* $[T_n < s \le T_{n+1}]$.

REMARK 5.6. That f_n is the quantity defined in (5.1) is the content of requirements (a) and (b). Since $\mathscr{P}_s(m)$ is optional, arguments similar to those in Jacod [(1975), Lemma 3.3], show that $\mathscr{P}_s(m)$ has a decomposition satisfying (5.3) for some functions f_n. Requirement (b) is then a reexpression of the fact that

$$(5.7)\qquad \begin{aligned} &P_s(N_t - N_s = m)I[T_n \le s < T_{n+1}] \\ &\qquad = P(N_t - N_s = m \mid Y, T_1,\dots,T_n, T_{n+1} > s), \end{aligned}$$

which can be established as in Brown [(1978), Lemma 7]. Requirement (c) arises because the optional projection of a right-continuous process is right-continuous [Dellacherie and Meyer (1982), Theorem VI.47]. Thus functions $\{f_n\}$ satisfying all of (a), (b) and (c) always exist. However, the proof the theorem as stated is slightly simpler than fleshing out the above arguments. It is also adequate for the applications later. Expression (5.1) is, however, important for finding f_n.

It is also worth remarking that the functions f_n can be expressed directly in terms of the compensator, although in such a complicated way that we have not pursued this in general, leaving it to examples to do the computations (see examples 5.8 and 5.20). The reason that f_n can be expressed in terms of the compensators is that by the uniqueness theorem of Jacod (1975) the conditional distribution of point times given past points and Y can be written in terms of the compensator. Hence so can the marginal distributions, and so also can the distribution of the number of points in an interval, on using the general relation that $[N_t \ge j] = [S_j \le t]$.

PROOF OF THEOREM 5.2. The probability attributed by the right-hand side of (5.3) to m is right-continuous in s and so, by Rogers and Williams [(1987), Theorem VI.7.10], this probability is $\mathscr{P}_s(m)$ if $P_s(N_t - N_s = m)$ is equal to the right-hand side of (5.3) for each s. But, by checking on generating sets, it is easily verified that $\mathscr{F}_s \cap [T_n \le s < T_{n+1}] = \sigma(T_1,\dots,T_n) \cap [T_n \le s < T_{n+1}]$. Thus, the desired equation comes from requirements (a) and (b) on f_n. Furthermore, the right-hand side of (5.3) *is* a probability distribution in $\mathbb{Z}^+$ for each s, thus completing the verification of (5.3).

We can write, for $s > 0$,

$$\mathscr{P}_s = \sum_{n=1}^{\infty} f_n(Y, s, T_1, \dots, T_{n-1}, s) I[T_n = s]$$

$$+ \sum_{n=0}^{\infty} f_n(Y, s, T_1, \dots, T_n) I[T_n < s < T_{n+1}]$$

$$= \left\{ \sum_{n=0}^{\infty} f_{n+1}(Y, s, T_1, \dots, T_n, s) I[T_n < s \le T_{n+1}] \right\} \Delta N(s)$$

$$+ \left\{ \sum_{n=0}^{\infty} f_n(Y, s, T_1, \dots, T_n) I[T_n < s \le T_{n+1}] \right\} (1 - \Delta N(s)).$$

We call the two processes in braces Y_1 and Y_0, respectively. By Jacod [(1975), Lemma (3.3)], these two processes are previsible. Now it is easy to check that the previsible projection of a raw process is equal to the previsible projection of its optional projection and hence

$$^pI[N_t - N_s = m] = {}^p\mathscr{P}_s(m)$$

and

$$^p(Y_1 \Delta N + Y_0(1 - \Delta N)) = Y_1{}^p\Delta N + Y_0{}^p(1 - \Delta N)$$

by Dellacherie and Meyer [(1982), page 106 (e) and page 104]. However, ΔA is the previsible projection of ΔN [Dellacherie and Meyer (1982), page 136]. Thus

$$\mathscr{P}_s^-(m) = \sum_{n=0}^{\infty} \{ f_{n+1}(Y, s, T_1, T_2, \dots, T_n, s)(m) I[T_n < s \le T_{n+1}] \Delta A_s$$

$$+ f_n(Y, s, T_1, \dots, T_n)(m) I[T_n < s \le T_{n+1}](1 - \Delta A_s) \}.$$

Equation (5.4) follows because the right-hand side *is* a probability distribution over m, because it is a convex combination of probability distributions for each path.

To obtain (5.5), suppose that $0 \le p \le 1$ and that $\mathscr{P}$ and $\mathscr{Q}$ are in $\mathscr{P}(\mathbb{N})$. Then by (2.8), for $I = \{0, 1, \dots, i\}$,

$$d_W(\mathscr{P}, p\mathscr{P} + (1-p)\mathscr{Q}) = \sum_{i=0}^{\infty} |\mathscr{P}(I) - (p\mathscr{P}(I) + (1-p)\mathscr{Q}(I))|$$

$$= (1-p) d_W(\mathscr{P}, \mathscr{Q}).$$

Thus on $[s = T_{n+1}]$, $n = 0, 1, 2, \dots,$

$$d_W(\mathscr{P}_s, \mathscr{P}_s^-) = (1 - \Delta A_s) d_W(f_{n+1}(Y, s, T_1, \dots, T_n, s) f_n(Y, s, T_1, \dots, T_n)),$$

which gives the first equation of (5.5). The second follows because d_s is previsible, by Jacod [(1975), Lemma 3.3]. □

The next two examples make use of Theorem 5.2 in concrete cases.

5.8. THE MARKOV POINT PROCESS. We consider a point process on the line which is Markovian with stationary transition rates. In language suitable for applying Theorem 3.7, we assume that N is a point process whose compensator is

$$A_t = \int_0^t \lambda_{N_s+1}\, ds, \tag{5.9}$$

where $\lambda_1, \lambda_2, \dots$ are positive constants, and, for the moment, $\mathscr{F}_0$ is trivial. As is well known, the times between points are then independent exponential random variables $X_1, X_2, \dots$ with parameters $\lambda_1, \lambda_2, \dots$. To obtain our bounds, we assume that the λ's are bounded away from 0 and ∞. We let $m_t = EN_t$, $v_t = \operatorname{Var} N_t$, $\lambda_{\inf}(\lambda_{\sup})$ be the infimum (supremum) of the λ's, $r = \lambda_{\sup}/\lambda_{\inf}$, $[\cdot]$ be the integer part function,

$$\bar{\lambda} = (m_t + 1) \Big/ \left\{ \int_0^{m_t+1} 1/\lambda_{[s]+1}\, ds \right\} \tag{5.10}$$

(so that in some sense $\bar{\lambda}$ is the "harmonic average" λ experienced on $[0, t]$) and ε be the supremum of $|1 - \bar{\lambda}/\lambda_i|$ over $i \in \mathbb{Z}^+$. We would expect a good Poisson approximation if r is close to 1 and ε is close to 0, because in this case the system has independent exponential interpoint times with nearly constant means.

We are able to prove that, for any $t \geq 0$,

$$\begin{aligned} d_{\mathrm{TV}}(N_t, \mathrm{Poisson}_{\bar{\lambda}t}) &\leq K_1(\bar{\lambda}t)\varepsilon\left\{\sqrt{m_t + 1} + \sqrt{v_t} + 1\right\} \\ &\quad + K_2(\bar{\lambda}t)\{\lambda_{\sup}t\}\{r(r-1)\}. \end{aligned} \tag{5.11}$$

In typical cases where the bound might be applied, m_t and v_t are both of order t as t increases, because otherwise a Poisson approximation with parameter $\bar{\lambda}t$ might not be sensible. In this case, the bound is of constant order in t. Previous bounds using compensators, such as (1.1), all have the bound increasing with the square root of t rather than being of constant order and would not reveal that the approximation is good whenever the rate that points appear is nearly constant. Thus Stein's technique enables one to get bounds that are an order of magnitude better than otherwise. Although the approximation by a Poisson distribution is perhaps of minor interest for this process, whose distributions are in principle simple to calculate, it *is* of interest in understanding the significance of (3.8). It also appears to the authors to be difficult to calculate the direct point process bound (3.2) for this process, although the compensator bound is relatively straightforward.

The first term of (5.11) arises because from (5.9)

$$E|A_t - \bar{\lambda}t| \leq E\left| \sum_{i=1}^{N_t} \lambda_i X_i + \lambda_{N_t+1}(t - T_{N_t}) - \bar{\lambda}t \right|. \tag{5.12}$$

We proceed to estimate the right-hand side of (5.12). Let τ be the discrete stopping time $N_t + 1$ for the sequence $X_1, X_2, \dots$. Using the triangle inequality, the right-hand side of (5.12) is bounded by

$$E\left|\sum_{i=1}^{\tau}\left(1-\bar{\lambda}/\lambda_i\right)E_i\right| + E\left|\int_0^{\tau}\left(1-\bar{\lambda}/\lambda_{[s]+1}\right)ds\right| \tag{5.13}$$
$$+ E\left|\left(1-\bar{\lambda}/\lambda_\tau\right)\lambda_\tau(T_\tau - t)\right|,$$

where $E_i = (\lambda_i X_i - 1)$ is an exponential(1) random variable centered at its mean. The reason for introducing E_i is that we may apply Wald's identity [in its second moment form; see, for example, Chow and Teicher (1988), Exercise 5.3.8] to calculate the variance of the first sum in (5.13). This yields

$$E\left(\left\{\sum_{i=1}^{\tau}\left(1-\bar{\lambda}/\lambda_i\right)E_i\right\}^2\right) = E\sum_{i=1}^{\tau}\left(1-\bar{\lambda}/\lambda_i\right)^2. \tag{5.14}$$

Using the fact that standard deviation exceeds mean absolute deviation, a bound for the first term of (5.13) is

$$\sup_i\left|1-\bar{\lambda}/\lambda_i\right|\sqrt{E\tau} = \varepsilon\sqrt{E\tau}. \tag{5.15}$$

Now $\bar{\lambda}$ has been defined precisely so that the next term can be estimated, using the identity

$$\int_0^{E\tau}\left(1-\bar{\lambda}/\lambda_{[s]+1}\right)ds = 0.$$

Thus the second term of (5.13) is bounded by

$$E\left|\int_\tau^{E\tau}\left(1-\bar{\lambda}/\lambda_{[s]+1}\right)ds\right| \le \varepsilon E|\tau - E\tau| \le \varepsilon\sqrt{\operatorname{Var}\tau}. \tag{5.16}$$

By the Markov property, $(T_\tau - t)$ has an exponential distribution with parameter λ_τ, conditional on the value of τ. Thus, a bound for the last term of (5.13) is ε times the mean absolute deviation of an exponential(1) random variable. This with the information in (5.12)–(5.16) substituted into the first term of (3.8) gives the first term of (5.11).

The second term of (3.8) is zero in this case because the compensator is everywhere continuous. To obtain the third term of (3.8), we use the following lemma, in conjunction with Theorem 5.2.

LEMMA 5.17. *Suppose that, for $n = 1, 2, \dots,$ $t \ge 0$ and an n-vector $\mathbf{x}$ of positive constants, $F(t, \mathbf{x})$ denotes the distribution function of the sum of n independent exponential random variables with means given by the elements of $\mathbf{x}$. Then, for any two vectors $\mathbf{x}_0, \mathbf{x}_1$ of the form $\mathbf{x}_0 = (x_0, x_1, \dots, x_{n-1})$ and $\mathbf{x}_1 = (x_1, \dots, x_n)$,*

$$F(t, \mathbf{x}_1) - F(t, \mathbf{x}_0) = \left[\{x_n - x_0\}/\{x_0 x_n\}\right] f(t, (x_0, x_1, \dots, x_n)),$$

where f is the density corresponding to F.

PROOF. We have

$$F(t,\mathbf{x}_0) = \int_0^t (1 - \exp(-x_0(t-s)) f(s,(x_1,\dots,x_{n-1}))\,ds)$$

and

$$F(t,\mathbf{x}_1) = \int_0^t (1 - \exp(-x_n(t-s)) f(s,(x_1,\dots,x_{n-1}))\,ds).$$

Since $\exp(-x_i(t-s))$ is $1/x_i$ times the density of an exponential(x_i) at $t-s$ $(i = 0, n)$, we have

$$F(t,\mathbf{x}_1) - F(t,\mathbf{x}_0) = f(t,(x_0,\dots,x_{n-1}))/x_0 - f(t,(x_1,\dots,x_n))/x_n.$$

Feller [(1971), page 40], gives an explicit formula for f in the case where the x's are all different (note the minor misprint). In this case, the lemma results from applying his formula in the last two equations, collecting together the terms which have the same exponential and then taking out a factor $\{x_n - x_0\}$. If some of the x's are the same, the lemma results from an obvious continuity argument. □

The relevance of Lemma 5.17 is that, in the terminology of Theorem 5.2, for $n, m \in \mathbb{Z}^+$, $0 \le s \le t$ and $t_1 < t_2 < \cdots < t_n \le s$ (here Y is deterministically 0 and the argument corresponding to it is omitted),

$$\begin{aligned} f_n(s,t_1,\dots,t_{n+1})(m) &= F(t-s,(\lambda_n,\dots,\lambda_{n+m})) \\ &\quad - F(t-s,(\lambda_{n+1},\dots,\lambda_{n+m+1})), \end{aligned}$$

an identity which is easy to check against the criteria (a), (b) and (c) using the fact that the interpoint times are independent exponential random variables. Hence at the jumps of N the random variable d_s in (5.5) is

$$\sum_{m=0}^{\infty} \left| F\big(t-s,(\lambda_{N_s},\dots,\lambda_{N_s+m})\big) - F\big(t-s,(\lambda_{N_s+1},\dots,\lambda_{N_s+m+1})\big)\right|,$$

using (2.8) and the fact that the terms in the previous expression for $f_n(\cdot)\,(j)$ telescope when added over $j = m, m+1, \dots$. From Lemma 5.17,

$$d_s \le \{\lambda_{\sup} - \lambda_{\inf}\}/\lambda_{\inf}^2 \sum_{m=0}^{\infty} f\big(t-s,(\lambda_{N_s},\dots,\lambda_{N_s+m+1})\big). \tag{5.18}$$

The sum on the right-hand side of (5.18) is the derivative at $t-s$ of the expected number of points in a Markov point process with rates $\lambda(N_s)$, $\lambda(N_s + 1),\dots,$ conditional on N_s. This is clearly bounded above by the derivative of the expected number of points in a Poisson process of rate $\lambda_{\sup}$, which gives $d_s \le \{r-1\}r$ and completes the proof of the bound (5.11), upon using (5.5) and the previous bound for the rate of points.

The bound (5.11) is for a Markov point process with fixed initial state. For a stationary Markov point process, it would be necessary to introduce an independent nonnegative random variable Y and to suppose that, conditional on Y, the rates of the process were $\lambda_{Y+1}, \lambda_{Y+2}, \dots$. The estimate for the second term of (3.8) remains 0 and the upper bound for d_s of $\{r-1\}r$ remains true. Hence the second term of (5.11) is unchanged. If λ is arbitrary and $\bar{\lambda}_Y$ is defined as in (5.10) but with rates $\lambda_{Y+1}, \lambda_{Y+2}, \dots$ and with m_t replaced by $E(N_t|Y)$, so that $\bar{\lambda}_Y$ is a random variable, then

$$E|A_t - \lambda t| \le E\Big\{E\big[|A_t - \bar{\lambda}_Y t|\,|Y\big]\Big\} + E|\bar{\lambda}_Y - \lambda|t. \tag{5.19}$$

Since the conditional expectation on the right-hand side of (5.19) is a special case of the expectation in (5.12), inequality (5.19) makes it possible to extend (5.11) to an arbitrary initial distribution. The resulting bound will still be of constant order in t if the second term of (5.19) increases no faster than $\sqrt{t}$.

A specific example follows.

5.20. THE ALTERNATING POISSON PROCESS. We consider a stationary Markov point process whose rates alternate between α and β with $\alpha < \beta$. We can think of this process as having a compensator whose derivative a is given by

$$a = \alpha(1-Z) + \beta Z,$$

where Z is a stationary Markov chain with state space $\{0,1\}$ and rates of transition α from 0 to 1 and β from 1 to 0. In this case $Y = Z_0$ and $\lambda_0 = \alpha$, $\lambda_1 = \beta$, $\lambda_2 = \alpha, \dots$. Now N_t has the same expectation as A_t, which, on using Fubini's theorem, is tEa_0. Thus, we choose

$$\lambda = Ea_0 = 2\alpha\beta/(\alpha+\beta),$$

using the fact that the stationary probability of Z being 0 is $\beta/(\alpha+\beta)$. In this example, it is easier to bound $E|A_t - \lambda t|$ directly rather than use (5.19). The result is that, if $r = \beta/\alpha$ and $t \ge 0$,

$$\begin{aligned} &d_{\mathrm{TV}}(\mathscr{L}N_t, \mathrm{Poisson}_{\lambda t}) \\ &\quad\le (r-1)\big[K_1(\lambda t)\{\sqrt{\lambda t + 2} + 1\}/(1+r) + K_2(\lambda t)\beta t r\big] \\ &\quad\le (r-1)[4/(1+r) + (1+r)r/2]. \end{aligned} \tag{5.21}$$

Note that the bound only depends on r, so that for any t, if r is close to 1, a good Poisson approximation is obtained.

The computation of an estimate for $E|A_t - \lambda t|$ follows the same idea as that for (5.12), but we use the cyclic nature of the λ's. The estimate is computed by first conditioning on Y, and as the argument is only notationally different for the two possible values of Y, we assume $Y = 0$. In this case, odd interevent times $(X_{2i-1}, i = 1,2,\dots)$ have exponential(α) distribution and even interevent times $(X_{2i}, i = 1,2,\dots)$ have exponential(β) distribution. Let S_i be the sum of $X_1, \dots, X_{2i}$, so that $S_1, S_2, \dots$ are the times of a renewal process with renewal distribution F, where F is the convolution of exponential(α) with exponential(β). Let σ be the index of the first renewal in this process after t and note that σ is a stopping time for the discrete filtration $\mathscr{G}_i =$

$\sigma(X_1, X_2, \ldots, X_{2i})$. Arguing as for equations (5.13)–(5.16),

$$E|A_t - \lambda t| \le E\left|\sum_{i=1}^{\sigma} (\alpha - \lambda) X_{2i-1} + (\beta - \lambda) X_{2i}\right| + E\left|(A(S_\sigma) - A_t) - \lambda(S_\sigma - t)\right|. \tag{5.22}$$

The first term of (5.22) equals

$$E\left|\sum_{i=1}^{\sigma} (\beta - \alpha)(\alpha X_{2i-1} - \beta X_{2i})/(\alpha + \beta)\right|. \tag{5.23}$$

Now, conditional on N_t even, the strong Markov property gives the distribution of $A(S_\sigma) - A_t$ as that of $\alpha X_1 + \beta X_2$ and the distribution of $\lambda(S_\sigma - t)$ as that of $\lambda(X_1 + X_2)$. The same results are true if N_t is odd. Hence the second term of (5.22) is dominated by the expected absolute value of one of the summands in (5.23). Thus, applying Wald's identity as before and using the fact that $\sigma \le N_t/2 + 1$,

$$E|A_t - \lambda t| \le (r - 1)\{\sqrt{\lambda t + 2} + 1\}/(r + 1). \tag{5.24}$$

Equation (5.24) and the penultimate paragraph in the Markov point process example 5.8 give the estimate (5.21).

Interestingly, for the alternating Poisson process the third term of (3.8) can actually be computed exactly using renewal theory. The distance d_s is related to an integral with respect to the renewal function associated to F, which can be evaluated from its Laplace transform. From this argument, whose details we omit, the bound (5.21) can be sharpened to

$$d_{\mathrm{TV}}(\mathscr{L}N_t, \mathrm{Poisson}_{\lambda t}) \le (r - 1)\left[K_1(\lambda t)\{\sqrt{\lambda t + 2} + 1\} + K_2(\lambda t)\{\lambda t - \gamma\}\right]/(r + 1), \tag{5.25}$$

where

$$\gamma = \lambda(1 - \exp(-(\alpha + \beta)t))/(\alpha + \beta).$$

Elementary calculus shows that for all values of r, α and $t > 0$, we have $\gamma > 0$. Moreover, for large mean, the bound (5.25) is asymptotically $2.4(r - 1)/(r + 1)$, by (2.9) and (2.11). The bound of K_1 for the function g in (2.10) can actually be improved: Barbour and Brown (1990) show that g must be in the range $\pm(1 \wedge 0.61\mu^{-1/2})$, thus improving the constant 1.4 in (2.9) to 0.61. This means that the asymptotic value of the bound (5.25), for large λt, can be sharpened to $1.6(r - 1)/(r + 1)$.

One of the reasons for including this example is that the Palm probability bound (3.2) can also be estimated in this case. We take $\mu = \lambda t$ so that (3.2) only has one term. Then it can be shown that

$$d_{\mathrm{TV}}(N_t, \mathrm{Poisson}_{\lambda t}) \le \tfrac{1}{2}K_2(\lambda t)\lambda t(r - 1)\{(1 + r)^2 - 1\}/(1 + r), \tag{5.26}$$

which can be compared with the compensator bound (5.25). In general, (5.26) is smaller for r close to 1: For large mean, (5.26) gives the bound $r(r-1)(2+r)/(2+2r)$, which improves on the compensator bound for $r < 1.29$. On the other hand, the two bounds are still remarkably close. For large λt and r close to 1, the ratio of the two bounds is $1.61/1.5$.

To demonstrate (5.26), in view of (3.2), it is enough to bound $d_W(\mathscr{L}N_t, \mathscr{L}(N^s - \delta_s)_t)$, for all s in $(0, t]$. This we do as follows. For fixed s, we construct $\hat{N}^{(1)}$ and $\hat{N}^{(2)}$, on a common probability space, in such a way that $\mathscr{L}\hat{N}^{(1)} = \mathscr{L}N$, $\mathscr{L}\hat{N}^{(2)} = \mathscr{L}N^s$ and $\hat{N}_t^{(1)} + 1$ is close to $\hat{N}_t^{(2)}$. To achieve this, $\hat{N}^{(1)}$ and $\hat{N}^{(2)}$ count the number of transitions in processes $\hat{Z}^{(1)}$ and $\hat{Z}^{(2)}$ which are suitably defined. Realise $\hat{Z}^{(1)}$ from the stationary Z-chain over the whole of $\mathbb{R}$, and independently choose $\hat{Z}_s^{(2)}$ to be 0 or 1 with probability $\frac{1}{2}$ [because the process N is stationary, the Palm probability can be calculated as in Baccelli and Brémaud (1980), (3.1.1), and it is then easy to see that $\frac{1}{2}$ is the correct value]. If $\hat{Z}_s^{(1)} = \hat{Z}_s^{(2)}$, we can define $\hat{Z}_u^{(1)} = \hat{Z}_u^{(2)}$, $u \geq s$, but this does not work for $u < s$ because α- and β-intervals alternate in the opposite order in the two processes. Instead, for $u < s$, let $V_k^{(i)}$, $i = 1, 2$, denote the time of the kth transition of $\hat{Z}^{(i)}$ *before* time s: We have defined $V_k^{(1)}$ by defining $\hat{Z}^{(1)}$, but we need to define $V_k^{(2)}$ for $k \geq 1$. For $j \geq 1$, set $V_{2j}^{(2)} = V_{2j}^{(1)}$ while constructing $V_{2j-1}^{(2)}$ in such a way that it is maximally coupled with $V_{2j-1}^{(1)}$, but has the correct conditional distribution given the common values of V_{2j-2} and V_{2j} [see Barbour, Holst and Janson (1992), Appendix, for the construction of maximal couplings]. If $\hat{Z}_s^{(1)} \neq \hat{Z}_s^{(2)}$, use an analogous coupling for the portion $u \geq s$ and define the processes to be the same for $u < s$. The construction clearly yields the required distributions for $\hat{N}^{(1)}$ and $\hat{N}^{(2)}$, and has the property that $|\hat{N}_t^{(1)} + 1 - \hat{N}_t^{(2)}| \leq 1$, equality being possible for $\hat{Z}_s^{(1)} = \hat{Z}_s^{(2)}$ if and only if $V_{2J-1}^{(1)} \neq V_{2J-1}^{(2)}$, where J is such that $(V_{2J}, V_{2J-2}) \ni 0$, or under analogous circumstances when $\hat{Z}_s^{(1)} \neq \hat{Z}_s^{(2)}$.

Now, direct calculation based on the exponential density function shows that

$$P\Big(V_{2j-1}^{(2)} \neq V_{2j-1}^{(1)} | \hat{Z}_s^{(1)} = \hat{Z}_s^{(2)}, V_{2j-2} - V_{2j} = l, J = j\Big) = \tanh(l(\beta - \alpha)/4)$$
$$\leq l(\beta - \alpha)/4.$$

In addition, the standard formula for the expected length of the interval containing zero in an equilibrium renewal process, evaluated for the renewal distribution F, shows that

$$E(V_{2J-2} - V_{2J}) = 2(1/\alpha + 1/\beta - 1/(\alpha + \beta)).$$

Hence

$$E|\hat{N}_t^{(1)} + 1 - \hat{N}_t^{(2)}| \leq \tfrac{1}{2}(\beta - \alpha)(1/\alpha + 1/\beta - 1/(\alpha + \beta)),$$

and estimate (5.26) follows.

5.27. A CHANGE HALFWAY. This example shows that both the first term and the third can be simultaneously of the critical order of magnitude. Let N be a simple point process on $\{1, 2, \ldots, 2m\}$ and write $I_j = N(\{j\})$. Let $I_1, \ldots, I_m$ be independent Be(p) random variables, and, conditional on $N_m = mp + x$, let $I_{m+1}, \dots, I_{2m}$ be independent Be($(p - \vartheta x)_+$) random variables. Write $W_1 = N_m$, $W_2 = N_{2m} - N_m$ and $\mu = EN_{2m}$. We consider the case in which m is large, $p = m^{-1/2}$ and $\vartheta = m^{-9/8}$.

To start with, we show that

$$d_{\mathrm{TV}}(\mathscr{L}N_{2m}, \mathrm{Poisson}_\mu) \ge cm\vartheta \asymp m^{-1/8},$$

for some $c > 0$, by comparing the two measures on the set $B = \{w: |w - \mu| \le (2mp)^{1/2}\}$. First, from elementary Chebyshev-type exponential inequalities for the Bi(m, p) distribution, it follows easily that

$$\mu = 2mp + o(m^{-1/8}),$$

so that it suffices to compare $\mathscr{L}N_{2m}$ with $\mathrm{Poisson}_{2mp}$, and that

$$P\big[|W_1 - mp| \ge (mp)^{1/2} \log m\big] = o(m^{-1/8}).$$

Next, the Berry–Esseen theorem shows that, because only the set B is being considered, replacing $\mathrm{Poisson}_{2mp}$, $\mathscr{L}W_1$ and $\mathscr{L}(W_2|W_1 = mp + x)$ by the normal distributions with the same means and variances introduces an error of at most $O((mp)^{-1/2}) = o(m^{-1/8})$ whenever $|x| \le (mp)^{1/2} \log m$. Finally, by considering the set $\{x: |x - \phi| < \sigma\}$, an elementary calculation shows that

$$d_{\mathrm{TV}}(\mathscr{N}(\phi, \sigma^2(1 + \eta)), \mathscr{N}(\phi, \sigma^2)) \asymp \eta(2\pi e)^{-1/2} \tag{5.28}$$

as $\eta \to 0$. This last implies that $\mathscr{L}(W_2|W_1 = mp + x)$ can be replaced by $\mathscr{N}(m(p - \vartheta x), mp(1 - p))$ with an error of

$$O(\vartheta x p^{-1}) = O(m^{-3/8} \log m) = o(m^{-1/8})$$

whenever $|x| \le (mp)^{1/2} \log m$. However, if $W_1^* \sim \mathscr{N}(mp, mp(1 - p))$ and, given $W_1^* = mp + x$, $W_2^* \sim \mathscr{N}(m(p - \vartheta x), mp(1 - p))$, the distribution of $W_1^* + W_2^*$ is $\mathscr{N}(2mp, mp(1 - p)[1 + (1 - m\vartheta)^2])$, so that, once again using the calculation which gave rise to (5.28),

$$|P[W_1^* + W_2^* \in B] - P[\mathscr{N}(2mp, 2mp) \in B]| \asymp p + m\vartheta \asymp m^{-1/8}.$$

Combining the various approximations leads to the desired result.

We now show that, in the bound on $d_{\mathrm{TV}}(\mathscr{L}N, \mathrm{Poisson}_\mu)$ given by Theorem 3.7, both the first and third terms are of order $m^{-1/8}$ (the second is clearly of order $m^{-1/2}$). For the first term, observe that

$$A_{2m} - 2mp = m(p - \vartheta(W_1 - mp))_+ - mp = -m\vartheta(W_1 - mp)$$

whenever $W_1 \le m^{1/2} + m^{5/8}$, from which it follows easily that

$$K_1(\mu)E|A_{2m} - \mu| \asymp (mp)^{-1/2} m\vartheta (mp)^{1/2} = m\vartheta = m^{-1/8}.$$

For the third, note that $d_W(\mathscr{P}_s, \mathscr{P}_s^-) = 0$ for $s > m$. Furthermore, since $d_W(\mathrm{Bi}(m, p), \mathrm{Bi}(m, q)) = m|p - q|$, and using (5.5),

$$K_2(\mu)E\int_0^m d_W(\mathscr{P}_s, \mathscr{P}_s^-)\, dN_s = K_2(\mu) \sum_{i=1}^m p(1 - p)(m\vartheta + o(m^{-1/8})),$$

where the order term arises only from the possibility that $W_1 > m^{1/2} + m^{5/8}$, and thus the third term is also of order $m\vartheta = m^{-1/8}$.

5.29. THE THIRD TERM CAN BE LARGEST. Let N be the simple point process with compensator A determined by the intensity

$$a(t) = \begin{cases} \lambda, & t \le 1, \\ \lambda I[N_t \in 2\mathbb{Z}], & t > 1, \end{cases}$$

so that $\mathscr{L}N_1 = \text{Poisson}_\lambda$, and $N_\infty = N_1 + I[N_1 \in 2\mathbb{Z}]$ is N_1 rounded up to the next odd integer. Then, if q denotes the probability that N_1 is even, $\mu = EN_\infty = \lambda + q$ and

$$d_{\mathrm{TV}}(\mathscr{L}N_\infty, \text{Poisson}_\mu) \ge \text{Poisson}_\mu(2\mathbb{Z}) \sim \tfrac{1}{2} \quad \text{as } \mu \to \infty.$$

Now the first term of (3.8) gives

$$K_1(\mu)E|A_\infty - \mu| = 2qe^{-q}K_1(\mu) \asymp \mu^{-1/2},$$

and the second term is zero. To compute the third term, observe that, for n even, f_n defined in Theorem 5.2 has support the odd integers and, for n odd, it has support the even integers. Now, from (5.5), d_s is always the distance between an f_n with n even and an f_n with n odd. Taking h to be the indicator of the even integers in (2.6), $d_W(\mathscr{P}_s, \mathscr{P}_s^-)$ is identically 1. Thus the third term is just

$$\mu K_2(\mu) \sim 1 \quad \text{as } \mu \to \infty,$$

and hence the factor $K_2(\mu)$ in the third term is sharp up to at most a factor of $\frac{1}{2}$.

Acknowledgments. The authors wish to thank the organisers of the 1989 Singapore Probability Conference for an opportunity to discuss this work. The first author gratefully acknowledges the hospitality of the University of Western Australia, and the second the hospitality of Universität Zürich.

REFERENCES

BACCELLI, F. and BRÉMAUD, P. (1980). *Palm Probabilities and Stationary Queues. Lecture Notes Statist.* **41**. Springer, New York.

BARBOUR, A. D. and BROWN, T. C. (1990). Stein's method and point process approximation. *Stochastic Process. Appl.* To appear.

BARBOUR, A. D. and EAGLESON, G. K. (1983). Poisson approximation for some statistics based on exchangeable trials. *Adv. in Appl. Probab.* **15** 585–600.

BARBOUR, A. D. and HALL, P. (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95** 473–480.

BARBOUR, A. D. and HOLST, L. (1989). Some applications of the Stein–Chen method for proving Poisson convergence. *Adv. in Appl. Probab.* **21** 74–90.

BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation*. Oxford Univ. Press.

BROWN, T. C. (1978). A martingale approach to the Poisson convergence of simple point processes. *Ann. Probab.* **6** 615–628.

BROWN, T. C. (1981). Compensators and Cox convergence. *Math. Proc. Cambridge Philos. Soc.* **90** 305–319.

BROWN, T. C. (1982). Poisson approximations and exchangeable random variables. In *Exchangeability in Probability and Statistics* 177–183. North-Holland, Amsterdam.

BROWN, T. C. (1983). Some Poisson approximations using compensators. *Ann. Probab.* **11** 726–744.

BROWN, T. C. and SILVERMAN, B. W. (1979). Rates of Poisson convergence for *U*-statistics. *J. Appl. Probab.* **16** 428–432.

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

CHOW, Y. S. and TEICHER, H. (1988). *Probability Theory: Independence, Interchangeability and Martingales*, 2nd ed. Springer, New York.

DELLACHERIE, C. and MEYER, P. A. (1978). *Probabilities and Potential*. North-Holland, Amsterdam.

DELLACHERIE, C. and MEYER, P. A. (1982). *Probabilities and Potential B*. North-Holland, Amsterdam.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

FREEDMAN, D. (1974). The Poisson approximation for dependent events. *Ann. Probab.* **2** 256–269.

JACOD, J. (1975). Multivariate point processes: Predictable projection, Radon–Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Gebiete* **31** 235–253.

KABANOV, Y. and LIPTSER, R. S. (1983). On convergence in variation of the distributions of multivariate point processes. *Z. Wahrsch. Verw. Gebiete* **63** 473–485.

KABANOV, Y., LIPTSER, R. S. and SHIRYAYEV, A. N. (1986). On the variation distance for probability measures defined on a filtered space. *Probab. Theory Related Fields* **71** 19–35.

KALLENBERG, O. (1976). *Random Measures*. Academic, London.

ROGERS, L. C. G. and WILLIAMS, D. (1987). *Diffusions, Markov Processes and Martingales* **2**. Wiley, Chichester.

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

VALKEILA, E. (1982). A general Poisson approximation theorem. *Stochastics* **7** 159–171.

VALKEILA, E. (1984). Studies in distributional properties of counting processes. Ph.D. dissertation, Dept. Mathematics, Univ. Helsinki.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT ZÜRICH
RÄMISTRASSE 74
CH-8001 ZÜRICH
SWITZERLAND

DEPARTMENT OF STATISTICS
UNIVERSITY OF MELBOURNE
PARKVILLE
VIC 3052
AUSTRALIA