Adv. Appl. Prob. **39,** 799–825 (2007)
Printed in Northern Ireland

r-SCAN STATISTICS OF A POISSON PROCESS WITH EVENTS TRANSFORMED BY DUPLICATIONS, DELETIONS, AND DISPLACEMENTS

CHINGFER CHEN [*] AND
SAMUEL KARLIN,[*] [**] *Stanford University*

Abstract

A stochastic model of a dynamic marker array in which markers could disappear, duplicate, and move relative to its original position is constructed to reflect on the nature of long DNA sequences. The sequence changes of deletions, duplications, and displacements follow the stochastic rules: (i) the original distribution of the marker array $\{\dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots\}$ is a Poisson process on the real line; (ii) each marker is replicated l times; replication or loss of marker points occur independently; (iii) each replicated point is independently and randomly displaced by an amount Y relative to its original position, with the Y displacements sampled from a continuous density $g(y)$. Limiting distributions for the maximal and minimal statistics of the r-scan lengths (collection of distances between $r + 1$ successive markers) for the l-shift model are derived with the aid of the Chen–Stein method and properties of Poisson processes.

Keywords: r-scan statistic; Poisson process; Chen–Stein method; Poisson approximation

2000 Mathematics Subject Classification: Primary 60H30
Secondary 60G70

1. Introduction

The motivation for the l-fold shift model analyzed in this paper stems from the dynamic and heterogenous nature of long DNA sequences. Genomic local and global compositional heterogeneity occurs on many scales. Examples of DNA heterogeneity include isochore compartments (regions dominated by either G+C or A+T nucleotides as determined by density-gradient centrifugation especially in mammalian species) (Bernardi *et al.* (1985), (1988)); mobile elements (DNA sequences that move around the genome such as *Alu elements* in human, *Ty* sequences in yeast, and *IS* segments in *Escherichia coli* (Berg and Howe (1989)); characteristic satellite centrometric tandem repeats (such as the 171-units of human alpha satellite DNA); characteristic telomeric sequences (at the chromosomal termini such as the TGTGGG tandem repeats in humans) (Willard and Waye (1987); Blackburn (1991)); CpG islands (human DNA sequences that occur generally upstream of genes and are abundant with unmethylated CG dinucleotides) (Bird (1986)); repetitive extragenic palindromes (REPs) found in the bacterial genomes of *Escherichia coli* and *Salmonella typhimurium*; recombinational hot spots (such as *chi* elements GCTGGTGG in *Escherichia coli*) (Krawiec and Riley (1990), Gilson *et al.* (1991)); almost universal under-representation of the dinucleotide TA; suppression of

Received 11 April 2006; revision received 21 June 2007.

[*] Postal address: Department of Mathematics, Stanford University, Stanford, CA 94305, USA.
[**] Email address: Karlin@math.stanford.edu

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

the dinucleotide CG in vertebrate species (Josse *et al.* (1961)); the rarity of the tetranucleotide CTAG in several proteobacterial and archaeal genomes (Burge *et al.* (1992), Reinert and Schbath (1998)). GNN periodicity in manmalian coding sequences (Ficket (1982)). Thus, genome organization is complex and variegated.

This paper develops a theoretical framework for ascertaining regions of clustering or overdispersion in a marker array (e.g. genes, oligonucletides, transposable elements, and nucleosomes) along a DNA sequence following long-term mutation events such as sequence deletions, duplications, displacements, and rearrangements. To this purpose we consider a Poisson point process model, where each marker is independently replicated a random number of times and the replicas are randomly displaced. The maximal and minimal r-scan lengths (r-scans consist of all distances between $r+1$ consecutive points of the marker array) are investigated to identify special inhomogeneous regions. We use multidimensional inhomogeneous Poisson processes in conjunction with the Chen–Stein methodology in characterizing extremal r-scans. Moreover, the Kingman mapping theorem (Kingman (1993, Chapter 5)) concerned with transformed Poisson processes in multidimensional spaces is used to achieve essential estimates.

The biological model discussed in this paper describes a stochastic version to these kinds of biological changes and obeys the following rules.

1. The original distribution of the marker array, $(\dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots)$, is a point process.
2. Each marker is independently replicated l times. Replications or loss of marker points occur independently.
3. Each replicated point is independently displaced by an amount Y $(-\infty < Y < \infty)$ relative to its original position, with the Y displacements sampled from the density $g(y)$, $-\infty < y < \infty$.

From the altered process, the r-scan statistics (see Dembo and Karlin (1992)), are the collection of the interval lengths between all $r+1$ successive marker points. The overdispersion and cluster regions of the marker array correspond to the regions containing the maximal and minimal r-scan lengths, respectively. The objective of this paper is to characterize the asymptotic distributions of the maximal and minimal r-scan lengths of the shift process. For previous literature and applications of r-scan statistics in molecular genetic analysis , see Karlin and Macken (1991), Dembo and Karlin (1992), Karlin and Brendel (1992), Karlin and Cardon (1994), Karlin *et al.* (1996), and Gerstein (1997). For studies of clustering in other domains with extensive bibliography, see Naus (1979), (1982) and the recent books of Barbour *et al.* (1992) and Glaz *et al.* (2001).

We will concentrate on the case in which A (the original ancestor marker array) is distributed as a homogeneous Poisson process of parameter 1. Let $\Pi_1^{(l)}$ denote the shift model constructed from a Poisson(1) process involving an l-fold replication and an independent displacement sampled from the density g applied to each replicated point. Thus, the $\Pi_1^{(l)}$ array consists of the points

$$\Pi_1^{(l)} = \{Z_i^k = X_i + Y_i^k; i = 0, \pm 1, \pm 2, \dots, k = 1, 2, \dots, l\}.$$

The asymptotic distributions of the extremal r-scan lengths descendant from the marker array $\Pi_1^{(l)}$ will be deduced from the l-dimensional inhomogeneous Poisson process Π_l^* which has the intensity $f_l(z_1, \dots, z_l) = \int_{-\infty}^{\infty} g(z_1 - s) \cdots g(z_l - s)\, \mathrm{d}s$. To clarify the ideas and constructions, the case in which $l = 2$ will be elaborated.

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

The analysis (lemmas and proofs) of the case in which $l > 2$ is omitted here, but is available online at http://math.stanford.edu/~karlin/ through the supplemental information link under the publications heading. Consider $l = 2$ duplications, producing

$$\ldots, \begin{pmatrix} X_{-1} \\ X_{-1} \end{pmatrix}, \begin{pmatrix} X_0 \\ X_0 \end{pmatrix}, \begin{pmatrix} X_1 \\ X_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ X_2 \end{pmatrix}, \ldots.$$

The displaced array is

$$\ldots, \begin{pmatrix} Z_{-1}^1 = X_{-1} + Y_{-1}^1 \\ Z_{-1}^2 = X_{-1} + Y_{-1}^2 \end{pmatrix}, \begin{pmatrix} Z_0^1 = X_0 + Y_0^1 \\ Z_0^2 = X_0 + Y_0^2 \end{pmatrix}, \begin{pmatrix} Z_1^1 = X_1 + Y_1^1 \\ Z_1^2 = X_1 + Y_1^2 \end{pmatrix}, \ldots,$$

where all $\{Y\}$ are independent and identically distributed samples from the density $g(y)$. We assume that $\int_{-\infty}^{\infty} |u| g(u)\,\mathrm{d}u < \infty$ and so $\int\int |u - v| g(u) g(v)\,\mathrm{d}u\,\mathrm{d}v < \infty$. It is convenient to introduce the two-dimensional process $(X_i + Y_i^1, X_i + Y_i^2)$, designated Π_2^*, which is an inhomogeneous Poisson process with the intensity rate $f_2(z_1, z_2) = \int_{-\infty}^{\infty} g(z_1 - s) g(z_2 - s)\,\mathrm{d}s$.

Theorem 1. (Asymptotic maximal r-scan for the 2-fold shift model.) *Let $\Pi_1^{(2)}$ be the 2-fold shift process. Assume that the shift length density $g(s)$ satisfies the condition $\int |s| g(s)\,\mathrm{d}s < \infty$. Then, for any fixed integer $r = 2p + 1$, where p is a nonnegative integer, the kth longest r-scan length of $\Pi_1^{(2)}$ in $(0, t)$, $M_{t,k}$, possesses the asymptotic distribution $(t \to \infty)$*

$$\lim_{t\to\infty} \Pr\left\{ M_{t,k} \le \ln t + \left\lfloor \frac{r-1}{2} \right\rfloor \ln\ln t + x \right\} = \sum_{j=0}^{k-1} \mathrm{e}^{-\lambda} \frac{\lambda^j}{j!},$$

where $\lfloor w \rfloor$ is the integer part of w and

$$\lambda = \frac{\exp\{-(x + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |u - v| g(u) g(v)\,\mathrm{d}u\,\mathrm{d}v)\}}{\lfloor (r-1)/2 \rfloor !}.$$

Theorem 2. (Asymptotic minimal r-scan for the 2-fold shift model.) *Let $\Pi_1^{(2)}$ be the 2-fold shift process. Assume that the shift density g is continuous. Then the kth smallest r-scan length of $\Pi_1^{(2)}$ in $(0, t)$, $m_{t,k}$, possesses the asymptotic distribution*

$$\lim_{t\to\infty} \Pr\left\{ m_{t,k} \ge \sqrt[r]{\frac{x}{t}} \right\} = \sum_{j=0}^{k-1} \mathrm{e}^{-\alpha} \frac{\alpha^j}{j!},$$

with

$$\alpha = (r+1)x \sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \frac{2^{r+1-2j} [f_2(0,0)]^j}{j!(r+1-2j)!}.$$

1.1. The l-fold shift model

For each marker duplicated l times and each replicate randomly displaced independently as before, we construct a corresponding l-dimensional inhomogeneous Poisson process Π_l^* with rate parameter

$$f_l(z_1, \ldots, z_l) = \int_{-\infty}^{\infty} g(z_1 - s) \cdots g(z_l - s)\,\mathrm{d}s.$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

For $1 \leq \nu \leq l$, let

$$\begin{aligned}
S_\nu(z) &= \{(x_1, \dots, x_l)\colon \text{ where the } \nu\text{th coordinate satisfies } 0 \leq x_\nu \leq z\},\\
\overline{S_\nu}(z) &= \text{the complement of } S_\nu(z),\\
T_\nu(z) &= \cap_{j=1}^{\nu} S_j(z) \cap_{j=\nu+1}^{l} \overline{S_j}(z).
\end{aligned}$$

We will establish the existence of the following limits

$$\begin{aligned}
c_\nu &= \lim_{z\to\infty} \int \cdots \int_{T_\nu(z)} f_l(x_1, \dots, x_l)\, \mathrm{d}x_1 \cdots \mathrm{d}x_l \quad \text{for } 1 \leq \nu \leq l-1,\\
c_l &= \lim_{z\to\infty} \biggl[z - \int \cdots \int_{T_l(z)} f_l(x_1, \dots, x_l)\, \mathrm{d}x_1 \cdots \mathrm{d}x_l \biggr].
\end{aligned}$$

Theorem 3. (Asymptotic maximal r-scan length of the l-fold shift model.) *Let $\Pi_1^{(l)}$ be the l-fold shift process. Suppose that $\int_{-\infty}^{\infty} |s| g(s)\, \mathrm{d}s < \infty$ and $r = lp + 1$, for p a nonnegative integer. Then the kth maximal r-scan length from $\Pi_1^{(l)}$ in $(0, t)$, $M_{t,k}^{(l)}$, possesses the asymptotic distribution*

$$\lim_{t\to\infty} \Pr\biggl\{ M_{t,k}^{(l)} \leq \ln t + \biggl\lfloor \frac{r-1}{l} \biggr\rfloor \ln\ln t + x \biggr\} = \sum_{j=0}^{k-1} \mathrm{e}^{-\lambda_{(l)}} \frac{\lambda_{(l)}^{j}}{j!},$$

with

$$\lambda_{(l)} = \exp\biggl\{ -\biggl(x + \sum_{\nu=1}^{l-1} \binom{l}{\nu} c_\nu - c_l \biggr) \biggr\} \Big/ \biggl\lfloor \frac{r-1}{l} \biggr\rfloor !.$$

Theorem 4. (Asymptotic minimal r-scan length of the l-fold shift model.) *Let $\Pi_1^{(l)}$ be the l-fold shift process. Suppose that the shift density g is continuous. Then the kth minimal r-scan length from $\Pi_1^{(l)}$ in $(0, t)$, $m_{t,k}^{(l)}$, possesses the asymptotic distribution*

$$\lim_{t\to\infty} \Pr\biggl\{ m_{t,k}^{(l)} \geq \sqrt[r]{\frac{x}{t}} \biggr\} = \sum_{j=0}^{k-1} \mathrm{e}^{-\alpha_{(l)}} \frac{(\alpha_{(l)})^j}{j!},$$

with

$$\alpha_{(l)} = (r+1)x \sum_{\substack{i_1, i_2, \dots, i_l \in \mathbb{Z}^+ \\ \sum_{\nu=1}^{l} \nu i_\nu = r+1}} \biggl(\prod_{\nu=1}^{l} \biggl(\binom{l}{\nu} f_\nu(0, \dots, 0) \biggr)^{i_\nu} \Big/ i_\nu! \biggr)$$

for $f_d(0, \dots, 0) = \int_{-\infty}^{\infty} [g(s)]^d\, \mathrm{d}s$.

2. The Chen–Stein method and transformed Poisson processes

We review first the Chen–Stein method (Chen (1975)), which provides the basic tool to determine the error bound between a sum of (dependent) Bernoulli random variables and its asymptotic Poisson law. In this paper we adopt the formulation of the Chen–Stein method from Arratia *et al.* (1989).

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Theorem 5. (Chen–Stein method.) *Let $\{Z_i\}$ be Bernoulli (p_i) random variables and $W = \sum_{i\in\Omega} Z_i$, where Ω is a finite or countable index set. Let $\mathcal{P}_\lambda$ denote the Poisson(λ) random variable and let $d(U, V)$ denote the total variation distance between the discrete distributions of U and V:*

$$d(U, V) = \sup_{\mathcal{A}}(\Pr\{U \in \mathcal{A}\} - \Pr\{V \in \mathcal{A}\}) \quad \textit{(where $\mathcal{A}$ is any measurable set)}$$
$$= \frac{1}{2}\sum_{k=0}^{\infty} |\Pr\{U = k\} - \Pr\{V = k\}|.$$

Then

$$d(W, \mathcal{P}_\lambda) \le (u_1 + u_2)\frac{1 - \mathrm{e}^{-\lambda}}{\lambda} + u_3 \min\left(1, \frac{\sqrt{2}}{\sqrt{\lambda}}\right),$$

where

$$\begin{aligned} \lambda &= \sum_{i\in\Omega} p_i, & u_1 &= \sum_{i\in\Omega}\sum_{j\in\mathcal{B}_i} p_i p_j, \\ u_2 &= \sum_{i\in\Omega}\sum_{j\in\mathcal{B}_i, j\neq i} \mathrm{E}[Z_i Z_j], & u_3 &= \sum_{i\in\Omega} \mathrm{E}[|\,\mathrm{E}[Z_i \mid \{Z_j\}_{j\notin B_i}] - p_i|], \end{aligned} \tag{1}$$

and $\{\mathcal{B}_i\}$ is an appropriate family of subsets indexed by Ω.

Theorems 6 and 7 are fundamental for inhomogeneous Poisson processes; see, e.g. Kingman (1993, Chapter 5).

Theorem 6. *Let Π be a Poisson process on a space S with rate measure μ. Suppose that, with each point X of Π, we associate a random variable m_X (the mark of X) taking values in some metric space M. The distribution of m_X may depend on X but not on other points of Π, and m_X for different X are independent.*

The pair (X, m_X) can be regarded as a random point X^* in the product space $S \times M$. Then, the ensemble of points X^* generate a Poisson process $\Pi^* = \{(X, m_X)\}_{X\in\Pi}$ on the direct product space $S \times M$ with rate measure μ^* given by

$$\mu^*(C) = \iint_{(x,m)\in C} \mu(\mathrm{d}x) p_x(\mathrm{d}m), \tag{2}$$

where $p_x(\mathrm{d}m)$ is the conditional distribution of m given x.

To adapt Theorem 6 to the 2-fold shift model, we determine $m_X \equiv (m_1, m_2) = (X + Y_X^1, X + Y_X^2)$, with the displacements Y_X^1 and Y_X^2 arising as independent, real valued, random variables sampled from the density $g(y)$. Then $p_x(\mathrm{d}m)$ applied in (2) is $p_x(\mathrm{d}m) = g(m_1 - x)g(m_2 - x)\,\mathrm{d}m_1\,\mathrm{d}m_2$. The Poisson process Π^* has the rate measure μ^* on $\mathbb{R} \times \mathbb{R}^2$ calculated as

$$\mu^*(C) = \iiint_{(x,m_1,m_2)\in C} g(m_1 - x)g(m_2 - x)\,\mathrm{d}m_1\,\mathrm{d}m_2\,\mathrm{d}x$$

for any measurable set C on $\mathbb{R}^3$.

Theorem 7. (Mapping theorem.) *Let Π be a Poisson process with a finite mean rate measure μ on the space S, and let $\Lambda\colon S \to T$ be a measurable mapping such that the induced measure of μ transferred to T is atomless. Then $\tilde{\Pi} = \Lambda(\Pi)$ is a Poisson process on T and has rate measure $\tilde{\mu} = \mu * \Lambda^{-1}$.*

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

In the 2-fold shift model, Λ in Theorem 7 is specified as the projection of $\mathbb{R}^3$ to $\mathbb{R}^2$ such that $\Lambda\colon (x, m_1, m_2) \to (m_1, m_2)$, and therefore the induced process

$$\Pi_2^* = \Lambda(\{X, X + Y_X^1, X + Y_X^2\}) = \{(X + Y_X^1, X + Y_X^2)\}$$

is a two-dimensional Poisson process with the rate measure $f_2(z_1, z_2) = \int_{-\infty}^{\infty} g(z_1 - s) \times g(z_2 - s)\, \mathrm{d}s$. We assume that the displacement distribution density g has finite mean. Then f_2 possesses the following three properties.

1. Symmetry: $f_2(z_1, z_2) = f_2(z_2, z_1)$.
2. Invariance under equal translation: $f_2(z_1 + a, z_2 + a) = f_2(z_1, z_2)$ for all real a.
3. $f_2(z, 0)$ (or $f_2(0, z)$) is a continous, symmetric density function of z such that

$$\int_{-\infty}^{\infty} |z| f_2(z, 0)\, \mathrm{d}z = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u - v| g(u) g(v)\, \mathrm{d}u\, \mathrm{d}v < \infty.$$

To construct the relevant one-dimensional shift process $\Pi_1^{(2)}$, we project the two-dimensional Poisson process Π_2^* separately to its z_1-axis and to its z_2-axis and concatenate the two one-dimensional point processes yielding the one-dimensional process

$$\Pi_1^{(2)} = \{\{X + Y_X^1\} \cup \{X + Y_X^2\}\}_{(X+Y_X^1, X+Y_X^2)\in\Pi_2^*}. \tag{3}$$

Theorems 6 and 7 will enable us to study the distributional properties of the two-dimensional Poisson process Π_2^* and, subsequently, to calculate the distribution of the one-dimensional shift process $\Pi_1^{(2)}$. For example, consider an interval (a, b), for $a < b$. The two events

$$\{\text{no one-dimensional } \Pi_1^{(2)} \text{ point occurs in an interval } (a, b)\}$$

and

$$\{\text{no two-dimensional } \Pi_2^* \text{ point occurs in a region of } \{(a, b) \times (-\infty, \infty)\} \cup \{(-\infty, \infty) \times (a, b)\}\}$$

are equivalent. The following notations indicate the regions in Π_2^* associated with the interval (a, b) in $\Pi_1^{(2)}$. Let '\' denote set subtraction and

$$A(a, b) = \{(a, b) \times (-\infty, \infty) \cup (-\infty, \infty) \times (a, b)\},$$
$$A_1(a, b) = \{(a, b) \times (a, b)\}, \quad \text{and} \quad A_2(a, b) = A(a, b) \setminus A_1(a, b); \tag{4}$$

see Figure 1.

Each point of Π_2^* in $A_1(a, b)$ corresponds to two points of $\Pi_1^{(2)}$ in (a, b) (precluding the points along the diagonal), whereas each point of Π_2^* in $A_2(a, b)$ generates a single point of $\Pi_1^{(2)}$ in (a, b). Let $V(C)$ equal the integration of $f_2(z_1, z_2)$ over a set C. Since the rate density $f_2(z_1, z_2)$ of Π_2^* is invariant along the diagonal, it is clear that

$$V(A_1(a, b)) = V(A_1(0, b - a)) \quad \text{and} \quad V(A_2(a, b)) = V(A_2(0, b - a)).$$

Actually, $\Pi_1^{(2)}$ is a stationary point process on the line.

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

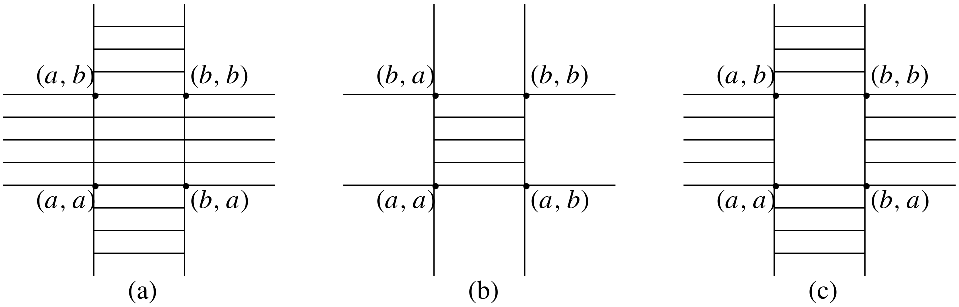

FIGURE 1: The stripped regions correspond to (a) $A(a, b)$, (b) $A_1(a, b)$, and (c) $A_2(a, b)$.

3. Estimates required for the multiple shift model

In this section we provide the estimates required to prove Theorems 1–4. The detailed proof for each estimate is presented in Section 5 of this paper. The analysis of the r-scan statistics for the observed marker array $\Pi_1^{(2)}$ in (3) is based on the counts of associated Bernoulli variables ((6) and (7), below, over the time horizon $(0, t)$). To study the maximum r-scan length, we partition $(0, t)$ with a small spacing $\triangle_t = 1/t$ and place a window of width

$$b_t = \ln t + \left\lfloor \frac{r-1}{2} \right\rfloor \ln \ln t + x \tag{5}$$

at each position $j\triangle_t,\ j = 0, 1, 2, \dots$. Now, let

$$Z_j^+(b_t) = \begin{cases} 1 & \text{if there is a single marker of } \Pi_1^{(2)} \text{ in } ((j-1)\triangle_t, j\triangle_t), \text{ and} \\ & \text{less than } r \text{ markers in the window of } (j\triangle_t, j\triangle_t + b_t), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Then $Z_j^+(b_t) = 1$ signifies the existence of an r-scan interval of length exceeding b_t, where its interval begins about $j\triangle_t$. We define $n_t^+(b_t) = \sum_{j=0}^{\lfloor (t-b_t)/\triangle_t \rfloor} Z_j^+(b_t)$ and prove that $n_t^+(b_t)$ is a good approximation of $N_t^+(b_t)$, the count of r-scan intervals in $(0, t)$ that exceed b_t. Theorem 5 can be applied to derive the asymptotic Poisson law for $n_t^+(b_t)$.

The distribution of the minimum r-scan length is studied in a similar manner by partitioning $(0, t)$ with a spacing of $\delta_t = 1/t^2$ and by putting a window of extent $a_t = \sqrt[r]{x/t}$ at each discrete position $j\delta_t,\ j = 0, 1, 2, \dots$. A Bernoulli random variable is specified at each position $j\delta_t$:

$$Z_j^-(a_t) = \begin{cases} 1 & \text{if there is a single marker of } \Pi_1^{(2)} \text{ in } ((j-1)\delta_t, j\delta_t), \text{ and} \\ & \text{at least } r \text{ markers in } (j\delta_t, (j-1)\delta_t + a_t), \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

When $Z_j^-(a_t) = 1$, there is an r-scan interval of length less than a_t with its initial marker about $j\delta_t$. Let $n_t^-(a_t) = \sum_{j=0}^{\lfloor (t-a_t)/\delta_t \rfloor} Z_j^-(a_t)$ and let $N_t^-(a_t)$ be the count of r-scan intervals in $(0, t)$ that do not exceed a_t. Then $n_t^-(a_t)$ is a candidate to represent $N_t^-(a_t)$ such that $n_t^-(a_t)$ converges to $N_t^-(a_t)$ in probability as $t \to \infty$. The asymptotic Poisson law of $n_t^-(a_t)$ can be ascertained by Theorem 5.

Before calculating probabilities of events of the process $\Pi_1^{(2)}$, we describe concisely the method. Since Π_2^* is a two-dimensional Poisson process, realizations in disjoint areas are

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

independent. But points occurring in disjoint intervals of the one-dimensional $\Pi_1^{(2)}$, say, (a, b) and (c, d), have some overlap through the rectangle areas $\{(a, b)\times(c, d)\}$ and $\{(c, d)\times(a, b)\}$ of Π_2^*, in which each point could project to two one-dimensional points in the intervals (a, b) and (c, d) each. For convenience, we use the notation $A(a, b)$, $A_1(a, b)$, and $A_2(a, b)$ (see Figure 1) to represent regions of Π_2^* relevant to an interval (a, b) in $\Pi_1^{(2)}$. We define $B((a, b)\times(c, d)) = \{\{(a, b) \times (c, d)\} \cup \{(c, d) \times (a, b)\}\}$, which is the area relative to both intervals (a, b) and (c, d). We use $||\ ||$ to indicate the count of points in regions of both one-dimensional or two-dimensional and higher dimensional; e.g. $||C||$ and $||(a, b)||$ are the count of points of Π_2^* in area C and the count of points of $\Pi_1^{(2)}$ in the interval (a, b), respectively.

In this section we state the upper bounds of errors when applying the Chen–Stein method to the shift processes. The proofs are given in Section 5. Lemmas 1 and 2, below, are necessary to calculate the values of $\{\mathrm{E}[Z_j^+(b_t)]\}_{j\geq 0}$, the expectations of the Bernoulli random variables of maximal r-scan lengths.

Lemma 1. *Assume that* $\int |s|g(s)\,\mathrm{d}s < \infty$. *Then, for each nonnegative integer* k *and* $z \to \infty$,

$$\begin{aligned}&\Pr\{\textit{There are at most } k \textit{ points of } \Pi_1^{(2)} \textit{ in } (0, z)\}\\&\quad = \exp\{-(z + m_2)\}\frac{(z - m_2)^{\lfloor k/2\rfloor}}{\lfloor k/2\rfloor!} I_2(k)(1 + o(1)),\end{aligned}$$

where $o(1)$ *converges to* 0 *as* $z \to \infty$, *and*

$$m_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u - v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v < \infty; \qquad I_2(k) = \begin{cases}1, & k \textit{ is even},\\ 1 + 2m_2, & k \textit{ is odd}.\end{cases} \tag{8}$$

Lemma 2. *Assume that* $\int |s|g(s)\,\mathrm{d}s < \infty$ *and* $r = 2p+1$, $p = 0, 1, 2, \ldots$. *Then, for* $\triangle_t = t^{-1}$ *and* $b_t = \ln t + \lfloor(r - 1)/2\rfloor \ln\ln t + x$,

$$\begin{aligned}\mathrm{E}[Z_1^+(b_t)] &= \Pr\{\textit{there is a single point of } \Pi_1^{(2)} \textit{ in } (0, \triangle_t) \textit{ and at most}\\&\qquad (r - 1) \textit{ points of } \Pi_1^{(2)} \textit{ in } (\triangle_t, \triangle_t + b_t)\}\\&= \frac{\triangle_t \exp\{-(x + m_2)\}}{t\lfloor(r - 1)/2\rfloor!}(1 + o(1)).\end{aligned}$$

The following two lemmas are required for the study of the maximal r-scan distribution. Lemma 3 is necessary to evaluate the probability of the event $\{n_t^+(b_t) \neq N_t^+(b_t)\}$. Lemma 4 is required for the calculation of the error bound of u_2 in (1) with the neighborhood sets $\mathcal{B}_{i\{i\geq 1\}}$ specified as $\mathcal{B}_i = \{j : |j - i| < \lfloor 2b_t/\triangle_t\rfloor\}$ when invoking the Chen–Stein method applied to the Bernoulli sum of $n_t^+(b_t) = \sum_{j=0}^{\lfloor(t-b_t)/\triangle_t\rfloor} Z_j^+(b_t)$. Here, and throughout the paper, we consider only the index i, j such that $i\triangle_t$, $j\triangle_t$ are within the interval $(0,\ t)$.

Lemma 3. *The following estimates assure convergence in probability of* $n_t^+(b_t)$ *to* $N_t^+(b_t)$.

$$\Pr\{||(0, \triangle_t)|| \geq 2, ||(\triangle_t, \triangle_t + b_t)|| \leq r - 1\} \leq O\bigg(\frac{\triangle_t^2}{t}\bigg), \tag{9}$$

$$\Pr\{\ ||(0, \triangle_t)|| = 1, ||(0, b_t)|| \leq r, ||(0, \triangle_t + b_t)|| \geq r + 1\} \leq O\bigg(\frac{\triangle_t^2}{t}\bigg). \tag{10}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Lemma 4. *The following estimate is necessary for evaluating the parameter u_2 in (1) when the Chen–Stein method is applied to $n_t^+(b_t)$ by setting $\triangle_t = 1/t$. We obtain*

$$\sum_{j\in B_i,\, j\neq i} \mathrm{E}[Z_i^+(b_t)Z_j^+(b_t)] \le 2\frac{\ln\ln t}{\triangle_t} O\biggl(\frac{\triangle_t^2}{t\ln t}\biggr) + 2\frac{b_t}{\triangle_t} O\biggl(\frac{\triangle_t^2(\ln t\ln t)^{\lfloor (r-1)/2\rfloor}}{t(\ln t)^2}\biggr) + 2\frac{2b_t}{\triangle_t} O\biggl(\frac{\triangle_t^2}{t^2}\biggr). \tag{11}$$

Lemmas 5, 6, and 7 are important for the proof of the asymptotic Poisson law when applying the Chen–Stein method to the Bernoulli sum $\sum_{i=1}^{\lfloor (t-a_t)/\delta_t\rfloor} Z_i^-(a_t)$. Lemma 5 provides the asymptotic value of $\mathrm{E}[Z_i^-(a_t)]$. Lemma 6 gives estimates relevant to the convergence of $n_t^-(a_t)$ to $N_t^-(a_t)$ in probability. Lemma 7 provides an upper bound of u_2 in (1) when applying the Chen–Stein method to $n_t^-(a_t)$.

Lemma 5. *Assume that the shift density g is continuous, then, for $\delta_t = t^{-2}$, $a_t = \sqrt[r]{x/t}$, and $t\to\infty$,*

$$\begin{aligned}\mathrm{E}[Z_1^-(a_t)] &\equiv \Pr\{||(0,\delta_t)|| = 1, ||(\delta_t,a_t)|| \ge r\}\\ &= \frac{(r+1)\delta_t x}{t}\Biggl(\sum_{j=0}^{\lfloor (r+1)/2\rfloor} \frac{2^{r+1-2j}(f_2(0,0))^j}{j!\,(r+1-2j)!}\Biggr)(1+o(1)).\end{aligned}$$

Lemma 6. *The probability bounds of the events $\{||(0,\delta_t)|| \ge 2, ||(\delta_t,a_t)|| \ge r-1\}$ and $\{||(0,\delta_t)|| = 1, ||(0,a_t)|| < r+1, ||(0,a_t+\delta_t)|| \ge r+1\}$ are respectively*

$$\Pr\{||(0,\delta_t)|| \ge 2, ||(\delta_t,a_t)|| \ge r-1\} \le O(\delta_t^2), \tag{12}$$

$$\Pr\{||(0,\delta_t)|| = 1, ||(0,a_t)|| < r+1, ||(0,a_t+\delta_t)|| \ge r+1\} \le O(\delta_t^2). \tag{13}$$

Lemma 7. *The following estimate is necessary for evaluating the parameter u_2 of Theorem 5 when the Chen–Stein method is applied to $n_t^-(a_t)$. For $2 \le i \le \lfloor a_t/\delta_t\rfloor + 1$,*

$$\mathrm{E}[Z_1^-(a_t)Z_i^-(a_t)] \le O(\delta_t\delta_t a_t^r). \tag{14}$$

With the preparations above, we are ready to validate the limiting theorems of the extremal r-scan lengths of the 2-fold shift model.

4. Theorems for extremal *r*-scan lengths

4.1. Extremal distribution of maximum and minimum *r*-scans

The asymptotic distribution of the kth largest r-scan length generated from the 2-fold shift model arises from the sum of Bernoulli random variables associated with the set of discrete times $\{j\triangle_t\}$, $j = 1,\dots,\lfloor (t-b_t)/\triangle_t\rfloor$, for the choices of $\triangle_t = t^{-1}$ and $b_t = \ln t + \lfloor (r-1)/2\rfloor \ln\ln t + x$. Explicitly, for $1 \le j \le \lfloor (t-b_t)/\triangle_t\rfloor$, we define the following Bernoulli random variables:

$$Z_j^+(b_t) = \begin{cases} 1 & \text{if } ||((j-1)\triangle_t, j\triangle_t)|| = 1 \text{ and } ||(j\triangle_t, j\triangle_t + b_t)|| < r,\\ 0 & \text{otherwise.}\end{cases}$$

We claim that the Bernoulli sum $n_t^+(b_t) = \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t\rfloor} Z_j^+(b_t)$ is a good approximation to the count $N_t^+(b_t)$ of the number of r-scan intervals in $(0,t)$ based on the points of $\Pi_1^{(2)}$ whose

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

r-scans exceed b_t. Let $\{X_{t,i}\}_{i\geq 1}$ be the ordered points of $\Pi_1^{(2)}$ in $(0, t)$ and set $X_{t,0} = 0$ for definiteness. Then $\{R_{t,i} = X_{t,i-1+r} - X_{t,i-1}\}_{i\geq 1}$ are the successive r-scan segments along the line. Let $\{M_{t,1}, M_{t,2}, M_{t,3}, \dots\}$ be the order statistics for $\{R_{t,i}\}_{i\geq 1}$ in decreasing order. That is, $M_{t,1}$ is the largest r-scan length of $\Pi_1^{(2)}$ in $(0, t)$ and $M_{t,k}$ is the kth largest r-scan length. The duality relation guarantees that

$$\{M_{t,k} \leq b_t\} = \{N_t^+(b_t) \leq k - 1\}. \tag{15}$$

If $\int |s| g(s)\,\mathrm{d}s < \infty$ and $r = 2p + 1$, for some nonnegative integer p, we will prove that the Bernoulli sum, $n_t^+(b_t)$, possesses the following two properties.

Property 1. $\lim_{t\to\infty} \Pr\{n_t^+(b_t) \neq N_t^+(b_t)\} = 0$.

Property 2. *$n_t^+(b_t)$ is asymptotically Poisson(λ) with*

$$\lambda = \frac{\exp\{-(x + m_2)\}}{\lfloor (r-1)/2 \rfloor!} \quad \textit{for } b_t \textit{ as defined in (5) and } m_2 \textit{ as defined in (8).}$$

With Properties 1 and 2, Theorem 1 can be proved as follows.

Proof of Theorem 1. The duality relation, (15), gives

$$\begin{aligned}\lim_{t\to\infty} \Pr\{M_{t,k} \leq b_t\} &= \lim_{t\to\infty} \Pr\{N_t^+(b_t) \leq k - 1\} \\ &= \lim_{t\to\infty} \Pr\{n_t^+(b_t) \leq k - 1\} \quad \text{(from Property 1)} \\ &= \sum_{j=0}^{k-1} \mathrm{e}^{-\lambda} \frac{\lambda^j}{j!} \quad \text{(from Property 2)},\end{aligned}$$

as described in Theorem 1. We now prove Properties 1 and 2.

Proof of Property 1. Under the condition of at least r points of $\Pi_1^{(2)}$ in $(t - b_t, t)$, a count $Z_j^+(b_t) = 1$ will not show an r-segment extending over the position t. $Z_j^+(b_t) = 1$ signifies a count of an r-scan segment in $(0, t)$ with a single point in $((j - 1)\triangle_t, j\triangle_t)$ and with length exceeding b_t. Therefore, $\{||(t - b_t, t)|| \geq r\} \subseteq \{n_t^+(b_t) \leq N_t^+(b_t)\}$, which implies that

$$\{n_t^+(b_t) > N_t^+(b_t)\} \subseteq \{||(t - b_t, t)|| \leq r - 1\}. \tag{16}$$

Thus,

$$\begin{aligned}\Pr\{n_t^+(b_t) > N_t^+(b_t)\} &\leq \Pr\{||(t - b_t, t)|| \leq r - 1\} \quad \text{(according to (16))} \\ &= \exp\{-(b_t + m_2)\} \frac{(b_t - m_2)^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r-1)/2 \rfloor!} (1 + o(1)) \\ &= O\left(\frac{1}{t}\right) \quad \text{(by Lemma 1 and substituting} \\ &\qquad\qquad b_t = \ln t + \lfloor (r-1)/2 \rfloor \ln \ln t + x).\end{aligned} \tag{17}$$

Conversely, the event $\{N_t^+(b_t) > n_t^+(b_t)\}$ occurs only if there exists a $j, 1 \leq j \leq \lfloor (t - b_t)/\triangle_t \rfloor$, such that one of the following two cases hold.

Case 1. $\|((j - 1)\triangle_t, j\triangle_t)\| \geq 2$ *and* $||(j\triangle_t, j\triangle_t + b_t)|| \leq r - 1$.

Case 2. $||((j-1)\triangle_t, j\triangle_t)|| = 1$ *and* $||(j\triangle_t, (j-1)\triangle_t + b_t)|| \le r-1$ *and* $||(j\triangle_t, j\triangle_t + b_t)|| \ge r$.

For Case 1, $Z_j^+(b_t) = 0$ and there is at least one r-scan segment starting within the interval $((j-1)\triangle_t, j\triangle_t)$ with length exceeding b_t. Case 2 occurs when $Z_j^+(b_t) = 0$ and there is an r-scan interval with its initial marker within the interval $((j-1)\triangle_t, j\triangle_t)$ and its last marker within the interval $((j-1)\triangle_t + b_t, j\triangle_t + b_t)$ and its length exceeds b_t.

As shown in (9) and (10), we have the following estimates:

$$\Pr\{\text{Case 1 occurs for index } j\} = O\left(\frac{\triangle_t^2}{t}\right), \qquad \Pr\{\text{Case 2 occurs for index } j\} = O\left(\frac{\triangle_t^2}{t}\right).$$

Thus,

$$\begin{aligned}\Pr\{n_t^+(b_t) < N_t^+(b_t)\} &= \Pr\{\text{Case 1 or Case 2 occurs for some index } j\}\\ &\le \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \Pr\{\text{Case 1 occurs for index } j\}\\ &\quad + \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \Pr\{\text{Case 2 occurs for index} j\}\\ &= O\left(\frac{t}{\triangle_t}\right)O\left(\frac{\triangle_t^2}{t}\right) + O\left(\frac{t}{\triangle_t}\right)O\left(\frac{\triangle_t^2}{t}\right)\\ &= O(\triangle_t)\\ &= O\left(\frac{1}{t}\right) \quad \text{(by (9) and (10)).} \end{aligned} \tag{18}$$

Therefore, by (17) and (18), we have

$$\Pr\{N_t^+(b_t) \ne n_t^+(b_t)\} = \Pr\{N_t^+(b_t) < n_t^+(b_t)\} + \Pr\{N_t^+(b_t) > n_t^+(b_t)\} = O\left(\frac{1}{t}\right).$$

This completes the proof of Property 1.

Proof of Property 2. Assuming that $\int |s| g(s)\,\mathrm{d}s < \infty$ and $r = 2p+1$, for some nonnegative integer p, we apply the Chen–Stein method to verify the asymptotic Poisson law of $n_t^+(b_t) = \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} Z_j^+(b_t)$. Following the result of Lemma 2 and the stationarity of $\Pi_1^{(2)}$, we have

$$\mathrm{E}[Z_j^+(b_t)] = \mathrm{E}[Z_1^+(b_t)] = \frac{\triangle_t \exp\{-(x+m_2)\}}{\lfloor (r-1)/2 \rfloor!\, t}(1+o(1)). \tag{19}$$

Therefore,

$$\lambda_t \equiv \mathrm{E}[n_t^+(b_t)] = \left\lfloor \frac{t-b_t}{\triangle_t} \right\rfloor \mathrm{E}[Z_1^+(b_t)] = \lambda(1+o(1)) \quad \text{for } \lambda = \frac{\exp\{-(x+m_2)\}}{\lfloor (r-1)/2 \rfloor!}. \tag{20}$$

To demonstrate the Poisson approximation, we construct an index neighborhood subset $\mathcal{B}_j$ for each j, $\mathcal{B}_j = \{i : |i-j| \le \lfloor 2b_t/\triangle_t \rfloor\}$. According to the Chen–Stein protocol

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

(see Barbour *et al.* (1992)), the upper bound for the total variational distance between $n_t^+(b_t)$ and a Poisson random variable $\mathrm{Po}(\lambda)$ is

$$\begin{aligned} d(n_t^+(b_t), \mathrm{Po}(\lambda)) &\le d(n_t^+(b_t), \mathrm{Po}(\lambda_t)) + d(\mathrm{Po}(\lambda_t), \mathrm{Po}(\lambda)) \\ &\le (u_1+u_2)\biggl(\frac{1-\mathrm{e}^{-\lambda_t}}{\lambda_t}\biggr) + u_3 \min(1, \sqrt{\lambda_t}) + |\lambda - \lambda_t|, \end{aligned}$$

where

$$u_1 = \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \sum_{i\in\mathcal{B}_j} \mathrm{E}[Z_j^+(b_t)]\,\mathrm{E}[Z_i^+(b_t)], \qquad u_2 = \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \sum_{i\in\mathcal{B}_j,\, i\ne j} \mathrm{E}[Z_j^+(b_t) Z_i^+(b_t)],$$

$$\text{and} \quad u_3 = 2 \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \mathrm{E}[Z_j^+(b_t)]\, d\biggl(\biggl(\sum_{i\notin\mathcal{B}_j} Z_i^+(b_t) \biggm| Z_j^+(b_t)=1\biggr), \sum_{i\notin\mathcal{B}_j} Z_i^+(b_t)\biggr).$$

Estimates of u_1, u_2, and u_3 are assessed by evaluation of corresponding areas of the two-dimensional process Π_2^*. We use Lemma 2 and Lemma 4 to provide the following estimates:

$$\begin{aligned} u_1 &= \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \sum_{i\in B_j} \mathrm{E}[Z_j^+(b_t)]\,\mathrm{E}[Z_i^+(b_t)] \\ &\le 2\biggl(\frac{t-b_t}{\triangle_t}\biggr)\biggl(\frac{2b_t}{\triangle_t}\biggr)(\mathrm{E}[Z_1^+(b_t)])^2 \quad \text{(by stationarity)} \\ &\le 4O\biggl(\frac{tb_t}{\triangle_t^2}\biggr)\biggl(\frac{\triangle_t}{t}\lambda\biggr)^2 \quad \biggl(\text{by (19), (20), and } \lambda = \frac{\exp\{-(x+m_2)\}}{\lfloor (r-1)/2 \rfloor !}\biggr) \\ &\le 4O\biggl(\frac{\ln t}{t}\biggr)\lambda^2, \end{aligned}$$

$$\begin{aligned} u_2 &= \sum_{j=1}^{\lfloor (t-b_t)/\triangle_t \rfloor} \sum_{i\in B_j,\, i\ne j} \mathrm{E}[Z_j^+(b_t) Z_i^+(b_t)] \\ &\le 2\frac{t}{\triangle_t} \sum_{i=2}^{\lfloor (2b_t)/\triangle_t \rfloor + 1} \mathrm{E}[Z_1^+(b_t) Z_i^+(b_t)] \quad \text{(by stationarity)} \\ &= 2\frac{t}{\triangle_t}\Biggl\{ \sum_{i=2}^{\lfloor \ln\ln t/\triangle_t \rfloor + 1} \mathrm{E}[Z_1^+(b_t) Z_i^+(b_t)] + \sum_{\lfloor \ln\ln t/\triangle_t \rfloor + 2}^{\lfloor b_t/\triangle_t \rfloor + 1} \mathrm{E}[Z_1^+(b_t) Z_i^+(b_t)] \\ &\qquad + \sum_{\lfloor b_t/\triangle_t \rfloor + 2}^{\lfloor 2b_t/\triangle_t \rfloor + 1} \mathrm{E}[Z_1^+(b_t) Z_i^+(b_t)] \Biggr\} \\ &\le 2\frac{t}{\triangle_t}\biggl(\frac{\ln\ln t}{\triangle_t} O\biggl(\frac{\triangle_t^2}{t\ln t}\biggr) + \frac{b_t}{\triangle_t} O\biggl(\frac{\triangle_t^2 (\ln\ln t)^{(r-1)/2}}{t(\ln t)^2}\biggr) + \frac{2b_t}{\triangle_t} O\biggl(\frac{\triangle_t^2}{t^2}\biggr)\biggr) \quad \text{(see (11))} \\ &= O\biggl(\frac{\ln\ln t}{\ln t} + \frac{(\ln\ln t)^{(r-1)/2}}{\ln t}\biggr). \end{aligned}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

To study the convergence property of u_3, for $j \le \lfloor 2b_t/\triangle_t \rfloor$, let $E_{j,1}$ and $E_{j,2}$ be sets in $\mathbb{R}^2$ determined as

$$E_{j,1} = \left\{((j-1)\triangle_t,\, j\triangle_t + b_t) \times \left(\left(j - 1 + \left\lfloor \frac{2b_t}{\triangle_t} \right\rfloor\right)\triangle_t, \infty\right)\right\},$$
$$E_{j,2} = \left\{\left(\left(j - 1 + \left\lfloor \frac{2b_t}{\triangle_t} \right\rfloor\right)\triangle_t, \infty\right) \times ((j-1)\triangle_t,\, j\triangle_t + b_t)\right\}.$$

And, for $\lfloor 2b_t/\triangle_t \rfloor < j \le \lfloor (t - b_t)/\triangle_t \rfloor$, let $E_{j,1}$ and $E_{j,2}$ be determined as above and let $E_{j,3}$ and $E_{j,4}$ be determined as

$$E_{j,3} = \left\{((j-1)\triangle_t,\, j\triangle_t + b_t) \times \left(0, \left(j - 1 - \left\lfloor \frac{2b_t}{\triangle_t} \right\rfloor\right)\triangle_t\right)\right\},$$
$$E_{j,4} = \left\{\left(0, \left(j - 1 - \left\lfloor \frac{2b_t}{\triangle_t} \right\rfloor\right)\triangle_t\right) \times ((j-1)\triangle_t,\, j\triangle_t + b_t)\right\}.$$

Then the random variables Z_j^+ and $\{Z_i^+\}_{\{i \notin \mathcal{B}_j\}}$ interact through these regions. Conditioning on the event that there is no point in $E_{j,1} \cup E_{j,2} \cup E_{j,3} \cup E_{j,4}$ (that is $\| E_{j,1} \cup E_{j,2} \cup E_{j,3} \cup E_{j,4} \| = 0$), the random variables Z_j^+ and $\{Z_i^+\}_{\{i \notin \mathcal{B}_j\}}$ are independent. A direct calculation shows that

$$\begin{aligned} V(E_{j,2}) = V(E_{j,1}) &= \int\!\!\int_{(x,y)\in E_{j,1}} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{(j-1)\triangle_t}^{j\triangle_t + b_t} \int_{(j-1+\lfloor 2b_t/\triangle_t \rfloor)\triangle_t}^{\infty} f_2(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{(j-1)\triangle_t}^{j\triangle_t + b_t} \int_{(j-1+\lfloor 2b_t/\triangle_t \rfloor)\triangle_t}^{\infty} f_2(0, y-x)\,\mathrm{d}y\,\mathrm{d}x \end{aligned}$$

and since f_2 is nonnegative and $(y - x, \infty) \subset (b_t - 2\triangle_t, \infty)$

$$\begin{aligned} &< \int_{(j-1)\triangle_t}^{j\triangle_t + b_t} \int_{b_t - 2\triangle_t}^{\infty} f_2(0,v)\,\mathrm{d}v\,\mathrm{d}x \\ &= \int_{b_t - 2\triangle_t}^{\infty} (\triangle_t + b_t) f_2(0,v)\,\mathrm{d}v \\ &\le \int_{b_t - 2\triangle_t}^{\infty} (3\triangle_t + v) f_2(0,v)\,\mathrm{d}v \quad (\text{since } v \ge b_t - 2\triangle_t \text{ and} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad 3\triangle_t + v \ge \triangle_t + b_t). \end{aligned}$$

Similarly,

$$V(E_{j,3}) = V(E_{j,4}) \le \int_{2b_t}^{\infty} (\triangle_t + b_t) f_2(0,v)\,\mathrm{d}v \le \int_{2b_t}^{\infty} v f_2(0,v)\,\mathrm{d}v.$$

Since $\int_0^\infty v f_2(0,v)\,\mathrm{d}v$ is finite, $V(E_{j,1})$, $V(E_{j,2})$, $V(E_{j,3})$, and $V(E_{j,4})$ all converge to 0 as $t \to \infty$.

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Under the event $\mathscr{E}\colon \{\Pi_2^* \cap \{E_{j,1} \cup E_{j,2} \cup E_{j,3} \cup E_{j,4}\} = \varnothing\}$, $Z_j^+(b_t)$ and $\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t)$ are independent, i.e. $(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \mid \mathscr{E})$ and $(Z_j^+(b_t) \mid \mathscr{E})$ are independent. Therefore,

$$\begin{aligned}
&d\bigg(\bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, Z_j^+(b_t) = 1 \,\bigg|\, \mathscr{E}\bigg), \bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, \mathscr{E}\bigg)\bigg) \\
&\quad = d\bigg(\bigg(\bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, \mathscr{E}\bigg) \,\bigg|\, (Z_j^+(b_t) = 1 \mid \mathscr{E})\bigg), \bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, \mathscr{E}\bigg)\bigg) \\
&\quad = 0.
\end{aligned}$$

Let $\mathscr{E}^{\mathrm{c}}$ denote the complementary event of $\mathscr{E}$. We then have

$$\begin{aligned}
&d\bigg(\bigg(\sum_{i\notin B_j} Z_i^+(b_t) \,\bigg|\, Z_j^+(b_t) = 1\bigg), \sum_{i\notin B_j} Z_i^+(b_t)\bigg) \\
&\quad \le d\bigg(\bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, Z_j^+(b_t) = 1 \,\bigg|\, \mathscr{E}\bigg), \bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, \mathscr{E}\bigg)\bigg) \Pr\{\mathscr{E}\} \\
&\qquad + d\bigg(\bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, Z_j^+(b_t) = 1 \,\bigg|\, \mathscr{E}^{\mathrm{c}}\bigg), \bigg(\sum_{i\notin\mathscr{B}_j} Z_i^+(b_t) \,\bigg|\, \mathscr{E}^{\mathrm{c}}\bigg)\bigg) \Pr\{\mathscr{E}^{\mathrm{c}}\} \\
&\quad \le \Pr\{\mathscr{E}^{\mathrm{c}}\} \\
&\quad = \Pr\{\Pi_2^* \cap \{E_1 \cup E_2 \cup E_3 \cup E_4\} \ne \varnothing\} \\
&\quad = 1 - \exp(-V(E_{j,1} \cup E_{j,2} \cup E_{j,3} \cup E_{j,4})) \\
&\quad \longrightarrow 0.
\end{aligned}$$

This proves the convergence of u_3 to 0. Therefore, the Poisson approximation of $n_t^+(b_t)$ is proved. It has been proved in (20) that $\lim_{t\to\infty} \lambda_t = \lambda$. Therefore, we have the Poisson distribution with parameter λ for $n_t^+(b_t)$ with $b_t = \ln t + \lfloor (r-1)/2 \rfloor \ln\ln t + x$. This completes the proof.

4.2. Asymtotic minimum r-scan distribution

For the asymptotic distribution of the minimum r-scan length of $\Pi_1^{(2)}$, we set the partition length $\delta_t = t^{-2}$ and $a_t = \sqrt[r]{x/t}$. The associated Bernoulli random variables, for $1 \le j \le \lfloor (t-a_t)/\delta_t \rfloor$, are

$$Z_j^-(a_t) = \begin{cases} 1 & \text{if } \|((j-1)\delta_t, j\delta_t)\| = 1 \text{ and } \|(j\delta_t, (j-1)\delta_t + a_t)\| \ge r, \\ 0 & \text{otherwise.} \end{cases}$$

We form the sum $n_t^-(a_t) = \sum_{j=1}^{\lfloor (t-a_t)/\delta_t \rfloor} Z_j^-(a_t)$, and define $N_t^-(a_t)$ as the count of r-scan intervals which do not exceed a_t in the interval $(0, t)$ generated from the points of $\Pi_1^{(2)}$.

If the shift density g is continuous, the asymptotic distribution of the kth smallest r-scan length of the 2-fold shift model will be based on the following two properties.

Property 3. $\lim_{t\to\infty} \Pr\{N_t^-(a_t) \ne n_t^-(a_t)\} = 0.$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Property 4. *$n_t^-(a_t)$ is distributed asymptotically Poisson with parameter*

$$\alpha = (r+1)x \sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \frac{2^{r+1-2j}[f_2(0,0)]^j}{j!\,(r+1-2j)!}.$$

From Properties 3 and 4, Theorem 2 can be proved as follows.

Proof of Theorem 2. The duality relation shows that $\Pr\{m_{t,k} \geq a_t\} = \Pr\{N_t^-(a_t) \leq k-1\}$. Also,

$$\begin{aligned}\lim_{t\to\infty} \Pr\{N_t^-(a_t) \leq k-1\} &= \lim_{t\to\infty} \Pr\{n_t^-(a_t) \leq k-1\} \quad \text{(from Property 3)}\\ &= \sum_{j=0}^{k-1} \mathrm{e}^{-\alpha} \frac{\alpha^j}{j!} \quad \text{(from Property 4)}.\end{aligned}$$

Therefore,

$$\lim_{t\to\infty} \Pr\{m_{t,k} \geq a_t\} = \sum_{j=0}^{k-1} \mathrm{e}^{-\alpha} \frac{\alpha^j}{j!},$$

as described in Theorem 2.

Proof of Property 3. The event of $Z_j^-(a_t) = 1$ indicates the existence of an r-scan interval with its initial marker in $((j-1)\delta_t, j\delta_t)$ and its length less than a_t. Therefore, we have $N_t^-(a_t) \geq n_t^-(a_t)$. Thus, $\Pr\{N_t^-(a_t) \neq n_t^-(a_t)\} = \Pr\{N_t^-(a_t) > n_t^-(a_t)\}$. The event of $\{N_t^-(a_t) > n_t^-(a_t)\}$ will occur only if there exists a j, $1 \leq j \leq \lfloor (t-a_t)/\delta_t \rfloor$, such that one of the following two events occur.

Case 3. $\|((j-1)\delta_t, j\delta_t)\| \geq 2$ *and* $\|(j\delta_t, j\delta_t + a_t)\| \geq r-1$.

Case 4. $||((j-1)\delta_t, j\delta_t)|| = 1$ *and* $||(j\delta_t, (j-1)\delta_t + a_t)|| < r$ *and* $||(j\delta_t, j\delta_t + a_t)|| \geq r$.

Case 3 occurs when $Z_j^-(a_t) = 0$ and there is an r-scan segment which qualifies for length less than a_t and with its first two markers very close together and occurring in $((j-1)\delta_t, j\delta_t)$. Here $Z_j^-(a_t) = 1$ indicates the existence of an r-scan interval with the first point in the subinterval $((j-1)\delta_t, j\delta_t))$ and the r-scan length less than a_t. The probability difference is bounded by

$$\Pr\{Z_j^-(a_t) = 0, \|((j-1)\delta_t, j\delta_t)\| = 1, \|(j\delta_t, j\delta_t + a_t)\| \geq r\} = \Pr\{\text{Case } 4\}.$$

As shown in (12) and (13), the estimates of $\Pr\{\text{Case 3 occurs for index } j\} \leq O(\delta_t^2)$ and $\Pr\{\text{Case 4 occurs for index } j\} \leq O(\delta_t^2)$ prevail. Therefore,

$$\begin{aligned}&\Pr\{N_t^-(a_t) \neq n_t^-(a_t)\}\\ &\quad = \Pr\{N_t^-(a_t) > n_t^-(a_t)\}\\ &\quad \leq \Pr\left\{\text{Case 3 occurs for some index } j,\ 1 \leq j \leq \left\lfloor \frac{t-a_t}{\delta_t} \right\rfloor\right\}\\ &\qquad + \Pr\left\{\text{Case 4 occurs for some index } j,\ 1 \leq j \leq \left\lfloor \frac{t-a_t}{\delta_t} \right\rfloor\right\}\\ &\quad \leq \sum_{j=1}^{\lfloor (t-a_t)/\delta_t \rfloor} \Pr\{\text{Case 3 occurs for index } j\} + \sum_{j=1}^{\lfloor (t-a_t)/\delta_t \rfloor} \Pr\{\text{Case 4 occurs for index } j\}\end{aligned}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

$$\begin{aligned}&\le O\left(\frac{t}{\delta_t}\right)O(\delta_t^2)+O\left(\frac{t}{\delta_t}\right)O(\delta_t^2)\\&=O(t^{-1}).\end{aligned}$$

This completes the proof.

Proof of Property 4. We will apply the Chen–Stein method to the Bernoulli sum $n_t^-(a_t)=\sum_{j=1}^{\lfloor (t-a_t)/\delta_t\rfloor} Z_j^-(a_t)$ for $a_t=\sqrt[r]{x/t}$ and $\delta_t=t^{-2}$. According to Lemma 5 and the stationary property of $\Pi_1^{(2)}$, we have

$$\mathrm{E}[Z_j^-(a_t)]=\mathrm{E}[Z_1^-(a_t)]=(r+1)\delta_t\frac{x}{t}\left(\sum_{j=0}^{\lfloor (r+1)/2\rfloor}\frac{2^{r+1-2j}[f_2(0,0)]^j}{j!\,(r+1-2j)!}\right)(1+o(1)).$$

Therefore,

$$\begin{aligned}\alpha_t:=\mathrm{E}[n_t^-(a_t)]&=\mathrm{E}\left[\sum_{j=1}^{\lfloor (t-a_t)/\delta_t\rfloor} Z_j^-(a_t)\right]\\&=\sum_{j=1}^{\lfloor (t-\sqrt[r]{x/t})/\delta_t\rfloor}\mathrm{E}[Z_j^-(a_t)]\\&=\left(\frac{t}{\delta_t}\right)\mathrm{E}[Z_1^-(a_t)](1+o(1))\\&=\alpha(1+o(1))\end{aligned}$$

for

$$\alpha=(r+1)x\sum_{j=0}^{\lfloor (r+1)/2\rfloor}\frac{2^{r+1-2j}[f_2(0,0)]^j}{(r+1-2j)!\,j!}. \tag{21}$$

To validate the Poisson approximation of $n_t^-(a_t)$, we construct the set of neighborhoods $\{\mathcal{D}_j\}$ for each j as $\mathcal{D}_j=\{i\colon |i-j|\le\lfloor a_t/\delta_t\rfloor\}$. Then, according to the Chen–Stein protocol (see Theorem 5), the upper bound for the total variational distance between the Bernoulli sum $n_t^-(a_t)$ and the Poisson $\mathrm{Po}(\alpha)$ random variable is as follows:

$$\begin{aligned}d(n_t^-(a_t),\mathrm{Po}(\alpha))&\le d(n_t^-(a_t),\mathrm{Po}(\alpha_t))+d(\mathrm{Po}(\alpha_t),\mathrm{Po}(\alpha))\\&\le(v_1+v_2)\left(\frac{1-\mathrm{e}^{-\alpha_t}}{\alpha_t}\right)+v_3\min(1,\alpha_t^{-1/2})+|\alpha-\alpha_t|,\end{aligned}$$

where

$$\begin{aligned}v_1&=\sum_{j=1}^{\lfloor (t-a_t)/\delta_t\rfloor}\sum_{i\in\mathcal{D}_j}\mathrm{E}[Z_j^-(a_t)]\,\mathrm{E}[Z_i^-(a_t)],\\v_2&=\sum_{j=1}^{\lfloor (t-a_t)/\delta_t\rfloor}\sum_{i\in\mathcal{D}_j,i\neq j}\mathrm{E}[Z_j^-(a_t)Z_i^-(a_t)],\quad\text{and}\\v_3&=2\sum_{j=1}^{\lfloor (t-a_t)/\delta_t\rfloor}\mathrm{E}[Z_j^-(a_t)]d\left(\left(\sum_{i\notin\mathcal{D}_j}Z_i^-(a_t)\,\middle|\,Z_j^-(a_t)=1\right),\sum_{i\notin\mathcal{D}_j}Z_i^-(a_t)\right).\end{aligned}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Detailed estimates of v_1, v_2, and v_3 are calculated by evaluating appropriate events of the two-dimensional Poisson process Π_2^*. We use Lemmas 5, 6, and 7 to provide the following estimates:

$$
\begin{aligned}
v_1 &= \sum_{j=1}^{\lfloor (t-a_t)/\delta_t \rfloor} \sum_{i \in \mathcal{D}_j} \mathrm{E}[Z_j^-(a_t)]\,\mathrm{E}[Z_i^-(a_t)] \\
&\le 2\left(\frac{t}{\delta_t}\right)\left(\frac{a_t}{\delta_t}\right)(\mathrm{E}[Z_1^-(a_t)])^2 \quad \text{(by the stationary property of } \Pi_1^{(2)}\text{)} \\
&= 2\left(\frac{t}{\delta_t}\right)\left(\frac{a_t}{\delta_t}\right)\left(\alpha\frac{\delta_t}{t}\right)^2(1+o(1)) \quad \text{(by Lemma 5, with } \alpha \text{ as defined in (21))} \\
&= O(t^{-(r+1)/r}), \\
v_2 &= \sum_{j=1}^{\lfloor (t-\sqrt[r]{x/t})/\delta_t \rfloor} \sum_{i \in \mathcal{D}_j,\, i \ne j} \mathrm{E}[Z_j^-(a_t) Z_i^-(a_t)] \\
&\le 2\left(\frac{t}{\delta_t}\right)\left(\frac{a_t}{\delta_t}\right) O(\delta_t^2 t^{-1}) \quad \text{(by (14) in Lemma 7)} \\
&= O(t^{-1/r}).
\end{aligned}
$$

To verify the convergence of v_3 to 0, let $\tilde{E}_{j,1}$, $\tilde{E}_{j,2}$, $\tilde{E}_{j,3}$, and $\tilde{E}_{j,4}$ be sets in $\mathbb{R}^2$ determined as follows. For $j \le \lfloor a_t/\delta_t \rfloor$,

$$
\begin{aligned}
\tilde{E}_{j,1} &= \{((j-1)\delta_t, (j-1)\delta_t + a_t) \times ((j-1)\delta_t + a_t, \infty)\}, \\
\tilde{E}_{j,2} &= \{((j-1)\delta_t + a_t, \infty) \times ((j-1)\delta_t, (j-1)\delta_t + a_t)\}.
\end{aligned}
$$

For $j \ge \lfloor a_t/\delta_t \rfloor + 1$, let $\tilde{E}_{j,1}$ and $\tilde{E}_{j,2}$ be as above, and let

$$
\begin{aligned}
\tilde{E}_{j,3} &= \{((j-1)\delta_t, (j-1)\delta_t + a_t) \times (0, (j-1)\delta_t - a_t)\}, \\
\tilde{E}_{j,4} &= \{(0, (j-1)\delta_t - a_t) \times ((j-1)\delta_t, (j-1)\delta_t + a_t)\}.
\end{aligned}
$$

Direct calculation shows that

$$
\begin{aligned}
V(\cup_{i=1}^4 \tilde{E}_{j,i}) &= \int\!\!\int_{\cup_{i=1}^4 \tilde{E}_i} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&< \int_{-\infty}^{\infty} \int_{(j-1)\delta_t}^{(j-1)\delta_t + a_t} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{(j-1)\delta_t}^{(j-1)\delta_t + a_t} \int_{-\infty}^{\infty} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&\quad + \int_{-\infty}^{\infty} \int_{(j-1)\delta_t}^{(j-1)\delta_t + a_t} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{(j-1)\delta_t}^{(j-1)\delta_t + a_t} \int_{-\infty}^{\infty} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= 4a_t \quad \text{(see the detailed proof of (22) in Section 5).}
\end{aligned}
$$

For each j, $\cup_{i=1}^4 \tilde{E}_{j,i}$ is the region of $A(Z_j^-(a_t)) \bigcap \cup_{i \notin \mathcal{D}_j} A(Z_i^-(a_t))$. Therefore, under the occurrence of the event of $\tilde{\mathcal{E}}\colon \Pi_2^* \bigcap \{\tilde{E}_{j,1} \cup \tilde{E}_{j,2} \cup \tilde{E}_{j,3} \cup \tilde{E}_{j,4}\} = \varnothing$, $Z_j^-(a_t)$ and $\sum_{i \notin \mathcal{D}_j} Z_i^-(a_t)$ are determined by the realizations of disjoint regions in the two-dimentional Poisson process, and are therefore independent. That is, $(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t) \mid \tilde{\mathcal{E}})$ and $(Z_j^-(a_t) \mid \tilde{\mathcal{E}})$ are independent. Therefore, the variation distance

$$
\begin{aligned}
&d\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm| Z_j^-(a_t)=1\biggm|\tilde{\mathcal{E}}\biggr),\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm|\tilde{\mathcal{E}}\biggr)\biggr)\\
&\qquad\equiv d\biggl(\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm|\tilde{\mathcal{E}}\biggr)\biggm|(Z_j^-(a_t)=1\mid\tilde{\mathcal{E}})\biggr),\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm|\tilde{\mathcal{E}}\biggr)\biggr)\\
&\qquad=0.
\end{aligned}
$$

Let $\tilde{\mathcal{E}}^{\mathrm{c}}$ denote the complement of $\tilde{\mathcal{E}}$. We have

$$
\begin{aligned}
&d\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm| Z_j^-(a_t)=1\biggr),\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggr)\\
&\qquad\le d\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm| Z_j^-(a_t)=1\biggm|\tilde{\mathcal{E}}\biggr),\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm|\tilde{\mathcal{E}}\biggr)\biggr)\Pr\{\tilde{\mathcal{E}}\}\\
&\qquad\quad+d\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm| Z_j^-(a_t)=1\biggm|\tilde{\mathcal{E}}^{\mathrm{c}}\biggr),\biggl(\sum_{i\notin D_j} Z_i^-(a_t)\biggm|\tilde{\mathcal{E}}^{\mathrm{c}}\biggr)\biggr)\Pr\{\tilde{\mathcal{E}}^{\mathrm{c}}\}\\
&\qquad\le \Pr\{\tilde{\mathcal{E}}^{\mathrm{c}}\}\\
&\qquad\le \Pr\{\Pi_2^*\cap\{\tilde{E}_1\cup\tilde{E}_2\cup\tilde{E}_3\cup\tilde{E}_4\}\ne\varnothing\}\\
&\qquad=1-\exp(-V(\cup_{i=1}^4\tilde{E}_{j,i}))\\
&\qquad=4a_t(1+o(1));
\end{aligned}
$$

and it follows that

$$
\begin{aligned}
v_3&=2\sum_{j=1}^{\lfloor(t-\sqrt[r]{x/t})/\delta_t\rfloor}\mathrm{E}[Z_j^-(a_t)]d\biggl(\biggl(\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggm| Z_j^-(a_t)=1\biggr),\sum_{i\notin\mathcal{D}_j} Z_i^-(a_t)\biggr)\\
&\le 2\biggl(\sum_{j=1}^{\lfloor(t-\sqrt[r]{x/t})/\delta_t\rfloor}\mathrm{E}[Z_j^-(a_t)]\biggr)4a_t(1+o(1))\\
&=8\alpha_t a_t(1+o(1))\\
&\to 0
\end{aligned}
$$

as $\lim_{t\to\infty}\alpha_t=\alpha$ and $\lim_{t\to\infty}a_t=0$. Therefore, the Poisson law with parameter α for $n_t^-(a_t)$ is established.

5. Details of estimates

In this section the proof of the estimates of Section 3 will be elaborated.

Proof of Lemma 1. Some simple manipulations give

$$
\int_0^z\int_{-\infty}^{\infty} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y=\int_0^z\int_{-\infty}^{\infty}\biggl(\int_{-\infty}^{\infty} g(x-s)g(y-s)\,\mathrm{d}x\biggr)\mathrm{d}s\,\mathrm{d}y=z, \qquad (22)
$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

and similarly for $\int_{-\infty}^{\infty}\int_0^z f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y = z$. Since

$$\begin{aligned}
V(A_1(0,z)) &= \int_0^z\int_0^z f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= 2\int_0^z\int_0^x f_2(x-y,0)\,\mathrm{d}y\,\mathrm{d}x \quad (\text{since } f_2(x,y)=f_2(y,x) \text{ and} \\
&\qquad\qquad\qquad\qquad\qquad\qquad f_2(x,y)=f_2(x-y,0)) \\
&= 2\int_0^z\int_0^x f_2(s,0)\,\mathrm{d}s\,\mathrm{d}x \\
&= 2\int_0^z\int_s^z f_2(s,0)\,\mathrm{d}x\,\mathrm{d}x \\
&= 2\int_0^z (z-s)f_2(s,0)\,\mathrm{d}s \\
&= z - 2\int_0^{\infty} sf_2(s,0)\,\mathrm{d}s + 2\int_z^{\infty}(s-z)f_2(s,0)\,\mathrm{d}s \\
&= z - \int_{-\infty}^{\infty}|s|f_2(s,0)\,\mathrm{d}s + 2\int_z^{\infty}(s-z)f_2(s,0)\,\mathrm{d}s \\
&= z - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v + 2\int_z^{\infty}(s-z)f_2(s,0)\,\mathrm{d}s,
\end{aligned}$$

we have

$$\begin{aligned}
2z &= \int_{-\infty}^{\infty}\int_0^z f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_0^z\int_{-\infty}^{\infty} f_2(x,y)\,\mathrm{d}x\,\mathrm{d}y \quad (\text{from (22)}) \\
&= V(A_2(0,z)) + 2V(A_1(0,z)) \quad (\text{from the defintions of } A_1 \text{ and } A_2 \text{ in Figure 1}).
\end{aligned}$$

Therefore, where $\int |s|g(s)\,\mathrm{d}s < \infty$, we have

$$\lim_{z\to\infty}(z - V(A_1(0,z))) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v, \tag{23}$$

$$\lim_{z\to\infty} V(A_2(0,z)) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v. \tag{24}$$

Now, we have

$$\begin{aligned}
&\Pr\{\text{there are at most } k \text{ points of } \Pi^{(2)} \text{ in } (0,z)\} \\
&\quad= \sum_{j=0}^{\lfloor k/2\rfloor}\Pr\{||A_1(0,z)|| = j;\ ||A_2(0,z)|| \le k-2j\}
\end{aligned}$$

and since $A_1(0,z)$ and $A_2(0,z)$ are disjoint regions, and $\Pi_2^{\star}$ is a two-dimensional Poisson process, we have

$$= \sum_{j=0}^{\lfloor k/2\rfloor} \mathrm{e}^{-V(A_1(0,z))}\frac{[V(A_1(0,z))]^j}{j!}\mathrm{e}^{-V(A_2(0,z))}\biggl(\sum_{i=0}^{k-2j}\frac{[V(A_2(0,z))]^i}{i!}\biggr)$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

$$= \mathrm{e}^{-V(A_1(0,z))}\frac{[V(A_1(0,z))]^{\lfloor k/2\rfloor}}{\lfloor k/2\rfloor!}\mathrm{e}^{-V(A_2(0,z))}I_2(k) \quad (\text{for } I_2(k) \text{ as defined in (8)})$$

$$+\sum_{j=0}^{\lfloor k/2\rfloor-1}\mathrm{e}^{-V(A_1(0,z))}\frac{[V(A_1(0,z))]^{j}}{j!}\mathrm{e}^{-V(A_2(0,z))}\bigg(\sum_{i=0}^{k-2j}\frac{[V(A_2(0,z))]^{i}}{i!}\bigg)$$

$$= \exp\{-(V(A_1(0,z))+V(A_2(0,z)))\}\frac{[V(A_1(0,z))]^{\lfloor k/2\rfloor}}{\lfloor k/2\rfloor!}I_2(k)$$

$$+ \text{ smaller-order terms of } z\text{, and using (23) and (24)}$$

$$= \exp\bigg\{-\bigg(z+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v\bigg)\bigg\}$$

$$\times\frac{[z-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v]^{\lfloor k/2\rfloor}}{\lfloor k/2\rfloor!}$$

$$\times\ I_2(k)(1+o(1)). \tag{25}$$

Proof of Lemma 2. Lemma 2 is proved by evaluating the requisite events of the two-dimensional Poisson process Π_2^*. Observe that

$$\begin{aligned}&\Pr\{\|(\triangle_t,\triangle_t+b_t)\|\le r\}-\Pr\{\|(0,\triangle_t+b_t)\|\le r\}\\&\quad=\Pr\{\|(\triangle_t,\triangle_t+b_t)\|\le r,\ \|(0,\triangle_t+b_t)\|>r\}\\&\quad=\sum_{k=1}^{r}\Pr\{\|(0,\triangle_t)\|=k;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-k\}\\&\quad=\Pr\{\|(0,\triangle_t)\|=1;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-1\}\\&\qquad+\sum_{k=2}^{r}\Pr\{\|(0,\triangle_t)\|=k;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-k\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\Pr\{\|(0,\triangle_t)\|=1;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-1\}\\&\quad=\Pr\{\|(\triangle_t,\triangle_t+b_t)\|\le r\}-\Pr\{\|(0,\triangle_t+b_t)\|\le r\}\\&\qquad-\sum_{k=2}^{r}\Pr\{\|(0,\triangle_t)\|=k;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-k\}\end{aligned}$$

applying (25) for $z=b_t=\ln t+\lfloor(r-1)/2\rfloor\ln\ln t+x$ and $z=\triangle_t+b_t$, and also knowing that $\Pr\{\|(0,\triangle_t)\|=k;\ \|(\triangle_t,\triangle_t+b_t)\|\le r-k\}\le O(\triangle_t^k)O(\exp(-b_t)b_t{}^{\lfloor(r-k)/2\rfloor})$, we obtain

$$=\frac{\triangle_t\exp\{-(x+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u-v|g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v)\}}{\lfloor(r-1)/2\rfloor!\,t}(1+o(1)),$$

$o(1)\to 0$ as $t\to\infty$.

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Proof of Lemma 3. First we consider the following two-dimensional disjoint sets:

$$\begin{aligned}C_1 &= \{(-\infty, 0) \times (0, \triangle_t)\} \cup \{(0, \triangle_t) \times (-\infty, 0)\},\\ C_2 &= \{(\triangle_t + b_t, \infty) \times (0, \triangle_t)\} \cup \{(0, \triangle_t) \times (\triangle_t + b_t, \infty)\},\\ C_3 &= \{(\triangle_t, \triangle_t + b_t) \times (0, \triangle_t)\} \cup \{(0, \triangle_t) \times (\triangle_t, \triangle_t + b_t)\},\\ C_4 &= \{(0, \triangle_t) \times (0, \triangle_t)\}.\end{aligned} \tag{26}$$

Consider the sets $A(\triangle_t, b_t + \triangle_t)$, $A_1(\triangle_t, b_t + \triangle_t)$, and $A_2(\triangle_t, b_t + \triangle_t)$ as defined in Figure 1 by setting $a = \triangle_t$ and $b = b_t + \triangle_t$. It should be noted that C_3 is a subset of $A_2(\triangle_t, b_t + \triangle_t)$. Based on (22), we have

$$\begin{aligned}V(C_1) + V(C_2) + V(C_3) + 2V(C_4) &= \int_0^{\triangle_t}\int_{-\infty}^{\infty} f_2(x, y)\,\mathrm{d}x\,\mathrm{d}y + \int_{-\infty}^{\infty}\int_0^{\triangle_t} f_2(x, y)\,\mathrm{d}x\,\mathrm{d}y\\ &= 2\triangle_t.\end{aligned} \tag{27}$$

Also,

$$\begin{aligned}V(C_4) &= \int_0^{\triangle_t}\int_0^{\triangle_t} f_2(x, y)\,\mathrm{d}x\,\mathrm{d}y\\ &= f_2(0, 0)\triangle_t^2(1 + o(1)) \quad \text{with } f_2(0, 0) = \int_{-\infty}^{\infty} g^2(\eta)\,\mathrm{d}\eta,\end{aligned} \tag{28}$$

and

$$V(C_3) = 2\int_0^{\triangle_t}\int_{\triangle_t}^{\triangle_t + b_t} f_2(x, y)\,\mathrm{d}x\,\mathrm{d}y = 2\int_0^{\triangle_t}\int_{\triangle_t}^{\triangle_t + b_t} f_2(x - y, 0)\,\mathrm{d}x\,\mathrm{d}y.$$

Since, for each fixed y, $0 < y < \triangle_t$,

$$\int_{\triangle_t}^{b_t} f_2(x, 0)\,\mathrm{d}x < \int_{\triangle_t}^{\triangle_t + b_t} f_2(x - y, 0)\,\mathrm{d}x < \int_0^{\triangle_t + b_t} f_2(x, 0)\,\mathrm{d}x,$$

we conclude, after integrating over y from 0 to $\triangle_t$, that

$$2\triangle_t\int_{\triangle_t}^{b_t} f_2(x, 0)\,\mathrm{d}x < V(C_3) < 2\triangle_t\int_0^{\triangle_t + b_t} f_2(x, 0)\,\mathrm{d}x,$$

which can be rewritten as

$$\triangle_t\biggl(\int_{-b_t}^{b_t} f_2(x, 0)\,\mathrm{d}x - \int_{-\triangle_t}^{\triangle_t} f_2(x, 0)\,\mathrm{d}x\biggr) < V(C_3) < \triangle_t\int_{-(\triangle_t + b_t)}^{\triangle_t + b_t} f_2(x, 0)\,\mathrm{d}x,$$

and therefore,

$$V(C_3) = \triangle_t(1 + o(1)). \tag{29}$$

Owing to (27)–(29), we have

$$V(C_1 \cup C_2) = 2\triangle_t - V(C_3) - 2V(C_4) = \triangle_t(1 + o(1)). \tag{30}$$

According to (23) and (24), in the proof of Lemma 1, with $z = b_t$, as $t \to \infty$,

$$V(A_1(\triangle_t, b_t + \triangle_t)) = b_t - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u - v| g(u)g(v)\,\mathrm{d}u\,\mathrm{d}v + o(1). \tag{31}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

Consider $\{C_i\}_{i=1}^4$ as defined in (26) and $A_1(\triangle_t, \triangle_t + b_t)$ and $A_2(\triangle_t, \triangle_t + b_t)$ as defined in Figure 1. The probability of the event $\{||(0, \triangle_t)|| \geq 2, ||(\triangle_t, \triangle_t + b_t)|| \leq r - 1\}$ can be bounded as follows.

$$\begin{aligned}
&\Pr\{||(0, \triangle_t)|| \geq 2, ||(\triangle_t, \triangle_t + b_t)|| \leq r - 1\} \\
&\quad = \Pr\{||C_1 \cup C_2 \cup C_3|| + 2||C_4|| \geq 2; 2||A_1(\triangle_t, \triangle_t + b_t)|| \\
&\qquad + ||A_2(\triangle_t, \triangle_t + b_t)|| \leq r - 1\} \\
&\quad = \sum_{j=0}^{\lfloor (r-1)/2 \rfloor} \Pr\{||C_1 \cup C_2 \cup C_3|| + 2||C_4|| \geq 2; ||A_1(\triangle_t, \triangle_t + b_t)|| = j; \\
&\qquad\qquad ||A_2(\triangle_t, \triangle_t + b_t)|| \leq r - 1 - 2j\} \\
&\quad \leq \sum_{j=0}^{\lfloor (r-1)/2 \rfloor} \Pr\{||C_1 \cup C_2 \cup C_3|| + 2||C_4|| \geq 2; ||A_1(\triangle_t, \triangle_t + b_t)|| = j\},
\end{aligned}$$

since C_1, C_2, C_3, C_4, and $A_1(\triangle_t, \triangle_t + b_t)$ are all disjoint regions, we have the bound

$$\begin{aligned}
&< (\Pr\{||C_1 \cup C_2 \cup C_3|| \geq 2\} + \Pr\{||C_4|| \geq 1\}) \\
&\quad \times \left(\sum_{j=0}^{\lfloor (r-1)/2 \rfloor} \Pr\{||A_1(\triangle_t, \triangle_t + b_t)|| = j\} \right),
\end{aligned}$$

and, according to the Poisson law and evaluations of $V(C_1 \cup C_2 \cup C_3)$, $V(C_4)$, and $V(A_1(\triangle_t, \triangle_t + b_t))$ in (28)–(31), we obtain

$$\begin{aligned}
&= O(\triangle_t^2) O\left(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r-1)/2 \rfloor} \right) \\
&= O\left(\frac{\triangle_t^2}{t} \right).
\end{aligned}$$

The probability of the event $\{||(0, \triangle_t)|| = 1, ||(0, b_t)|| \leq r, ||(0, \triangle_t + b_t)|| \geq r + 1\}$ can be bounded as follows.

$$\begin{aligned}
&\Pr\{||(0, \triangle_t)|| = 1, ||(0, b_t)|| \leq r, ||(0, \triangle_t + b_t)|| \geq r + 1\} \\
&\quad < \Pr\{||(0, \triangle_t)|| = 1, ||(\triangle_t, b_t)|| \leq r - 1, ||(b_t, \triangle_t + b_t)|| \geq 1\},
\end{aligned}$$

since $A_1(\triangle_t, b_t)$ is disjoint from both $A(0, \triangle_t)$ and $A(b_t, b_t + \triangle_t)$, we have the bound

$$\begin{aligned}
&\leq \Pr\{||(0, \triangle_t)|| = 1, ||(b_t, \triangle_t + b_t)|| \geq 1\} \Pr\left\{ ||A_1(\triangle_t, b_t)|| \leq \left\lfloor \frac{r-1}{2} \right\rfloor \right\} \\
&= (\Pr\{||(0, \triangle_t)|| = 1\} - \Pr\{||(0, \triangle_t)|| = 1, ||(b_t, \triangle_t + b_t)|| = 0\}) \\
&\quad \times \Pr\left\{ ||A_1(\triangle_t, b_t)|| \leq \left\lfloor \frac{r-1}{2} \right\rfloor \right\},
\end{aligned}$$

then, according to the Poisson law, we obtain

$$= O(\mathrm{e}^{-2\triangle_t}(2\triangle_t) - \mathrm{e}^{-2\triangle_t}(2\triangle_t)\mathrm{e}^{-2\triangle_t}) O\left(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r-1)/2 \rfloor !} \right)$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

$$
\begin{aligned}
&= O(\triangle_t^2)O(\mathrm{e}^{-\ln t}) \\
&= O\bigg(\frac{\triangle_t^2}{t}\bigg).
\end{aligned}
$$

This completes the proof.

Proof of Lemma 4. If $r = 2p+1$ and p is a nonnegative integer, we will verify the estimates, (35), (36), and (37), below, of $\mathrm{E}[Z_j^+(b_t)Z_i^+(b_t)]$, $j \neq i$, which is involved in the calculation of the error parameter u_2 of the Chen–Stein method in (1), where the neighborhood subset B_i is specified as

$$
\begin{aligned}
B_i &= \bigg\{j : |j-i| < \bigg\lfloor \frac{2b_t}{\triangle_t} \bigg\rfloor \bigg\} \\
&= \bigg\{j : |j-i| \le \bigg\lfloor \frac{\ln\ln t}{\triangle_t} \bigg\rfloor \bigg\} \cup \bigg\{j : \bigg\lfloor \frac{\ln\ln t}{\triangle_t} \bigg\rfloor < |j-i| \le \bigg\lfloor \frac{b_t}{\triangle_t} \bigg\rfloor \bigg\} \\
&\quad \cup \bigg\{j : \bigg\lfloor \frac{b_t}{\triangle_t} \bigg\rfloor < |j-i| \le \bigg\lfloor \frac{2b_t}{\triangle_t} \bigg\rfloor \bigg\} \\
&= (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).
\end{aligned} \tag{32}
$$

For the case in which $p = 0$ (1-scan case), we have, for $1 \le |i-j| \le \lfloor b_t/\triangle_t \rfloor$,

$$
\begin{aligned}
\mathrm{E}[Z_j^+(b_t)Z_i^+(b_t)] &= \mathrm{E}[Z_1^+(b_t)Z_{i-j+1}^+(b_t)] \\
&= \Pr\{\|(0,\triangle_t)\| = 1, \|(\triangle_t, \triangle_t + b_t)\| = 0, \\
&\qquad \|((i-j)\triangle_t, (i-j+1)\triangle_t)\| = 1, \\
&\qquad \|((i-j+1)\triangle_t, (i-j+1)\triangle_t + b_t)\| = 0\} \quad \text{(directly from the definitions of } Z_j^+ \text{ and } Z_i^+\text{)},
\end{aligned} \tag{33}
$$

since $\|((i-j)\triangle_t, (i-j+1)\triangle_t)\| = 1$ and $\|(\triangle_t, \triangle_t + b_t)\| = 0$ are disjoint events for $1 \le |i-j| \le \lfloor b_t/\triangle_t \rfloor$ the last equation equals 0.

For $\lfloor b_t/\triangle_t \rfloor + 1 \le |j-i| \le \lfloor 2b_t/\triangle_t \rfloor$, we have

$$
\begin{aligned}
\mathrm{E}[Z_j^+(b_t)Z_i^+(b_t)] &= \mathrm{E}[Z_1^+(b_t)Z_{i-j+1}^+(b_t)] \\
&= \Pr\{||(0,\triangle_t)|| = 1, ||(\triangle_t, \triangle_t + b_t)|| = 0, \\
&\qquad ||((i-j)\triangle_t, (i-j+1)\triangle_t)|| = 1, \\
&\qquad ||((i-j+1)\triangle_t, (i-j+1)\triangle_t + b_t)|| = 0\}.
\end{aligned}
$$

Because of the disjoint nature of $A_1(a,b)$ to other regions of $A(c,d)$ in the two-dimensional space when (a,b) and (c,d) are nonoverlapping intervals of the line, we have

$$
\begin{aligned}
&\le \Pr\{||(0,\triangle_t)|| = 1, ||((i-j)\triangle_t, (i-j+1)\triangle_t)|| = 1, \} \\
&\quad \times \Pr\{||A_1(\triangle_t, \triangle_t + b_t)|| = 0\} \\
&\quad \times \Pr\{||A_1((i-j+1)\triangle_t, (i-j+1)\triangle_t + b_t)|| = 0, \} \\
&= O(\triangle_t^2)O(\mathrm{e}^{-b_t})O(\mathrm{e}^{-b_t}) \\
&\le O\bigg(\frac{\triangle^2}{t^2}\bigg).
\end{aligned} \tag{34}
$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

To sum up, from (33) and (34) we obtain, when $p = 0$ $(r = 1)$,

$$\sum_{j \in B_i, j \neq i} \mathrm{E}[Z_i^+(b_t) Z_j^+(b_t)] \leq 2 \frac{b_t}{\Delta_t} O\left(\frac{\Delta_t^2}{t^2}\right) \leq O\left(\frac{\ln t}{t^3}\right).$$

For the case in which $p \geq 1$ $(r \geq 3)$, the bounds of (I), (II), and (III) in (32) are evaluated next. For (I) in (32) and $1 \leq |i - j| \leq \lfloor \ln \ln t / \Delta_t \rfloor$,

$$\begin{aligned} \mathrm{E}[Z_j^+(b_t) Z_i^+(b_t)] &= \mathrm{E}[Z_1^+(b_t) Z_{i-j+1}^+(b_t)] \quad \text{(by the stationary property of } \Pi_1^{(2)}\text{)} \\ &< \Pr\{\|(0, \Delta_t)\| = 1, \|((i - j)\Delta_t, (i - j + 1)\Delta_t)\| = 1, \\ &\qquad \|(\Delta_t, \Delta_t + b_t) \setminus ((i - j)\Delta_t, (i - j + 1)\Delta_t)\| \leq r - 2\} \\ &\leq \Pr\{\|(0, \Delta_t)\| = 1, \|((i - j)\Delta_t, (i - j + 1)\Delta_t)\| = 1\} \\ &\quad \times \Pr\left\{\|A_1(\Delta_t, \Delta_t + b_t) \setminus A((i - j)\Delta_t, (i - j + 1)\Delta_t)\| \leq \left\lfloor \frac{r - 2}{2} \right\rfloor\right\} \\ &= O(\Delta_t^2) O\left(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-2)/2 \rfloor !}}{\lfloor (r - 2)/2 \rfloor !}\right) \\ &= O\left(\frac{\Delta_t^2}{t \ln t}\right). \end{aligned} \tag{35}$$

For (II) in (32) and $\lfloor \ln \ln t / \Delta_t \rfloor + 1 \leq |i - j| \leq \lfloor b_t / \Delta_t \rfloor$,

$$\begin{aligned} \mathrm{E}[Z_j^+(b_t) Z_i^+(b_t)] &= \mathrm{E}[Z_1^+(b_t) Z_{i-j+1}^+(b_t)] \\ &\leq \Pr\{\|(0, \Delta_t)\| = 1, \|((i - j)\Delta_t, (i - j + 1)\Delta_t)\| = 1, \\ &\qquad \|(\Delta_t, \Delta_t + b_t) \setminus ((i - j)\Delta_t, (i - j + 1)\Delta_t)\| \leq r - 2, \\ &\qquad \|(\Delta_t + b_t, \Delta_t + b_t + \ln \ln t)\| \leq r - 1\} \\ &\leq \Pr\{\|(0, \Delta_t)\| = 1, \|((i - j)\Delta_t, (i - j + 1)\Delta_t)\| = 1\} \\ &\quad \times \Pr\{\|A_1(\Delta_t, \Delta_t + b_t) \setminus A((i - j)\Delta_t, (i - j + 1)\Delta_t)\| \leq r - 2\} \\ &\quad \times \Pr\{\|A_1(\Delta_t + b_t, \Delta_t + b_t + \ln \ln t)\| \leq r - 1\} \\ &= O(\Delta_t^2) O\left(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-2)/2 \rfloor}}{\lfloor (r - 2)/2 \rfloor !}\right) O\left(\mathrm{e}^{-\ln \ln t} \frac{(\ln \ln t)^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r - 1)/2 \rfloor !}\right) \\ &= O\left(\frac{\Delta_t^2 (\ln \ln t)^{\lfloor (r-1)/2 \rfloor}}{t (\ln t)^2}\right). \end{aligned} \tag{36}$$

For (III) in (32) and $\lfloor b_t / \Delta_t \rfloor + 1 \leq i - j \leq \lfloor 2b_t / \Delta_t \rfloor$, using an argument similar to the one used above for the case in which $p = 0$, we have

$$\begin{aligned} \mathrm{E}[Z_j^+(b_t) Z_i^+(b_t)] &= \mathrm{E}[Z_1^+(b_t) Z_{i-j+1}^+(b_t)] \\ &= \Pr\{\|(0, \Delta_t)\| = 1, \|(\Delta_t, \Delta_t + b_t)\| \leq r - 1, \\ &\qquad \|((i - j)\Delta_t, (i - j + 1)\Delta_t)\| = 1, \\ &\qquad \|((i - j + 1)\Delta_t, (i - j + 1)\Delta_t + b_t)\| \leq r - 1\} \end{aligned}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

$$
\begin{aligned}
&\leq \Pr\{||(0, \triangle_t)|| = 1, ||((i-j)\triangle_t, (i-j+1)\triangle_t)|| = 1, \} \\
&\quad \times \Pr\{||A_1(\triangle_t, \triangle_t + b_t)|| \leq r-1\} \\
&\quad \times \Pr\{||A_1((i-j+1)\triangle_t, (i-j+1)\triangle_t + b_t)|| \leq r-1\} \\
&= O(\triangle_t^2) O\biggl(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r-1)/2 \rfloor !}\biggr) O\biggl(\mathrm{e}^{-b_t} \frac{b_t^{\lfloor (r-1)/2 \rfloor}}{\lfloor (r-1)/2 \rfloor !}\biggr) \\
&= o\biggl(\frac{\triangle_t^2}{t^2}\biggr). \qquad (37)
\end{aligned}
$$

Therefore, from (35), (36), and (37) we have, for $p \geq 1$,

$$
\begin{aligned}
\sum_{j \in B_i, j \neq i} \mathrm{E}[Z_i^+(b_t) Z_j^+(b_t)] &= \sum_{1 \leq |j-i| \leq \lfloor \ln\ln t/\triangle_t \rfloor} \mathrm{E}[Z_i^+(b_t) Z_j^+(b_t)] \\
&\quad + \sum_{\lfloor \ln\ln t/\triangle_t \rfloor < |j-i| \leq \lfloor b_t/\triangle_t \rfloor} \mathrm{E}[Z_i^+(b_t) Z_j^+(b_t)] \\
&\quad + \sum_{\lfloor b_t/\triangle_t \rfloor \leq |i-j| \leq \lfloor 2b_t/\triangle_t \rfloor} \mathrm{E}[Z_i^+(b_t) Z_j^+(b_t)] \\
&\leq 2 \biggl\lfloor \frac{\ln\ln t}{\triangle_t} \biggr\rfloor O\biggl(\frac{\triangle_t^2}{t \ln t}\biggr) + 2 \biggl\lfloor \frac{b_t}{\triangle_t} \biggr\rfloor O\biggl(\frac{\triangle_t^2 (\ln\ln t)^{\lfloor (r-1)/2 \rfloor}}{t (\ln t)^2}\biggr) \\
&\quad + 2 \biggl\lfloor \frac{2b_t}{\triangle_t} \biggr\rfloor o\biggl(\frac{\triangle_t^2}{t^2}\biggr).
\end{aligned}
$$

This completes the proof.

Proof of Lemma 5. For $y \to 0$,

$$
\begin{aligned}
&\Pr\{\|(0, y)\| \geq r+1\} \\
&\quad = \sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \Pr\{\|A_1(0, y)\| = j, \|A_2(0, y)\| \geq r+1-2j\} \\
&\qquad + \Pr\biggl\{\|A_1(0, y)\| \geq \biggl\lfloor \frac{r+1}{2} \biggr\rfloor + 1\biggr\},
\end{aligned}
$$

then by substituting $V(A_1(0, y)) = f_2(0, 0) y^2 (1 + o(1))$ ($o(1) \to 0$ as $y \to 0$), and $V(A_2(0, y)) = 2y - V(A_1(0, y))$, we obtain

$$
\begin{aligned}
&= \sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \biggl\{\exp\bigl(-\{f_2(0, 0) y^2\}\bigr) \frac{(f_2(0, 0) y^2)^j}{j!} \exp(-2y) \frac{(2y)^{r+1-2j}}{(r+1-2j)!}\biggr\} \\
&\quad + \exp\{-f_2(0, 0) y^2\} \frac{(f_2(0, 0) y^2)^{\lfloor (r+1)/2 \rfloor + 1}}{(\lfloor (r+1)/2 \rfloor + 1)!} + \text{smaller-order terms of } y \\
&= \biggl\{\sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \frac{f_2(0, 0)^j}{j!} \frac{2^{r+1-2j}}{(r+1-2j)!}\biggr\} y^{r+1} + \text{smaller-order terms of } y. \qquad (38)
\end{aligned}
$$

Also,

$$\begin{aligned}
&\Pr\{\|(0, a_t)\| \geq r+1\} - \Pr\{\|(\delta_t, a_t)\| \geq r+1\} \\
&\quad = \Pr\{\|(0, a_t)\| \geq r+1,\ \|(\delta_t, a_t)\| \leq r\} \\
&\quad = \Pr\{\|(0, \delta_t)\| \geq 1,\ \|(\delta_t, a_t)\| = r\} + \sum_{j=0}^{r-1} \Pr\{\|(0, \delta_t)\| \geq r+1-j,\ \|(\delta_t, a_t)\| = j\} \\
&\quad = \Pr\{\|(0, \delta_t)\| = 1,\ \|(\delta_t, a_t)\| = r\} + \Pr\{\|(0, \delta_t)\| \geq 2,\ \|(\delta_t, a_t)\| = r\} \\
&\qquad + \sum_{j=0}^{r-1} \Pr\{\|(0, \delta_t)\| \geq r+1-j,\ \|(\delta_t, a_t)\| = j\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Pr\{\|(0, \delta_t)\| = 1,\ \|(\delta_t, a_t)\| = r\} \\
&\quad = \Pr\{\|(0, a_t)\| \geq r+1\} - \Pr\{\|(\delta_t, a_t)\| \geq r+1\} \\
&\qquad - \Pr\{\|(0, \delta_t)\| \geq 2,\ \|(\delta_t, a_t)\| = r\} - \sum_{j=0}^{r-1} \Pr\{\|(0, \delta_t)\| \geq r+1-j,\ \|(\delta_t, a_t)\| = j\},
\end{aligned}$$

then, using (38) for $y = a_t = \sqrt[r]{x/t}$ and $\delta_t = 1/t^2$, and knowing that $\Pr\{\|(0, \delta_t)\| \geq 2,\ \|(\delta_t, a_t)\| = r\} + \sum_{j=0}^{r_1} \Pr\{\|(0, \delta_t)\| \geq r+1-j,\ \|(\delta_t, a_t)\| = j\} = O(\delta_t^2)$, we obtain

$$= \frac{(r+1)\delta_t x}{t} \sum_{j=0}^{\lfloor (r+1)/2 \rfloor} \frac{2^{r+1-2j} f_2(0,0)^j}{j!\,(r+1-2j)!}(1+o(1)). \tag{39}$$

This completes the proof of Lemma 5.

For the analysis of the minimum r-scan length, Lemmas 6 and 7 provide the necessary estimates under the continuity assumption of the density g. Lemma 6 gives the results relevant to Property 1, describing the convergence in probability of $n_t^-(a_t)$ to $N_t^-(a_t)$. Lemma 7 is required as an error bound of u_2 in (1) when applying the Chen–Stein method to $n_t^-(a_t)$.

Proof of Lemma 6. The estimates (12) and (13) give the bounds of events

$$\begin{aligned}
&\{||(0, \delta_t)|| \geq 2,\ ||(\delta_t, a_t)|| \geq r-1\} \quad \text{and} \\
&\{||(0, \delta_t)|| = 1,\ ||(0, a_t)|| < r+1,\ ||(0, a_t + \delta_t)|| \geq r+1\},
\end{aligned}$$

respectively. As discussed in Property 1, their value contributes to the estimate of the probability of the event $\{n_t^-(a_t) \neq N_t^-(a_t)\}$,

$$\begin{aligned}
&\Pr\{||(0, \delta_t)|| \geq 2,\ ||(\delta_t, a_t)|| \geq r-1\} \\
&\quad \leq \Pr\{||(0, \delta_t)|| \geq 2\} \\
&\quad = \Pr\{||A_1(0, \delta_t)|| \geq 1\} + \Pr\{||A_1(0, \delta_t)|| = 0,\ ||A_1(0, \delta_t)|| \geq 2\} \\
&\quad = O(\delta^2).
\end{aligned}$$

https://doi.org/10.1239/aap/1189518639 Published online by Cambridge University Press

and this proves the bound in (12). Paraphrasing the argument above, we can also prove the bound in (13),

$$\begin{aligned}
&\Pr\{||(0,\delta_t)|| = 1, ||(0,a_t)|| < r+1, ||(0,a_t+\delta_t)|| \geq r+1\} \\
&\quad \leq \Pr\{||(0,\delta_t)|| = 1, ||(\delta_t,a_t)|| < r, ||(a_t,a_t+\delta_t)|| \geq 1\} \\
&\quad < \Pr\{||(0,\delta_t)|| = 1, ||(a_t,a_t+\delta_t)|| \geq 1\} \\
&\quad = O(\delta_t^2).
\end{aligned}$$

Proof of Lemma 7. Paraphrasing the argument of (39), for $2 \leq i \leq \lfloor a_t/\delta_t \rfloor + 1$, we have

$$\begin{aligned}
&\mathrm{E}[Z_1^-(a_t)Z_i^-(a_t)] \\
&\quad = \Pr\{||(0,\delta_t)|| = 1, ||(\delta_t,a_t)|| \geq r||((i-1)\delta_t, i\delta_t)|| = 1, ||(i\delta_t,(i-1)\delta_t+a_t)|| \geq r\} \\
&\quad \leq \Pr\{||(0,\delta_t)|| = 1, ||((i-1)\delta_t, i\delta_t)|| = 1, ||(i\delta_t,(i-1)\delta_t+a_t)|| \geq r\} \\
&\quad = O(\delta_t\delta_t a_t^r).
\end{aligned}$$

This proves the bound in (14).

References

Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Prob.* **17,** 9–25.

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

Berg, D. E. and Howe, M. M. (1989). *Mobile DNA*. American Society for Microbiology, Washington, DC.

Bernardi, G. *et al.* (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958.

Bernardi, G., Mouchiroud, D., Gautier, C., Bernardi, G. (1988). Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Molec. Evol.* **28,** 7–18.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321,** 209–213.

Blackburn, E. H. (1991). Structure and function of telomeres. *Nature* **350**, 569–573.

Burge, C., Campbell, A. and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Nat. Acad. Sci. USA* **89,** 1358–1362.

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

Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2,** 329–337.

Ficket, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10,** 5303–5318.

Gerstein, M. (1997). A structure census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Molec. Biol.* **274,** 562–576.

Gilson, E. *et al.* (1991). Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.* **19,** 1375–1383.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. Springer, New York.

Josse, J., Kaiser, A. D. and Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. *J. Biol. Chem.* **236,** 864–875.

Karlin, S. and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* **257,** 39–49.

Karlin, S. and Cardon, L. R. (1994). Computational DNA sequence analysis. *Ann. Rev. Microbiol.* **48,** 619–654.

Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86,** 27–35.

Karlin, S., Mrázek, J. and Campbell, A. (1996). Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Res.* **24,** 4263–4272.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press.

Krawiec, S. and Riley, M. (1990). Organization of the bacterial chromosome. *Microbiol. Rev.* **54,** 502–539.

Naus, J. I. (1979). An indexed bibliography of clusters, clumps and coincidences. *Internat. Statist. Rev.* **47,** 47–78.

Naus, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77,** 177–183.

Reinert, G. and Schbath, S. (1998). Compound Poisson and Poisson approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.* **5,** 223–254.

Willard, H. F. and Waye, J. S. (1987). Hierachical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3,** 192–198.