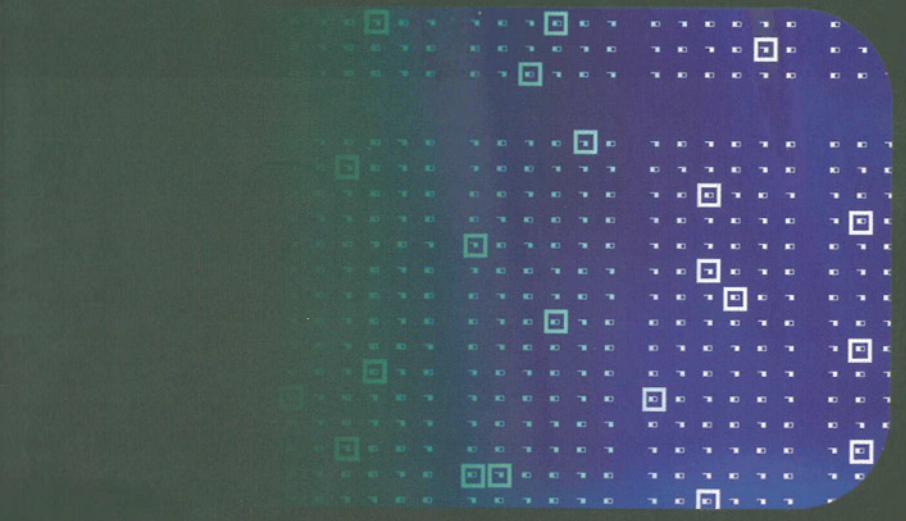
Lecture Notes Series, Institute for Mathematical Sciences,
National University of Singapore
Vol. 4



A D Barbour
Louis H Y Chen

AN INTRODUCTION TO STEIN'S METHOD

AN INTRODUCTION TO STEIN'S METHOD

LECTURE NOTES SERIES
Institute for Mathematical Sciences, National University of Singapore

Series Editors: Louis H. Y. Chen and Denny Leung
Institute for Mathematical Sciences
National University of Singapore

Published

Vol. 1 Coding Theory and Cryptology
edited by Harald Niederreiter

Vol. 2 Representations of Real and p-Adic Groups
edited by Eng-Chye Tan & Chen-Bo Zhu

Vol. 3 Selected Topics in Post-Genome Knowledge Discovery
edited by Limsoon Wong & Louxin Zhang

Vol. 4 An Introduction to Stein's Method
edited by A. D. Barbour & Louis H. Y. Chen

Lecture Notes Series, Institute for Mathematical Sciences,
National University of Singapore
Vol. 4

AN INTRODUCTION TO STEIN'S METHOD

A D Barbour
University of Zürich, Switzerland

Louis H Y Chen
National University of Singapore, Singapore

SINGAPORE UNIVERSITY PRESS
NATIONAL UNIVERSITY OF SINGAPORE

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

Singapore University Press
Yusof Ishak House, National University of Singapore
31 Lower Kent Ridge Road, Singapore 119078

and

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

AN INTRODUCTION TO STEIN'S METHOD

ISBN 981-256-280-X
ISBN 981-256-330-X (pbk)

Printed in Singapore by World Scientific Printers (S) Pte Ltd

CONTENTS

FOREWORD

The Institute for Mathematical Sciences at the National University of Singapore was established on 1 July 2000 with funding from the Ministry of Education and the University. Its mission is to provide an international center of excellence in mathematical research and, in particular, to promote within Singapore and the region active research in the mathematical sciences and their applications. It seeks to serve as a focal point for scientists of diverse backgrounds to interact and collaborate in research through tutorials, workshops, seminars and informal discussions.

The Institute organizes thematic programs of duration ranging from one to six months. The theme or themes of each program will be in accordance with the developing trends of the mathematical sciences and the needs and interests of the local scientific community. Generally, for each program there will be tutorial lectures on background material followed by workshops at the research level.

As the tutorial lectures form a core component of a program, the lecture notes are usually made available to the participants for their immediate benefit during the period of the tutorial. The main objective of the Institute's Lecture Notes Series is to bring these lectures to a wider audience. Occasionally, the Series may also include the proceedings of workshops and expository lectures organized by the Institute. The World Scientific Publishing Company and the Singapore University Press have kindly agreed to publish jointly the Lecture Notes Series. This volume, "An introduction to Stein's method," is the 4th of this Series. We hope that through regular publication of lecture notes the Institute will achieve, in part, its objective of promoting research in the mathematical sciences and their applications.

December 2004

Louis H. Y. Chen
Denny Leung
Series Editors

PREFACE

Probability theory in the first half of the twentieth century was substantially dominated by the formulation and proof of the classical limit theorems — laws of large numbers, central limit theorem, law of the iterated logarithm — for sums of independent random variables. The central limit theorem in particular has found regular application in statistics, and forms the basis of the distribution theory of many test statistics. However, the classical approach to the CLT relied heavily on Fourier methods, which are not naturally suited to providing estimates of the accuracy of limits such as the CLT as approximations in pre-limiting circumstances, and it was only in 1940 that Berry and Esseen, by means of the smoothing inequality, first obtained the correct rate of approximation in the form of an explicit, universal bound. Curiously enough, the comparable theorem for the conceptually simpler Poisson law of small numbers was not proved until 26 years later, by Le Cam.

These theorems all concerned sums of independent random variables. However, dependence is the rule rather than the exception in applications, and had been increasingly studied since 1950. Without independence, Fourier methods are much more difficult to apply, and bounds for the accuracy of approximations become correspondingly more difficult to find; even for such frequently occurring settings as sums of stationary, mixing random variables or the combinatorial CLT, establishing good rates seemed to be intractable.

It was into this situation that Charles Stein introduced his startling technique for normal approximation. Now known simply as Stein's method, the technique relies on an indirect approach, involving a differential operator and a cleverly chosen exchangeable pair of random variables, which are combined in almost magical fashion to deliver explicit estimates of approximation error, with or without independence. This latter feature, in particular, has led to the wide range of application of the method.

Stein originally developed his method to provide a new proof of the combinatorial CLT for use in a lecture course, and its first published application, in the *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* in 1972, was to give bounds for the accuracy of the CLT for sums of stationary, mixing random variables. Since then, the scope of his discovery has expanded rapidly. Poisson approximation was studied in 1975; the correct Lyapounov error bound in the combinatorial CLT was obtained in 1984; the method was extended to the approximation of the distributions of whole random processes in 1988; its importance in the theoretical underpinning of molecular sequence comparison algorithms was recognized in 1989; rates of convergence in the multivariate CLT were derived in 1991; good general bounds in the multivariate CLT, when dependence is expressed in terms of neighborhoods of possibly very general structure, were given in 1996; and Stein's idea of arguing by way of a concentration inequality was developed in 2001 to a point where it can be put to very effective use.

Despite the progress made over the last 30 years, the reasons for the effectiveness of Stein's method still remain something of a mystery. There are still many open problems, even at a rather basic level. Controlling the behavior of the solutions of the Stein equation, fundamental to the success of the method, is at present a difficult task, if the probabilistic approach cannot be used. The field of multivariate discrete distributions is almost untouched. There is a relative of the concentration inequality approach, involving the comparison of a distribution with its translations, which promises much, but is at present in its early stages. Point process approximation, other than in the Poisson context, is largely unexplored: the list goes on.

Due to its broad range of application, Stein's method has become particularly important, not only in the future development of probability theory, but also in a wide range of other fields, some theoretical, some extremely practical. These include spatial statistics, computer science, the theory of random graphs, computational molecular biology, interacting particle systems, the bootstrap, the mathematical theory of epidemics, algebraic analogues of probabilistic number theory, insurance and financial mathematics, population ecology and the combinatorics of logarithmic structures. Many, in their turn, because of their particular structure, have led to the development of variants of Stein's original approach, with their own theoretical importance, one such being the coupling method.

This volume contains an introduction to Stein's method in four chapters, corresponding to the tutorial lectures given during the meeting

STEIN'S METHOD AND APPLICATIONS:
A PROGRAM IN HONOR OF CHARLES STEIN,

held in Singapore at the Institute for Mathematical Sciences, from 28 July to 31 August 2003. The material provides a detailed introduction to the theory and application of Stein's method, in a form suitable for graduate students who want to acquaint themselves with the method. The accompanying volume, consisting of papers given at the workshop held during the same meeting, provides a cross-section of the research work currently being undertaken in this area.

To get a flavour of the magic and mystery of Stein's method, take the following elementary setting: $X_1, X_2, \ldots, X_n$ are independent 0–1 random variables, with $\mathbb{P}[X_i = 1] = 1 - \mathbb{P}[X_i = 0] = p_i$, and W denotes their sum. How close is the distribution $\mathcal{L}(W)$ to the Poisson distribution $\mathrm{Po}(\lambda)$ with mean $\lambda = \sum_{i=1}^n p_i$? A good answer can be obtained in three small steps.

(1) For any $A \subset \mathbb{Z}_+$, recursively define the function $g = g_{\lambda,A}$ on $\mathbb{Z}_+$ by setting $g(0) = 0$ and then

$$\lambda g(j+1) = jg(j) + \mathbf{1}_A(j) - \mathrm{Po}(\lambda)\{A\} \tag{0.1}$$

for $j = 0, 1, 2, \ldots$. Then, by taking expectations, it follows that

$$\mathbb{P}[W \in A] - \mathrm{Po}(\lambda)\{A\} = \mathbb{E}\{\lambda g(W+1) - Wg(W)\}, \tag{0.2}$$

as long as $jg(j)$ is bounded in j (it is).

(2) Then note that $X_i g(W) = X_i g(W_i + 1)$, where $W_i = \sum_{j \neq i} X_j$, because X_i takes only the values 0 and 1. Since also W_i is *independent* of X_i, it thus follows that $\mathbb{E}\{X_i g(W)\} = p_i \mathbb{E} g(W_i + 1)$, and hence that

$$\mathbb{E}\{Wg(W)\} = \sum_{i=1}^n p_i \mathbb{E} g(W_i + 1). \tag{0.3}$$

(3) Combining (0.2) and (0.3), we have

$$\begin{aligned} |\mathbb{P}[W \in A] - \mathrm{Po}(\lambda)\{A\}| &= \left| \sum_{i=1}^n p_i \mathbb{E}[g(W+1) - g(W_i+1)] \right| \\ &= \left| \sum_{i=1}^n p_i \mathbb{E}[g(W_i + X_i + 1) - g(W_i + 1)] \right|, \end{aligned}$$

from which it follows that

$$|\mathbb{P}[W \in A] - \mathrm{Po}\,(\lambda)\{A\}| \leq k(\lambda) \sum_{i=1}^{n} p_i^2 \tag{0.4}$$

for all $A \subset \mathbb{Z}_+$, where

$$k(\lambda) := \sup_{A \subset \mathbb{Z}_+} \sup_{j \geq 1} |g_{\lambda,A}(j+1) - g_{\lambda,A}(j)|,$$

since X_i differs from 0 only with probability p_i, when it takes the value 1. And it can also be shown that $k(\lambda) \leq (1 - e^{-\lambda})/\lambda$.

The upshot of this argument is that the difference between the probability given by $\mathcal{L}(W)$ to *any* set A and that assigned to it by $\mathrm{Po}\,(\lambda)$ is at most

$$\lambda^{-1}(1 - e^{-\lambda}) \sum_{i=1}^{n} p_i^2 \leq \max_{1 \leq i \leq n} p_i, \tag{0.5}$$

a remarkably neat and surprisingly sharp result. This volume shows how the simple argument that led to it (the Stein–Chen method) fits into the much more general and powerful framework of Stein's method. Reasons are advanced for choosing equation (0.1) in connection with the Poisson distribution $\mathrm{Po}\,(\lambda)$. Some rules are given for constructing analogous equations for other distributions, both on the line and on more elaborate spaces, such as measure spaces, and some help is also provided with bounding the counterparts of $k(\lambda)$ that emerge. Finally, ways of modifying (0.3) when W is a sum of dependent random elements are also proposed.

The material is arranged in four chapters, successively addressing the normal distribution, Poisson and compound Poisson distributions, Poisson point processes, and then quite general distributions. Each chapter is written by an expert in the field. We hope that the resulting tutorial survey will encourage the reader to become as enthusiastic about Stein's method as we are.

December 2004

A. D. Barbour
Louis H. Y. Chen
Program Co-Chairs

Stein's method for normal approximation

Louis H. Y. Chen and Qi-Man Shao
Institute for Mathematical Sciences, National University of Singapore
3 Prince George's Park, Singapore 118402
E-mail: lhychen@ims.nus.edu.sg
and
Department of Statistics and Applied Probability
National University of Singapore
6 Science Drive 2, Singapore 117543;
Department of Mathematics, University of Oregon
Eugene, OR 97403, USA
E-mail: qmshao@darkwing.uoregon.edu

Stein's method originated in 1972 in a paper in the Proceedings of the Sixth Berkeley Symposium. In that paper, he introduced the method in order to determine the accuracy of the normal approximation to the distribution of a sum of dependent random variables satisfying a mixing condition. Since then, many developments have taken place, both in extending the method beyond normal approximation and in applying the method to problems in other areas. In these lecture notes, we focus on univariate normal approximation, with our main emphasis on his approach exploiting an *a priori* estimate of the concentration function. We begin with a general explanation of Stein's method as applied to the normal distribution. We then go on to consider expectations of smooth functions, first for sums of independent and locally dependent random variables, and then in the more general setting of exchangeable pairs. The later sections are devoted to the use of concentration inequalities, in obtaining both uniform and non-uniform Berry–Esseen bounds for independent random variables. A number of applications are also discussed.

Contents

1. Introduction

Stein's method is a way of deriving explicit estimates of the accuracy of the approximation of one probability distribution by another. This is accomplished by comparing expectations, as indicated in the title of Stein's (1986) monograph, *Approximate computation of expectations*. An upper bound is computed for the difference between the expectations of any one of a (large) family of test functions under the two distributions, each family of test functions determining an associated metric. Any such bound in turn implies a corresponding upper bound for the distance between the two distributions, measured with respect to the associated metric.

Thus, if the family of test functions consists of the indicators of all (measurable) subsets, then accuracy is expressed in terms of the total variation

distance d_{TV} between the two distributions:

$$d_{TV}(P,Q) := \sup_{h\in\mathcal{H}} \left| \int h\,dP - \int h\,dQ \right| = \sup_A |P(A) - Q(A)|,$$

where $\mathcal{H} = \{\mathbf{1}_A;\ A \text{ measurable}\}$. If the distributions are on $\mathbb{R}$, and the test functions are the indicators of all half-lines, then accuracy is expressed using Kolmogorov distance, as is customary in Berry–Esseen theorems:

$$d_K(P,Q) := \sup_{h\in\mathcal{H}} \left| \int h\,dP - \int h\,dQ \right| = \sup_{z\in\mathbf{R}} |P(-\infty,z] - Q(-\infty,z]|,$$

where $\mathcal{H} = \{\mathbf{1}_{(-\infty,z]};\ z \in \mathbb{R}\}$. If the test functions consist of all uniformly Lipschitz functions h with constant bounded by 1, then the Wasserstein distance results:

$$d_W(P,Q) := \sup_{h\in\mathcal{H}} \left| \int h\,dP - \int h\,dQ \right|,$$

where $\mathcal{H} = \{h\colon \mathbb{R} \to \mathbb{R};\ \|h'\| \le 1\} =: \mathrm{Lip}(1)$, and where, for a function $g\colon \mathbb{R} \to \mathbb{R}$, $\|g\|$ denotes $\sup_{x\in\mathbb{R}} |g(x)|$. If the test functions are uniformly bounded and uniformly Lipschitz, bounded Wasserstein distances result:

$$d_{BW(k)}(P,Q) := \sup_{h\in\mathcal{H}} \left| \int h\,dP - \int h\,dQ \right|,$$

where $\mathcal{H} = \{h\colon \mathbb{R} \to \mathbb{R};\ \|h\| \le 1, \|h'\| \le k\}$. Stein's method applies to all of these distances, and to many more.

For normal approximation on $\mathbb{R}$, Stein began with the observation that

$$\mathbb{E}\{f'(Z) - Zf(Z)\} = 0 \tag{1.1}$$

for *any* bounded function f with bounded derivative, if Z has the standard normal distribution $\mathcal{N}(0,1)$. This can be verified by partial integration:

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x) e^{-x^2/2}\,dx \\
&\quad = \left[\frac{1}{\sqrt{2\pi}} f(x) e^{-x^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x f(x) e^{-x^2/2}\,dx \\
&\quad = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x f(x) e^{-x^2/2}\,dx.
\end{aligned}$$

However, the same partial integration can also be used to solve the differential equation

$$f'(x) - xf(x) = g(x), \qquad \lim_{x\downarrow-\infty} f(x) e^{-x^2/2} = 0, \tag{1.2}$$

for g any bounded function, giving

$$\begin{aligned}\int_{-\infty}^{y} g(x)e^{-x^2/2}\,dx &= \int_{-\infty}^{y} \{f'(x) - xf(x)\}e^{-x^2/2}\,dx \\ &= \int_{-\infty}^{y} \frac{d}{dx}\{f(x)e^{-x^2/2}\}\,dx \\ &= f(y)e^{-y^2/2},\end{aligned}$$

and hence

$$f(y) = e^{y^2/2}\int_{-\infty}^{y} g(x)e^{-x^2/2}\,dx.$$

Note that this f actually satisfies $\lim_{y\downarrow-\infty} f(y) = 0$, because

$$\Phi(y) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-x^2/2}\,dx \sim (y\sqrt{2\pi})^{-1}e^{-y^2/2}$$

as $y \downarrow -\infty$, and that f is bounded (and then, by the same argument, has $\lim_{y\uparrow\infty} f(y) = 0$) if and only if $\int_{-\infty}^{\infty} g(x)e^{-x^2/2}\,dx = 0$, or, equivalently, if $\mathbb{E}g(Z) = 0$.

Hence, taking any bounded function h, we observe that the function f_h, defined by

$$f_h(x) := e^{x^2/2}\int_{-\infty}^{x} \{h(t) - \mathbb{E}h(Z)\}e^{-t^2/2}\,dt, \tag{1.3}$$

satisfies (1.2) for $g(x) = h(x) - \mathbb{E}h(Z)$; substituting any random variable W for x in (1.2) and taking expectations then yields

$$\mathbb{E}\{f_h'(W) - Wf_h(W)\} = \mathbb{E}h(W) - \mathbb{E}h(Z). \tag{1.4}$$

Thus the characterization (1.1) of the standard normal distribution also delivers an upper bound for normal approximation with respect to any of the distances introduced above: for any class $\mathcal{H}$ of (bounded) test functions h,

$$\sup_{h\in\mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| = \sup_{h\in\mathcal{H}} |\mathbb{E}\{f_h'(W) - Wf_h(W)\}|. \tag{1.5}$$

The curious fact is that the supremum on the right hand side of (1.5), which contains only the random variable W, is frequently much simpler to bound than that on the left hand side. The differential characterization (1.1) of the normal distribution is reflected in the fact that the quantity $\mathbb{E}\{f'(W) - Wf(W)\}$ is often relatively easily shown to be small, when the structure of W is such as to make normal approximation seem plausible; for instance, when W is a sum of individually small and only weakly dependent

random variables. This is illustrated in the following elementary argument for the classical central limit theorem.

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, with mean zero and unit variance, and such that $\mathbb{E}|X_1|^3 < \infty$; let $W = n^{-1/2} \sum_{j=1}^n X_j$. We evaluate the quantity $\mathbb{E}\{f'(W) - Wf(W)\}$ for a smooth and very well-behaved function f. Since W is a sum of identically distributed components, we have

$$\mathbb{E}\{Wf(W)\} = n\mathbb{E}\{n^{-1/2}X_1 f(W)\},$$

and $W = n^{-1/2}X_1 + W_1$, where $W_1 = n^{-1/2} \sum_{j=2}^n X_j$ is independent of X_1. Hence, by Taylor's expansion,

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= n^{1/2}\mathbb{E}\{X_1 f(W_1 + n^{-1/2}X_1)\} \\ &= n^{1/2}\mathbb{E}\{X_1(f(W_1) + n^{-1/2}X_1 f'(W_1))\} + \eta_1, \quad (1.6)\end{aligned}$$

where $|\eta_1| \leq \frac{1}{2}n^{-1/2}\mathbb{E}|X_1|^3 \, \|f''\|$. On the other hand, again by Taylor's expansion,

$$\mathbb{E}\{f'(W) = \mathbb{E}f'(W_1 + n^{-1/2}X_1) = \mathbb{E}f'(W_1) + \eta_2, \quad (1.7)$$

where $|\eta_2| \leq n^{-1/2}\mathbb{E}|X_1| \, \|f''\|$. But now, since $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$, and since X_1 and W_1 are independent, it follows that

$$|\mathbb{E}\{f'(W) - Wf(W)\}| \leq n^{-1/2}\{1 + \tfrac{1}{2}\mathbb{E}|X_1|^3\}\|f''\|,$$

for any f with bounded second derivative. Hence, if $\mathcal{H}$ is any class of test functions and

$$C_{\mathcal{H}} := \sup_{h \in \mathcal{H}} \|f_h''\|,$$

then it follows that

$$\sup_{h \in \mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq C_{\mathcal{H}} \, n^{-1/2}\{1 + \tfrac{1}{2}\mathbb{E}|X_1|^3\}. \quad (1.8)$$

The inequality (1.8) thus establishes an accuracy of order $O(n^{-1/2})$ for the normal approximation of W, with respect to the distance $d_{\mathcal{H}}$ on probability measures on $\mathbb{R}$ defined by

$$d_{\mathcal{H}}(P, Q) = \sup_{h \in \mathcal{H}} \left| \int h \, dP - \int h \, dQ \right|,$$

provided only that $C_{\mathcal{H}} < \infty$. For example, as observed by Erickson (1974), the class of test functions h for the Wasserstein distance d_W is just the Lipschitz functions Lip(1) with constant no greater than 1, and, for this class, $C_{\mathcal{H}} = 2$: see (2.13) of Lemma 2.3. The distinction between the bounds

for different distances is then reflected in the differences between the values of $C_{\mathcal{H}}$. Computing good estimates of $C_{\mathcal{H}}$ involves studying the properties of the solutions f_h to the Stein equation given in (1.3) for $h \in \mathcal{H}$. Fortunately, for normal approximation, a lot is known.

It sadly turns out that this simple argument is not enough to prove the Berry–Esseen theorem, which states that

$$\sup_{z \in \mathbf{R}} |\mathbb{P}[W \le z] - \Phi(z)| \le C n^{-1/2} \mathbb{E}|X_1|^3,$$

for a universal constant C. This is because, for $\mathcal{H}$ the set of indicators of half-lines, $C_{\mathcal{H}} = \infty$. However, it is possible to refine the argument, by making less crude estimates of $|\eta_1|$ and $|\eta_2|$. Broadly speaking, if $h = \mathbf{1}_{(-\infty,z]}$, and writing f_z for the corresponding solution f_h, then f_z' is smooth, except for a jump discontinuity at z, as can be seen from (1.2). Hence, taking $|\eta_2|$ for illustration,

$$\begin{aligned} &|f_z'(W_1 + n^{-1/2}X_1) - f_z'(W_1)| \\ &\quad \le n^{-1/2}|X_1| M(z, f_z)(|W_1| + n^{-1/2}|X_1| + 1), \end{aligned} \tag{1.9}$$

where $M(z,f) := \sup_{x \ne z}\{|f''(x)|/(|x|+1)\}$, provided that z does not lie between $W_1 + n^{-1/2}X_1$ and W_1. If it does, the difference is still bounded by $2\|f_z'\|$. Hence, since both $\sup_{z \in \mathbf{R}} \|f_z'\|$ and $\sup_{z \in \mathbf{R}} M(z, f_z)$ are finite, explicit bounds of order $O(n^{-1/2})$ can still be deduced for the Berry–Esseen theorem, provided that it can be shown that

$$\sup_{z \in \mathbf{R}} \mathbb{P}(W_1 \in [z, z+h]) \le C(h + n^{-1/2}\mathbb{E}|X_1|^3) \tag{1.10}$$

for some constant C (the corresponding estimates for $|\eta_1|$ are a little more involved, but the argument can be carried through analogously). Inequalities of this form are precisely the concentration inequalities that play a central part in these notes, and the above discussion illustrates their importance.

If the random variables X_i are dependent, these simple arguments are no longer valid. However, they can be adapted to work effectively, if less cleanly, for many weak dependence structures. For instance, if individual summands have little influence on W, as is typical when a normal approximation is expected to be good, then it may be possible to decompose W in the form

$$W = n^{-1/2}(X_i + U_i) + W_i,$$

where W_i is 'almost' independent of X_i, and $|U_i|$ is not too 'large', enabling Taylor expansions analogous to (1.6) and (1.7) to be attempted.

However, especially if Kolmogorov distance is to be considered, it pays to be rather less brutal, making direct use of (1.4), and rewriting the errors in exact, integral form; this will appear time and again in what follows. Better bounds can be achieved as a result, and concentration arguments also become feasible: note that finding sharp bounds for a probability such as $\mathbb{P}(a \leq W_i \leq a + n^{-1/2}(X_i + U_i))$ is not an easy task when W_i, X_i and U_i are dependent.

In these lecture notes, we follow Stein's original theme, by presenting his ideas in the context of normal approximation. We shall focus on the approach using concentration inequalities. This approach dates back to Stein's lectures around 1970 (see Ho & Chen (1978)). It was later extended by Chen (1986) to dependent and non-identically distributed random variables with arbitrary index set. A proof of the Berry–Esseen theorem for independent and non-identically distributed random variables using the concentration inequality approach is given in Section 2 of Chen (1998). The approach has further been shown to be effective in obtaining not only uniform but also non-uniform Berry–Esseen bounds for the accuracy of normal approximation for independent random variables (Chen & Shao, 2001) and for locally dependent random fields as well (Chen & Shao, 2004a). As an extension, a randomized concentration inequality was used to establish uniform and non-uniform Berry–Esseen bounds for non-linear statistics in Chen & Shao (2004b). In view of its current successes, the technique seems well worth further exploration.

We also briefly discuss the exchangeable pairs coupling, and use it to obtain a Berry–Esseen bound for the number of 1's in the binary expansion of a random integer. The use of exchangeable pairs was discussed in Stein's monograph (Stein, 1986), and is now widely known (see, for example, Diaconis, Holmes & Reinert (2004)). Other proofs of the Berry–Esseen theorem which will not be included in these lectures are the inductive argument of Barbour & Hall (1984), Stroock (1993) and Bolthausen (1984), the size bias coupling used by Goldstein & Rinott (1996) and the zero bias transformation introduced by Goldstein & Reinert (1997). The inductive argument works well for uniform bounds when the conditional distribution of the sum under consideration, given any one of the variables in the sum, has a form similar to that of the corresponding unconditional distribution of the sum. The size biasing method works well for many combinatorial problems, such as counting the number of vertices in a random graph; theoretically speaking, the zero bias transformation approach works for arbitrary random variables W with mean zero and finite variance. More on these approaches

can be found in Sections 4.5, 5.2 and 8 of Chapter 4 of this volume. We also refer to Goldstein & Reinert (1997) for some of the important features of the zero bias transformation and to Goldstein (2005) for Berry–Esseen bounds for combinatorial central limit theorems and pattern occurrences using zero and size biasing approaches. Short surveys of normal approximation by Stein's method were given in Rinott & Rotar (2000) and in Chen (1998). Stein's method has also been used to establish bounds on the approximation error in the multivariate central limit theorem, in particular by Götze (1991) and by Rinott & Rotar (1996). A collection of expository lectures on a variety of aspects of Stein's method and its applications has recently been edited by Diaconis & Holmes (2004).

These notes are organized as follows. We begin with a more detailed account of Stein's method than the sketch provided in this section. In Section 3, we consider the expectation of smooth functions under independence and local dependence, and also in the setting of an exchangeable pair having the linear regression property. Some applications are given. Section 4 discusses sums of uniformly bounded random variables. Here, it is shown that Berry–Esseen bounds for distribution functions can be obtained without the use of concentration inequalities. Both the independent case and the problem of counting the number of ones in the binary expansion of a random integer are studied, where in the latter case the use of an exchangeable pair is demonstrated. In Sections 5 and 6, concentration inequalities are invoked to obtain both uniform and non-uniform Berry–Esseen bounds for sums for independent random variables. The last section is devoted to obtaining a uniform Berry–Esseen bound under local dependence.

2. Fundamentals of Stein's method

In this section, we follow Stein (1986), and give a more detailed account of the basic method sketched in the introduction.

2.1. *Characterization*

Let Z be a standard normally distributed random variable, and let $\mathcal{C}_{bd}$ be the set of continuous and piecewise continuously differentiable functions $f\colon\ R \to R$ with $\mathbb{E}|f'(Z)| < \infty$. Stein's method rests on the following characterization, which is a slight strengthening of (1.1).

Lemma 2.1: *Let W be a real valued random variable. Then W has a standard normal distribution if and only if*

$$\mathbb{E}f'(W) = \mathbb{E}\{Wf(W)\}, \tag{2.1}$$

for all $f \in \mathcal{C}_{bd}$.

Proof: *Necessity.* If W has a standard normal distribution, then for $f \in \mathcal{C}_{bd}$

$$\begin{aligned}\mathbb{E}f'(W) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(w) e^{-w^2/2}\, dw \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(w) \Big(\int_{-\infty}^{w} (-x) e^{-x^2/2}\, dx \Big)\, dw \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(w) \Big(\int_{w}^{\infty} x e^{-x^2/2}\, dx \Big)\, dw.\end{aligned}$$

By Fubini's theorem, it thus follows that

$$\begin{aligned}\mathbb{E}f'(W) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \Big(\int_{x}^{0} f'(w)\, dw \Big)(-x) e^{-x^2/2}\, dx \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \Big(\int_{0}^{x} f'(w)\, dw \Big) x e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [f(x) - f(0)] x e^{-x^2/2}\, dx \\ &= \mathbb{E}Wf(W).\end{aligned}$$

Sufficiency. For fixed $z \in \mathbb{R}$, let $f(w) := f_z(w)$ denote the solution of the equation

$$f'(w) - wf(w) = \mathbf{1}_{(-\infty,z]}(w) - \Phi(z). \tag{2.2}$$

Multiplying by $-e^{-w^2/2}$ on both sides of (2.2) yields

$$\left(e^{-w^2/2} f(w)\right)' = -e^{-w^2/2} \left(\mathbf{1}_{(-\infty,z]}(w) - \Phi(z)\right).$$

Thus f_z is given by

$$\begin{aligned}f_z(w) &= e^{w^2/2} \int_{-\infty}^{w} [\mathbf{1}_{(-\infty,z]}(x) - \Phi(z)] e^{-x^2/2}\, dx \\ &= -e^{w^2/2} \int_{w}^{\infty} [\mathbf{1}_{(-\infty,z]}(x) - \Phi(z)] e^{-x^2/2}\, dx \\ &= \begin{cases} \sqrt{2\pi} e^{w^2/2} \Phi(w)[1 - \Phi(z)] & \text{if } w \le z, \\ \sqrt{2\pi} e^{w^2/2} \Phi(z)[1 - \Phi(w)] & \text{if } w \ge z. \end{cases}\end{aligned} \tag{2.3}$$

By Lemma 2.2 below, f_z is a bounded continuous and piecewise continuously differentiable function. Suppose that (2.1) holds for all $f \in \mathcal{C}_{bd}$. Then it holds for f_z, and hence, by (2.2)

$$0 = \mathbb{E}[f_z'(W) - Wf_z(W)] = \mathbb{E}[\mathbf{1}_{(-\infty,z]}(w) - \Phi(z)] = P(W \le z) - \Phi(z).$$

Thus W has a standard normal distribution. ∎

Equation (2.2) is a particular case of the more general Stein equation

$$f'(w) - wf(w) = h(w) - \mathbb{E}h(Z), \tag{2.4}$$

which is to be solved for f, given a real valued measurable function h with $\mathbb{E}|h(Z)| < \infty$: c.f. (1.2). As for (2.3), the solution $f = f_h$ is given by

$$\begin{aligned} f_h(w) &= e^{w^2/2} \int_{-\infty}^{w} [h(x) - \mathbb{E}h(Z)]e^{-x^2/2}\,dx \\ &= -e^{w^2/2} \int_{w}^{\infty} [h(x) - \mathbb{E}h(Z)]e^{-x^2/2}\,dx. \end{aligned} \tag{2.5}$$

2.2. *Properties of the solutions*

We now list some basic properties of the solutions (2.3) and (2.5) to the Stein equations (2.2) and (2.4). The reasons why we need them have already been indicated in (1.8) and (1.9), where estimates of $\sup_{h\in\mathcal{H}} \|f_h''\|$, $\sup_{z\in\mathbf{R}} \|f_z'\|$ and $\sup_{z\in\mathbf{R}} \sup_{x\neq z} |f_z'(x)|$ were required, in order to determine our error bounds for the various approximations. For the more detailed arguments to come, further properties are also needed. We defer the proofs to the appendix, since they only involve careful real analysis.

We begin by considering the solution f_z to (2.2), given in (2.3).

Lemma 2.2: *The function f_z defined by (2.3) is such that*

$$wf_z(w) \text{ is an increasing function of } w. \tag{2.6}$$

Moreover, for all real w, u and v,

$$|wf_z(w)| \le 1, \quad |wf_z(w) - uf_z(u)| \le 1 \tag{2.7}$$

$$|f_z'(w)| \le 1, \ |f_z'(w) - f_z'(v)| \le 1 \tag{2.8}$$

$$0 < f_z(w) \le \min(\sqrt{2\pi}/4, 1/|z|) \tag{2.9}$$

and

$$|(w+u)f_z(w+u) - (w+v)f_z(w+v)| \le (|w| + \sqrt{2\pi}/4)(|u| + |v|). \tag{2.10}$$

We mostly use (2.8) and (2.9) for our approximations. If one does not care about the constants, one can easily obtain

$$|f_z'(w)| \le 2 \quad \text{and} \quad 0 < f_z(w) \le \sqrt{\pi/2}$$

by using the well-known inequality

$$1 - \Phi(w) \le \min\left(\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right) e^{-w^2/2}, \qquad w > 0.$$

Next, we consider the solution f_h to the Stein equation (2.4), as given in (2.5), for any bounded absolutely continuous function h.

Lemma 2.3: *For any absolutely continuous function* $h\colon \mathbb{R} \to \mathbb{R}$, *the solution* f_h *given in* (2.5) *satisfies*

$$\|f_h\| \le \min\left(\sqrt{\pi/2}\|h(\cdot) - \mathbb{E}h(Z)\|, 2\|h'\|\right), \tag{2.11}$$

$$\|f_h'\| \le \min\left(2\|h(\cdot) - \mathbb{E}h(Z)\|, 4\|h'\|\right) \tag{2.12}$$

and

$$\|f_h''\| \le 2\|h'\|. \tag{2.13}$$

2.3. *Construction of the Stein identities*

In this section, we return to the elementary use of Stein's method which led to the simple bound (1.8), but express the remainders η_1 and η_2 in a form better suited to deriving more advanced results. We now take $\xi_1, \xi_2, \cdots, \xi_n$ to be independent random variables satisfying $\mathbb{E}\xi_i = 0$, $1 \le i \le n$, and such that $\sum_{i=1}^n \mathbb{E}\xi_i^2 = 1$. Thus ξ_i corresponds to the normalized random variable $n^{-1/2}X_i$ of the introduction; here, however, we no longer require the ξ_i to be identically distributed. Much as before, we set

$$W = \sum_{i=1}^n \xi_i \quad \text{and} \quad W^{(i)} = W - \xi_i, \tag{2.14}$$

and define

$$K_i(t) = \mathbb{E}\{\xi_i(I_{\{0 \le t \le \xi_i\}} - I_{\{\xi_i \le t < 0\}})\}. \tag{2.15}$$

It is easy to check that $K_i(t) \ge 0$ for all real t, that

$$\int_{-\infty}^{\infty} K_i(t)\,dt = \mathbb{E}\xi_i^2 \quad \text{and that} \quad \int_{-\infty}^{\infty} |t| K_i(t)\,dt = \tfrac{1}{2}\mathbb{E}|\xi_i|^3. \tag{2.16}$$

Let h be a measurable function with $\mathbb{E}|h(Z)| < \infty$, and let $f = f_h$ be the corresponding solution of the Stein equation (2.4). Our goal is to estimate

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}\{f'(W) - Wf(W)\}.$$

The following argument is fundamental to the approach, and many of the tricks reappear repeatedly in what follows.

Since ξ_i and $W^{(i)}$ are independent for each $1 \le i \le n$, we have

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= \sum_{i=1}^{n} \mathbb{E}\{\xi_i f(W)\} \\ &= \sum_{i=1}^{n} \mathbb{E}\{\xi_i [f(W) - f(W^{(i)})]\},\end{aligned}$$

where the last equality follows because $\mathbb{E}\xi_i = 0$. Writing the final difference in integral form, we thus have

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= \sum_{i=1}^{n} \mathbb{E}\left\{\xi_i \int_0^{\xi_i} f'(W^{(i)} + t)\, dt\right\} \\ &= \sum_{i=1}^{n} \mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W^{(i)} + t)\xi_i (I_{\{0 \le t \le \xi_i\}} - I_{\{\xi_i \le t < 0\}})\, dt\right\} \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\left\{f'(W^{(i)} + t)\right\} K_i(t)\, dt, \qquad (2.17)\end{aligned}$$

from the definition of K_i and again using independence. However, because

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_i(t)\, dt = \sum_{i=1}^{n} \mathbb{E}\xi_i^2 = 1,$$

it follows that

$$\mathbb{E}f'(W) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\{f'(W)\} K_i(t)\, dt. \qquad (2.18)$$

Thus, by (2.17) and (2.18),

$$\mathbb{E}\{f'(W) - Wf(W)\} = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\{f'(W) - f'(W^{(i)} + t)\} K_i(t)\, dt. \qquad (2.19)$$

Equations (2.17) and (2.19) play a key role in proving good normal approximations. Note in particular that (2.19) is an *equality*, replacing the clumsier bounds for $|\eta_1|$ and $|\eta_2|$ which led to (1.8). This more careful treatment of the errors pays big dividends later on. Note also that (2.17) and (2.19) hold for all bounded absolute continuous f.

3. Normal approximation for smooth functions

Our goal in this section is to estimate $\mathbb{E}h(W) - \mathbb{E}h(Z)$ for various classes of random variables W, when h is a smooth function satisfying

$$\|h'\| := \sup_x |h'(x)| < \infty. \tag{3.1}$$

Such estimates lead naturally to bounds on the accuracy of the standard normal approximation to W in terms of the Wasserstein distance d_W, and hence suffice to prove the central limit theorem. The next theorem highlights this, and also shows that Wasserstein bounds imply bounds, albeit rather weaker, with respect to the Kolmogorov distance d_K.

Theorem 3.1: *Assume that there exists a δ such that, for any uniformly Lipschitz function h,*

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq \delta \|h'\|. \tag{3.2}$$

Then

$$d_W(\mathcal{L}(W), \mathcal{N}(0,1)) := \sup_{h \in Lip(1)} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq \delta; \tag{3.3}$$

$$d_K(\mathcal{L}(W), \mathcal{N}(0,1)) := \sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq 2\delta^{1/2}. \tag{3.4}$$

Proof: The first statement is immediate from the definition of d_W. For the second, we can assume that $\delta \leq 1/4$, since otherwise (3.4) is trivial. Let $\alpha = \delta^{1/2}(2\pi)^{1/4}$, and define for fixed z

$$h_\alpha(w) = \begin{cases} 1 & \text{if } w \leq z, \\ 0 & \text{if } w \geq z + \alpha, \\ \text{linear} & \text{if } z \leq w \leq z + \alpha. \end{cases}$$

Then $\|h'\| = 1/\alpha$, and hence, by (3.2),

$$\begin{aligned} \mathbb{P}(W \leq z) - \Phi(z) &\leq \mathbb{E}h_\alpha(W) - \mathbb{E}h_\alpha(Z) + \mathbb{E}h_\alpha(Z) - \Phi(z) \\ &\leq \frac{\delta}{\alpha} + \mathbb{P}\{z \leq Z \leq z + \alpha\} \\ &\leq \frac{\delta}{\alpha} + \frac{\alpha}{\sqrt{2\pi}}, \end{aligned}$$

and hence

$$\mathbb{P}(W \leq z) - \Phi(z) \leq 2(2\pi)^{-1/4}\delta^{1/2} \leq 2\delta^{1/2}. \tag{3.5}$$

Similarly, we have

$$\mathbb{P}(W \le z) - \Phi(z) \ge -2\delta^{1/2}, \tag{3.6}$$

proving (3.4). ∎

In the next three sections, we show that (3.2) is satisfied with suitably small δ, when W is a sum of (i) independent random variables, or (ii) locally dependent random variables. We also show that (3.2) is satisfied when (iii) W is such that an exchangeable pair (W, W') can be constructed having the linear regression property:

$$\mathbb{E}\{W' \mid W\} = (1-\lambda)W \qquad \text{for some } 0 < \lambda < 1. \tag{3.7}$$

3.1. *Independent random variables*

In this section, we use (2.19) to prove (3.2) for W a sum of independent random variables with zero means and finite third moments, extending (1.8) to non-identically distributed random variables.

Theorem 3.2: *Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables satisfying $\mathbb{E}\xi_i = 0$ and $\mathbb{E}|\xi_i|^3 < \infty$ for each $1 \le i \le n$, and such that $\sum_{i=1}^n \mathbb{E}\xi_i^2 = 1$. Then Theorem 3.1 can be applied with*

$$\delta = 3\sum_{i=1}^n \mathbb{E}|\xi_i|^3. \tag{3.8}$$

In particular, we have

$$\left|\mathbb{E}|W| - \sqrt{\frac{2}{\pi}}\right| \le 3\sum_{i=1}^n \mathbb{E}|\xi_i|^3.$$

Proof: It follows from Lemma 2.3 that $\|f_h''\| \le 2\|h'\|$. Therefore, by (2.19) and the mean value theorem,

$$\begin{aligned}|\mathbb{E}\{f_h'(W) - Wf_h(W)\}| &\le \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}|f_h'(W) - f_h'(W^{(i)} + t)|K_i(t)\,dt \\ &\le 2\|h'\| \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}(|t| + |\xi_i|)K_i(t)\,dt.\end{aligned}$$

Using (2.16), it thus follows that

$$|\mathbb{E}\{f_h'(W) - Wf_h(W)\}| \leq 2\|h'\| \sum_{i=1}^{n} (\mathbb{E}|\xi_i|^3/2 + \mathbb{E}|\xi_i|\mathbb{E}\xi_i^2)$$
$$\leq 3\|h'\| \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3. \qquad \blacksquare$$

It is actually not necessary to assume the existence of finite third moments in Theorem 3.2. The quantity δ can be computed in terms of the elements appearing in the statement of the Lindeberg central limit theorem.

Theorem 3.3: *Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables satisfying $\mathbb{E}\xi_i = 0$ for each $1 \leq i \leq n$ and such that $\sum_{i=1}^{n} \mathbb{E}\xi_i^2 = 1$. Then Theorem 3.1 can be applied with*

$$\delta = 4(4\beta_2 + 3\beta_3), \tag{3.9}$$

where

$$\beta_2 = \sum_{i=1}^{n} \mathbb{E}\xi_i^2 I_{\{|\xi_i|>1\}} \quad \text{and} \quad \beta_3 = \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3 I_{\{|\xi_i|\leq 1\}}. \tag{3.10}$$

Proof: We use (2.12) and (2.13) to show that

$$|f_h'(W) - f_h'(W^{(i)} + t)| \leq \|h'\| \min\big(8, 2(|t| + |\xi_i|)\big)$$
$$\leq 8\|h'\|(|t| \wedge 1 + |\xi_i| \wedge 1),$$

where $a \wedge b$ denotes $\min(a, b)$. Substituting this bound into (2.19), we obtain

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq 8\|h'\| \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}(|t| \wedge 1 + |\xi_i| \wedge 1) K_i(t)\, dt. \tag{3.11}$$

Now

$$\int_{-\infty}^{\infty} (|t| \wedge 1)\{\mathbf{1}_{[0,x]}(t) - \mathbf{1}_{[-x,0)}(t)\}\, dt = \begin{cases} \frac{1}{2}|x|t^2 + |x|(|x| - 1) & \text{if } |x| > 1; \\ \frac{1}{2}|x|^3, & \text{if } |x| \leq 1, \end{cases}$$

so that

$$\int_{-\infty}^{\infty} \mathbb{E}(|t| \wedge 1 + |\xi_i| \wedge 1) K_i(t)\, dt$$
$$= \mathbb{E}\{|\xi_i|(|\xi_i| - 1) I_{\{|\xi_i|>1\}}\} + \tfrac{1}{2}\mathbb{E}\{|\xi_i|(|\xi_i| \wedge 1)^2\} + \mathbb{E}\{\xi_i^2 \mathbb{E}(|\xi_i| \wedge 1)\}$$
$$= \mathbb{E}\{\xi_i^2 I_{\{|\xi_i|>1\}}\} - \tfrac{1}{2}\mathbb{E}\{|\xi_i| I_{\{|\xi_i|>1\}}\}$$
$$+ \tfrac{1}{2}\mathbb{E}\{|\xi_i|^3 I_{\{|\xi_i|\leq 1\}}\} + \mathbb{E}\{\xi_i^2 \mathbb{E}(|\xi_i| \wedge 1)\}.$$

Adding over i, it follows that

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq 8\|h'\| \left\{ \beta_2 + \tfrac{1}{2}\beta_3 + \sum_{i=1}^{n} \mathbb{E}\xi_i^2 \mathbb{E}(|\xi_i| \wedge 1) \right\}. \quad (3.12)$$

However, since both x^2 and $(x \wedge 1)$ are increasing functions of $x \geq 0$, it follows that, for any random variable ξ,

$$\mathbb{E}\xi^2 \mathbb{E}(|\xi| \wedge 1) \leq \mathbb{E}\{\xi^2(|\xi| \wedge 1)\} = \mathbb{E}|\xi|^3 I_{\{|\xi| \leq 1\}} + \mathbb{E}\xi^2 I_{\{|\xi| > 1\}}, \quad (3.13)$$

so that the final sum is no greater than $\beta_3 + \beta_2$, completing the bound. ∎

Theorem 3.1 and Theorem 3.3 together yield the Lindeberg central limit theorem, as follows. Let $X_1, X_2, \cdots, X_n$ be independent random variables with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 < \infty$ for each $1 \leq i \leq n$. Put

$$S_n = \sum_{i=1}^{n} X_i \quad \text{and} \quad B_n^2 = \sum_{i=1}^{n} \mathbb{E}X_i^2.$$

To apply Theorems 3.1 and 3.3, let

$$\xi_i = X_i / B_n \quad \text{and} \quad W = S_n / B_n. \quad (3.14)$$

Define β_2 and β_3 as in (3.10), and observe that, for any $0 < \varepsilon < 1$,

$$\begin{aligned} \beta_2 + \beta_3 &= \frac{1}{B_n^2} \sum_{i=1}^{n} \mathbb{E}\{X_i^2 I_{\{|X_i| > B_n\}}\} + \frac{1}{B_n^3} \sum_{i=1}^{n} \mathbb{E}\{|X_i|^3 I_{\{|X_i| \leq B_n\}}\} \\ &\leq \frac{1}{B_n^2} \sum_{i=1}^{n} \mathbb{E}\{X_i^2 I_{\{|X_i| > B_n\}}\} + \frac{1}{B_n^3} \sum_{i=1}^{n} B_n \mathbb{E}\{X_i^2 I_{\{\varepsilon B_n \leq |X_i| \leq B_n\}}\} \\ &\quad + \frac{1}{B_n^3} \sum_{i=1}^{n} \varepsilon B_n \mathbb{E}\{X_i^2 I_{\{|X_i| < \varepsilon B_n\}}\} \\ &\leq \frac{1}{B_n^2} \sum_{i=1}^{n} \mathbb{E}\{X_i^2 I_{\{|X_i| > \varepsilon B_n\}}\} + \varepsilon. \end{aligned} \quad (3.15)$$

If the Lindeberg condition holds, that is if

$$\frac{1}{B_n^2} \sum_{i=1}^{n} \mathbb{E}\{X_i^2 I_{\{|X_i| > \varepsilon B_n\}}\} \to 0 \quad \text{as} \quad n \to \infty \quad (3.16)$$

for all $\varepsilon > 0$, then (3.15) implies $\beta_2 + \beta_3 \to 0$ as $n \to \infty$, since ε is arbitrary. Hence, by Theorems 3.1 and 3.3,

$$\sup_z |\mathbb{P}(S_n / B_n \leq z) - \Phi(z)| \leq 8(\beta_2 + \beta_3)^{1/2} \to 0 \quad \text{as} \quad n \to \infty.$$

This proves the Lindeberg central limit theorem.

3.2. *Locally dependent random variables*

An m-dependent sequence of random variables ξ_i, $i \in \mathbb{Z}$, is one with the property that, for each i, the sets of random variables $\{\xi_j, j \leq i\}$ and $\{\xi_j, j > i + m\}$ are independent. As a special case, sequences of independent random variables are 0-dependent. Local dependence generalizes the notion of m-dependence to random variables with arbitrary index set. It is applicable, for instance, to random variables indexed by the vertices of a graph, and such that the collections $\{\xi_i, i \in I\}$ and $\{\xi_j, j \in J\}$ are independent whenever $I \cap J = \emptyset$ and the graph contains no edges $\{i, j\}$ with $i \in I$ and $j \in J$.

Let $\mathcal{J}$ be a finite index set of cardinality n, and let $\{\xi_i, i \in \mathcal{J}\}$ be a random field with zero means and finite variances. Define $W = \sum_{i \in \mathcal{J}} \xi_i$, and assume that $\mathrm{Var}(W) = 1$. For $A \subset \mathcal{J}$, let ξ_A denote $\{\xi_i, i \in A\}$ and $A^c = \{j \in \mathcal{J} : j \notin A\}$. We introduce the following two assumptions, defining different strengths of local dependence.

(LD1) For each $i \in \mathcal{J}$ there exists $A_i \subset \mathcal{J}$ such that ξ_i and $\xi_{A_i^c}$ are independent.

(LD2) For each $i \in \mathcal{J}$ there exist $A_i \subset B_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{A_i^c}$ and ξ_{A_i} is independent of $\xi_{B_i^c}$.

We then define $\eta_i = \sum_{j \in A_i} \xi_j$ and $\tau_i = \sum_{j \in B_i} \xi_j$. Note that, for independent random variables ξ_i, we can take $A_i = B_i = \{i\}$, for which then $\eta_i = \tau_i = \xi_i$.

Theorem 3.4: *Theorem 3.1 can be applied with*

$$\delta = 4\mathbb{E}\left|\sum_{i \in \mathcal{J}} \{\xi_i \eta_i - \mathbb{E}(\xi_i \eta_i)\}\right| + \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i \eta_i^2| \tag{3.17}$$

under (LD1), and with

$$\delta = 2 \sum_{i \in \mathcal{J}} (\mathbb{E}|\xi_i \eta_i \tau_i| + |\mathbb{E}(\xi_i \eta_i)| \mathbb{E}|\tau_i|) + \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i \eta_i^2| \tag{3.18}$$

under (LD2).

Remark: For independent random variables, the value of δ in (3.18) is $5 \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i|^3$, somewhat larger than the direct bound given in Theorem 3.2.

Proof: We first derive Stein identities similar to (2.17) and (2.19). Let $f = f_h$ be the solution of the Stein equation (2.4). Then

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= \sum_{i\in\mathcal{J}} \mathbb{E}\xi_i f(W) \\ &= \sum_{i\in\mathcal{J}} \mathbb{E}\xi_i [f(W) - f(W-\eta_i)],\end{aligned}$$

by the independence of ξ_i and $W - \eta_i$. Hence

$$\begin{aligned}\mathbb{E}\{Wf(W)\} = \sum_{i\in\mathcal{J}} \mathbb{E}\{\xi_i [f(W) - f(W-\eta_i) - \eta_i f'(W)]\} \\ + \mathbb{E}\left\{\left(\sum_{i\in\mathcal{J}} \xi_i\eta_i\right) f'(W)\right\}. \qquad (3.19)\end{aligned}$$

Now, because $\mathbb{E}\xi_i = 0$ for all i, and from (LD1), it follows that

$$1 = \mathbb{E}W^2 = \sum_{i\in J}\sum_{j\in J} \mathbb{E}\{\xi_i\xi_j\} = \sum_{i\in J} \mathbb{E}\{\xi_i\eta_i\},$$

giving

$$\begin{aligned}\mathbb{E}\{f'(W) - Wf(W)\} = -\mathbb{E}\Big(\sum_{i\in\mathcal{J}} \{\xi_i\eta_i - \mathbb{E}(\xi_i\eta_i)\} f'(W)\Big) \qquad (3.20) \\ - \sum_{i\in\mathcal{J}} \mathbb{E}\left\{\xi_i [f(W) - f(W-\eta_i) - \eta_i f'(W)]\right\}.\end{aligned}$$

By (2.12) and (2.13), $\|f'\| \le 4\|h'\|$ and $\|f''\| \le 2\|h'\|$. Therefore it follows from (3.20) and the Taylor expansion that

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \le \|h'\| \left\{ 4\mathbb{E}\left|\sum_{i\in\mathcal{J}} \{\xi_i\eta_i - \mathbb{E}(\xi_i\eta_i)\}\right| + \sum_{i\in\mathcal{J}} \mathbb{E}|\xi_i\eta_i^2| \right\}.$$

This proves (3.17).

When (LD2) is satisfied, $f'(W - \tau_i)$ and $\xi_i\eta_i$ are independent for each $i \in \mathcal{J}$. Hence, using (3.20), we can write

$$\begin{aligned}&|\mathbb{E}h(W) - \mathbb{E}h(Z)| \\ &\le \left|\mathbb{E}\sum_{i\in\mathcal{J}} \{\xi_i\eta_i - \mathbb{E}(\xi_i\eta_i)\}(f'(W) - f'(W-\tau_i))\right| + \|h'\| \sum_{i\in\mathcal{J}} \mathbb{E}|\xi_i\eta_i^2| \\ &\le \|h'\| \left\{ 2\sum_{i\in\mathcal{J}} (\mathbb{E}|\xi_i\eta_i\tau_i| + |\mathbb{E}(\xi_i\eta_i)|\mathbb{E}|\tau_i|) + \sum_{i\in\mathcal{J}} \mathbb{E}|\xi_i\eta_i^2| \right\},\end{aligned}$$

as desired. ∎

Here are two examples of locally dependent random fields. We refer to Baldi & Rinott (1989), Rinott (1994), Baldi, Rinott & Stein (1989), Dembo & Rinott (1996), and Chen & Shao (2004) for more details.

Example 1. *Graphical dependence.*

Consider a set of random variables $\{X_i, i \in \mathcal{V}\}$ indexed by the vertices of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{G}$ is said to be a dependency graph if, for any pair of disjoint sets Γ_1 and Γ_2 in $\mathcal{V}$ such that no edge in $\mathcal{E}$ has one endpoint in Γ_1 and the other in Γ_2, the sets of random variables $\{X_i, i \in \Gamma_1\}$ and $\{X_i, i \in \Gamma_2\}$ are independent. Let D denote the maximal degree of G; that is, the maximal number of edges incident to a single vertex. Let

$$A_i = \{i\} \cup \{j \in \mathcal{V}\text{: there is an edge connecting } j \text{ and } i\}$$

and $B_i = \bigcup_{j \in A_i} A_j$. Then $\{X_i,\ i \in \mathcal{V}\}$ satisfies (LD2). Hence (3.18) holds.

Example 2. *The number of local maxima on a graph.*

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (which is not necessarily a dependency graph) and independent and identically distributed continuous random variables $\{Y_i, i \in \mathcal{V}\}$. For $i \in \mathcal{V}$ define the 0-1 indicator variable

$$X_i = \begin{cases} 1 & \text{if } Y_i > Y_j \text{ for all } j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

where $N_i = \{j \in \mathcal{V}\text{: } \{i, j\} \in \mathcal{E}\}$, so that $X_i = 1$ indicates that Y_i is a local maximum. Let $W = \sum_{i \in \mathcal{V}} X_i$ be the number of local maxima. Let

$$A_i = \{i\} \cup \left\{ \bigcup_{j \in N_i} N_j \right\} \quad \text{and} \quad B_i = \bigcup_{j \in A_i} A_j.$$

Then $\{X_i,\ i \in \mathcal{V}\}$ satisfies (LD2), and (3.18) holds.

One can refer to Dembo & Rinott (1996) and Rinott & Rotar (1996) for more examples and results under local dependence.

3.3. *Exchangeable pairs*

Let W be a random variable that is not necessarily the partial sum of independent random variables. Suppose that W is approximately normal, and that we want to find how accurate the approximation is. Another basic approach to Stein's method is to introduce a second random variable W' on the same probability space, in such a way that (W, W') is an exchangeable pair; that is, such that (W, W') and (W', W) have the same distribution.

The approach makes essential use of the elementary fact that, if (W, W') is an exchangeable pair, then

$$\mathbb{E}g(W, W') = 0 \tag{3.21}$$

for all *antisymmetric* measurable functions $g(x, y)$ such that the expected value exists.

A key identity is the following lemma (see Stein 1986).

Lemma 3.5: *Let (W, W') be an exchangeable pair of real random variables with finite variance, having the linear regression property*

$$\mathbb{E}(W' \mid W) = (1 - \lambda)W \tag{3.22}$$

for some $0 < \lambda < 1$. Then

$$\mathbb{E}W = 0 \quad \text{and} \quad \mathbb{E}(W' - W)^2 = 2\lambda\mathbb{E}W^2, \tag{3.23}$$

and, for every piecewise continuous function f satisfying the growth condition $|f(w)| \le C(1 + |w|)$, we have

$$\mathbb{E}\{Wf(W)\} = \frac{1}{2\lambda}\mathbb{E}\{(W - W')(f(W) - f(W'))\}. \tag{3.24}$$

Proof: The proof exploits (3.21) with $g(x, y) = (x - y)(f(y) + f(x))$, for which $\mathbb{E}g(W, W')$ exists, because of the assumption on f. Then (3.21) gives

$$\begin{aligned} 0 &= \mathbb{E}\{(W - W')(f(W') + f(W))\} \\ &= \mathbb{E}\{(W - W')(f(W') - f(W))\} + 2\mathbb{E}\{f(W)(W - W')\} \\ &= \mathbb{E}\{(W - W')(f(W') - f(W))\} + 2\mathbb{E}\{f(W)\mathbb{E}(W - W' \mid W)\} \\ &= \mathbb{E}\{(W - W')(f(W') - f(W))\} + 2\lambda\mathbb{E}\{Wf(W)\}, \end{aligned} \tag{3.25}$$

this last by (3.22), and this is just (3.24). ∎

Using this lemma, we can deduce the following theorem.

Theorem 3.6: *If (W, W') is exchangeable and (3.22) holds, it follows that Theorem 3.1 can be applied with*

$$\delta = 4\mathbb{E}\left|1 - \frac{1}{2\lambda}\mathbb{E}((W' - W)^2 \mid W)\right| + \frac{1}{2\lambda}\mathbb{E}|W - W'|^3. \tag{3.26}$$

Remark: In the first term, note that

$$\mathbb{E}\left\{1 - \frac{1}{2\lambda}\mathbb{E}((W' - W)^2 \mid W)\right\} = 1 - \frac{1}{2\lambda}\mathbb{E}(W' - W)^2 = 1 - \mathbb{E}W^2,$$

so that the bound is unlikely to be useful unless $\mathbb{E}W^2$ is close to 1.

Proof: Let $f = f_h$ be the solution (2.5) to the Stein equation, and define

$$\widehat{K}(t) = (W - W')\big(I_{\{-(W-W')\le t\le 0\}} - I_{\{0<t\le -(W-W')\}}\big) \ge 0,$$

noting that

$$\int_{-\infty}^{\infty} \widehat{K}(t)\,dt = (W - W')^2. \tag{3.27}$$

By (3.24),

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= \frac{1}{2\lambda}\mathbb{E}\left\{\int_{-(W-W')}^{0} f'(W+t)(W-W')\,dt\right\} \\ &= \frac{1}{2\lambda}\mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W+t)\widehat{K}(t)\,dt\right\}\end{aligned}$$

and

$$\mathbb{E}f'(W) = \mathbb{E}\left\{f'(W)\left(1 - \frac{1}{2\lambda}(W-W')^2\right)\right\} + \frac{1}{2\lambda}\mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W)\widehat{K}(t)\,dt\right\},$$

from (3.27). Putting the two together gives

$$\begin{aligned}&|\mathbb{E}h(W) - \mathbb{E}h(Z)| = |\mathbb{E}\{f'(W) - Wf(W)\}| \\ &= \left|\mathbb{E}f'(W)\Big(1 - \frac{1}{2\lambda}(W-W')^2\Big) + \frac{1}{2\lambda}\mathbb{E}\int_{-\infty}^{\infty}(f'(W) - f'(W+t))\widehat{K}(t)\,dt\right| \\ &\le \left|\mathbb{E}\Big\{f'(W)\Big(1 - \frac{1}{2\lambda}\mathbb{E}((W-W')^2 \mid W)\Big)\Big\}\right| \\ &\qquad + \frac{1}{2\lambda}\mathbb{E}\left|\int_{-\infty}^{\infty}(f'(W) - f'(W+t))\widehat{K}(t)\,dt\right|,\end{aligned}$$

and now the bounds in Lemma 2.3 give

$$\begin{aligned}&|\mathbb{E}h(W) - \mathbb{E}h(Z)| \\ &\quad\le \|h'\|\left\{4\mathbb{E}\Big|1 - \frac{1}{2\lambda}\mathbb{E}((W-W')^2 \mid W)\Big| + \frac{1}{\lambda}\mathbb{E}\int_{-\infty}^{\infty}|t|\widehat{K}(t)\,dt\right\} \\ &\quad= \|h'\|\left\{4\mathbb{E}\Big|1 - \frac{1}{2\lambda}\mathbb{E}\big((W'-W)^2 \mid W\big)\Big| + \frac{1}{2\lambda}\mathbb{E}|W-W'|^3\right\},\end{aligned}$$

from (3.27), as desired. ∎

We use the following example to show how to apply the bound in the above theorem. Let ξ_i be independent random variables with zero means and $\sum_{i=1}^n \mathbb{E}\xi_i^2 = 1$, and put $W = \sum_{i=1}^n \xi_i$. Let $\{\xi_i^*, 1 \le i \le n\}$ be an independent copy of $\{\xi_i, 1 \le i \le n\}$, and let I have uniform distribution

on $\{1,2,\ldots,n\}$, independent of $\{\xi_i\}$ and $\{\xi_i^*\}$. Define $W' = W - \xi_I + \xi_I^*$. Then (W, W') is an exchangeable pair, and

$$\mathbb{E}(W' \mid W) = \left(1 - \frac{1}{n}\right) W,$$

so that (3.22) is satisfied with $\lambda = 1/n$. Direct calculation also gives

$$\mathbb{E}|W - W'|^3 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|\xi_i - \xi_i^*|^3 \le (8/n)\sum_{i=1}^{n} \mathbb{E}|\xi_i|^3$$

and

$$\mathbb{E}((W - W')^2 \mid W) = \frac{1}{n}\left(1 + \sum_{i=1}^{n} \mathbb{E}(\xi_i^2 \mid W)\right); \tag{3.28}$$

from the latter, we have

$$\begin{aligned}\mathbb{E}\left|1 - \frac{1}{2\lambda}\mathbb{E}((W' - W)^2 \mid W)\right| &= \tfrac{1}{2}\mathbb{E}\left|1 - \mathbb{E}(\sum_{i=1}^{n} \xi_i^2 \mid W)\right| \\ &\le \tfrac{1}{2}\mathbb{E}\left|\sum_{i=1}^{n}(\xi_i^2 - \mathbb{E}\xi_i^2)\right|.\end{aligned}$$

Thus, if the ξ_i have finite fourth moments, Theorem 3.1 can be applied with

$$\delta = 2\sqrt{\sum_{i=1}^{n} \mathrm{Var}(\xi_i^2)} + 4\sum_{i=1}^{n} \mathbb{E}|\xi_i|^3 \;\le\; 2\sqrt{\sum_{i=1}^{n} \mathbb{E}|\xi_i|^4} + 4\sum_{i=1}^{n} \mathbb{E}|\xi_i|^3.$$

In particular, if $\xi_i = n^{-1/2}X_i$, where the random variables X_i are independent and identically distributed random variables with finite fourth moments, then the bound is of the correct order $O(n^{-1/2})$.

On the other hand, if we keep the original form

$$\left|\mathbb{E}\left\{f'(W)\left(1 - \frac{1}{2\lambda}\mathbb{E}((W - W')^2 \mid W)\right)\right\}\right|$$

which gave rise to the first term on the right hand side of (3.26), then we only need to assume finite third moments. To see this, by (3.28).

$$\begin{aligned}\left|\mathbb{E}\left\{f'(W)\left(1 - \frac{1}{2\lambda}\mathbb{E}((W - W')^2 \mid W)\right)\right\}\right| &= \frac{1}{2}\left|\mathbb{E}\left\{f'(W)\sum_{i=1}^{n}(\mathbb{E}\xi_i^2 - \xi_i^2)\right\}\right| \\ &= \frac{1}{2}\left|\sum_{i=1}^{n}\mathbb{E}\left\{(f'(W) - f'(W - \xi_i))(\mathbb{E}\xi_i^2 - \xi_i^2)\right\}\right|,\end{aligned}$$

because $W - \xi_i$ and ξ_i are independent. This is now bounded using Lemma 2.3 by

$$\|h'\| \sum_{i=1}^{n} \mathbb{E}|\xi_i(\mathbb{E}\xi_i^2 - \xi_i^2)| \leq 2\|h'\| \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3,$$

and Theorem 3.1 can be applied with

$$\delta = 6 \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3.$$

We therefore end this section with the following alternative to (3.26): if (W, W') is exchangeable and (3.22) holds, then

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq \left|\mathbb{E}f_h'(W)(1 - \tfrac{1}{2\lambda}(W - W')^2)\right| + \tfrac{1}{2\lambda}\|h'\|\mathbb{E}|W - W'|^3. \tag{3.29}$$

4. Uniform Berry–Esseen bounds: the bounded case

In the previous section, the sharpest bounds in Theorem 3.1 were of order $O(\delta)$, and were obtained for the Wasserstein distance d_W. Those for the Kolmogorov distance d_K were only of the larger order $O(\delta^{1/2})$. Here, we turn to deriving bounds for d_K which are of comparable order to those for d_W. We begin with the simplest case of independent summands. The method of proof motivates the formulation of a rather general theorem, which is then applied in a dependent setting, that of the number of 1's in the binary expansion of a randomly chosen integer, using an exchangeable pair.

4.1. *Independent random variables*

Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables with zero means and with $\sum_{i=1}^{n} \mathbb{E}\xi_i^2 = 1$. We use the notation of Section 2.3:

$$W = \sum_{i=1}^{n} \xi_i, \quad W^{(i)} = W - \xi_i \quad \text{and} \quad K_i(t) = \mathbb{E}\left\{\xi_i(I_{\{0 \leq t \leq \xi_i\}} - I_{\{\xi_i \leq t < 0\}})\right\}.$$

Let f_z be the solution of the Stein equation (2.1). For bounded ξ_i, we are ready to apply (2.17) to obtain the following Berry–Esseen bound.

Theorem 4.1: *If* $|\xi_i| \leq \delta_0$ *for* $1 \leq i \leq n$, *then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq 3.3\,\delta_0. \tag{4.1}$$

Proof: Write $f = f_z$. It follows from (2.17) and because f satisfies the Stein equation (2.1) that

$$\begin{aligned}\mathbb{E}\{Wf(W)\} &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\{f'(W^{(i)}+t)\}K_i(t)\,dt \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\big\{(W^{(i)}+t)f(W^{(i)}+t) + I_{\{W^{(i)}+t\le z\}} - \Phi(z)\big\}K_i(t)\,dt.\end{aligned}$$

Reorganizing this, and recalling that

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_i(t)\,dt = \sum_{i=1}^{n} \mathbb{E}\xi_i^2 = 1,$$

we have

$$\begin{aligned}&\sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{P}(W^{(i)}+t\le z)K_i(t)\,dt - \Phi(z) \\ &\quad = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\big\{Wf(W) - (W^{(i)}+t)f(W^{(i)}+t)\big\}K_i(t)\,dt. \qquad (4.2)\end{aligned}$$

Now, by (2.10),

$$\begin{aligned}&\sum_{i=1}^{n} \mathbb{E} \int_{-\infty}^{\infty} \big|Wf(W) - (W^{(i)}+t)f(W^{(i)}+t)\big|K_i(t)\,dt \\ &\le \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{E}\left\{(|W^{(i)}| + \sqrt{2\pi}/4)(|\xi_i| + |t|)\right\} K_i(t)\,dt \\ &\le (1+\sqrt{2\pi}/4) \sum_{i=1}^{n} \int_{-\infty}^{\infty} (\mathbb{E}|\xi_i| + |t|)K_i(t),\end{aligned}$$

since $\mathbb{E}\{W^{(i)}\}^2 \le 1$ and ξ_i and $W^{(i)}$ are independent. Hence, recalling (2.16), we have

$$\begin{aligned}&\left|\sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{P}(W^{(i)}+t\le z)K_i(t)\,dt - \Phi(z)\right| \\ &\le (1+\sqrt{2\pi}/4) \sum_{i=1}^{n} \{\mathbb{E}|\xi_i|\mathbb{E}\xi_i^2 + \tfrac{1}{2}\mathbb{E}|\xi_i|^3\} \\ &\le \tfrac{3}{2}(1+\sqrt{2\pi}/4) \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3. \qquad (4.3)\end{aligned}$$

Hence we would be finished if $\mathbb{P}(W^{(i)} + t \le z)$ could be replaced by $\mathbb{P}(W \le z)$, since we have $\sum_{i=1}^n \int_{-\infty}^{\infty} K_i(t)\,dt = 1$. Clearly, in view of the fact that $\mathbb{P}(W \le z) = \mathbb{P}(W^{(i)} \le z - \xi_i)$, we have

$$\begin{aligned} &|\mathbb{P}(W^{(i)} + t \le z) - \mathbb{P}(W \le z)| \\ &\qquad \le \mathbb{P}(z - \max\{\xi_i, t\} \le W^{(i)} \le z - \min\{\xi_i, t\}), \end{aligned} \tag{4.4}$$

and the difference should be small if both $|t|$ and $|\xi_i|$ are.

Since the $|\xi_i|$ are uniformly bounded by δ_0, proving that the difference is small is particularly simple. First, we note that $|\xi_i| \le \delta_0$ implies also that $K_i(t) = 0$ for $|t| > \delta_0$, so that we only need to consider (4.4) with both $|t|$ and $|\xi_i|$ bounded by δ_0. But then

$$\mathbb{P}(W^{(i)} + t \le z) = \mathbb{P}(W - \xi_i + t \le z) \begin{cases} \ge \mathbb{P}(W \le z - 2\delta_0) \\ \le \mathbb{P}(W \le z + 2\delta_0). \end{cases} \tag{4.5}$$

Taking $z + 2\delta_0$ for z in the first inequality in (4.5) and substituting it into (4.3) gives

$$\begin{aligned} \mathbb{P}(W \le z) - \Phi(z) &\le \Phi(z + 2\delta_0) - \Phi(z) + \tfrac{3}{2}(1 + \sqrt{2\pi}/4) \sum_{i=1}^n \mathbb{E}|\xi_i|^3 \\ &\le \frac{2\delta_0}{\sqrt{2\pi}} + \frac{3}{2}(1 + \sqrt{2\pi}/4)\delta_0 \;\le\; 3.3\,\delta_0, \end{aligned} \tag{4.6}$$

and the corresponding lower bound follows from the second inequality in (4.5) and (4.3) with $z - 2\delta_0$ for z, completing the proof. ∎

One can see from the above approach that the key ingredient of the proof is to rewrite $\mathbb{E}\{Wf(W)\}$ in terms of a functional of f'. We formulate this in abstract form as follows.

Theorem 4.2: *Let W be a real valued random variable having $\mathbb{E}|W| \le 1$, and let f_z be the solution of the Stein equation (2.1). Suppose that there exist random variables R_1, R_2 and $M(t) \ge 0$, and constants δ_0, δ_1 and δ_2 that do not depend on z, such that*

$$\int_{|t| \le \delta_0} M(t)\,dt = 1, \tag{4.7}$$

$$|R_1| \le \delta_1, \quad |\mathbb{E}(R_2)| \le \delta_2 \tag{4.8}$$

and

$$\mathbb{E}\{Wf_z(W)\} = \mathbb{E}\left\{\int_{|t|\le\delta_0} f_z'(W+R_1+t)M(t)\,dt\right\} + \mathbb{E}(R_2). \tag{4.9}$$

Then it follows that

$$\sup_z |\mathbb{P}(W\le z) - \Phi(z)| \le 2.1(\delta_0+\delta_1) + \delta_2. \tag{4.10}$$

Remark: Note that, although the function M is random, its integral always takes the fixed value 1.

Proof: Since f_z satisfies the Stein equation (2.1), we have

$$\begin{aligned}
&\mathbb{E}\left\{\int_{|t|\le\delta_0} f_z'(W+R_1+t)M(t)\,dt\right\} \\
&\quad= \mathbb{E}\left\{\int_{|t|\le\delta_0} (I_{\{W+R_1+t\le z\}} - \Phi(z))M(t)\,dt\right\} \\
&\qquad+ \mathbb{E}\left\{\int_{|t|\le\delta_0} (W+R_1+t)f_z(W+R_1+t)M(t)\,dt\right\} \\
&\quad\le \mathbb{E}\left\{\int_{|t|\le\delta_0} (I_{\{W\le z+\delta_0+\delta_1\}} - \Phi(z))M(t)\,dt\right\} \\
&\qquad+ \mathbb{E}\left\{\int_{|t|\le\delta_0} (W+R_1+t)f_z(W+R_1+t)M(t)\,dt\right\},
\end{aligned}$$

where the last line follows because $-R_1 - t \le \delta_0 + \delta_1$. Hence, and since $\int_{|t|\le\delta_0} M(t)\,dt = 1$, we find that

$$\begin{aligned}
&\mathbb{E}\left\{\int_{|t|\le\delta_0} f_z'(W+R_1+t)M(t)\,dt\right\} \\
&\quad\le \mathbb{P}(W\le z+\delta_0+\delta_1) - \Phi(z) \\
&\qquad+ \int_{|t|\le\delta_0} \mathbb{E}\{(W+R_1+t)f_z(W+R_1+t)M(t)\}\,dt \\
&\quad\le \mathbb{P}(W\le z+\delta_0+\delta_1) - \Phi(z+\delta_0+\delta_1) \\
&\qquad+ \frac{(\delta_0+\delta_1)}{\sqrt{2\pi}} + \int_{|t|\le\delta_0} \mathbb{E}\{(W+R_1+t)f_z(W+R_1+t)M(t)\}\,dt.
\end{aligned}$$

Thus by (4.9), (4.8) and (4.7)

$$\begin{aligned}
&\mathbb{P}(W \le z + \delta_0 + \delta_1) - \Phi(z + \delta_0 + \delta_1) \\
&\ge -\frac{(\delta_0 + \delta_1)}{\sqrt{2\pi}} - \mathbb{E}R_2 \\
&\quad + \int_{|t| \le \delta_0} \mathbb{E}\{[Wf_z(W) - (W + R_1 + t)f_z(W + R_1 + t)]M(t)\}\, dt \\
&\ge -\frac{(\delta_0 + \delta_1)}{\sqrt{2\pi}} - \delta_2 - \int_{|t| \le \delta_0} \mathbb{E}\{(|W| + \sqrt{2\pi}/4)(|R_1| + |t|)M(t)\}\, dt,
\end{aligned}$$

this last by (2.10), and so

$$\begin{aligned}
&\mathbb{P}(W \le z + \delta_0 + \delta_1) - \Phi(z + \delta_0 + \delta_1) \\
&\ge -\frac{(\delta_0 + \delta_1)}{\sqrt{2\pi}} - \delta_2 - \mathbb{E}\left\{\int_{|t| \le \delta_0} (|W| + \sqrt{2\pi}/4)(\delta_0 + \delta_1)M(t)\, dt\right\} \\
&= -\frac{(\delta_0 + \delta_1)}{\sqrt{2\pi}} - \delta_2 - (\mathbb{E}|W| + \sqrt{2\pi}/4)(\delta_0 + \delta_1) \\
&\ge -2.1(\delta_0 + \delta_1) - \delta_2. \qquad (4.11)
\end{aligned}$$

A similar argument gives

$$\mathbb{P}(W \le z - \delta_0 - \delta_1) - \Phi(z - \delta_0 - \delta_1) \le 2.1(\delta_0 + \delta_1) + \delta_2, \qquad (4.12)$$

and this proves (4.10). ∎

To see why Theorem 4.1 is a special case of Theorem 4.2, let I be independent of $\{\xi_i, 1 \le i \le n\}$ with $\mathbb{P}(I = i) = \mathbb{E}\xi_i^2$ for $i = 1, 2, \cdots, n$. Then we can rewrite (2.17) as

$$\begin{aligned}
\mathbb{E}Wf(W) &= \mathbb{E}\int_{|t| \le \delta_0} f'(W^{(I)} + t)\widetilde{K}_I(t)\, dt \\
&= \mathbb{E}\int_{|t| \le \delta_0} f'(W + W^{(I)} - W + t)\widetilde{K}_I(t)\, dt,
\end{aligned}$$

where $\widetilde{K}_i(t) = K_i(t)/\mathbb{E}\xi_i^2$. Set $R_1 = W^{(I)} - W$, $R_2 = 0$ and $M(t) = \widetilde{K}_I(t)$ in Theorem 4.2. It is easy to see that conditions (4.7) – (4.9) are satisfied with $\delta_1 = \delta_0$ and $\delta_2 = 0$. Hence (4.10) holds with a bound of $4.2\,\delta_0$.

In the next section, we illustrate how to use Theorem 4.2 to get a Berry–Esseen bound for the number of ones in the binary expansion of a random integer, using an exchangeable pair approach.

4.2. *Binary expansion of a random integer*

Let $n \geq 2$ be a natural number and X be a random variable uniformly distributed over the set $\{0, 1, \cdots, n-1\}$. Let k be such that $2^{k-1} < n \leq 2^k$. Write the binary expansion of X as

$$X = \sum_{i=1}^{k} X_i 2^{k-i}$$

and let $S = X_1 + \cdots + X_k$ be the number of ones in the binary expansion of X. When $n = 2^k$, the distribution of S is the binomial distribution for k trials with probability $1/2$, and hence can be approximated by a normal distribution. We shall show that the normal approximation is good for *any* large n.

Theorem 4.3: *Let k be such that $2^{k-1} < n \leq 2^k$, and set $W = \frac{S-(k/2)}{\sqrt{k/4}}$. Then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq 6.2k^{-1/2}. \tag{4.13}$$

Proof: We use the exchangeable pair approach. Let I be a random variable uniformly distributed over the set $\{1, 2, \cdots, k\}$ and independent of X, and let the random variable X' be defined by

$$X' = \sum_{i=1}^{k} X_i' 2^{k-i},$$

where

$$X_i' = \begin{cases} X_i & \text{if } i \neq I \\ 1 - X_I & \text{if } i = I \text{ and } X + 2^{k-I} < n \\ 0 & \text{if } X_I = 0 \text{ and } X + 2^{k-I} \geq n. \end{cases}$$

Also let $S' = \sum_{i=1}^{k} X_i'$ and $W' = \frac{S'-(k/2)}{\sqrt{k/4}}$. The ordered pair (X, X') of random variables is exchangeable, and thus the pairs (S, S') and (W, W') are both exchangeable. Hence, for any function f on $\{0, 1, \ldots, k\}$, we have

$$\begin{aligned} 0 &= \mathbb{E}\{(S' - S)(f(S) + f(S'))\} \\ &= 2\mathbb{E}\{(S' - S)f(S)\} + \mathbb{E}\{(S' - S)(f(S') - f(S))\} \\ &= 2\mathbb{E}\{f(S)\mathbb{E}(S' - S|X)\} + \mathbb{E}\{\mathbb{E}((S' - S)(f(S') - f(S))|X)\}. \end{aligned} \tag{4.14}$$

Observe that

$$\begin{aligned}
\mathbb{E}(S'-S|X) &= \mathbb{P}(S'-S=1|X) - \mathbb{P}(S'-S=-1|X) \\
&= \mathbb{P}(X_I=0, X_I'=1|X) - \mathbb{P}(X_I=1|X) \\
&= \mathbb{P}(X_I=0|X) - \mathbb{P}(X_I=0, X_I'=0|X) - \mathbb{P}(X_I=1|X) \\
&= \mathbb{E}(1-X_I|X) - \mathbb{P}(X_I=0, X_I'=0|X) - \mathbb{P}(X_I=1|X) \\
&= \frac{1}{k}\sum_{i=1}^{k}(1-X_i) - \frac{1}{k}\sum_{i=1}^{k} I_{\{X_i=0, X+2^{k-i}\geq n\}} - \frac{1}{k}\sum_{i=1}^{k} X_i \\
&= 1 - \frac{2S}{k} - \frac{Q}{k}, \qquad (4.15)
\end{aligned}$$

where $Q = \sum_{i=1}^{k} I_{\{X_i=0, X+2^{k-i}\geq n\}}$. Now rewrite (4.14), with a re-definition of the (arbitrary) function f, as

$$k^{1/2}\mathbb{E}\{f(W)\mathbb{E}(S-S'|X)\} = \tfrac{1}{2}k^{1/2}\mathbb{E}\{\mathbb{E}((f(W')-f(W))(S'-S)|X)\}. \qquad (4.16)$$

The left hand side of (4.16) is

$$\begin{aligned}
k^{1/2}\mathbb{E}\left\{f(W)\Big(-1+\frac{2S}{k}+\frac{Q}{k}\Big)\right\} &= \mathbb{E}\left\{f(W)\Big(\frac{S-k/2}{k^{1/2}/2}+\frac{Q}{k^{1/2}}\Big)\right\} \\
&= \mathbb{E}\{Wf(W)\} + k^{-1/2}\mathbb{E}Qf(W), \qquad (4.17)
\end{aligned}$$

and the right hand side of (4.16) can be written as

$$\begin{aligned}
&\tfrac{1}{2}k^{1/2}\mathbb{E}\Big\{\mathbb{E}\{(f(W')-f(W))I_{\{S'-S=1\}}|X\} \\
&\qquad - \mathbb{E}\{(f(W')-f(W))I_{\{S'-S=-1\}}|X\}\Big\} \\
&= \tfrac{1}{2}k^{1/2}\mathbb{E}\Big\{\mathbb{E}\{((f(W+2k^{-1/2})-f(W))I_{\{S'-S=1\}}|X\} \\
&\qquad - \mathbb{E}\{((f(W-2k^{-1/2})-f(W))I_{\{S'-S=-1\}}|X\}\Big\} \\
&= \tfrac{1}{2}k^{1/2}\mathbb{E}\Big\{(f(W+2k^{-1/2})-f(W))\mathbb{P}(S'-S=1|X) \\
&\qquad - (f(W-2k^{-1/2})-f(W))\mathbb{P}(S'-S=-1|X)\Big\} \\
&= \tfrac{1}{2}k^{1/2}\mathbb{E}\Big\{(f(W+2k^{-1/2})-f(W))\Big(1-\frac{S}{k}-\frac{Q}{k}\Big) \\
&\qquad - (f(W-2k^{-1/2})-f(W))\frac{S}{k}\Big\}.
\end{aligned}$$

The latter expression can in turn be written as

$$\begin{aligned}
&\mathbb{E}\Big\{(f(W+2k^{-1/2})-f(W))\Big(\frac{k-S}{2k^{1/2}}\Big)\Big\} \\
&\qquad -\mathbb{E}\Big\{(f(W-2k^{-1/2})-f(W))\Big(\frac{S}{2k^{1/2}}\Big)\Big\} \\
&\qquad -\mathbb{E}\Big\{(f(W+2k^{-1/2})-f(W))\frac{Q}{2k^{1/2}}\Big\} \qquad (4.18)\\
&= \mathbb{E}\int_{|t|\le 2k^{-1/2}} f'(W+t)M(t)\,dt - \mathbb{E}\Big\{(f(W+2k^{-1/2})-f(W))\frac{Q}{2k^{1/2}}\Big\},
\end{aligned}$$

where

$$M(t)=\begin{cases}\frac{k-S}{2k^{1/2}} & \text{for } 0\le t\le \frac{2}{k^{1/2}}\\[2ex] \frac{S}{2k^{1/2}} & \text{for } -\frac{2}{k^{1/2}}\le t<0.\end{cases}$$

Note that $M(t)\ge 0$ and $\int_{|t|\le 2k^{-1/2}} M(t)\,dt=1$. So, taking $f=f_z$ as given in (2.2), we have

$$\mathbb{E}\{Wf_z(W)\}=\int_{|t|\le 2k^{-1/2}}\mathbb{E}\{f_z'(W+t)M(t)\}\,dt+\mathbb{E}(R_2), \qquad (4.19)$$

of the form (4.9) with $R_1=0$ and $\delta_0=2k^{-1/2}$, where

$$R_2=-\Big\{(f(W+2k^{-1/2})-f(W))\frac{Q}{2k^{1/2}}\Big\}-k^{-1/2}Qf(W).$$

Then, by (2.9), it follows that

$$|\mathbb{E}(R_2)|\le\Big(\frac{\sqrt{2\pi}}{8}+\frac{\sqrt{2\pi}}{4}\Big)\frac{\mathbb{E}Q}{\sqrt{k}}\le 2k^{-1/2},$$

since $\mathbb{E}Q\le 2$, as shown below. It thus follows from Theorem 4.2 that

$$\sup_z|\mathbb{P}(W\le z)-\Phi(z)|\le 2.1(2k^{-1/2})+2k^{-1/2}=6.2k^{-1/2},$$

as desired.

To show that $\mathbb{E}Q\le 2$, simply observe that, from its definition,

$$\begin{aligned}
\mathbb{E}Q&=\sum_{i=1}^{k}\mathbb{P}(X_i=0,X+2^{k-i}\ge n)\ \le\ \sum_{i=1}^{k}\mathbb{P}(X\ge n-2^{k-i})\\
&=\sum_{i=1}^{k}\frac{2^{k-i}}{n}\ =\ \frac{2^k-1}{n}\ \le\ \frac{2^k-1}{2^{k-1}}\ \le\ 2.
\end{aligned}$$

This completes the proof. ∎

The binary expansion of a random integer has previously studied by a number of authors. Diaconis (1977) and Stein (1986) also proved that the distribution of S, the number of ones in the binary expansion, is only order $O(k^{-1/2})$ away from the $B(k, 1/2)$, whereas Barbour & Chen (1992) further proved that, if a mixture of the $B(k-1, 1/2)$ and $B(k, 1/2)$ is used as an approximation, the error can be reduced to $O(k^{-1})$.

5. Uniform Berry–Esseen bounds: the independent case

Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables with zero means and $\sum_{i=1}^n \mathbb{E}\xi_i^2 = 1$. Let

$$\gamma = \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3 \tag{5.1}$$

and $W = \sum_{i=1}^n \xi_i$. If $\mathbb{E}|\xi_i|^3 < \infty$, then we have the uniform Berry–Esseen inequality

$$\sup_z |\mathbb{P}(W \le z) - \Phi(z)| \le C_0\gamma \tag{5.2}$$

and the non-uniform Berry-Esseen inequality

$$\forall\, z \in R^1, \quad |\mathbb{P}(W \le z) - \Phi(z)| \le C_1(1+|z|)^{-3}\gamma, \tag{5.3}$$

where both C_0 and C_1 are absolute constants. One can take $C_0 = 0.7975$ [van Beeck (1972)] and $C_1 = 31.935$ [Paditz (1989)]. We shall use the concentration inequality approach to give a direct proof of (5.2) in this section and (5.3) in next section, albeit with different constants. We refer to Chen & Shao (2001) for more details.

5.1. *The concentration inequality approach*

As indicated in (1.10) and by (4.4) in the proof of Theorem 4.1, a key step in proving the Berry–Esseen bound is to have a good bound for the probability $\mathbb{P}(a \le W^{(i)} \le b)$, where $W^{(i)} = \sum_{j \ne i} \xi_j$. In this section, we prove such a concentration inequality. The idea is once again to use the Stein identity. More precisely, we use the fact that $\mathbb{E}f'(W)$ is close to $\mathbb{E}\{Wf(W)\}$ when W is close to the normal. If f' is taken to be the indicator $\mathbf{1}_{[a,b]}$, then $\mathbb{E}f'(W) = \mathbb{P}(a \le W \le b)$; on the other hand, choosing $f(\frac{1}{2}(b-a)) = 0$, it follows that $\|f\| = \frac{1}{2}(b-a)$, so that

$$|\mathbb{E}\{Wf(W)\}| \le \tfrac{1}{2}(b-a)\mathbb{E}|W| \le \tfrac{1}{2}(b-a)$$

if $\mathbb{E}W^2 = 1$. Thus, if $\mathbb{E}f'(W)$ and $\mathbb{E}\{Wf(W)\}$ are close, it follows that $\mathbb{P}(a \le W \le b)$ is close to $\frac{1}{2}(b-a)$. The proof of the inequality below makes this heuristic precise.

Proposition 5.1: *We have*

$$\mathbb{P}(a \le W^{(i)} \le b) \le \sqrt{2}(b-a) + (1+\sqrt{2})\gamma \tag{5.4}$$

for all real $a < b$ and for every $1 \le i \le n$.

Proof: Define $\delta = \gamma/2$ and take

$$f(w) = \begin{cases} -\frac{1}{2}(b-a) - \delta & \text{if } w < a-\delta, \\ w - \frac{1}{2}(b+a) & \text{if } a-\delta \le w \le b+\delta, \\ \frac{1}{2}(b-a) + \delta & \text{for } w > b+\delta, \end{cases} \tag{5.5}$$

so that $f' = \mathbf{1}_{[a-\delta, b+\delta]}$ and $\|f\| = \frac{1}{2}(b-a) + \delta$. Set

$$\begin{aligned} \widehat{M}_j(t) &= \xi_j(I_{\{-\xi_j \le t \le 0\}} - I_{\{0 < t \le -\xi_j\}}) \ \ge\ 0, \\ \widehat{M}(t) &= \sum_{1 \le j \le n} \widehat{M}_j(t), \quad M(t) \ =\ \mathbb{E}\widehat{M}(t). \end{aligned} \tag{5.6}$$

Since ξ_j and $W^{(i)} - \xi_j$ are independent for $j \ne i$, ξ_i is independent of $W^{(i)}$ and $\mathbb{E}\xi_j = 0$ for all j, we have

$$\begin{aligned} &\mathbb{E}\{W^{(i)} f(W^{(i)})\} - \mathbb{E}\{\xi_i f(W^{(i)} - \xi_i)\} \\ &= \sum_{j=1}^{n} \mathbb{E}\{\xi_j [f(W^{(i)}) - f(W^{(i)} - \xi_j)]\} \ =\ \sum_{j=1}^{n} \mathbb{E}\left\{\xi_j \int_{-\xi_j}^{0} f'(W^{(i)} + t)\, dt\right\}. \end{aligned}$$

Hence, using (5.6), we have

$$\begin{aligned} &\mathbb{E}\{W^{(i)} f(W^{(i)})\} - \mathbb{E}\{\xi_i f(W^{(i)} - \xi_i)\} \\ &= \sum_{j=1}^{n} \mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W^{(i)} + t)\widehat{M}_j(t)\, dt\right\} = \mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W^{(i)} + t)\widehat{M}(t)\, dt\right\}. \end{aligned}$$

At this point, we have reached a precise replacement for the statement '$\mathbb{E}f'(W)$ is close to $\mathbb{E}\{Wf(W)\}$' of the heuristic, with $W^{(i)}$ for W. To exploit it, we first note that

$$\mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W^{(i)} + t)\widehat{M}(t)\, dt\right\} \ \ge\ \mathbb{E}\left\{\int_{|t| \le \delta} f'(W^{(i)} + t)\widehat{M}(t)\, dt\right\},$$

because $f'(t) \geq 0$ and $\widehat{M}(t) \geq 0$. But now the definition of f implies that

$$\mathbb{E}\left\{\int_{|t|\leq\delta} f'(W^{(i)}+t)\widehat{M}(t)\,dt\right\} \geq \mathbb{E}\left\{I_{\{a\leq W^{(i)}\leq b\}}\int_{|t|\leq\delta}\widehat{M}(t)\,dt\right\}$$

$$= \mathbb{E}\left\{I_{\{a\leq W^{(i)}\leq b\}}\sum_{j=1}^{n}|\xi_j|\min(\delta,|\xi_j|)\right\} \geq H_{1,1} - H_{1,2}, \tag{5.7}$$

where

$$H_{1,1} = \mathbb{P}(a \leq W^{(i)} \leq b)\sum_{j=1}^{n}\mathbb{E}|\xi_j|\min(\delta,|\xi_j|)$$

and

$$H_{1,2} = \mathbb{E}\left|\sum_{j=1}^{n}\{|\xi_j|\min(\delta,|\xi_j|) - \mathbb{E}|\xi_j|\min(\delta,|\xi_j|)\}\right|.$$

A direct calculation yields

$$\min(x,y) \geq x - x^2/(4y) \tag{5.8}$$

for $x > 0$ and $y > 0$, implying that

$$\sum_{j=1}^{n}\mathbb{E}|\xi_j|\min(\delta,|\xi_j|) \geq \sum_{j=1}^{n}\left\{\mathbb{E}\xi_j^2 - \frac{\mathbb{E}|\xi_j|^3}{4\delta}\right\} = \frac{1}{2}, \tag{5.9}$$

by choice of δ, and hence that

$$H_{1,1} \geq \tfrac{1}{2}\mathbb{P}(a \leq W^{(i)} \leq b). \tag{5.10}$$

By the Hölder inequality,

$$H_{1,2} \leq \left(\mathrm{Var}\left\{\sum_{j=1}^{n}|\xi_j|\min(\delta,|\xi_j|)\right\}\right)^{1/2}$$

$$\leq \left(\sum_{j=1}^{n}\mathbb{E}\xi_j^2\min(\delta,|\xi_j|)^2\right)^{1/2}$$

$$\leq \delta\left(\sum_{j=1}^{n}\mathbb{E}\xi_j^2\right)^{1/2} = \delta. \tag{5.11}$$

Hence

$$\mathbb{E}\left\{\int_{-\infty}^{\infty} f'(W^{(i)}+t)\widehat{M}(t)\,dt\right\} \geq \tfrac{1}{2}\,\mathbb{P}(a \leq W^{(i)} \leq b) - \delta, \tag{5.12}$$

to be compared with the equation '$\mathbb{E}f'(W) = \mathbb{P}(a \le W \le b)$' of the heuristic.

On the other hand, recalling that $\|f\| \le \frac{1}{2}(b-a) + \delta$, we have

$$\begin{aligned}
&\mathbb{E}\{W^{(i)} f(W^{(i)})\} - \mathbb{E}\{\xi_i f(W^{(i)} - \xi_i)\} \\
&\quad \le \{\tfrac{1}{2}(b-a) + \delta\}(\mathbb{E}|W^{(i)}| + \mathbb{E}|\xi_i|) \\
&\quad \le \frac{1}{\sqrt{2}} \left((\mathbb{E}|W^{(i)}|)^2 + (\mathbb{E}|\xi_i|)^2 \right)^{1/2} (b-a+2\delta) \\
&\quad \le \frac{1}{\sqrt{2}} \left(\mathbb{E}|W^{(i)}|^2 + \mathbb{E}|\xi_i|^2 \right)^{1/2} (b-a+2\delta) \\
&\quad = \frac{1}{\sqrt{2}} (b-a+2\delta), \qquad (5.13)
\end{aligned}$$

the inequality to be compared with '$|\mathbb{E}\{Wf(W)\}| \le \frac{1}{2}(b-a)$' from the heuristic. Combining (5.12) and (5.13) thus gives

$$\mathbb{P}(a \le W^{(i)} \le b) \le \sqrt{2}(b-a) + 2(1+\sqrt{2})\delta \;=\; \sqrt{2}(b-a) + (1+\sqrt{2})\gamma$$

as desired. ∎

5.2. *Proving the Berry–Esseen theorem*

We are now ready to prove the classical Berry–Esseen theorem, with a constant of 7.

Theorem 5.2: *We have*

$$\sup_z |\mathbb{P}(W \le z) - \Phi(z)| \le 7\gamma. \qquad (5.14)$$

Proof: As discussed in the previous section, we need to find a way to replace $\mathbb{P}(W^{(i)} + t \le z)$ by $\mathbb{P}(W \le z)$ in (4.3). However, it follows from (5.4) that

$$\begin{aligned}
&\Big| \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{P}(W^{(i)} + t \le z) K_i(t)\, dt - \mathbb{P}(W \le z) \Big| \\
&\quad \le \sum_{i=1}^n \int_{-\infty}^{\infty} |\mathbb{P}(W^{(i)} + t \le z) - \mathbb{P}(W \le z)| K_i(t)\, dt \\
&\quad = \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}\big\{ \mathbb{P}(z - \max(t, \xi_i) \le W^{(i)} \le z - \min(t, \xi_i) \mid \xi_i) \big\} K_i(t)\, dt.
\end{aligned}$$

Proposition 5.1 thus gives the bound

$$\begin{aligned}
&\Big|\sum_{i=1}^{n}\int_{-\infty}^{\infty}\mathbb{P}(W^{(i)}+t\le z)K_i(t)\,dt-\mathbb{P}(W\le z)\Big| \\
&\quad\le \sum_{i=1}^{n}\int_{-\infty}^{\infty}\mathbb{E}\{\sqrt{2}(|t|+|\xi_i|)+(1+\sqrt{2})\gamma\}K_i(t)\,dt \\
&\quad= (1+\sqrt{2})\gamma+\sqrt{2}\sum_{i=1}^{n}(\tfrac{1}{2}\mathbb{E}|\xi_i|^3+\mathbb{E}|\xi_i|\mathbb{E}\xi_i^2) \\
&\quad\le (1+2.5\sqrt{2})\gamma. \qquad (5.15)
\end{aligned}$$

Now by (4.2)

$$|\mathbb{P}(W\le z)-\Phi(z)|\le(1+2.5\sqrt{2}+1.5(1+\sqrt{2\pi}/4))\gamma\le 7\gamma,$$

which is (4.1). ∎

We remark that following the above lines of proof, one can prove that

$$\sup_z|\mathbb{P}(W\le z)-\Phi(z)|\le 7\sum_{i=1}^{n}(\mathbb{E}\xi_i^2 I_{\{|\xi_i|>1\}}+\mathbb{E}|\xi_i|^3 I_{\{|\xi_i|\le 1\}}),$$

dispensing with the third moment assumption. We leave the proof to the reader. With a more refined concentration inequality, the constant 7 can be reduced to 4.1 (see Chen & Shao (2001)).

5.3. *A lower bound*

Let X_i, $i\ge 1$, be independent random variables with zero means and finite variances, and define $B_n=\sum_{i=1}^{n}\mathrm{Var}X_i$. It is known that if the Feller condition

$$\max_{1\le i\le n}\mathbb{E}X_i^2/B_n^2\to 0, \qquad (5.16)$$

is satisfied, then the Lindeberg condition is necessary for the central limit theorem. Barbour & Hall (1984) used Stein's method to provide not only a nice proof of the necessity, but also a lower bound for the Kolmogorov distance between the distribution of W and the normal, which is as close as can be expected to being a multiple of one of Lindeberg's sums $B_n^{-2}\sum_{i=1}^{n}\mathbb{E}\{X_i^2 I_{\{B_n^{-1}|X_i|>\varepsilon\}}\}$. Note that no lower bound could simply be a multiple of a Lindeberg sum, since a sum of n identically distributed *normal* random variables is itself normally distributed, but the corresponding Lindeberg sum is not zero. The next theorem and its proof are due to Barbour & Hall (1984).

Theorem 5.3: *Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables which have zero means and finite variances $\mathbb{E}\xi_i^2 = \sigma_i^2$, $1 \le i \le n$, and satisfy $\sum_{i=1}^n \sigma_i^2 = 1$. Then there exists an absolute constant C such that for all $\varepsilon > 0$*

$$(1 - e^{-\varepsilon^2/4}) \sum_{i=1}^n \mathbb{E}\{\xi_i^2 I_{\{|\xi_i| > \varepsilon\}}\} \le C\Big(\sup_z |\mathbb{P}(W \le z) - \Phi(z)| + \sum_{i=1}^n \sigma_i^4 \Big). \tag{5.17}$$

Remark: For the sequence of independent random variables $\{X_i, i \ge 1\}$, we take $\xi_i = X_i/B_n$. Clearly, Feller's negligibility condition (5.16) implies that $\sum_{i=1}^n \sigma_i^4 \le \max_{1 \le i \le n} \sigma_i^2 \to 0$ as $n \to \infty$. Therefore, if S_n/B_n is asymptotically normal, then

$$\sum_{i=1}^n \mathbb{E}\{\xi_i^2 I_{\{|\xi_i| > \varepsilon\}}\} \to 0$$

as $n \to \infty$ for every $\varepsilon > 0$, and the Lindeberg condition is satisfied.

Proof: Once again, the argument starts with the Stein identity

$$\mathbb{E}\{f_h'(W) - W f_h(W)\} = \mathbb{E}h(W) - \mathbb{E}h(Z),$$

for h yet to be chosen. Integrating by parts, the right hand side is bounded by

$$\begin{aligned} |\mathbb{E}h(W) - \mathbb{E}h(Z)| &= \left| \int_{-\infty}^{\infty} h'(w)\{\mathbb{P}(W \le w) - \Phi(w)\}\, dw \right| \\ &\le \Delta \int_{-\infty}^{\infty} |h'(w)|\, dw, \end{aligned} \tag{5.18}$$

where $\Delta = \sup_z |\mathbb{P}(W \le z) - \Phi(z)|$. For the left hand side, in the usual way, because ξ_i and $W^{(i)}$ are independent and $\mathbb{E}\xi_i = 0$, we have

$$\begin{aligned} \mathbb{E}\{W f_h(W)\} &= \sum_{i=1}^n \mathbb{E}\{\xi_i^2 f_h'(W^{(i)})\} \\ &\quad + \sum_{i=1}^n \mathbb{E}\{\xi_i (f_h(W^{(i)} + \xi_i) - f_h(W^{(i)}) - \xi_i f_h'(W^{(i)}))\}, \end{aligned}$$

and, because $\sum_{i=1}^n \sigma_i^2 = 1$,

$$\begin{aligned} \mathbb{E}f_h'(W) &= \sum_{i=1}^n \sigma_i^2 \mathbb{E} f_h(W^{(i)} + \xi_i) \\ &= \sum_{i=1}^n \sigma_i^2 \mathbb{E} f_h'(W^{(i)}) + \sum_{i=1}^n \sigma_i^2 \mathbb{E}\{f_h'(W) - f_h'(W^{(i)})\}, \end{aligned}$$

with the last term easily bounded by $\frac{1}{2}\|f_h'''\| \sum_{i=1}^n \sigma_i^4$. Hence

$$\left| \mathbb{E}\{f_h'(W) - Wf_h(W)\} - \sum_{i=1}^n \mathbb{E}\{\xi_i^2 g(W^{(i)}, \xi_i)\} \right| \leq \tfrac{1}{2}\|f_h'''\| \sum_{i=1}^n \sigma_i^4, \quad (5.19)$$

where

$$g(w,y) = g_h(w,y) = -y^{-1}\{f_h(w+y) - f_h(w) - yf_h'(w)\}.$$

Intuitively, if W is close to being normally distributed, then

$$R_1 := \sum_{i=1}^n \mathbb{E}\{\xi_i^2 g(W^{(i)}, \xi_i)\} \quad \text{and} \quad R := \sum_{i=1}^n \mathbb{E}\{\xi_i^2 g(Z, \xi_i)\},$$

for $Z \sim \mathcal{N}(0,1)$ independent of the ξ_i's, should be close to one another.

Taking (5.18) and (5.19) together, it follows that we have a lower bound for Δ, provided that a function h can be found with $\int_{-\infty}^{\infty} |h'(w)|\,dw < \infty$ for which $\mathbb{E}g_h(Z,y)$ is of constant sign, and provided also that $\|f_h'''\| < \infty$. In practice, it is easier to look for a suitable f, and then define h by setting $h(w) = f'(w) - wf(w)$. The function g is zero for any linear function f, and, for even functions f, it follows that $\mathbb{E}g(Z,y)$ is antisymmetric in y about zero, but odd functions are possibilities; for instance, $f(x) = x^3$ has $\mathbb{E}g(Z,y) = -y^2$, of constant sign. Unfortunately, this f fails to have finite $\int_{-\infty}^{\infty} |h'(w)|\,dw$, but the choice $f(w) = we^{-w^2/2}$ is a good one; it behaves much like the sum of a linear and a cubic function for those values of w where $\mathcal{N}(0,1)$ puts most of its mass, yet dies away to zero fast when $|w|$ is large. Making the computations,

$$\begin{aligned}
\mathbb{E}g(Z,y) &= -\frac{1}{y\sqrt{2\pi}} \int_{-\infty}^{\infty} \Big\{ (w+y)e^{-(w+y)^2/2} \\
&\qquad\qquad - we^{-w^2/2} - ye^{-w^2/2}(1-w^2) \Big\}\, e^{-w^2/2}\,dw \\
&= -y^{-1} \left\{ \frac{y}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}(w^2 + (w+y)^2)\}\,dw - \frac{y}{2}\sqrt{\frac{\pi}{2}} \right\} \\
&= \frac{1}{2\sqrt{2}}(1 - e^{-y^2/4}) \;\geq\; 0, \qquad (5.20)
\end{aligned}$$

and $\mathbb{E}g(Z,y) \geq \frac{1}{2\sqrt{2}}(1 - e^{-\varepsilon^2/4})$ whenever $|y| \geq \varepsilon$. Hence, for this choice of f, we have

$$R \;\geq\; \frac{1}{2\sqrt{2}}(1 - e^{-\varepsilon^2/4}) \sum_{i=1}^n \mathbb{E}\{\xi_i^2 I_{\{|\xi_i| > \varepsilon\}}\} \qquad (5.21)$$

for any $\varepsilon > 0$. It thus remains to show that R and R_1 are close enough, after which (5.18), (5.19) and (5.21) complete the proof.

To show this, note first that, taking this choice of f, and setting $h(w) = f'(w) - wf(w)$, we have

$$c_1 := \int_{-\infty}^{\infty} |h'(w)|\,dw \le 5;\ c_2 := \int_{-\infty}^{\infty} |f''(w)|\,dw \le 4;\ c_3 := \sup_w |f'''(w)| = 3.$$

Now define an intermediate R_2 between R_1 and R, by

$$R_2 := \sum_{i=1}^{n} \mathbb{E}\{\xi_i^2 g(W^*, \xi_i)\},$$

where W^* has the same distribution as W, but is independent of the ξ_i's. Then

$$\begin{aligned} R_1 &= -\sum_{i=1}^{n} \mathbb{E}\Big\{\xi_i^2 \int_0^1 [f'(W^{(i)} + t\xi_i) - f'(W^{(i)})]\,dt\Big\} \\ &= R_2 + \sum_{i=1}^{n} \mathbb{E}\Big\{\xi_i^2 \int_0^1 [f'(W^* + t\xi_i) - f'(W^{(i)} + t\xi_i)]\,dt\Big\} \\ &\quad - \sum_{i=1}^{n} \mathbb{E}\Big\{\xi_i^2 \int_0^1 [f'(W^*) - f'(W^{(i)})]\,dt\Big\}. \end{aligned} \tag{5.22}$$

Now, for any θ, and because ξ_i and $W^{(i)}$ are independent, with $\mathbb{E}\xi_i = 0$,

$$\begin{aligned} &|\mathbb{E}\big(f'(W^* + \theta) - f'(W^{(i)} + \theta)\big)| \\ &\quad = |\mathbb{E}\big(f'(W^{(i)} + \xi_i + \theta) - f'(W^{(i)} + \theta)\big)| \\ &\quad = |\mathbb{E}\big(f'(W^{(i)} + \xi_i + \theta) - f'(W^{(i)} + \theta) - \xi_i f''(W^{(i)} + \theta)\big)| \\ &\quad \le \tfrac{1}{2} c_3 \sigma_i^2, \end{aligned}$$

by Taylor's theorem. Hence, from (5.22) and because $\sum_{i=1}^{n} \sigma_i^2 = 1$,

$$R_1 \ \ge\ R_2 - c_3 \sum_{i=1}^{n} \sigma_i^4. \tag{5.23}$$

Similarly,

$$\begin{aligned} R_2 = R &+ \sum_{i=1}^{n} \mathbb{E}\Big\{\xi_i^2 \int_0^1 [f'(Z) + t\xi_i) - f'(W^* + t\xi_i)]\,dt\Big\} \\ &- \sum_{i=1}^{n} \mathbb{E}\Big\{\xi_i^2 \int_0^1 [f'(Z) - f'(W^*)]\,dt\Big\}, \end{aligned}$$

and, for any θ,

$$\begin{aligned}&|\mathbb{E}f'(W^*+\theta) - \mathbb{E}f'(Z+\theta)| \\ &\quad = \left|\int_{-\infty}^{\infty} f''(w)\big(\mathbb{P}(W^* \le w-\theta) - \Phi(w-\theta)\big)\,dw\right| \;\le\; c_2\Delta,\end{aligned}$$

so that

$$R_2 \ge R - 2c_2\Delta. \tag{5.24}$$

Combining (5.18) and (5.19) with (5.23) and (5.24), it follows that

$$c_1\Delta \;\ge\; R_1 - \tfrac{1}{2}c_3\sum_{i=1}^n \sigma_i^4 \;\ge\; R - \tfrac{3}{2}c_3\sum_{i=1}^n \sigma_i^4 - 2c_2\Delta.$$

In view of (5.21), collecting terms, it follows that

$$\Delta(c_1+2c_2) + \tfrac{3}{2}c_3\sum_{i=1}^n \sigma_i^4 \ge \frac{1}{2\sqrt{2}}(1-e^{-\varepsilon^2/4})\sum_{i=1}^n \mathbb{E}\{\xi_i^2 I_{\{|\xi_i|>\varepsilon\}}\} \tag{5.25}$$

for any $\varepsilon > 0$. This proves (5.17), with $C \le 40$. ∎

6. Non-uniform Berry–Esseen bounds: the independent case

In this section, we prove a non-uniform Berry–Esseen bound similar to (5.3), following the proof in Chen & Shao (2001). To do this, as in the previous section, we first need a concentration inequality; it must itself now be non-uniform.

6.1. *A non-uniform concentration inequality*

Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables satisfying $\mathbb{E}\xi_i = 0$ for every $1 \le i \le n$ and $\sum_{i=1}^n \mathbb{E}\xi_i^2 = 1$. Let

$$\bar{\xi}_i = \xi_i I_{\{\xi_i \le 1\}}, \quad \overline{W} = \sum_{i=1}^n \bar{\xi}_i, \quad \overline{W}^{(i)} = \overline{W} - \bar{\xi}_i.$$

Proposition 6.1: *We have*

$$\mathbb{P}(a \le \overline{W}^{(i)} \le b) \le e^{-a/2}(5(b-a)+7\gamma) \tag{6.1}$$

for all real $b > a$ and for every $1 \le i \le n$, where $\gamma = \sum_{i=1}^n \mathbb{E}|\xi_i|^3$.

To prove it, we first need to have the following Bennett–Hoeffding inequality.

Lemma 6.2: *Let $\eta_1, \eta_2, \cdots, \eta_n$ be independent random variables satisfying $\mathbb{E}\eta_i \le 0$, $\eta_i \le \alpha$ for $1 \le i \le n$, and $\sum_{i=1}^n \mathbb{E}\eta_i^2 \le B_n^2$. Put $S_n = \sum_{i=1}^n \eta_i$. Then*

$$\mathbb{E}e^{tS_n} \le \exp\Big(\alpha^{-2}(e^{t\alpha} - 1 - t\alpha)B_n^2\Big) \tag{6.2}$$

for $t > 0$,

$$\mathbb{P}(S_n \ge x) \le \exp\Big(-\frac{B_n^2}{\alpha^2}\Big[(1 + \frac{\alpha x}{B_n^2})\ln(1 + \frac{\alpha x}{B_n^2}) - \frac{\alpha x}{B_n^2}\Big]\Big) \tag{6.3}$$

and

$$\mathbb{P}(S_n \ge x) \le \exp\Big(-\frac{x^2}{2(B_n^2 + \alpha x)}\Big) \tag{6.4}$$

for $x > 0$.

Proof: It is easy to see that $(e^s - 1 - s)/s^2$ is an increasing function of $s \in \mathbb{R}$, from which it follows that

$$e^{ts} \le 1 + ts + (ts)^2(e^{t\alpha} - 1 - t\alpha)/(t\alpha)^2 \tag{6.5}$$

for $s \le \alpha$, if $t > 0$. Using the properties of the η_i's, we thus have

$$\begin{aligned}
\mathbb{E}e^{tS_n} &= \prod_{i=1}^n \mathbb{E}e^{t\eta_i} \\
&\le \prod_{i=1}^n \big(1 + t\mathbb{E}\eta_i + \alpha^{-2}(e^{t\alpha} - 1 - t\alpha)\mathbb{E}\eta_i^2\big) \\
&\le \prod_{i=1}^n \big(1 + \alpha^{-2}(e^{t\alpha} - 1 - t\alpha)\mathbb{E}\eta_i^2\big) \\
&\le \exp\Big(\alpha^{-2}(e^{t\alpha} - 1 - t\alpha)B_n^2\Big).
\end{aligned}$$

This proves (6.2).

To prove inequality (6.3), let

$$t = \frac{1}{\alpha}\ln\Big(1 + \frac{\alpha x}{B_n^2}\Big).$$

Then, by (6.2),

$$\begin{aligned}\mathbb{P}(S_n \geq x) &\leq e^{-tx}\mathbb{E}e^{tS_n} \\ &\leq \exp\Big(-tx + \alpha^{-2}(e^{t\alpha} - 1 - t\alpha)B_n^2\Big) \\ &= \exp\Big(-\frac{B_n^2}{\alpha^2}\Big[(1 + \frac{\alpha x}{B_n^2})\ln(1 + \frac{\alpha x}{B_n^2}) - \frac{\alpha x}{B_n^2}\Big]\Big).\end{aligned}$$

In view of the fact that

$$(1+s)\ln(1+s) - s \;\geq\; \frac{s^2}{2(1+s)}$$

for $s > 0$, (6.4) follows from (6.3). ∎

Proof of Proposition 6.1. It follows from (6.2) with $\alpha = 1$ and $B_n^2 = 1$ that

$$\begin{aligned}\mathbb{P}(a \leq \overline{W}^{(i)} \leq b) &\leq e^{-a/2}\mathbb{E}e^{\overline{W}^{(i)}/2} \\ &\leq e^{-a/2}\exp(e^{1/2} - \tfrac{3}{2}) \;\leq\; 1.19\, e^{-a/2}.\end{aligned}$$

Thus, (6.1) holds if $7\gamma \geq 1.19$.

We now assume that $\gamma \leq 1.19/7 = 0.17$. Much as in the proof of Proposition 5.1, define $\delta = \gamma/2$ (≤ 0.085), and set

$$f(w) = \begin{cases} 0 & \text{if } w < a - \delta, \\ e^{w/2}(w - a + \delta) & \text{if } a - \delta \leq w \leq b + \delta, \\ e^{w/2}(b - a + 2\delta) & \text{if } w > b + \delta. \end{cases} \tag{6.6}$$

Put

$$\overline{M}_i(t) = \xi_i(I_{\{-\bar{\xi}_i \leq t \leq 0\}} - I_{\{0 < t \leq -\bar{\xi}_i\}}), \quad \overline{M}^{(i)}(t) = \sum_{j \neq i} \overline{M}_j(t).$$

Clearly, $\overline{M}^{(i)}(t) \geq 0$, $f'(w) \geq 0$ and $f'(w) \geq e^{w/2}$ for $a - \delta \leq w \leq b + \delta$.

Arguing much as in the derivation of (5.7), we obtain

$$\begin{aligned}\mathbb{E}\{W^{(i)} f(\overline{W}^{(i)})\} &= \sum_{j \neq i} \mathbb{E}\Big\{\xi_j[f(\overline{W}^{(i)}) - f(\overline{W}^{(i)} - \bar{\xi}_j)]\Big\} \\ &= \sum_{j \neq i} \mathbb{E}\Big\{\int_{-\infty}^{\infty} f'(\overline{W}^{(i)} + t)\overline{M}_j(t)\, dt\Big\} \\ &= \mathbb{E}\Big\{\int_{-\infty}^{\infty} f'(\overline{W}^{(i)} + t)\overline{M}^{(i)}(t)\, dt\Big\}.\end{aligned}$$

This expression is now bounded below, giving

$$\begin{aligned}
\mathbb{E}\{W^{(i)} f(\overline{W}^{(i)})\} &\geq \mathbb{E}\left\{ I_{\{a \leq \overline{W}^{(i)} \leq b\}} \int_{|t| \leq \delta} f'(\overline{W}^{(i)} + t) \overline{M}^{(i)}(t)\, dt \right\} \\
&\geq \mathbb{E}\left\{ e^{(\overline{W}^{(i)} - \delta)/2} I_{\{a \leq \overline{W}^{(i)} \leq b\}} \int_{|t| \leq \delta} \overline{M}^{(i)}(t)\, dt \right\} \\
&\geq \mathbb{E}\left\{ e^{(\overline{W}^{(i)} - \delta)/2} I_{\{a \leq \overline{W}^{(i)} \leq b\}} \sum_{j \neq i} |\xi_j| \min(\delta, |\bar{\xi}_j|) \right\} \\
&\geq e^{-\delta/2}(H_{2,1} - H_{2,2}), \qquad (6.7)
\end{aligned}$$

where

$$\begin{aligned}
H_{2,1} &= \mathbb{E}\Big\{ e^{\overline{W}^{(i)}/2} I_{\{a \leq \overline{W}^{(i)} \leq b\}} \Big\} \sum_{j \neq i} \mathbb{E}\{|\xi_j| \min(\delta, |\bar{\xi}_j|)\}, \\
H_{2,2} &= \mathbb{E}\Big\{ e^{\overline{W}^{(i)}/2} \Big| \sum_{j \neq i} |\xi_j| \min(\delta, |\bar{\xi}_j|) - \mathbb{E}|\xi_j| \min(\delta, |\bar{\xi}_j|) \Big| \Big\}.
\end{aligned}$$

Noting that $\delta \leq .085$ and that $\gamma \leq .17$, and following the proof of (5.9), we have

$$\begin{aligned}
\sum_{j \neq i} \mathbb{E}\{|\xi_j| \min(\delta, |\bar{\xi}_j|)\} &= \sum_{j \neq i} \mathbb{E}(|\xi_j| [\min(\delta, |\xi_j|) - \delta I_{\{\xi_j > 1\}}]) \\
&\geq -\delta \mathbb{E}|\xi_i| + \sum_{j=1}^{n} \mathbb{E}\{|\xi_j| \min(\delta, |\xi_j|)\} - \delta\gamma \\
&\geq -\delta\gamma^{1/3} + 0.5 - \delta\gamma \\
&\geq -0.085(0.17)^{1/3} + 0.5 - 0.085(0.17) \;\geq\; 0.43. \qquad (6.8)
\end{aligned}$$

Hence

$$H_{2,1} \geq 0.43\, e^{a/2}\, \mathbb{P}(a \leq \overline{W}^{(i)} \leq b). \qquad (6.9)$$

By the Bennett inequality (6.2) again, we have

$$\mathbb{E} e^{\overline{W}^{(i)}} \leq \exp(e - 2)$$

and hence, as for (5.11),

$$\begin{aligned}
H_{2,2} &\leq \Big(\mathbb{E} e^{\overline{W}^{(i)}} \Big)^{1/2} \Big(\mathrm{Var}\Big(\sum_{j \neq i} |\xi_j| \min(\delta, |\bar{\xi}_j|) \Big) \Big)^{1/2} \\
&\leq \exp(\tfrac{e}{2} - 1)\delta \;\leq\; 1.44\, \delta. \qquad (6.10)
\end{aligned}$$

As to the left hand side of (6.7), we have

$$\begin{aligned}\mathbb{E}\{W^{(i)}f(\overline{W}^{(i)})\} &\le (b-a+2\delta)\mathbb{E}\left\{|W^{(i)}|e^{\overline{W}^{(i)}/2}\right\}\\ &\le (b-a+2\delta)(\mathbb{E}|W^{(i)}|^2)^{1/2}\left(\mathbb{E}e^{\overline{W}^{(i)}}\right)^{1/2}\\ &\le (b-a+2\delta)\exp(e-2) \;\le\; 2.06(b-a+2\delta).\end{aligned}$$

Combining the above inequalities yields

$$\begin{aligned}\mathbb{P}(a \le \overline{W}^{(i)} \le b) &\le \frac{e^{-a/2}}{0.43}\left(e^{\delta/2}2.06(b-a+2\delta)+1.44\,\delta\right)\\ &\le \frac{e^{-a/2}}{0.43}\left(e^{.0425}2.06(b-a+2\delta)+1.44\,\delta\right)\\ &\le e^{-a/2}(5(b-a)+13.4\,\delta)\\ &\le e^{-a/2}(5(b-a)+7\gamma).\end{aligned}$$

This proves (6.1). ∎

Before proving the main theorem, we need the following moment inequality.

Lemma 6.3: *Let $2 < p \le 3$, and let $\{\eta_i, 1 \le i \le n\}$ be independent random variables with $\mathbb{E}\eta_i = 0$ and $\mathbb{E}|\eta_i|^p < \infty$. Put $S_n = \sum_{i=1}^n \eta_i$ and $B_n^2 = \sum_{i=1}^n \mathbb{E}\eta_i^2$. Then*

$$\mathbb{E}|S_n|^p \le (p-1)B_n^p + \sum_{i=1}^n \mathbb{E}|\eta_i|^p \tag{6.11}$$

Remark: Moment inequalities of this kind were first proved by Rosenthal (1970), in a more general martingale setting.

Proof: Let $S_n^{(i)} = S_n - \eta_i$. Then

$$\begin{aligned}\mathbb{E}|S_n|^p &= \sum_{i=1}^n \mathbb{E}\eta_i S_n|S_n|^{p-2}\\ &= \sum_{i=1}^n \mathbb{E}\eta_i(S_n|S_n|^{p-2} - S_n^{(i)}|S_n|^{p-2})\\ &\qquad + \sum_{i=1}^n \mathbb{E}\eta_i(S_n^{(i)}|S_n|^{p-2} - S_n^{(i)}|S_n^{(i)}|^{p-2}),\end{aligned}$$

once again because η_i and $S_n^{(i)}$ are independent, and $\mathbb{E}\eta_i = 0$. Thus we have

$$\begin{aligned}\mathbb{E}|S_n|^p &\le \sum_{i=1}^n \mathbb{E}\eta_i^2 |S_n|^{p-2} + \sum_{i=1}^n \mathbb{E}|\eta_i||S_n^{(i)}|\{(|S_n^{(i)}| + |\eta_i|)^{p-2} - |S_n^{(i)}|^{p-2}\} \\ &\le \sum_{i=1}^n \mathbb{E}\eta_i^2 (|\eta_i|^{p-2} + |S_n^{(i)}|^{p-2}) \\ &\qquad + \sum_{i=1}^n \mathbb{E}|\eta_i||S_n^{(i)}|^{p-1}\{(1 + |\eta_i|/|S_n^{(i)}|)^{p-2} - 1\}.\end{aligned}$$

Since $(1+x)^{p-2} - 1 \le (p-2)x$ in $x \ge 0$, we thus have

$$\begin{aligned}\mathbb{E}|S_n|^p &\le \sum_{i=1}^n \mathbb{E}|\eta_i|^p + \sum_{i=1}^n \mathbb{E}\eta_i^2 \mathbb{E}|S_n^{(i)}|^{p-2} \\ &\qquad + \sum_{i=1}^n \mathbb{E}|\eta_i||S_n^{(i)}|^{p-1}(p-2)|\eta_i|/|S_n^{(i)}| \\ &= \sum_{i=1}^n \mathbb{E}|\eta_i|^p + (p-1)\sum_{i=1}^n \mathbb{E}\eta_i^2 \mathbb{E}|S_n^{(i)}|^{p-2},\end{aligned}$$

and Hölder's inequality now gives

$$\begin{aligned}\mathbb{E}|S_n|^p &\le \sum_{i=1}^n \mathbb{E}|\eta_i|^p + (p-1)\sum_{i=1}^n \mathbb{E}\eta_i^2 (\mathbb{E}|S_n^{(i)}|^2)^{(p-2)/2} \\ &\le \sum_{i=1}^n \mathbb{E}|\eta_i|^p + (p-1)B_n^p,\end{aligned}$$

as desired. ∎

6.2. *The final result*

We are now ready to prove the non-uniform Berry–Esseen inequality.

Theorem 6.4: *There exists an absolute constant C such that for every real number z,*

$$|\mathbb{P}(W \le z) - \Phi(z)| \le \frac{C\gamma}{1 + |z|^3}. \tag{6.12}$$

Proof: Without loss of generality, assume $z \ge 0$. By (6.11),

$$\mathbb{P}(W \ge z) \le \frac{1 + \mathbb{E}|W|^3}{1 + z^3} \le \frac{1 + 2 + \gamma}{1 + z^3}.$$

Thus (6.12) holds if $\gamma \geq 1$, and we can now assume $\gamma < 1$. Let

$$\bar{\xi}_i = \xi_i I_{\{\xi_i \leq 1\}}, \quad \overline{W} = \sum_{i=1}^{n} \bar{\xi}_i, \quad \overline{W}^{(i)} = \overline{W} - \bar{\xi}_i.$$

Observing that

$$\begin{aligned} \{W \geq z\} &= \{W \geq z, \max_{1 \leq i \leq n} \xi_i > 1\} \cup \{W \geq z, \max_{1 \leq i \leq n} \xi_i \leq 1\} \\ &\subset \{W \geq z, \max_{1 \leq i \leq n} \xi_i > 1\} \cup \{\overline{W} \geq z\}, \end{aligned}$$

we have

$$\mathbb{P}(W > z) \;\leq\; \mathbb{P}(\overline{W} > z) + \mathbb{P}(W > z, \max_{1 \leq i \leq n} \xi_i > 1), \tag{6.13}$$

and, since clearly $W \geq \overline{W}$,

$$\mathbb{P}(\overline{W} > z) \;\leq\; \mathbb{P}(W > z). \tag{6.14}$$

Note that

$$\begin{aligned} \mathbb{P}(W > z, \max_{1 \leq i \leq n} \xi_i > 1) \;&\leq\; \sum_{i=1}^{n} \mathbb{P}(W > z, \xi_i > 1) \\ &\leq\; \sum_{i=1}^{n} \mathbb{P}(\xi_i > \max(1, z/2)) + \sum_{i=1}^{n} \mathbb{P}(W^{(i)} > z/2, \xi_i > 1) \\ &=\; \sum_{i=1}^{n} \mathbb{P}(\xi_i > \max(1, z/2)) + \sum_{i=1}^{n} \mathbb{P}(W^{(i)} > z/2)\mathbb{P}(\xi_i > 1) \\ &\leq\; \frac{\gamma}{\max(1, z/2)^3} + \sum_{i=1}^{n} \frac{(1 + \mathbb{E}|W^{(i)}|^3)}{1 + (z/2)^3} \mathbb{E}|\xi_i|^3 \;\leq\; \frac{C\gamma}{1 + z^3}; \end{aligned}$$

here, and in what follows, C denotes a generic absolute constant, whose value may be different at each appearance. Thus, to prove (6.12), it suffices to show that

$$|\mathbb{P}(\overline{W} \leq z) - \Phi(z)| \;\leq\; Ce^{-z/2}\gamma. \tag{6.15}$$

Let f_z be the solution to the Stein equation (2.2), and define

$$\overline{K}_i(t) = \mathbb{E}\{\bar{\xi}_i(I_{\{0 \leq t \leq \bar{\xi}_i\}} - I_{\{\bar{\xi}_i \leq t < 0\}})\}.$$

We follow the proof of (2.17), noting that $\bar{\xi}_i \leq 1$, and that $\mathbb{E}\bar{\xi}_i$ need no longer in general be equal to zero, but may be negative. This gives

$$\mathbb{E}\{\overline{W} f_z(\overline{W})\} = \sum_{i=1}^{n} \int_{-\infty}^{1} \mathbb{E} f_z'(\overline{W}^{(i)} + t) \overline{K}_i(t)\, dt + \sum_{i=1}^{n} \mathbb{E}\bar{\xi}_i \mathbb{E} f_z(\overline{W}^{(i)}).$$

From

$$\sum_{i=1}^n \int_{-\infty}^1 \overline{K}_i(t)\,dt = \sum_{i=1}^n \mathbb{E}\bar{\xi}_i^2 = 1 - \sum_{i=1}^n \mathbb{E}\xi_i^2 I_{\{\xi_i>1\}},$$

we obtain that

$$\begin{aligned}
\mathbb{P}(\overline{W} \le z) - \Phi(z) &= \mathbb{E}f_z'(\overline{W}) - \mathbb{E}\{\overline{W} f_z(\overline{W})\} \\
&= \sum_{i=1}^n \mathbb{E}\{\xi_i^2 I_{\{\xi_i>1\}}\}\mathbb{E}f_z'(\overline{W}) \\
&\quad + \sum_{i=1}^n \int_{-\infty}^1 \mathbb{E}[f_z'(\overline{W}^{(i)} + \bar{\xi}_i) - f_z'(\overline{W}^{(i)} + t)]\overline{K}_i(t)\,dt \\
&\quad + \sum_{i=1}^n \mathbb{E}\{\xi_i I_{\{\xi_i>1\}}\}\mathbb{E}f_z(\overline{W}^{(i)}) \\
&:= R_1 + R_2 + R_3. \qquad (6.16)
\end{aligned}$$

By (8.3), (2.8) and (6.2),

$$\begin{aligned}
\mathbb{E}|f_z'(\overline{W})| &= \mathbb{E}\{|f_z'(\overline{W})|I_{\{\overline{W}\le z/2\}}\} + \mathbb{E}\{|f_z'(\overline{W})|I_{\{\overline{W}>z/2\}}\} \\
&\le (1 + \sqrt{2\pi}(z/2)e^{z^2/8})(1 - \Phi(z)) + \mathbb{P}(\overline{W} > z/2) \\
&\le (1 + \sqrt{2\pi}(z/2)e^{z^2/8})(1 - \Phi(z)) + e^{-z/2}\mathbb{E}e^{\overline{W}} \\
&\le Ce^{-z/2},
\end{aligned}$$

and hence

$$|R_1| \le C\gamma e^{-z/2}. \qquad (6.17)$$

Similarly, we have $\mathbb{E}f_z(\overline{W}^{(i)}) \le Ce^{-z/2}$ and

$$|R_3| \le C\gamma e^{-z/2}. \qquad (6.18)$$

To estimate R_2, write

$$R_2 = R_{2,1} + R_{2,2},$$

where

$$\begin{aligned}
R_{2,1} &= \sum_{i=1}^n \int_{-\infty}^1 \mathbb{E}[I_{\{\overline{W}^{(i)}+\bar{\xi}_i\le z\}} - I_{\{\overline{W}^{(i)}+t\le z\}}]\,\overline{K}_i(t)\,dt, \\
R_{2,2} &= \sum_{i=1}^n \int_{-\infty}^1 \mathbb{E}[(\overline{W}^{(i)} + \bar{\xi}_i) f_z(\overline{W}^{(i)} + \bar{\xi}_i) \\
&\qquad\qquad - (\overline{W}^{(i)} + t) f_z(\overline{W}^{(i)} + t)]\,\overline{K}_i(t)\,dt.
\end{aligned}$$

By Proposition 6.1,

$$\begin{aligned} R_{2,1} &\le \sum_{i=1}^{n} \int_{-\infty}^{1} \mathbb{E}\{I_{\{\bar{\xi}_i \le t\}} \mathbb{P}(z - t < \overline{W}^{(i)} \le z - \bar{\xi}_i \mid \bar{\xi}_i)\} \overline{K}_i(t)\, dt \\ &\le C \sum_{i=1}^{n} \int_{-\infty}^{1} e^{-(z-t)/2} \mathbb{E}(|\bar{\xi}_i| + |t| + \gamma) \overline{K}_i(t)\, dt \\ &\le C e^{-z/2} \gamma. \end{aligned} \tag{6.19}$$

From Lemma 6.5, proved below, it follows that

$$\begin{aligned} R_{2,2} &\le \sum_{i=1}^{n} \int_{-\infty}^{1} \mathbb{E}\Big\{ I_{\{t \le \bar{\xi}_i\}} [\mathbb{E}(\{\overline{W}^{(i)} + \bar{\xi}_i\} f_z(\overline{W}^{(i)} + \bar{\xi}_i) \mid \bar{\xi}_i) \\ &\qquad - \mathbb{E}(\overline{W}^{(i)} + t) f_z(\overline{W}^{(i)} + t)] \Big\} \overline{K}_i(t)\, dt \\ &\le C e^{-z/2} \sum_{i=1}^{n} \int_{-\infty}^{1} \mathbb{E}(|\bar{\xi}_i| + |t|) \overline{K}_i(t)\, dt \\ &\le C e^{-z/2} \gamma. \end{aligned} \tag{6.20}$$

Therefore

$$R_2 \le C e^{-z/2} \gamma. \tag{6.21}$$

Similarly, we have

$$R_2 \ge -C e^{-z/2} \gamma. \tag{6.22}$$

This proves the theorem. ∎

It remains to prove the following lemma.

Lemma 6.5: *For $s < t \le 1$, we have*

$$\begin{aligned} &\mathbb{E}\{(\overline{W}^{(i)} + t) f_z(\overline{W}^{(i)} + t)\} - \mathbb{E}\{(\overline{W}^{(i)} + s) f_z(\overline{W}^{(i)} + s)\} \\ &\qquad \le\ C e^{-z/2} (|s| + |t|). \end{aligned} \tag{6.23}$$

Proof: Let $g(w) = (w f_z(w))'$. Then

$$\mathbb{E}\{(\overline{W}^{(i)} + t) f_z(\overline{W}^{(i)} + t)\} - \mathbb{E}\{(\overline{W}^{(i)} + s) f_z(\overline{W}^{(i)} + s)\} = \int_s^t \mathbb{E} g(\overline{W}^{(i)} + u) du.$$

From the definition of g and f_z, we get

$$g(w) = \begin{cases} \Big(\sqrt{2\pi}(1 + w^2) e^{w^2/2} (1 - \Phi(w)) - w\Big) \Phi(z), & w \ge z \\ \Big(\sqrt{2\pi}(1 + w^2) e^{w^2/2} \Phi(w) + w\Big) (1 - \Phi(z)), & w < z. \end{cases}$$

By (2.6), $g(w) \geq 0$ for all real w. A direct calculation shows that

$$\sqrt{2\pi}(1+w^2)e^{w^2/2}\Phi(w) + w \leq 2 \quad \text{for } w \leq 0. \tag{6.24}$$

Thus, we have

$$g(w) \leq \begin{cases} 4(1+z^2)e^{z^2/8}(1-\Phi(z)) & \text{if } w \leq z/2 \\ 4(1+z^2)e^{z^2/2}(1-\Phi(z)) & \text{if } z/2 < w \leq z, \end{cases}$$

and this latter bound holds also for $w > z$, as can be seen by applying (6.24) with $-w$ for w to the formula for g in $w \geq z$. Hence, for any $u \in [s,t]$, we have

$$\begin{aligned} &\mathbb{E}g(\overline{W}^{(i)} + u) \\ &= \mathbb{E}\Big\{g(\overline{W}^{(i)} + u)I_{\{\overline{W}^{(i)}+u \leq z/2\}}\Big\} + \mathbb{E}\Big\{g(\overline{W}^{(i)} + u)I_{\{\overline{W}^{(i)}+u > z/2\}}\Big\} \\ &\leq 4(1+z^2)e^{z^2/8}(1-\Phi(z)) + 4(1+z^2)e^{z^2/2}(1-\Phi(z))\mathbb{P}(\overline{W}^{(i)} + u > z/2) \\ &\leq Ce^{-z/2} + C(1+z)e^{-z+2u}\mathbb{E}e^{2\overline{W}^{(i)}}. \end{aligned}$$

But $u \leq t \leq 1$, and so

$$\mathbb{E}g(\overline{W}^{(i)} + u) \leq Ce^{-z/2} + C(1+z)e^{-z}\,\mathbb{E}e^{2\overline{W}^{(i)}} \leq Ce^{-z/2},$$

by (6.2). This gives

$$\mathbb{E}\{(\overline{W}^{(i)} + t)f_z(\overline{W}^{(i)} + t)\} - \mathbb{E}\{(\overline{W}^{(i)} + s)f_z(\overline{W}^{(i)} + s)\} \leq Ce^{-z/2}(|s| + |t|),$$

proving (6.23). ∎

7. Uniform and non-uniform bounds under local dependence

In this section, we extend the discussion of normal approximation under local dependence using Stein's method, which was begun in Section 3.2. Our aim is to establish optimal uniform and non-uniform Berry–Esseen bounds under local dependence.

Throughout this section, let $\mathcal{J}$ be an index set of cardinality n and let $\{\xi_i, i \in \mathcal{J}\}$ be a random field with zero means and finite variances. Define $W = \sum_{i\in\mathcal{J}} \xi_i$ and assume that $\mathrm{Var}(W) = 1$. For $A \subset \mathcal{J}$, let ξ_A denote $\{\xi_i, i \in A\}$, $A^c = \{j \in \mathcal{J} : j \notin A\}$ and $|A|$ the cardinality of A. We introduce four dependence assumptions, the first two of which appeared in Section 3.2

(LD1) For each $i \in \mathcal{J}$ there exists $A_i \subset \mathcal{J}$ such that ξ_i and $\xi_{A_i^c}$ are independent.

(LD2) For each $i \in \mathcal{J}$ there exist $A_i \subset B_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{A_i^c}$ and ξ_{A_i} is independent of $\xi_{B_i^c}$.

(LD3) For each $i \in \mathcal{J}$ there exist $A_i \subset B_i \subset C_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{A_i^c}$, ξ_{A_i} is independent of $\xi_{B_i^c}$, and ξ_{B_i} is independent of $\xi_{C_i^c}$.

(LD4*) For each $i \in \mathcal{J}$ there exist $A_i \subset B_i \subset B_i^* \subset C_i^* \subset D_i^* \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{A_i^c}$, ξ_{A_i} is independent of $\xi_{B_i^c}$, and then ξ_{A_i} is independent of $\{\xi_{A_j}, j \in B_i^{*c}\}$, $\{\xi_{A_l}, l \in B_i^*\}$ is independent of $\{\xi_{A_j}, j \in C_i^{*c}\}$, and $\{\xi_{A_l}, l \in C_i^*\}$ is independent of $\{\xi_{A_j}, j \in D_i^{*c}\}$.

It is clear that (LD4*) implies (LD3), (LD3) yields (LD2) and (LD1) is the weakest assumption. Roughly speaking, (LD4*) is a version of (LD3) for $\{\xi_{A_i}, i \in \mathcal{J}\}$. On the other hand, (LD1) in many cases actually implies (LD2), (LD3) and (LD4*) and B_i, C_i, B_i^*, C_i^* and D_i^* could be chosen as: $B_i = \cup_{j \in A_i} A_j$, $C_i = \cup_{j \in B_i} A_j$, $B_i^* = \cup_{j \in A_i} B_j$, $C_i^* = \cup_{j \in B_i^*} B_j$ and $D_i^* = \cup_{j \in C_i^*} B_j$.

We first present a general uniform Berry–Esseen bound under assumption (LD2).

Theorem 7.1: *Let* $N(B_i) = \{j \in \mathcal{J} : B_j \cap B_i \neq \emptyset\}$ *and* $2 < p \leq 4$. *Assume that (LD2) is satisfied with* $|N(B_i)| \leq \kappa$. *Then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \tag{7.1}$$

$$\leq (13 + 11\kappa) \sum_{i \in \mathcal{J}} (\mathbb{E}|\xi_i|^{3 \wedge p} + \mathbb{E}|Y_i|^{3 \wedge p}) + 2.5 \Big(\kappa \sum_{i \in \mathcal{J}} (\mathbb{E}|\xi_i|^p + \mathbb{E}|Y_i|^p) \Big)^{1/2},$$

where $Y_i = \sum_{j \in A_i} \xi_j$. *In particular, if* $\mathbb{E}|\xi_i|^p + \mathbb{E}|Y_i|^p \leq \theta^p$ *for some* $\theta > 0$ *and for each* $i \in \mathcal{J}$, *then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq (13 + 11\kappa)\, n\, \theta^{3 \wedge p} + 2.5 \theta^{p/2} \sqrt{\kappa n}, \tag{7.2}$$

where $n = |\mathcal{J}|$.

Note that in many cases κ is bounded and θ is of order of $n^{-1/2}$. In those cases, $\kappa\, n\, \theta^{3 \wedge p} + \theta^{p/2} \sqrt{\kappa n} = O(n^{-(p-2)/4})$, which is of the best possible order of $n^{-1/2}$ when $p = 4$. However, the cost is the existence of fourth moments. To reduce the assumption on moments, we need the stronger condition (LD3).

Theorem 7.2: *Let $2 < p \leq 3$. Assume that (LD3) is satisfied with $|N(C_i)| \leq \kappa$, where $N(C_i) = \{j \in \mathcal{J} : C_i B_j \neq \emptyset\}$. Then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq 75\kappa^{p-1} \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i|^p. \tag{7.3}$$

We now present a general non-uniform bound for locally dependent random fields $\{\xi_i, i \in \mathcal{J}\}$ under (LD4*).

Theorem 7.3: *Assume that $\mathbb{E}|\xi_i|^p < \infty$ for $2 < p \leq 3$ and that (LD4*) is satisfied. Let $\kappa = \max_{i \in \mathcal{J}} \max(|D_i^*|, |\{j : i \in D_j^*\}|)$. Then*

$$|\mathbb{P}(W \leq z) - \Phi(z)| \leq C\kappa^p (1 + |z|)^{-p} \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i|^p, \tag{7.4}$$

where C is an absolute constant.

The above results can immediately be applied to m-dependent random fields. Let $d \geq 1$ and Z^d denote the d-dimensional space of positive integers. The distance between two points $i = (i_1, \cdots, i_d)$ and $j = (j_1, \cdots, j_d)$ in Z^d is defined by $|i - j| = \max_{1 \leq l \leq d} |i_l - j_l|$ and the distance between two subsets A and B of Z^d is defined by $\rho(A, B) = \inf\{|i - j| : i \in A, j \in B\}$. For a given subset $\mathcal{J}$ of Z^d, a set of random variables $\{\xi_i, i \in \mathcal{J}\}$ is said to be an m-dependent random field if $\{\xi_i, i \in A\}$ and $\{\xi_j, j \in B\}$ are independent whenever $\rho(A, B) > m$, for any subsets A and B of $\mathcal{J}$. Thus, choosing

$$A_i = \{j : |j - i| \leq m\} \cap \mathcal{J}, \quad B_i = \{j : |j - i| \leq 2m\} \cap \mathcal{J},$$
$$C_i = \{j : |j - i| \leq 3m\} \cap \mathcal{J}, \quad B_i^* = \{j : |j - i| \leq 3m\} \cap \mathcal{J},$$
$$C_i^* = \{j : |j - i| \leq 4m\} \cap \mathcal{J} \quad \text{and} \quad D_i^* = \{j : |j - i| \leq 5m\} \cap \mathcal{J}$$

in Theorems 7.2 and 7.3 yields a uniform and a non-uniform bound.

Theorem 7.4: *Let $\{\xi_i, i \in \mathcal{J}\}$ be an m-dependent random fields with zero means and finite $\mathbb{E}|\xi_i|^p < \infty$ for $2 < p \leq 3$. Then*

$$\sup_z |\mathbb{P}(W \leq z) - \Phi(z)| \leq 75(10m + 1)^{(p-1)d} \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i|^p \tag{7.5}$$

and

$$|\mathbb{P}(W \leq z) - \Phi(z)| \leq C(1 + |z|)^{-p} 11^{pd} (m + 1)^{(p-1)d} \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i|^p, \tag{7.6}$$

where C is an absolute constant.

The main idea of the proof is similar to that in Sections 3 and 4, first deriving a Stein identity and then uniform and non-uniform concentration inequalities. We outline some main steps in the proof and refer to Chen & Shao (2004a) for details.

Define

$$\begin{aligned} &\widehat{K}_i(t) = \xi_i\{I(-Y_i \le t < 0) - I(0 \le t \le -Y_i)\}, \quad K_i(t) = \mathbb{E}\widehat{K}_i(t), \\ &\widehat{K}(t) \ = \textstyle\sum_{i\in\mathcal{J}} \widehat{K}_i(t), \quad K(t) = \mathbb{E}\widehat{K}(t) = \sum_{i\in\mathcal{J}} K_i(t). \end{aligned} \tag{7.7}$$

We first derive a Stein identity for W. Let f be a bounded absolutely continuous function. Then

$$\begin{aligned} &\mathbb{E}\{Wf(W)\} \\ &= \sum_{i\in\mathcal{J}} \mathbb{E}\{\xi_i(f(W) - f(W - Y_i))\} \ = \ \sum_{i\in\mathcal{J}} \mathbb{E}\Big\{\xi_i \int_{-Y_i}^{0} f'(W+t)\,dt\Big\} \\ &= \sum_{i\in\mathcal{J}} \mathbb{E}\Big\{\int_{-\infty}^{\infty} f'(W+t)\widehat{K}_i(t)\,dt\Big\} \ = \ \mathbb{E}\Big\{\int_{-\infty}^{\infty} f'(W+t)\widehat{K}(t)\,dt\Big\}, \end{aligned} \tag{7.8}$$

and hence, by the fact that $\int_{-\infty}^{\infty} K(t)\,dt = \mathbb{E}W^2 = 1$,

$$\begin{aligned} \mathbb{E}\{f'(W) - Wf(W)\} \ &= \ \mathbb{E}\int_{-\infty}^{\infty} f'(W)K(t)\,dt - \mathbb{E}\int_{-\infty}^{\infty} f'(W+t)\widehat{K}(t)\,dt \\ &= \mathbb{E}\int_{-\infty}^{\infty} (f'(W) - f'(W+t))K(t)\,dt \\ &\quad + \mathbb{E}f'(W)\int_{-\infty}^{\infty} (K(t) - \widehat{K}(t))\,dt \\ &\quad + \mathbb{E}\int_{-\infty}^{\infty} (f'(W+t) - f'(W))(K(t) - \widehat{K}(t))\,dt \\ &:= R_1 + R_2 + R_3. \end{aligned}$$

Now let $f = f_z$ be the Stein solution (2.3). Then

$$\begin{aligned} |R_1| \ &\le \ \mathbb{E}\int_{-\infty}^{\infty} (|W|+1)|tK(t)|\,dt + \Big|\mathbb{E}\int_{-\infty}^{\infty} (I_{\{W\le z\}} - I_{\{W+t\le z\}})K(t)\,dt\Big| \\ &\le \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}(|W|+1)|\xi_i|Y_i^2 \\ &\qquad + \int_{-\infty}^{\infty} \mathbb{P}(z - \max(t,0) \le W \le z - \min(t,0))K(t)\,dt \\ &:= R_{1,1} + R_{1,2}. \end{aligned}$$

Estimating $R_{1,1}$ is not so difficult, while $R_{1,2}$ can be estimated by using a concentration inequality given below.

Observe that

$$R_2 = \mathbb{E}\left\{ f'(W) \sum_{i=1}^{n} (\xi_i Y_i - \mathbb{E}(\xi_i Y_i)) \right\},$$

which can also be estimated easily. The main difficulty arises from estimating R_3. The reader may refer to Chen & Shao (2004a) for details.

To conclude this section, we give the simplest concentration inequality in the paper of Chen & Shao (2004a), and provide a detailed proof to illustrate the difficulty for dependent variables.

Proposition 7.5: *Assume (LD1). Then for any real numbers $a < b$,*

$$\mathbb{P}(a \le W \le b) \le 0.625(b-a) + 4r_1 + 4r_2, \tag{7.9}$$

where $r_1 = \sum_{i \in \mathcal{J}} \mathbb{E}|\xi_i| Y_i^2$ and $r_2 = \int_{-\infty}^{\infty} Var(\widehat{K}(t))\, dt$.

Proof: Let $\alpha = r_1$ and define

$$f(w) = \begin{cases} -(b-a+\alpha)/2 & \text{for } w \le a-\alpha \\ \frac{1}{2\alpha}(w-a+\alpha)^2 - (b-a+\alpha)/2 & \text{for } a-\alpha < w \le a \\ w - (a+b)/2 & \text{for } a < w \le b \\ -\frac{1}{2\alpha}(w-b-\alpha)^2 + (b-a+\alpha)/2 & \text{for } b < w \le b+\alpha \\ (b-a+\alpha)/2 & \text{for } w > b+\alpha. \end{cases} \tag{7.10}$$

Then f' is a continuous function given by

$$f'(w) = \begin{cases} 1, & \text{for } a \le w \le b \\ 0, & \text{for } w \le a-\alpha \text{ or } w \ge b+\alpha, \\ \text{linear}, & \text{for } a-\alpha \le w \le a \text{ or } b \le w \le b+\alpha. \end{cases}$$

Clearly $|f(w)| \le (b-a+\alpha)/2$. With this f, and with $\widehat{K}(t)$ and $K(t)$ as defined in (7.7), we have, by (7.8),

$$\begin{aligned} (b-a+\alpha)/2 \ge \mathbb{E}Wf(W) &= \mathbb{E}\int_{-\infty}^{\infty} f'(W+t)\widehat{K}(t)\, dt \\ &:= \mathbb{E}f'(W)\int_{-\infty}^{\infty} K(t)\, dt + \mathbb{E}\int_{-\infty}^{\infty} (f'(W+t) - f'(W))K(t)\, dt \\ &\quad + \mathbb{E}\int_{-\infty}^{\infty} f'(W+t)(\widehat{K}(t) - K(t))\, dt \\ &:= H_1 + H_2 + H_3. \end{aligned} \tag{7.11}$$

Clearly,

$$H_1 = \mathbb{E}f'(W) \ge \mathbb{P}(a \le W \le b). \tag{7.12}$$

By the Cauchy inequality,

$$|H_3| \le (1/8)\mathbb{E}\int_{-\infty}^{\infty}[f'(W+t)]^2\,dt + 2\mathbb{E}\int_{-\infty}^{\infty}(\widehat{K}(t)-K(t))^2\,dt$$
$$\le (b-a+2\alpha)/8 + 2r_2. \tag{7.13}$$

To bound H_2, let

$$L(\alpha) = \sup_{x\in R}\mathbb{P}(x \le W \le x+\alpha).$$

Then, by writing

$$H_2 = \mathbb{E}\int_0^{\infty}\int_0^t f''(W+s)\,ds\,K(t)\,dt - \mathbb{E}\int_{-\infty}^0\int_t^0 f''(W+s)\,ds\,K(t)\,dt$$
$$= \alpha^{-1}\int_0^{\infty}\int_0^t\{\mathbb{P}(a-\alpha \le W+s \le a) - \mathbb{P}(b \le W+s \le b+\alpha)\}ds\,K(t)\,dt$$
$$-\alpha^{-1}\int_{-\infty}^0\int_t^0\{\mathbb{P}(a-\alpha \le W+s \le a) - \mathbb{P}(b \le W+s \le b+\alpha)\}ds\,K(t)\,dt,$$

we have

$$|H_2| \le \alpha^{-1}\int_0^{\infty}\int_0^t L(\alpha)ds|K(t)|\,dt + \alpha^{-1}\int_{-\infty}^0\int_t^0 L(\alpha)ds|K(t)|\,dt$$
$$= \alpha^{-1}L(\alpha)\int_{-\infty}^{\infty}|tK(t)|\,dt \le \tfrac{1}{2}\alpha^{-1}r_1L(\alpha) = \tfrac{1}{2}L(\alpha). \tag{7.14}$$

It follows from (7.11) – (7.14) that

$$\mathbb{P}(a \le W \le b) \le 0.625(b-a) + 0.75\alpha + 2r_2 + 0.5L(\alpha). \tag{7.15}$$

Substituting $a = x$ and $b = x+\alpha$ in (7.15), we obtain

$$L(\alpha) \le 1.375\alpha + 2r_2 + 0.5L(\alpha)$$

and hence

$$L(\alpha) \le 2.75\alpha + 4r_2. \tag{7.16}$$

Finally combining (7.15) and (7.16), we obtain (7.9). ∎

8. Appendix

Here we give detailed proofs of the basic properties of the solutions to the Stein equations (2.2) and (2.4), given in Lemmas 2.2 and 2.3. The proofs of Lemma 2.2 and part of Lemma 2.3 follow Stein (1986), and part of the proof of Lemma 2.3 is due to Stroock (1993).

Proof of Lemma 2.2: Since $f_z(w) = f_{-z}(-w)$, we need only consider the case $z \geq 0$. Note that for $w > 0$

$$\int_w^\infty e^{-x^2/2}\,dx \leq \int_w^\infty \frac{x}{w} e^{-x^2/2}\,dx = \frac{e^{-w^2/2}}{w},$$

which also yields

$$(1+w^2)\int_w^\infty e^{-x^2/2}\,dx \geq we^{-w^2/2},$$

by comparing the derivatives of the two functions. Thus

$$\frac{we^{-w^2/2}}{(1+w^2)\sqrt{2\pi}} \leq 1 - \Phi(w) \leq \frac{e^{-w^2/2}}{w\sqrt{2\pi}}. \tag{8.1}$$

It follows from (2.3) that

$$\begin{aligned}(wf_z(w))' &= \begin{cases} \sqrt{2\pi}[1-\Phi(z)]\Big((1+w^2)e^{w^2/2}\Phi(w) + \frac{w}{\sqrt{2\pi}}\Big) & \text{if } w < z; \\ \sqrt{2\pi}\Phi(z)\Big((1+w^2)e^{w^2/2}(1-\Phi(w)) - \frac{w}{\sqrt{2\pi}}\Big) & \text{if } w > z, \end{cases} \\ &\geq 0,\end{aligned}$$

by (8.1). This proves (2.6).

In view of the fact that

$$\lim_{w\to-\infty} wf_z(w) = \Phi(z) - 1 \text{ and } \lim_{w\to\infty} wf_z(w) = \Phi(z), \tag{8.2}$$

(2.7) follows by (2.6).

By (2.2), we have

$$\begin{aligned} f_z'(w) &= wf_z(w) + I_{\{w\leq z\}} - \Phi(z) \\ &= \begin{cases} wf_z(w) + 1 - \Phi(z) & \text{for } w < z, \\ wf_z(w) - \Phi(z) & \text{for } w > z; \end{cases} \\ &= \begin{cases} (\sqrt{2\pi}we^{w^2/2}\Phi(w) + 1)(1-\Phi(z)) & \text{for } w < z, \\ (\sqrt{2\pi}we^{w^2/2}(1-\Phi(w)) - 1)\Phi(z) & \text{for } w > z. \end{cases} \end{aligned} \tag{8.3}$$

Since $wf_z(w)$ is an increasing function of w, by (8.1) and (8.2),

$$0 < f_z'(w) \leq zf_z(z) + 1 - \Phi(z) < 1 \quad \text{for} \quad w < z \tag{8.4}$$

and

$$-1 < zf_z(z) - \Phi(z) \leq f_z'(w) < 0 \quad \text{for} \quad w > z\,. \tag{8.5}$$

Hence, for any w and v,

$$|f_z'(w) - f_z'(v)| \leq \max(1, zf_z(z) + 1 - \Phi(z) - (zf_z(z) - \Phi(z)) = 1.$$

This proves (2.8).

Observe that, by (8.4) and (8.5), f_z attains its maximum at z. Thus

$$0 < f_z(w) \le f_z(z) = \sqrt{2\pi} e^{z^2/2} \Phi(z)(1 - \Phi(z)). \tag{8.6}$$

By (8.1), $f_z(z) \le 1/z$. To finish the proof of (2.9), let

$$g(z) = \Phi(z)(1 - \Phi(z)) - e^{-z^2/2}/4 \quad \text{and} \quad g_1(z) = \frac{1}{\sqrt{2\pi}} + \frac{z}{4} - \frac{2\Phi(z)}{\sqrt{2\pi}}.$$

Observe that $g'(z) = e^{-z^2/2} g_1(z)$ and that

$$g_1'(z) = \frac{1}{4} - \frac{1}{\pi} e^{-z^2} \begin{cases} < 0 & \text{if } 0 \le z < z_0, \\ = 0 & \text{if } z = z_0, \\ > 0 & \text{if } z > z_0, \end{cases}$$

where $z_0 = (2\ln(4/\pi))^{1/2}$. Thus, $g_1(z)$ is decreasing on $[0, z_0)$ and increasing on (z_0, ∞). Since $g_1(0) = 0$ and $g_1(\infty) = \infty$, there exists $z_1 > 0$ such that $g_1(z) < 0$ for $0 < z < z_1$ and $g_1(z) > 0$ for $z > z_1$. Therefore, $g(z)$ attains its maximum at either $z = 0$ or $z = \infty$, that is

$$g(z) \le \max(g(0), g(\infty)) = 0,$$

which is equivalent to $f_z(z) \le \sqrt{2\pi}/4$. This completes the proof of (2.9). The last inequality (2.10) is a consequence of (2.8) and (2.9) by rewriting

$$\begin{aligned} &(w+u) f_z(w+u) - (w+v) f_z(w+v) \\ &\qquad = w(f_z(w+u) - f_z(w+v)) + u f_z(w+u) - v f_z(w+v) \end{aligned}$$

and using the Taylor expansion. ∎

Proof of Lemma 2.3: We define $\tilde{h}(w) = h(w) - \mathbb{E}h(Z)$, and then put $c_0 = \sup_w |\tilde{h}(w)|$, $c_1 = \sup_w |h'(w)|$. Since $\tilde{h}$ and f_h are unchanged when h is replaced by $h - h(0)$, we may assume that $h(0) = 0$. Therefore $|h(t)| \le c_1|t|$ and $|\mathbb{E}h(Z)| \le c_1 \mathbb{E}|Z| = c_1\sqrt{2/\pi}$.

First we verify (2.11). From the definition (2.5) of f_h, it follows that

$$\begin{aligned} |f_h(w)| &\le \begin{cases} e^{w^2/2} \int_{-\infty}^{w} |\tilde{h}(x)| e^{-x^2/2}\, dx & \text{if } w \le 0, \\ e^{w^2/2} \int_{w}^{\infty} |\tilde{h}(x)| e^{-x^2/2}\, dx & \text{if } w \ge 0 \end{cases} \\ &\le e^{w^2/2} \min\left(c_0 \int_{|w|}^{\infty} e^{-x^2/2})\, dx, c_1 \int_{|w|}^{\infty} (|x| + \sqrt{2/\pi}) e^{-x^2/2}\, dx \right) \\ &\le \min(\sqrt{\pi/2}, 2c_1), \end{aligned}$$

where in the last inequality we used the fact that

$$e^{w^2/2}\int_{|w|}^{\infty} e^{-x^2/2}\,dx \le \sqrt{\pi/2}.$$

Next we prove (2.12). By (2.4), for $w \ge 0$,

$$\begin{aligned}|f_h'(w)| &\le |h(w)-\mathbb{E}h(Z)| + we^{w^2/2}\int_w^{\infty} |h(x)-\mathbb{E}h(Z)|e^{-x^2/2}\,dx\\ &\le |h(w)-\mathbb{E}h(Z)| + c_0 we^{w^2/2}\int_w^{\infty} e^{-x^2/2}\,dx \le 2c_0,\end{aligned}$$

by (8.1). It follows from (2.5) again that

$$f''(w) - wf'(w) - f(w) = h'(w),$$

or equivalently that

$$(e^{-w^2/2}f'(w))' = e^{-w^2/2}(f(w)+h'(w)).$$

Therefore

$$f'(w) = -e^{w^2/2}\int_w^{\infty}(f(x)+h'(x))e^{-x^2/2}\,dx$$

and, by (2.11),

$$|f'(w)| \le 3c_1 e^{w^2/2}\int_w^{\infty} e^{-x^2/2}\,dx \le 3c_1\sqrt{\pi/2} \le 4c_1.$$

Thus we have

$$\sup_{w\ge 0}|f'(w)| \le \min(2c_0, 4c_1).$$

Similarly, the above bound holds for $\sup_{w\le 0}|f'(w)|$. This proves (2.12).

Now we prove (2.13). Differentiating (2.4) gives

$$\begin{aligned}f_h''(w) &= wf_h'(w) + f_h(w) + h'(w)\\ &= (1+w^2)f_h(w) + w(h(w)-\mathbb{E}h(Z)) + h'(w). \qquad (8.7)\end{aligned}$$

From

$$\begin{aligned}h(x)-\mathbb{E}h(Z) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}[h(x)-h(s)]e^{-s^2/2}\,ds\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\int_s^x h'(x)\,dte^{-s^2/2}\,ds - \frac{1}{\sqrt{2\pi}}\int_x^{\infty}\int_x^s h'(t)\,dte^{-s^2/2}\,ds\\ &= \int_{-\infty}^{x} h'(t)\Phi(t)\,dt - \int_x^{\infty} h'(t)(1-\Phi(t))\,dt, \qquad (8.8)\end{aligned}$$

it follows that

$$\begin{aligned} f_h(w) &= e^{w^2/2} \int_{-\infty}^{w} [h(x) - \mathbb{E}h(Z)] e^{-x^2/2}\, dx \\ &= e^{w^2/2} \int_{-\infty}^{w} \Big(\int_{-\infty}^{x} h'(t)\Phi(t)\, dt - \int_{x}^{\infty} h'(t)(1-\Phi(t))\, dt \Big) e^{-x^2/2}\, dx \\ &= -\sqrt{2\pi} e^{w^2/2}(1-\Phi(w)) \int_{-\infty}^{w} h'(t)\Phi(t)\, dt \\ &\quad - \sqrt{2\pi} e^{w^2/2}\Phi(w) \int_{w}^{\infty} h'(t)[1-\Phi(t)]\, dt. \end{aligned} \tag{8.9}$$

From (8.7) – (8.9) and (8.1) we now obtain

$$\begin{aligned} |f_h''(w)| &\le |h'(w)| + |(1+w^2)f_h(w) + w(h(w) - \mathbb{E}h(Z))| \\ &\le |h'(w)| + \Big| \Big(w - \sqrt{2\pi}(1+w^2)e^{w^2/2}(1-\Phi(w)) \Big) \int_{-\infty}^{w} h'(t)\Phi(t)\, dt \Big| \\ &\quad + \Big| \Big(-w - \sqrt{2\pi}(1+w^2)e^{w^2/2}\Phi(w) \Big) \int_{w}^{\infty} h'(t)(1-\Phi(t))\, dt \Big| \\ &\le |h'(w)| + c_1 \Big(-w + \sqrt{2\pi}(1+w^2)e^{w^2/2}(1-\Phi(w)) \Big) \int_{-\infty}^{w} \Phi(t)\, dt \\ &\quad + c_1 \Big(w + \sqrt{2\pi}(1+w^2)e^{w^2/2}\Phi(w) \Big) \int_{w}^{\infty} (1-\Phi(t))\, dt. \end{aligned} \tag{8.10}$$

Hence it follows that

$$\begin{aligned} |f_h''(w)| &\le |h'(w)| \\ &\quad + c_1 \Big(-w + \sqrt{2\pi}(1+w^2)e^{w^2/2}(1-\Phi(w)) \Big) \Big(w\Phi(w) + \frac{e^{-w^2/2}}{\sqrt{2\pi}} \Big) \\ &\quad + c_1 \Big(w + \sqrt{2\pi}(1+w^2)e^{w^2/2}\Phi(w) \Big) \Big(-w(1-\Phi(w)) + \frac{e^{-w^2/2}}{\sqrt{2\pi}} \Big) \\ &= |h'(w)| + c_1 \le 2c_1, \end{aligned} \tag{8.11}$$

as desired. ∎

Acknowledgement: We are grateful to Andrew Barbour for his careful editing of the paper and, in particular, for substantially improving the exposition in the Introduction.

References

1. P. Baldi & Y. Rinott (1989) On normal approximation of distributions in terms of dependency graph. *Ann. Probab.* **17**, 1646–1650.
2. P. Baldi, Y. Rinott & C. Stein (1989) A normal approximation for the number of local maxima of a random function on a graph. In: *Probability, statistics, and mathematics. Papers in honor of Samuel Karlin,* Eds: T. W. Anderson, K. B. Athreya, D. L. Iglehart, pp. 59–81. Academic Press, Boston.
3. A. D. Barbour & L. H. Y. Chen (1992) On the binary expansion of a random integer. *Statist. Probab. Lett.* **14**, 235–241.
4. A. D. Barbour & P. Hall (1984) Stein's method and the Berry–Esseen theorem. *Austral. J. Statist.* **26**, 8–15.
5. P. van Beeck (1972) An application of Fourier methods to the problem of sharpening the Berry–Esseen inequality. *Z. Wahrsch. verw. Geb.* **23**, 187–196.
6. A. Bikelis (1966) Estimates of the remainder in the central limit theorem. *Litovsk. Mat. Sb.* **6**, 323–346 (in Russian).
7. E. Bolthausen (1984) An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. verw. Geb.* **66**, 379–386.
8. E. Bolthausen & F. Götze (1993) The rate of convergence for multivariate sampling statistics. *Ann. Statist.* **21**, 1692–1710.
9. L. H. Y. Chen (1986) The rate of convergence in a central limit theorem for dependent random variables with arbitrary index set. IMA Preprint Series #243, Univ. Minnesota.
10. L. H. Y. Chen (1998) Stein's method: some perspectives with applications. In: *Probability towards 2000,* Eds: L. Accardi & C. C. Heyde, Lecture Notes in Statistics **128**, pp. 97–122. Springer-Verlag , Berlin.
11. L. H. Y. Chen (2000) Non-uniform bounds in probability approximations using Stein's method. In: *Probability and statistical models with applications: a volume in honor of Theophilos Cacoullos,* Eds: N. Balakrishnan, Ch. A. Charalambides & M. V. Koutras, pp. 3–14. Chapman & Hall/CRC Press, Boca Raton, Florida.
12. L. H. Y. Chen & Q. M. Shao (2001) A non-uniform Berry–Esseen bound via Stein's method. *Probab. Theory Rel. Fields* **120**, 236–254.
13. L. H. Y. Chen & Q. M. Shao (2004a) Normal approximation under local dependence. *Ann. Probab.* **32**, 1985–2028.
14. L. H. Y. Chen & Q. M. Shao (2004b) Normal approximation for non-linear statistics using a concentration inequality approach. Preprint.
15. A. Dembo & Y. Rinott (1996) Some examples of normal approximations by Stein's method. In: *Random Discrete Structures,* Eds: D. Aldous & R. Pemantle, pp. 25–44. Springer, New York.
16. P. Diaconis (1977) The distribution of leading digits and uniform distribution mod 1. *Ann. Probab.* **5**, 72–81.
17. P. Diaconis & S. Holmes (2004) *Stein's method: expository lectures and applications.* IMS Lecture Notes, Vol. **46**, Hayward, CA.

18. P. DIACONIS, S. HOLMES & G. REINERT (2004) Use of exchangeable pairs in the analysis of simulations. In: *Stein's method: expository lectures and applications,* Eds: P. Diaconis & S. Holmes, pp. 1–26. IMS Lecture Notes Vol. **46**, Hayward, CA.
19. R. V. ERICKSON (1974) L_1 bounds for asymptotic normality of m-dependent sums using Stein's technique. *Ann. Probab.* **2**, 522–529.
20. C. G. ESSEEN (1945) Fourier analysis of distribution functions: a mathematical study of the Laplace-Gaussian law. *Acta Math.* **77**, 1–125.
21. W. FELLER (1968) On the Berry–Esseen theorem. *Z. Wahrsch. verw. Geb.* **10**, 261–268.
22. L. GOLDSTEIN (2005) Berry–Esseen bounds for combinatorial central limit theorems and pattern occurrences, using zero and size biasing. *J. Appl. Probab.* **42** (to appear).
23. L. GOLDSTEIN & G. REINERT (1997) Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* **7**, 935–952.
24. L. GOLDSTEIN & Y. RINOTT (1996) Multivariate normal approximations by Stein's method and size bias couplings. *Adv. Appl. Prob.* **33**, 1–17.
25. F. GÖTZE (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19**, 724–739.
26. S. T. HO & L. H. Y. CHEN (1978) An L_p bound for the remainder in a combinatorial central limit theorem. *Ann. Probab.* **6**, 231–249.
27. S. V. NAGAEV (1965) Some limit theorems for large deviations. *Theory Probab. Appl.* **10**, 214–235.
28. L. PADITZ (1989) On the analytical structure of the constant in the nonuniform version of the Esseen inequality. *Statistics* **20**, 453–464.
29. V. V. PETROV (1995) *Limit theorems of probability theory: sequences of independent random variables.* Oxford Studies in Probability 4, Clarendon Press, Oxford.
30. Y. RINOTT (1994) On normal approximation rates for certain sums of dependent random variables. *J. Comput. Appl. Math.* **55**, 135–143.
31. Y. RINOTT & V. ROTAR (1996) A multivariate CLT for local dependence with $n^{-1/2} \log n$ rate and applications to multivate graph related statistics. *J. Multiv. Anal.* **56**, 333–350.
32. Y. RINOTT & V. ROTAR (2000) Normal approximations by Stein's method. *Decis. Econ. Finance* **23**, 15–29.
33. H. P. ROSENTHAL (1970) On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.* **8**, 273–303.
34. C. STEIN (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **2**, 583–602. Univ. California Press, Berkeley, CA.
35. C. STEIN (1986) *Approximation computation of expectations.* IMS Lecture Notes Vol. **7**, Hayward, CA.
36. D. W. STROOCK (1993) *Probability theory: an analytic view.* Cambridge Univ. Press, Cambridge, U.K.

Stein's method for Poisson and compound Poisson approximation

Torkel Erhardsson
Department of Mathematics, KTH
S-100 44 Stockholm, Sweden
E-mail: ter@math.kth.se

We describe how Stein's method is used to obtain error estimates in Poisson and compound Poisson approximation (in terms of bounds on the total variation distance) for sums of nonnegative integer valued random variables with finite means. The most important elements are bounds on the first differences of the solutions of the corresponding Stein equations, and the construction of error estimates from the Stein equations using the local or coupling approaches. Proofs are included for the most important results, as well as examples of applications. Some related topics are also treated, notably error estimates in approximations with translated signed discrete compound Poisson measures.

Contents

1. Introduction

What is Stein's method?

To answer this, consider the following situation. Let $(S, \mathcal{S}, \mu)$ be a probability space, let χ be the set of measurable functions $h : S \to \mathbb{R}$, and let $\chi_0 \subset \chi$ be a set of μ-integrable functions; e.g., χ_0 could be (a subset of) the set of indicator functions $\{I_A; A \in \mathcal{S}\}$. We want to compute $\int_S h \, d\mu$ for all $h \in \chi_0$, but μ has such a complicated structure that exact computation is not feasible; e.g., μ could be the distribution of a sum of a large number of dependent random variables. A natural idea is then to replace μ with a simpler and better known probability measure μ_0 close to μ such that the integrals $\int_S h \, d\mu_0$ are easy to compute, and try to estimate the approximation error. The estimates might, but need not, be uniform over all $h \in \chi_0$.

Stein's method is an attempt to construct such approximations, and to estimate the corresponding errors, in a systematic way. The method was first proposed and used by Stein (1972) in the context of normal approximation. Chen, Barbour and others have shown how the method can be adapted to approximation with a number of other probability distributions, notably the Poisson and compound Poisson distributions and the distribution of a Poisson point process; the latter topic is the subject of Chapter 3, and a very general treatment for arbitrary distributions is given in Chapter 4. The main purpose of this chapter is to survey the results that have so far been obtained for the Poisson and compound Poisson distributions.

In general terms, Stein's method is described in Stein (1986), Barbour (1997) and Chen (1998). The following is a brief summary. Let $(S, \mathcal{S}, \mu)$ be a probability space, and let $\chi_0 \subset \chi$ be a set of μ-integrable functions, as above. Choose a probability measure μ_0 on $(S, \mathcal{S})$ such that all $h \in \chi_0$ are μ_0-integrable, and all $\int_S h \, d\mu_0$ are easily computed. Find a set of functions $\mathcal{F}_0$ and a mapping $T_0 : \mathcal{F}_0 \to \chi$, such that, for each $h \in \chi_0$, the

equation

$$T_0 f = h - \int_S h \, d\mu_0 \tag{1.1}$$

has a solution $f \in \mathcal{F}_0$. Then, clearly,

$$\int_S (T_0 f) d\mu = \int_S h \, d\mu - \int_S h \, d\mu_0.$$

T_0 is called a *Stein operator* for the distribution μ_0, (1.1) is called a *Stein equation*, and the solution f of (1.1) is called a *Stein transform* of h. The idea is to choose a Stein operator in such a way that good estimates of $|\int_S (T_0 f) d\mu|$ can be found.

To construct a Stein operator T_0 for μ_0, the following procedure is proposed in Stein (1986).

(1) Choose a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ containing an exchangeable pair of random variables (X, Y) (i.e., their joint distribution should be permutation invariant) with marginal distribution μ_0.
(2) Choose a mapping $\alpha : \mathcal{F}_0 \to \mathcal{F}$, where $\mathcal{F}$ is the space of measurable antisymmetric functions $F : S^2 \to \mathbb{R}$ such that $\mathbb{E}(|F(X, Y)|) < \infty$.
(3) Take $T_0 = T \circ \alpha$, where $T : \mathcal{F} \to \chi$ is defined, for some version of the conditional expectation, by

$$(TF)(x) = \mathbb{E}(F(X, Y) | X = x) \qquad \forall x \in S.$$

It is then easy to see that

$$\int_S (T_0 f) d\mu_0 = \int_S (TF) d\mu_0 = \mathbb{E}(F(X, Y)) \qquad \forall f \in \mathcal{F}_0,$$

where $F = \alpha f$. Hence, by the antisymmetry of F,

$$\int_S (T_0 f) d\mu_0 = 0 \qquad \forall f \in \mathcal{F}_0, \tag{1.2}$$

which is a necessary property for a Stein operator. (1.2) is called a *Stein identity* for the distribution μ_0.

An alternative way to construct a Stein operator T_0 for μ_0 is proposed in Chen (1998). Let $\mathcal{F}_0 \subset L^2(S, \mu_0)$. Choose a linear mapping $A : \mathcal{F}_0 \to L^2(S, \mu_0)$ so that the constant function 1 on S belongs to the domain of the adjoint A^*. Then,

$$\int_S (Af) d\mu_0 = \int_S (A^* 1) f d\mu_0 \qquad \forall f \in \mathcal{F}_0,$$

and taking $T_0 = A - (A^*1)I$, we get

$$\int_S (T_0 f) d\mu_0 = \int_S (Af) d\mu_0 - \int_S (A^*1) f d\mu_0 \qquad \forall f \in \mathcal{F}_0,$$

so (1.2) holds also in this case.

The two above procedures can be used to construct virtually all Stein operators in use today; we shall see an example in Section 2.2. However, neither of them is guaranteed to produce Stein operators for which good estimates of $|\int_S (T_0 f) d\mu|$ can be found, without additional ad hoc considerations. Whether the procedures could be strengthened into automatically producing Stein operators with good properties is still unclear.

The rest of the chapter consists of two sections, Section 2 on Poisson approximation, and Section 3 on compound Poisson approximation. The Stein equation for the Poisson distribution is treated in Sections 2.1–2.2; error estimates in Poisson approximation for sums of indicator variables, in terms of bounds on the total variation distance, in Sections 2.3–2.4; Poisson approximation for the total mass of a point process in Section 2.5; Poisson-Charlier approximation for sums of independent indicator variables, and Poisson approximation for unbounded functions of such sums, in Sections 2.6–2.7. The compound Poisson distribution is defined, and its use for approximation purposes motivated, in Sections 3.1–3.2. The Stein equation for the compound Poisson distribution is treated in Section 3.3; error estimates in compound Poisson approximation, again in terms of bounds on the total variation distance, and some improved error estimates due to Barbour and Utev, in Sections 3.4–3.5; error estimates in terms of bounds on the Kolmogorov distance in Section 3.6; translated signed discrete compound Poisson measure approximation in Section 3.7; compound Poisson approximation via Poisson process approximation in Section 3.8; and, finally, an alternative Stein equation due to Chen, in Section 3.9.

2. Poisson approximation

2.1. *The Stein equation for* $\mathrm{Po}(\lambda)$ *and its solutions*

We first consider Poisson approximation, which is very natural for the distribution of a sum of indicator (0–1) random variables, if each variable has small probability of taking the value 1 , and they are not too strongly dependent. The Stein operator (2.1) is due to Chen (1975), who also demonstrated how it could be used for error estimation in Poisson approximation; hence, Stein's method for Poisson approximation is also known as the Stein–Chen

(or Chen–Stein) method. The method was further developed in a number of papers by Barbour and others; the results of this work were collected in the monograph by Barbour, Holst & Janson (1992).

To make a connection with the general setting in Section 1, let $(S, \mathcal{S}, \mu) = (\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+}, \mu)$, where $(\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+})$ is the set of nonnegative integers equipped with the power σ-algebra, and let $\mu_0 = \mathrm{Po}(\lambda)$. Also, let $\mathcal{F}_0 = \chi$, i.e., the set of all real-valued functions on $\mathbb{Z}_+$. Define the Stein operator $T_0 : \chi \to \chi$ by

$$(T_0 f)(k) = \lambda f(k+1) - k f(k) \qquad \forall k \in \mathbb{Z}_+. \tag{2.1}$$

In Section 2, we show how T_0 can be obtained from Stein's general procedure, described in Section 1. We begin in this section by listing some important properties of T_0.

Theorem 2.1: *The Stein equation*

$$T_0 f = h - \int_{\mathbb{Z}_+} h \, d\mu_0$$

has a solution f for each μ_0-integrable $h \in \chi$. The solution f is unique except for $f(0)$, which can be chosen arbitrarily. f can be computed recursively from the Stein equation, and is explicitly given by

$$\begin{aligned} f(k) &= \frac{(k-1)!}{\lambda^k} \sum_{i=0}^{k-1} \left(h(i) - \int_{\mathbb{Z}_+} h \, d\mu_0 \right) \frac{\lambda^i}{i!} \\ &= -\frac{(k-1)!}{\lambda^k} \sum_{i=k}^{\infty} \left(h(i) - \int_{\mathbb{Z}_+} h \, d\mu_0 \right) \frac{\lambda^i}{i!}, \qquad \forall k \in \mathbb{N}. \end{aligned}$$

Also: if h is bounded, then f is bounded.

Proof: The first assertions follow immediately from the Stein equation. The explicit representations of f are easily verified by direct computation. Finally, if h is bounded, it is immediate from the second representation that f is bounded. ∎

The following characterization of $\mathrm{Po}(\lambda)$ is similar to the characterization of the standard normal distribution given in Lemma 1 in Chapter II of Stein (1986) (the celebrated "Stein's lemma").

Theorem 2.2: *A probability measure μ on $(\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+})$ is $\mathrm{Po}(\lambda)$ if and only if*

$$\int_{\mathbb{Z}_+} (T_0 f) d\mu = 0$$

for all bounded $f : \mathbb{Z}_+ \to \mathbb{R}$.

Proof: The necessity part is easily shown by direct computation. For the sufficiency part, let f_A be the unique bounded solution of the Stein equation with $h = I_A$. Integrating the Stein equation with respect to μ gives

$$\mu(A) - \mu_0(A) = \int_S (T_0 f_A) d\mu = 0 \qquad \forall A \subset \mathbb{Z}_+. \qquad \blacksquare$$

For efficient error estimation in Poisson approximation it is essential to have good bounds on the supremum norm of the first difference of the solution of the Stein equation, and (optionally) on the supremum norm of the solution itself. The following bounds are due to Barbour & Eagleson (1983). The coefficients

$$k_1(\lambda) := \left(1 \wedge \sqrt{\frac{2}{e\lambda}}\right) \quad \text{and} \quad k_2(\lambda) := \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \tag{2.2}$$

appearing on the right hand side have been nicknamed *magic factors*, but are also known as *Stein factors*. Note that they both become smaller as λ increases, and that $k_2(\lambda)$ becomes smaller much faster than $k_1(\lambda)$.

Theorem 2.3: *Let* f *be the unique bounded solution of the Stein equation for* h *bounded. Then*

$$\|f\| \leq k_1(\lambda)\big(\sup_{i \in \mathbb{Z}_+} h(i) - \inf_{i \in \mathbb{Z}_+} h(i)\big);$$

$$\|\Delta f\| \leq k_2(\lambda)\big(\sup_{i \in \mathbb{Z}_+} h(i) - \inf_{i \in \mathbb{Z}_+} h(i)\big).$$

Proof: We prove only the second bound, which is most important to us. First, consider the case $h = I_{\{k\}}$, where $k \in \mathbb{Z}_+$. Denote the bounded solution in this case by $f_{\{k\}}$. We see from the explicit expressions in Theorem 2.1 that $f_{\{k\}}(i)$ is negative and decreasing for $1 \leq i \leq k$, and positive and decreasing for $i \geq k + 1$. Hence, the only positive value taken by $\Delta f_{\{k\}}(i) = f_{\{k\}}(i+1) - f_{\{k\}}(i)$ is

$$f_{\{k\}}(k+1) - f_{\{k\}}(k) = \frac{e^{-\lambda}}{\lambda}\Big(\sum_{r=k+1}^{\infty} \frac{\lambda^r}{r!} + \frac{r}{k}\sum_{r=1}^{k} \frac{\lambda^r}{r!}\Big) \leq \frac{1 - e^{-\lambda}}{\lambda}.$$

Turning to the general case, we can replace h by $h_+ = h - \inf_{i \in \mathbb{Z}_+} h(i)$, since this does not change the right hand side of the Stein equation. By insertion into the Stein equation we can show that

$$f(i) = \sum_{k=0}^{\infty} h_+(k) f_{\{k\}}(i) \qquad \forall i \geq 1.$$

From this, and because $h_+ \geq 0$ and $\Delta f_{\{k\}}(i)$ is positive only for $i = k$, it follows that

$$\begin{aligned} f(i+1) - f(i) &= \sum_{k=0}^{\infty} h_+(k)(f_{\{k\}}(i+1) - f_{\{k\}}(i)) \\ &\leq \left(\frac{1-e^{-\lambda}}{\lambda}\right)\left(\sup_{i\in\mathbb{Z}_+} h(i) - \inf_{i\in\mathbb{Z}_+} h(i)\right), \qquad \forall i \geq 1. \end{aligned}$$

Similarly, since $-f$ is the unique bounded solution of the Stein equation with h replaced by $h_- = \sup_{i\in\mathbb{Z}_+} h(i) - h$,

$$\begin{aligned} -\big(f(i+1) - f(i)\big) &= \sum_{k=0}^{\infty} h_-(k)(f_{\{k\}}(i+1) - f_{\{k\}}(i)) \\ &\leq \left(\frac{1-e^{-\lambda}}{\lambda}\right)\left(\sup_{i\in\mathbb{Z}_+} h(i) - \inf_{i\in\mathbb{Z}_+} h(i)\right), \qquad \forall i \geq 1. \end{aligned}$$

This gives the second bound. For the first, Barbour & Eagleson (1983) prove, using the explicit representations in Theorem 2.1 and Stirling's formula, that

$$\|f\| \leq \left(1 \wedge \frac{1.4}{\sqrt{\lambda}}\right)\left(\sup_{i\in\mathbb{Z}_+} h(i) - \inf_{i\in\mathbb{Z}_+} h(i)\right).$$

In Remark 10.2.4 in Barbour, Holst & Janson (1992), a completely different argument, given here in Chapter 3, Proposition 5.7, is used to prove that the constant 1.4 can be replaced by $\sqrt{2/e}$. ∎

2.2. *The generator interpretation*

We here demonstrate how the Stein operator (2.1) can be obtained through the general procedure of Stein described in Section 1. We use the approach of Barbour (1988), which is of particular importance, since it has proved rather easy to generalize to other approximating distributions; see the discussion in Chapter 4, Section 2 for more detail. It also has the advantage of giving, as a by-product, a probabilistic representation of the solution of the Stein equation.

Let $\{Z_t; t \in \mathbb{R}_+\}$ be a stationary immigration-death process on $\mathbb{Z}_+$, with immigration intensity λ and death intensity $\delta_i = i$ for each $i \in \mathbb{Z}_+$. It is well-known that this process is reversible, and it is easy to prove that the stationary distribution is $\mu_0 = \mathrm{Po}(\lambda)$. It follows that (Z_0, Z_u) is an exchangeable pair with marginal distribution μ_0.

Choose the mapping $\alpha : \chi \to \mathcal{F}$ defined (at least for functions g which do not grow too fast) by

$$(\alpha g)(k,l) = g(k) - g(l) \qquad \forall (k,l) \in \mathbb{Z}_+^2,$$

and the mapping $T : \mathcal{F} \to \chi$ defined by

$$(TF)(k) = \mathbb{E}(F(Z_0, Z_u)|Z_0 = k) \qquad \forall k \in \mathbb{Z}_+.$$

Then

$$\lim_{u \downarrow 0} \frac{1}{u}(T \circ \alpha g)(k) = \lim_{u \downarrow 0} \frac{1}{u}\mathbb{E}(g(Z_u) - g(k)|Z_0 = k) = (\mathcal{A}g)(k)$$
$$= \lambda g(k+1) + kg(k-1) - (\lambda + k)g(k) = (T_0 f)(k),$$

where $f(k) = \nabla g(k) = g(k) - g(k-1)$, and the operator $\mathcal{A}$ is the generator of $\{Z_t; t \in \mathbb{R}_+\}$. Thus, we have shown that the Stein operator T_0 may be interpreted as the generator $\mathcal{A}$ of a Markov process with stationary distribution μ_0, or more precisely that $\mathcal{A} = T_0 \circ \nabla$. This has proved fruitful as a way of constructing Stein operators for other distributions; see, in particular, the discussion in Chapter 3, Section 5.1, for the procedure in the context of Poisson process approximation.

Theorem 2.4: *If $h \in \chi$ is bounded, then the* Poisson's equation *corresponding to $\mathcal{A}$, i.e.,*

$$-\mathcal{A}g = h - \int_{\mathbb{Z}_+} h \, d\mu_0, \tag{2.3}$$

has the solution

$$g(k) = \int_0^\infty \left(\mathbb{E}(h(Z_t)|Z_0 = k) - \int_{\mathbb{Z}_+} h \, d\mu_0 \right) dt \qquad \forall k \in \mathbb{Z}_+.$$

Moreover, $f = -\nabla g$ is the unique bounded solution of the corresponding Stein equation.

Proof: To show that g is well defined, we use a coupling. First, we let $\{Z_t^{(0)}; t \in \mathbb{R}_+\}$ be an immigration-death process on $\mathbb{Z}_+$, with immigration intensity λ and death intensity $\delta_i = i$ for each $i \in \mathbb{Z}_+$, such that $Z^{(0)} = 0$. We then let $\{D_t; t \in \mathbb{R}_+\}$ and $\{\widetilde{D}_t; t \in \mathbb{R}_+\}$ be pure death processes on $\mathbb{Z}_+$, independent of each other and of $Z^{(0)}$, with death intensity $\delta_i = i$ for each $i \in \mathbb{Z}_+$, such that $D_0 = k$ and $\mathcal{L}(\widetilde{D}_0) = \mu_0$. We also set

$\tau_k = \inf\{u \geq 0; D_u = \widetilde{D}_u = 0\}$, and note that

$$\mathbb{E}(\tau_k) = \mathbb{E}\Big(\sum_{i=0}^{k+\widetilde{D}_0-1} \frac{1}{k+\widetilde{D}_0-i}\Big)$$
$$\leq \mathbb{E}\big(1+\log(k+\widetilde{D}_0+1)\big) = \log^+ k + C, \qquad \forall k \in \mathbb{Z}_+, \quad (2.4)$$

where $\log^+ k = \max(\log k, 0)$, and $C < \infty$. Then

$$\Big|\int_{t_1}^{t_2} \Big(\mathbb{E}(h(Z_t)|Z_0 = k) - \int_{\mathbb{Z}_+} h\, d\mu_0\Big)\, dt\Big|$$
$$\leq \int_{t_1}^{t_2} \big|\mathbb{E}(h(Z_t^{(0)} + D_t)) - \mathbb{E}(h(Z_t^{(0)} + \widetilde{D}_t))\big|\, dt$$
$$\leq 2\|h\| \int_{t_1}^{t_2} \mathbb{P}(\tau_k > t)\, dt, \qquad \forall\, 0 < t_1 < t_2 < \infty.$$

Hence, g is well defined (and grows at most logarithmically). Furthermore, using Fubini's theorem,

$$\mathbb{E}(g(Z_u)|Z_0 = k) = \mathbb{E}\left(\int_0^\infty \Big(\mathbb{E}(h(Z_{t+u})|Z_u) - \int_{\mathbb{Z}_+} h\, d\mu_0\Big)\, dt \,\Big|\, Z_0 = k\right)$$
$$= \int_0^\infty \left(\mathbb{E}(h(Z_{t+u})|Z_0 = k) - \int_{\mathbb{Z}_+} h\, d\mu_0\right) dt$$
$$= \int_u^\infty \left(\mathbb{E}(h(Z_t)|Z_0 = k) - \int_{\mathbb{Z}_+} h\, d\mu_0\right) dt,$$

implying that, using bounded convergence,

$$-\lim_{u \downarrow 0} \frac{1}{u} \mathbb{E}(g(Z_u) - g(k)|Z_0 = k)$$
$$= \lim_{u \downarrow 0} \frac{1}{u} \int_0^u \big(\mathbb{E}(h(Z_t)|Z_0 = k) - \int_{\mathbb{Z}_+} h\, d\mu_0\big) dt \;=\; h(k) - \int_{\mathbb{Z}_+} h\, d\mu_0.$$

Finally, using another coupling similar to the one above we can show that $f = -\nabla g$ is bounded, so f is the unique bounded solution of the Stein equation. ∎

We shall not use this very nice probabilistic representation of the solution of the Stein equation. However, an analogous result holds (under some additional conditions) in the compound Poisson case, and this result is very valuable, since it can be used to obtain magic factors similar to those in Theorem 2.3.

2.3. *Error estimates in Poisson approximation*

We now turn to the actual construction of error estimates in Poisson approximation. We shall focus on the case when the distribution to be approximated, μ, is the distribution of a sum of indicator variables.

For the rest of the section, unless explicitly stated otherwise, we use the following notation. Γ is the index set, which is finite. In most cases $\Gamma = \{1, \ldots, n\}$. $\{X_i; i \in \Gamma\}$ are indicator variables; $p_i = \mathbb{E}(X_i)$ for each $i \in \Gamma$; $W = \sum_{i \in \Gamma} X_i$; $\lambda = \mathbb{E}(W)$; $\mu = \mathcal{L}(W)$; and $\mu_0 = \mathrm{Po}(\lambda)$.

Our goal will be to bound the total variation distance between μ and μ_0, defined by

$$d_{TV}(\mu, \mu_0) = \sup_{A \subset \mathbb{Z}_+} |\mu(A) - \mu_0(A)|.$$

The total variation distance is a metric on the space of probability measures on $(\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+})$. Obviously, a bound for this quantity gives us an error estimate in a strong sense. Our starting point is the Stein equation, which gives, for each $A \subset \mathbb{Z}_+$,

$$\mu(A) - \mu_0(A) = \int_{\mathbb{Z}_+} (T_0 f_A) d\mu = \mathbb{E}(\lambda f_A(W+1) - W f_A(W)),$$

where f_A is the solution of the Stein equation with $h = I_A$. From this we get

$$d_{TV}(\mu, \mu_0) = \sup_{A \subset \mathbb{Z}_+} |\mathbb{E}(\lambda f_A(W+1) - W f_A(W))|. \tag{2.5}$$

To bound the right hand side in (2.5), a local and a coupling approach have been suggested. The first one was used by Chen (1975), and is convenient when the dependence structure of the indicator variables is local (meaning that each indicator is independent of "most" of the others).

Theorem 2.5: (The local approach). *Let $W = \sum_{i \in \Gamma} X_i$, where $\{X_i; i \in \Gamma\}$ are indicator variables. For each $i \in \Gamma$, divide $\Gamma \setminus \{i\}$ into two subsets Γ_i^s and Γ_i^w, so that, informally,*

$$\Gamma_i^s = \{j \in \Gamma \setminus \{i\}; X_j \text{ "strongly" dependent on } X_i\}.$$

Let $Z_i = \sum_{j \in \Gamma_i^s} X_j$ and $W_i = \sum_{j \in \Gamma_i^w} X_j$. Then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq k_2(\lambda) \sum_{i \in \Gamma} \big(p_i \mathbb{E}(X_i + Z_i) + \mathbb{E}(X_i Z_i)\big) + k_1(\lambda) \sum_{i \in \Gamma} \mathbb{E}\big|p_i - \mathbb{E}(X_i | W_i)\big|,$$

where $k_1(\cdot)$ and $k_2(\cdot)$ are as defined in (2.2).

Proof: We shall bound the right hand side in (2.5). For each $A \subset \mathbb{Z}_+$, we have

$$\begin{aligned}\mathbb{E}(\lambda f_A(W+1) - W f_A(W)) &= \sum_{i\in\Gamma} \mathbb{E}(p_i f_A(W+1) - X_i f_A(W)) \\ &= \sum_{i\in\Gamma} \mathbb{E}\{p_i f_A(W+1) - p_i f_A(W_i+1) + p_i f_A(W_i+1) \\ &\qquad - X_i f_A(W_i+1) + X_i f_A(W_i+1) - X_i f_A(W)\}.\end{aligned}$$

Since, for each $i \in \Gamma$, we have the elementary bounds

$$\begin{aligned}\big|f_A(W+1) - f_A(W_i+1)\big| &\le \|\Delta f_A\|(X_i + Z_i); \\ \big|X_i f_A(W_i+1) - X_i f_A(W)\big| &\le \|\Delta f_A\| X_i Z_i; \\ \big|\mathbb{E}\big(p_i f_A(W_i+1) - X_i f_A(W_i+1)\big)\big| &\le \|f_A\| \mathbb{E}\big|p_i - \mathbb{E}(X_i|W_i)\big|,\end{aligned}$$

it follows that

$$\begin{aligned}&|\mathbb{E}(\lambda f_A(W+1) - W f_A(W))| \\ &\le \|\Delta f_A\| \sum_{i\in\Gamma} \big(p_i \mathbb{E}(X_i + Z_i) + \mathbb{E}(X_i Z_i)\big) + \|f_A\| \sum_{i\in\Gamma} \mathbb{E}\big|p_i - \mathbb{E}(X_i|W_i)\big|,\end{aligned}$$

and the result follows from Theorem 2.3 and (2.5), since $A \subset \mathbb{Z}_+$ is arbitrary. ∎

We give a couple of examples of applications of Theorem 2.5, beginning with the independent case.

Example 2.6: (Independent indicators). Let $\{X_i; i \in \Gamma\}$ be independent. Choosing $\Gamma_i^s = \emptyset$ for each $i \in \Gamma$ in Theorem 2.5 gives

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le \frac{1 - e^{-\lambda}}{\lambda} \sum_{i\in\Gamma} p_i^2. \tag{2.6}$$

In the independent case it is possible to derive total variation distance bounds using other means than Stein's method, so a comparison can be made. Le Cam (1960) used Fourier methods to prove that

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le 4.5 \max_{i\in\Gamma} p_i,$$

and that, if $\max_{i\in\Gamma} p_i \le \frac{1}{4}$,

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le \Big(1 \wedge \frac{8}{\lambda}\Big) \sum_{i\in\Gamma} p_i^2.$$

The constant 8 in the second bound was improved to 1.05 by Kerstan (1964) and to 0.71 by Daley & Vere-Jones (1988). Note that (2.6) is smaller than the first of these bounds, and does not require the condition $\max_{i\in\Gamma} p_i \leq \frac{1}{4}$ to hold.

Moreover, (2.6) is in fact the right order of the total variation distance, not just an upper bound. We know this since Barbour & Hall (1984) proved the lower bound

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) \geq \frac{1}{32}\Big(1 \wedge \frac{1}{\lambda}\Big) \sum_{i\in\Gamma} p_i^2.$$

If the indicator variables are dependent, it seems very difficult to derive total variation distance bounds by any other means than Stein's method. Other methods appear to break down, so this is where Stein's method shows its real strength. The following example is taken from Arratia, Goldstein & Gordon (1989).

Example 2.7: (Classical birthday problem). n balls (people) are thrown independently into d equiprobable boxes (days of the year). Let W be the number of pairs of balls that go into the same box. Then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) \leq \frac{8\lambda(1-e^{-\lambda})}{n-1},$$

where $\lambda = \mathbb{E}(W) = \binom{n}{2} d^{-1}$.

Proof: Take Γ to be the set of all 2-subsets of $\{1,\ldots,n\}$, so that we have $\Gamma = \{i \subset \{1,\ldots,n\}; |i| = 2\}$. Let X_i, where $i = \{i_1, i_2\}$, be the indicator of the event "the balls i_1 and i_2 go into the same box". Clearly $W = \sum_{i\in\Gamma} X_i$, and $\{X_i; i \in \Gamma\}$ are *dissociated*, meaning that for any two subsets $\Gamma_1 \subset \Gamma$ and $\Gamma_2 \subset \Gamma$ such that $(\cup_{i\in\Gamma_1} i) \cap (\cup_{i\in\Gamma_2} i) = \emptyset$, the collections of random variables $\{X_i; i \in \Gamma_1\}$ and $\{X_i; i \in \Gamma_2\}$ are independent. We now choose $\Gamma_i^s = \{j \in \Gamma \setminus \{i\}; i \cap j \neq \emptyset\}$, so that the last term in the bound of Theorem 2.5 vanishes. Since also $\mathbb{E}(X_i) = d^{-1}$ for all $i \in \Gamma$, and $\mathbb{E}(X_i X_j) = \mathbb{E}(X_i)^2 = d^{-2}$ for all $i \neq j$, it follows that

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) &\leq k_2(\lambda) \sum_{i\in\Gamma} \big(p_i \mathbb{E}(X_i + Z_i) + \mathbb{E}(X_i Z_i)\big) \\ &= k_2(\lambda) \binom{n}{2} \left(\frac{2(n-1)+1}{d^2} + \frac{2(n-1)}{d^2} \right) \\ &= k_2(\lambda) \binom{n}{2} \frac{1}{d^2} (4n-3) \;\leq\; \frac{8\lambda(1-e^{-\lambda})}{n-1}, \end{aligned}$$

where $\lambda = \mathbb{E}(W) = \binom{n}{2} d^{-1}$. ∎

We next consider the coupling approach to bounding the right hand side in (2.5), which often works well even if the dependence structure of the indicators is non-local. The idea of combining the Stein equation with couplings was first generally stated in Stein (1986), Chapter VIII, page 92, though it appears earlier in particular contexts.

Theorem 2.8: (The coupling approach). *Let* $W = \sum_{i \in \Gamma} X_i$, *where the* $\{X_i; i \in \Gamma\}$ *are indicator variables. For each* $i \in \Gamma$, *divide* $\Gamma \setminus \{i\}$ *into two subsets* Γ_i^s *and* Γ_i^w. *Let* $Z_i = \sum_{j \in \Gamma_i^s} X_j$ *and* $W_i = \sum_{j \in \Gamma_i^w} X_j$. *Let two random variables* $\widetilde{W}_i^1$ *and* W_i^1 *such that*

$$\mathcal{L}(\widetilde{W}_i^1) = \mathcal{L}(W_i | X_i = 1) \quad \text{and} \quad \mathcal{L}(W_i^1) = \mathcal{L}(W_i)$$

be defined on the same probability space. Then

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq k_2(\lambda) \sum_{i \in \Gamma} \big(p_i \mathbb{E}(X_i + Z_i) + \mathbb{E}(X_i Z_i)\big) \\ + k_2(\lambda) \sum_{i \in \Gamma} p_i \mathbb{E}\big|W_i^1 - \widetilde{W}_i^1\big|, \end{aligned}$$

where $k_2(\cdot)$ *is as defined in (2.2).*

Proof: The proof is the same as for Theorem 2.5, except that the quantity $\mathbb{E}\big(p_i f_A(W_i + 1) - X_i f_A(W_i + 1)\big)$ is bounded differently. We now make use of the couplings, to write, for each $i \in \Gamma$,

$$\begin{aligned} &\mathbb{E}\big(p_i f_A(W_i + 1) - X_i f_A(W_i + 1)\big) \\ &= p_i\big(\mathbb{E}(f_A(W_i + 1)) - \mathbb{E}(f_A(W_i + 1) | X_i = 1)\big) \\ &= p_i \mathbb{E}\big(f_A(W_i^1 + 1) - f_A(\widetilde{W}_i^1 + 1)\big), \end{aligned}$$

implying that

$$\big|\mathbb{E}\big(p_i f_A(W_i + 1) - X_i f_A(W_i + 1)\big)\big| \leq \|\Delta f_A\| p_i \mathbb{E}\big|W_i^1 - \widetilde{W}_i^1\big|. \quad ∎$$

In the case when the indicators $\{X_i; i \in \Gamma\}$ are independent, by choosing $\Gamma_i^s = \emptyset$ and $\widetilde{W}_i^1 = W_i^1 = W_i$ in Theorem 2.8, we again get the bound (2.6). The following example, where the indicators are dependent, is taken from Barbour & Holst (1989).

Example 2.9: (Classical occupancy problem). r balls are thrown independently into n equiprobable boxes. Let W be the number of empty boxes.

Then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq \left(1 - \exp\left\{-n\left(\frac{n-1}{n}\right)^r\right\}\right)\left(n\left(\frac{n-1}{n}\right)^r - (n-1)\left(\frac{n-2}{n-1}\right)^r\right).$$

If $r = na_n$ with $\lim_{n\to\infty} a_n = \infty$, then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) = O(a_n e^{-a_n}) \quad \text{as } n \to \infty.$$

If $a_n = \log n - \log c$, then $\lim_{n\to\infty} \lambda = c$.

Proof: For each $i \in \Gamma = \{1, \ldots, n\}$, let X_i be the indicator for the event "the ith box is empty", so $W = \sum_{i\in\Gamma} X_i$. Let $\Gamma_i^s = \emptyset$ and $\Gamma_i^w = \Gamma \setminus \{i\}$ in Theorem 2.8. Define $\{\widetilde{X}_{i,j}^1; j \in \Gamma_i^w\}$ in the following way. Take those balls which have landed in the ith box, throw them independently into other boxes, and let $\widetilde{X}_{i,j}^1$ be the indicator for the event "the jth box is empty after this". Then

$$\mathcal{L}(\widetilde{X}_{i,j}^1; j \in \Gamma_i^w) = \mathcal{L}(X_j; j \in \Gamma_i^w | X_i = 1),$$

since, for each ball, the probability of ending up in a particular box is given by $\frac{1}{n} + \frac{1}{n(n-1)} = \frac{1}{n-1}$, implying that, for each $\Gamma' \subset \Gamma_i^w$,

$$\mathbb{P}(\widetilde{X}_{i,j}^1 = 1 \ \forall j \in \Gamma') = \left(\frac{n - |\Gamma'| - 1}{n-1}\right)^r = \mathbb{P}(X_j = 1 \ \forall j \in \Gamma' | X_i = 1).$$

Let $\widetilde{W}_i^1 = \sum_{j\in\Gamma_i^w} \widetilde{X}_{i,j}^1$ and $W_i^1 = W_i$. Observing that $\widetilde{X}_{i,j}^1 \leq X_j$ for each index $j \in \Gamma_i^w$, we get

$$\begin{aligned}
d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) &\leq k_2(\lambda) \sum_{i\in\Gamma} p_i \left(p_i + \mathbb{E}\left|W_i^1 - \widetilde{W}_i^1\right|\right) \\
&= k_2(\lambda) \sum_{i\in\Gamma} p_i \mathbb{E}\left(X_i + \left|\sum_{j\in\Gamma_i^w} (X_j - \widetilde{X}_{i,j}^1)\right|\right) \\
&= k_2(\lambda) \sum_{i\in\Gamma} p_i \mathbb{E}\left(W - \sum_{j\in\Gamma_i^w} \widetilde{X}_{i,j}^1\right) \\
&= k_2(\lambda)\left(\mathbb{E}(W)^2 - \sum_{i\in\Gamma}\sum_{j\in\Gamma_i^w} \mathbb{E}(X_i X_j)\right) \\
&= k_2(\lambda)\left(\mathbb{E}(W)^2 - \mathbb{E}(W^2) + \mathbb{E}(W)\right) \\
&= \left(1 - \exp\left\{-\left(\frac{n-1}{n}\right)^r\right\}\right)\left(n\left(\frac{n-1}{n}\right)^r - (n-1)\left(\frac{n-2}{n-1}\right)^r\right). \qquad \blacksquare
\end{aligned}$$

We finally mention, without proof, a slight extension of the coupling approach, which is sometimes useful.

Theorem 2.10: (The detailed coupling approach). *Let* $W = \sum_{i\in\Gamma} X_i$, *where* $\{X_i; i \in \Gamma\}$ *are indicator variables. For each* $i \in \Gamma$, *divide* $\Gamma\setminus\{i\}$ *into two subsets* Γ_i^s *and* Γ_i^w. *Let* $Z_i = \sum_{j\in\Gamma_i^s} X_j$ *and* $W_i = \sum_{j\in\Gamma_i^w} X_j$. *Let a random variable* σ_i *be defined on the same probability space as* X_i *and* W_i, *and let, for each* $x \in \mathbb{R}$, *two random variables* $\widetilde{W}_i^{1,x}$ *and* $W_i^{1,x}$ *such that*

$$\mathcal{L}(\widetilde{W}_i^{1,x}) = \mathcal{L}(W_i|X_i = 1, \sigma_i = x) \quad \text{and} \quad \mathcal{L}(W_i^{1,x}) = \mathcal{L}(W_i)$$

be defined on the same probability space. Then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) \le k_2(\lambda)\sum_{i\in\Gamma}\big(p_i\mathbb{E}(X_i + Z_i) + \mathbb{E}(X_iZ_i)\big) + k_2(\lambda)\sum_{i\in\Gamma}\mathbb{E}\Big(X_i\mathbb{E}\big|W_i^{1,x} - \widetilde{W}_i^{1,x}\big|_{x=\sigma_i}\Big),$$

where $k_2(\cdot)$ *is as defined in (2.2).*

2.4. *Monotone couplings*

In Example 2.9 the task of explicitly computing the total variation distance bound was simplified by the fact that $\widetilde{X}_{i,j}^1 \le X_j$ for each $i \in \Gamma$ and for each $j \in \Gamma_i^w$. Couplings with this property are called monotone couplings. They were introduced in Barbour & Holst (1989) and further developed in Barbour, Holst & Janson (1992). We give the most important results here.

Theorem 2.11: *Let* $W = \sum_{i\in\Gamma} X_i$, *where* $\{X_i; i \in \Gamma\}$ *are indicator variables. For each* $i \in \Gamma$, *let two collections of random variables* $\{\widetilde{X}_{i,j}^1; j \ne i\}$ *and* $\{X_{i,j}^1; j \ne i\}$ *such that*

$$\mathcal{L}(\widetilde{X}_{i,j}^1; j \ne i) = \mathcal{L}(X_j; j \ne i|X_i = 1);$$
$$\mathcal{L}(X_{i,j}^1; j \ne i) = \mathcal{L}(X_j; j \ne i),$$

be defined on the same probability space. If $\Gamma\setminus\{i\}$ *can be divided into three subsets* Γ_i^+, Γ_i^-, *and* Γ_i^0, *such that* $\widetilde{X}_{i,j}^1 \ge X_j$ *for* $j \in \Gamma_i^+$ *and* $\widetilde{X}_{i,j}^1 \le X_j$ *for* $j \in \Gamma_i^-$, *then*

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) \le k_2(\lambda)\left(\sum_{i\in\Gamma} p_i^2 + \sum_{i\in\Gamma}\sum_{j\in\Gamma_i^+}\mathrm{Cov}(X_i, X_j) + \sum_{i\in\Gamma}\sum_{j\in\Gamma_i^-}|\mathrm{Cov}(X_i, X_j)| + \sum_{i\in\Gamma}\sum_{j\in\Gamma_i^0}(\mathbb{E}(X_iX_j) + p_ip_j)\right),$$

where $k_2(\cdot)$ *is as defined in (2.2).*

Proof: For each $i \in \Gamma$, let $\Gamma_i^s = \emptyset$ and $\Gamma_i^w = \Gamma \setminus \{i\}$ in Theorem 2.8, and let $\widetilde{W}_i^1 = \sum_{j \neq i} \widetilde{X}_{i,j}^1$. Then

$$p_i \mathbb{E}\left|W_i^1 - \widetilde{W}_i^1\right| = p_i \mathbb{E}\left|\sum_{j \in \Gamma_i^w} (X_j - \widetilde{X}_{i,j}^1)\right|$$

$$\leq p_i \mathbb{E}\left(\sum_{j \in \Gamma_i^+} (\widetilde{X}_{i,j}^1 - X_j)\right) + p_i \mathbb{E}\left(\sum_{j \in \Gamma_i^-} (X_j - \widetilde{X}_{i,j}^1)\right)$$

$$+ p_i \mathbb{E}\left(\sum_{j \in \Gamma_i^0} (\widetilde{X}_{i,j}^1 + X_j)\right)$$

$$= \sum_{j \in \Gamma_i^+} \mathrm{Cov}(X_i, X_j) + \sum_{j \in \Gamma_i^-} |\mathrm{Cov}(X_i, X_j)| + \sum_{j \in \Gamma_i^0} (\mathbb{E}(X_i X_j) + p_i p_j). \qquad \blacksquare$$

Definition 2.12: The indicator variables $\{X_i; i \in \Gamma\}$ are called *positively related* if the conditions of Theorem 2.11 hold with $\Gamma_i^- = \Gamma_i^0 = \emptyset$, and *negatively related* if they hold with $\Gamma_i^+ = \Gamma_i^0 = \emptyset$.

It follows immediately from Theorem 2.11 that if $\{X_i; i \in \Gamma\}$ are positively related, then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq k_2(\lambda)\big(\mathrm{Var}(W) - \lambda + 2\sum_{i \in \Gamma} p_i^2\big),$$

while if they are negatively related, then

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq k_2(\lambda)\big(\lambda - \mathrm{Var}(W)\big).$$

In view of these simple formulæ, it is natural to ask whether one could find sufficient conditions for a collection of indicator variables to be positively or negatively related, which are reasonably often satisfied and not too difficult to verify. The answer is yes, and the key is the following result.

Theorem 2.13: *The indicator variables $\{X_i; i \in \Gamma\}$ are positively (negatively) related if and only if, for each $i \in \Gamma$ and each increasing function $\phi : \{0,1\}^{n-1} \to \{0,1\}$,*

$$\mathbb{E}(\phi(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) | X_i = 1)$$
$$\geq (\leq)\ \mathbb{E}(\phi(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)).$$

Proof: The necessity part is immediate. For the sufficiency part we use Strassen's theorem, which says the following. Let S be a partially ordered

topological space (the topology should satisfy a couple of technical conditions), and let X and Y be random variables taking values in S with distributions ν_1 and ν_2 respectively. Then X and Y can be constructed on the same probability space in such a fashion that $\mathbb{P}(X \geq Y) = 1$ if and only if $\mathbb{E}(\phi(X)) \geq \mathbb{E}(\phi(Y))$ for each increasing function $\phi : S \to \{0,1\}$. For a proof of Strassen's theorem, see e.g. Liggett (1985), page 72. Here, we simply take $S = \{0,1\}^{n-1}$. ∎

Using Theorem 2.13 it is possible to connect positive relatedness to the better known property of *association.* Introduced in Esary, Proschan & Walkup (1967), association has found uses in statistical mechanics, reliability theory, and many other areas.

Definition 2.14: The random variables $\{X_i; i \in \Gamma\}$ are said to be associated if they satisfy the FKG inequality: if f and g are bounded increasing functions, then

$$\mathbb{E}(f(X_i; i \in \Gamma)g(X_i; i \in \Gamma)) \geq \mathbb{E}(f(X_i; i \in \Gamma))\mathbb{E}(g(X_i; i \in \Gamma)).$$

Theorem 2.15: *Associated indicator variables are positively related.*

Proof: For each $i \in \Gamma$ and each increasing function $\phi : \{0,1\}^{n-1} \to \{0,1\}$, use the FKG inequality with

$$f(X_i; i \in \Gamma) = \phi(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$$

and $g(X_i; i \in \Gamma) = X_i$. ∎

The proofs of the next two theorems are omitted. A proof of Theorem 2.16 can be found in Liggett (1985), page 78, while Theorem 2.17 is an easy consequence of the definition of association. However, we give an example of how they can be applied, taken from Barbour, Holst & Janson (1992).

Theorem 2.16: *Independent random variables are associated.*

Theorem 2.17: *Increasing functions of associated random variables are associated.*

Example 2.18: (Extremes of moving average processes). Let $\{Z_i; i \in \mathbb{Z}\}$ be i.i.d., and let $\eta_i = \sum_{k=0}^{q} c_k Z_{i-k}$, where $c_k \geq 0$ for each $k = 0, \ldots, q$. Let $X_i = I\{\eta_i > a\}$, and let $W = \sum_{i=1}^{n} X_i$. From Theorems 2.16–2.17, $\{X_i; i = 1, \ldots, n\}$ are associated, so

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq k_2(\lambda)\Big(\mathrm{Var}(W) - \lambda + 2\sum_{i=1}^{n} p_i^2\Big).$$

It is easy to see that

$$\frac{\mathrm{Var}(W)}{\lambda} - 1 = -\frac{\lambda}{n} + 2\sum_{i=1}^{n} \frac{n-i}{n}\big(\mathbb{P}(\eta_i > a | \eta_0 > a) - \mathbb{P}(\eta_i > a)\big).$$

In the special case $Z_i \sim U(0,1)$ and $\eta_i = Z_i + Z_{i-1}$, let $a = 2 - \sqrt{2\lambda/n}$ (where $n \geq 2\lambda$). Then, $p_i = \lambda/n$, and

$$\frac{\mathrm{Var}(W)}{\lambda} - 1 = -\frac{\lambda}{n} + \frac{2(n-1)}{n}\Big(\frac{\sqrt{8}}{3}\sqrt{\frac{\lambda}{n}} - \frac{\lambda}{n}\Big),$$

implying that

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq (1 - e^{-\lambda})\Big\{\frac{\lambda}{n} + \frac{2(n-1)}{n}\Big(\frac{\sqrt{8}}{3}\sqrt{\frac{\lambda}{n}} - \frac{\lambda}{n}\Big)\Big\}.$$

In a very similar fashion, negative relatedness can be connected to the property of *negative association*, introduced in Joag-Dev & Proschan (1983).

Definition 2.19: The random variables $\{X_i; i \in \Gamma\}$ are said to be negatively associated if, whenever f and g are bounded increasing functions and Γ_1 and Γ_2 are disjoint subsets of Γ,

$$\mathbb{E}(f(X_i; i \in \Gamma_1) g(X_i; i \in \Gamma_2)) \leq \mathbb{E}(f(X_i; i \in \Gamma_1))\mathbb{E}(g(X_i; i \in \Gamma_2)).$$

Theorem 2.20: *Negatively associated indicator variables are negatively related.*

Proof: Analogous to Theorem 2.15. ∎

The following theorems, giving sufficient conditions for a collection of random variables to be negatively associated, are stated and proved, together with others, in Joag-Dev & Proschan (1983).

Theorem 2.21: *Independent random variables are negatively associated.*

Theorem 2.22: *Random variables $X_1, \ldots, X_n$ generated by a uniformly distributed random permutation of a sequence of real numbers $a_1, \ldots, a_n$ are negatively associated.*

Theorem 2.23: *Let $\{X_i; i \in \Gamma\}$ be independent with log concave densities, and let $I \subset \mathbb{R}$ be an interval such that $\mathbb{P}(\sum_{i \in \Gamma} X_i \in I) > 0$. Random variables with joint distribution $\mathcal{L}(X_i; i \in \Gamma | \sum_{i \in \Gamma} X_i \in I)$ are negatively associated.*

Theorem 2.24: *Let $\{X_i; i \in \Gamma\}$ be negatively associated. For each $i = 1, \ldots, n$, let $Y_i = g_i(X_j; j \in \Gamma_i)$, where $\Gamma_1, \ldots, \Gamma_n$ are disjoint subsets of Γ, and the functions $g_1, \ldots, g_n$ are all increasing (decreasing). Then $Y_1, \ldots, Y_n$ are negatively associated.*

2.5. *The total mass of a point process*

In Section 2.3, Poisson approximation for the random variable $W = \sum_{i \in \Gamma} X_i$ is considered, where $\{X_i; i \in \Gamma\}$ are indicator variables, and the index set Γ is finite. Clearly, W can be interpreted as the *total mass* of a point process on the carrier set Γ. Barbour & Brown (1992b) generalize Theorem 2.8, with the choice $\Gamma_i^s = \emptyset$, to the setting in which W is the total mass of a point process on a compact second countable Hausdorff topological carrier space Γ with a locally finite expectation measure. We give this theorem below; the corresponding generalization of Theorem 2.5 is given in Chapter 3, Theorem 5.8.

Theorem 2.25: *Let Ξ be a point process on $(\Gamma, \mathcal{B}_\Gamma)$, where Γ is a locally compact second countable Hausdorff topological space, with a locally finite expectation measure ν. For each $x \in \Gamma$, let a Palm process Ξ_x be defined on the same probability space as Ξ. Then*

$$d_{TV}(\mathcal{L}(\Xi(\Gamma)), \mathrm{Po}\,(\nu(\Gamma))) \;\leq\; k_2(\nu(\Gamma)) \int_\Gamma \mathbb{E}\big|\Xi(\Gamma) - \Xi_x(\Gamma) + 1\big| \nu(dx),$$

where $k_2(\cdot)$ is as defined in (2.2).

Proof: The proof consists of noting that

$$\begin{aligned} d_{TV}(\mathcal{L}(\Xi(\Gamma)), \mathrm{Po}\,(\nu(\Gamma))) &= \sup_{A \subset \mathbb{Z}_+} \big|\mathbb{E}(\Xi(\Gamma) f_A(\Xi(\Gamma)) - \nu(\Gamma) f_A(\Xi(\Gamma) + 1))\big| \\ &= \sup_{A \subset \mathbb{Z}_+} \left| \int_\Gamma \mathbb{E}(f_A(\Xi_x(\Gamma)) - f_A(\Xi(\Gamma) + 1)) \nu(dx) \right| \\ &\leq k_2(\nu(\Gamma)) \int_\Gamma \mathbb{E}\big|\Xi(\Gamma) - \Xi_x(\Gamma) + 1\big| \nu(dx), \end{aligned}$$

where in the first equality we used the definition of total variation distance and the Stein equation, in the second equality the definition of a Palm process, and in the final inequality Theorem 2.3. ∎

In the case when Ξ is a simple point process on $\Gamma = \mathbb{R}_+$ with compensator $\{A_t; t \geq 0\}$, Barbour & Brown (1992b) combine Stein's method with

methods from stochastic calculus to derive the following bound, presented here without proof.

$$d_{TV}(\mathcal{L}(\Xi((0,t])), \mathrm{Po}(\lambda)) \leq k_1(\lambda)\mathbb{E}|A_t - \lambda| + k_2(\lambda)\mathbb{E}\Big(\sum_{0<s\leq t}\{\Delta A_s\}^2\Big) + k_2(\lambda)\mathbb{E}\Big(\int_{(0,t]} d_2(\mathcal{P}_s, \mathcal{P}_s^-)\Xi(ds)\Big),$$

valid for any $\lambda > 0$. Here, $\Delta A_s = A_s - A_{s-}$, d_2 is the Wasserstein d_2 distance, and $\mathcal{P}_s$ and $\mathcal{P}_s^-$ are random probability measures on $\mathbb{Z}_+$ such that, for each $i \in \mathbb{Z}_+$ $\mathcal{P}_s(i)$ and $\mathcal{P}_s^-(i)$ are the optional and predictable projections of the process $I\{\Xi((s,t]) = i\}$. We refer to Barbour & Brown (1992b) for details.

We remark that, in the setting of Theorem 2.25, it is natural also to try to bound the distance between the distributions of the whole point process Ξ and a Poisson point process with intensity measure ν. Barbour & Brown (1992a) use Stein's method to derive such bounds; it turns out that the total variation distance is not the appropriate metric to use, since no "magic factors" as sharp as those in Theorem 2.3 can be derived, and that weaker metrics such as the Wasserstein d_2–metric are more appropriate. This theory is discussed in detail in Chapter 3, Section 5.

2.6. *Poisson-Charlier approximation*

Although a Poisson distribution is adequate as an approximation in many cases, more refined approximations are sometimes needed. The Poisson-Charlier signed measures $\{\mathcal{Q}_l; l \in \mathbb{N}\}$ on $\mathbb{Z}_+$, where $\mathcal{Q}_1 = \mathrm{Po}(\lambda)$, have been suggested for this purpose. They are similar in structure to the Edgeworth signed measures, but the Hermite polynomials have been replaced by Charlier polynomials. Barbour (1987) uses Stein's method to derive explicit error estimates in Poisson-Charlier approximation of the distribution of a sum of *independent* indicator variables; his main result is Theorem 2.26 below.

It should be pointed out, however, that in this case error estimates can be obtained also by other metods. Deheuvels & Pfeifer (1988) obtain error estimates for the same kind of approximations which are in most respects sharper than those in Theorem 2.26, using a combination of operator theoretic and complex analytic methods. Also, there exist promising alternatives to the Poisson-Charlier signed measures, some of which are described in Section 3.7 below.

Theorem 2.26: *Let* $W = \sum_{i\in\Gamma} X_i$, *where* $\{X_i; i \in \Gamma\}$ *are independent indicator variables. For each* $n \in \mathbb{Z}_+$, *let* $C_n(\lambda, x)$ *be the* n*th order Charlier*

polynomial,

$$C_n(\lambda, x) = \sum_{r=0}^{n} \binom{n}{r} (-1)^{n-r} \lambda^{-r} x(x-1)\dots(x-r+1). \tag{2.7}$$

For each $l \geq 1$, define the lth order Poisson-Charlier signed measure on $\mathbb{Z}_+$ by

$$\mathcal{Q}_l(i) = \left(\frac{e^{-\lambda}\lambda^i}{i!}\right)\left(1 + \sum_{s=1}^{l-1} \sum_{[s]} \prod_{j=1}^{s} \left[\frac{1}{r_j!}\left(\frac{(-1)^j \lambda_{j+1}}{j+1}\right)^{r_j}\right] C_{R+s}(\lambda, i)\right),$$

where $\lambda_{j+1} = \sum_{i \in \Gamma} p_i^{j+1}$, and $\sum_{[s]}$ denotes the sum over all s-tuples $(r_1, \dots, r_s) \in \mathbb{Z}_+^s$ such that $\sum_{j=1}^{s} j r_j = s$, and $R = \sum_{j=1}^{s} r_j$. Then, for each $h \in \chi$ and $l \geq 1$,

$$\left|\mathbb{E}(h(W)) - \int_{\mathbb{Z}_+} h \, d\mathcal{Q}_l\right| \leq k_2(\lambda) \lambda_{l+1} 2^{2l-1} \|h\|,$$

where $k_2(\cdot)$ is as defined in (2.2).

Proof: [Sketch of proof.] Let X be an indicator variable with $\mathbb{E}(X) = p$. Then, for each $f \in \chi$ and $j \in \mathbb{Z}_+$, it is easily verified that

$$\sum_{s=1}^{l-1} (-1)^s p^{s+1} \mathbb{E}(\Delta^s f(X+j+1)) = (-1)^{l+1} p^{l+1} \Delta^l f(j+1) - p^2 \Delta f(j+1),$$

where $\Delta^l f$ is the lth forward difference of f. Moreover,

$$\mathbb{E}(p f(X+j+1) - X f(X+j)) = p^2 \Delta f(j+1).$$

Letting $X = X_i$ and $j = W - X_i$, taking expectations, and summing over i, it follows that

$$\left|\mathbb{E}(\lambda f(W+1) - W f(W)) - \sum_{s=1}^{l-1} (-1)^{s+1} \lambda_{s+1} \mathbb{E}(\Delta^s f(W+1))\right| \leq \lambda_{l+1} \|\Delta^l f\|.$$

Choosing $f = S_0 h$ as the solution of the Stein equation for h, we get

$$\left|\mathbb{E}(h(W)) - \int_{\mathbb{Z}_+} h \, d\mu_0 - \sum_{s=1}^{l-1} (-1)^{s+1} \lambda_{s+1} \mathbb{E}(\Delta^s S_0 h(W+1))\right| \leq \lambda_{l+1} \|\Delta^l f\|.$$

Using this expression iteratively, we get

$$\mathbb{E}(h(W)) = \sum_{(l)} \left(\prod_{j=1}^{k} (-1)^{s_j+1} \lambda_{s_j+1} \right) \mathbb{E}\left\{ \left(\prod_{j=1}^{k} (\Delta^{s_j} S_0) h \right)(Z) \right\} + \eta,$$

where $Z \sim \mathrm{Po}(\lambda)$, $\sum_{(l)}$ denotes the sum over

$$\left\{ (s_1, \ldots, s_{k+1}) \in \mathbb{N}^{k+1};\ k \in \{0, \ldots, l-1\},\ \sum_{j=1}^{k+1} s_j = l \right\},$$

and

$$|\eta| \leq \sum_{(l)} \left(\prod_{j=1}^{k+1} \lambda_{s_j+1} \right) \left\| \prod_{j=1}^{k+1} (\Delta^{s_j} S_0) h) \right\|.$$

Rewriting, using the identities

$$\begin{aligned} \mathbb{E}(C_n(\lambda, Z)(\Delta f)(Z)) &= \mathbb{E}(C_{n+1}(\lambda, Z) f(Z)); \\ \mathbb{E}(C_n(\lambda, Z) S_0 h(Z)) &= -\frac{1}{n+1} \mathbb{E}(C_{n+1}(\lambda, Z) h(Z)), \end{aligned}$$

valid for each $n \in \mathbb{Z}_+$, and using Theorem 2.3 to bound η, the result is obtained. ∎

If our goal is to find a good approximation of a single, very small, probability, we need to choose a Poisson-Charlier signed measure of a very high order to obtain a small *relative* size in the error estimate. Barbour & Jensen (1989) circumvents this problem by considering, instead of the original indicator variables, a collection of indicator variables with a "tilted" probability distribution. For details, see Barbour, Holst & Janson (1992), Chapter 9.

2.7. *Poisson approximation for unbounded functions*

So far we have constructed estimates for $|\mathbb{E}(h(W)) - \int_{\mathbb{Z}_+} h \, d\mu_0|$, where W is a sum of indicator variables, $\mu_0 = \mathrm{Po}(\mathbb{E}(W))$, and h is a bounded function. If the indicator variables are independent, Stein's method can be used to find such estimates for a large class of unbounded functions. This is done in Barbour (1987), Chen & Choi (1992), and Barbour, Chen & Choi (1995), from which the following theorem is taken.

Here also it must be pointed out that estimates have recently been found using other methods which are in most respects sharper than those obtained using Stein's method; see Borisov (2002).

Theorem 2.27: *Let $W = \sum_{i\in\Gamma} X_i$, where $\{X_i; i \in \Gamma\}$ are independent indicator variables. Let h be such that $\mathbb{E}(Z^2 h(Z)) < \infty$, where $Z \sim \mathrm{Po}(\lambda)$. Then*

$$\left|\mathbb{E}(h(W)) - \int_{\mathbb{Z}_+} h\,d\mu_0\right| \le C_W \frac{\lambda_2}{2}\,(4k_2(\lambda)\mathbb{E}|h(Z+1)| + \{\mathbb{E}|h(Z+2)| - 2\mathbb{E}|h(Z+1)| + \mathbb{E}|h(Z)|\}),$$

where $\lambda_2 = \sum_{i\in\Gamma} p_i^2$, $k_2(\cdot)$ is as in (2.2), and

$$C_W = \begin{cases} (1 - \max_{i\in\Gamma} p_i)^{-1} e^{\lambda}, & \text{if } 0 < \lambda < 1; \\ (1 - \max_{i\in\Gamma} p_i)^{-1} e^{13/12}\sqrt{2\pi}(1 - \lambda_2/\lambda)^{-1/2}, & \text{if } \lambda \ge 1. \end{cases}$$

Proof: [Sketch of proof.] Denote by f the unique solution of the Stein equation for h. Let $W_i = W - X_i$ for each $i \in \Gamma$. Then

$$\begin{aligned} \mathbb{E}(h(W)) - \mathbb{E}(h(Z)) &= \lambda\mathbb{E}(f(W+1)) - \mathbb{E}(Wf(W)) \\ &= \sum_{i\in\Gamma} p_i \mathbb{E}(f(W+1)) - f(W_i+1)) \\ &= \sum_{i\in\Gamma} p_i \mathbb{E}\big(X_i(f(W_i+2) - f(W_i+1))\big) \\ &= \sum_{i\in\Gamma} p_i^2 \mathbb{E}\big(f(W_i+2)) - f(W_i+1)\big). \end{aligned}$$

In Barbour, Chen & Choi (1995), Proposition 2.1, it is shown that

$$\sup\left\{\frac{\mathbb{P}(W_i = r)}{\mathbb{P}(Z = r)};\ r \in \mathbb{Z}_+, i \in \Gamma\right\} \le C_W,$$

and in their Lemma 3.5 it is shown that

$$\begin{aligned} \mathbb{E}\big|f_{\{r\}}(Z+2) - f_{\{r\}}(Z+1)\big| &\le 2\,k_2(\lambda)\,\mathbb{P}(Z = r-1) \\ &+ \frac{1}{2}\big(\mathbb{P}(Z = r-2) - 2\mathbb{P}(Z = r-1) + \mathbb{P}(Z = r)\big), \qquad \forall r \in \mathbb{Z}_+, \end{aligned}$$

where $f_{\{r\}}$ is the solution of the Stein equation for $h = I_{\{r\}}$. This completes the proof, since

$$\begin{aligned} |\mathbb{E}(h(W)) - \mathbb{E}(h(Z))| &\le C_W\lambda_2\mathbb{E}\big|f(Z+2) - f(Z+1)\big| \\ &\le C_W\lambda_2\mathbb{E}\left|\sum_{r=0}^{\infty} h(r)(f_{\{r\}}(Z+2) - f_{\{r\}}(Z+1))\right| \\ &\le C_W\lambda_2 \sum_{r=0}^{\infty} |h(r)|\mathbb{E}\big|f_{\{r\}}(Z+2) - f_{\{r\}}(Z+1)\big|. \quad \blacksquare \end{aligned}$$

3. Compound Poisson approximation

3.1. *The* CP(π) *distribution*

In the remaining sections, we consider Stein's method for compound Poisson approximation. This research area is comparatively young, but has seen a rapid development in the last few years, including improvements on some of the results in the original paper by Barbour, Chen & Loh (1992).

We first recall the definition of the compound Poisson distribution, henceforth denoted by $\mathrm{CP}(\pi)$. $\mathrm{CP}(\pi)$ is the probability distribution on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$ which has the characteristic function

$$\varphi(t) = \exp\Big(-\int_0^\infty (1 - e^{itx}) d\pi(x)\Big),$$

where the measure π satisfies

$$\int_0^\infty (x \wedge 1)\, d\pi(x) < \infty.$$

If $\|\pi\| = \pi(\mathbb{R}_+) < \infty$, then $\mathrm{CP}(\pi) = \mathcal{L}(\sum_{i=1}^U T_i)$, where all random variables are independent, $\mathcal{L}(T_i) = \overline{\pi} = \pi/\|\pi\|$ for each $i \geq 1$, and $\mathcal{L}(U) = \mathrm{Po}(\|\pi\|)$. We call π the compounding measure, and $\overline{\pi}$ the compounding distribution.

We shall be concerned only with the case when $\|\pi\| < \infty$. Likewise, with the exception of Theorems 3.1–3.2, we shall be concerned only with the case when π is supported on the positive integers. We can then express the distribution $\mathrm{CP}(\pi)$ as $\mathcal{L}(\sum_{k=1}^\infty k Z_k)$, where $\{Z_k; k \geq 1\}$ are independent, and $Z_k \sim \mathrm{Po}(\pi_k)$ for each $k \geq 1$.

3.2. *Why compound Poisson approximation?*

$\mathrm{CP}(\pi)$ is a generalization of the Poisson distribution, but is it an interesting generalization for approximation purposes? When should we use a compound Poisson approximation rather than a simple Poisson approximation?

We answer this question by means of an example. Let $\{\eta_i; i \in \mathbb{Z}\}$ be a sequence of independent and identically distributed indicator variables such that $\mathbb{E}(\eta_i) = p$. Let $X_i = I\{\eta_i = \eta_{i-1} = \ldots = \eta_{i-r+1} = 1\}$ be the indicator of a run of r consecutive 1s occurring between indices $i - r + 1$ and i, where $r \geq 2$. Let $W = \sum_{i=1}^n X_i$. If r is large, $\mathbb{P}(X_i = 1) = p^r$ will be small, so a Poisson approximation seems natural for $\mathcal{L}(W)$. It can be

shown using Theorem 2.8 that

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(np^r)) \leq \frac{2p}{1-p} + p^r.$$

This is a special case of Theorem 8.H in Barbour, Holst & Janson (1992). The result leaves us dissatisfied, since the error estimate is of order p, and is thus much larger than the approximation error of at most p^r, which would be the case, were the indicators X_i independent; see (2.6). It seems that a Poisson approximation is not, after all, appropriate in this situation. Why?

Heuristically, even though the probability $\mathbb{P}(X_i = 1) = p^r$ is small, the conditional probability $\mathbb{P}(X_i = 1 | X_{i-1} = 1) = p$ is (comparatively) large. Thus, even though the events $\{X_i = 1\}$ are rare, when they do occur they tend to occur in clumps, rather than isolated from one another. This is what impairs the accuracy of the Poisson approximation, and since clumping of rare events is a common phenomenon, we should expect this to happen quite often.

A natural alternative to Poisson approximation in such cases is the following. We consider not the number of rare events, but the number of clumps of rare events, as approximately Poisson distributed, and we consider the clump sizes as independent and identically distributed. The distribution of the number of rare events is then approximately $\mathrm{CP}\,(\pi)$, where $\|\pi\|$ is the mean number of clumps, and $\overline{\pi}$ is the clump size distribution. It is this idea, sometimes called the "Poisson clumping heuristic", that we shall henceforth pursue. For aspects of the idea other than those mentioned here, see Aldous (1989).

3.3. *The Stein equation for* $\mathrm{CP}(\pi)$ *and its solutions*

We next study the Stein operator for the $\mathrm{CP}\,(\pi)$ distribution proposed in Barbour, Chen & Loh (1992), and the solutions of the corresponding Stein equation. Our first two theorems concern the general case, but all the results that follow concern only the discrete case, where π is supported on the positive integers. Let $(S, \mathcal{S}, \mu) = (\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+}, \mu)$, where $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$ is the set of nonnegative real numbers equipped with the Borel σ-algebra. Let χ be the set of all measurable functions $f : \mathbb{R}_+ \to \mathbb{R}$. Let $\mu_0 = \mathrm{CP}\,(\pi)$, where $\|\pi\| < \infty$. Let $\mathcal{F}_0 = \{f \in \chi; \sup_{x>0} |xf(x)| < \infty\}$, and define the Stein operator $T_0 : \mathcal{F}_0 \to \chi$ by

$$(T_0 f)(x) = \int_{\mathbb{R}'_+} tf(x+t)\, d\pi(t) - xf(x) \qquad \forall x \in \mathbb{R}_+. \tag{3.1}$$

Theorem 3.1: *If $h : \mathbb{R}_+ \to \mathbb{R}$ is bounded, the Stein equation*

$$T_0 f = h - \int_{\mathbb{Z}_+} h \, d\mu_0$$

has a solution $f \in \mathcal{F}_0$. The solution f is unique except for $f(0)$, which can be chosen arbitrarily. For each $x \in \mathbb{R}'_+$, $f(x)$ is given by

$$f(x) = -\sum_{k=0}^{\infty} \int_{\mathbb{R}_+^k} \frac{t_1 \cdots t_k \big(h(x + t_1 + \cdots + t_k) - \int_{\mathbb{R}_+} h \, d\mu_0\big)}{x(x + t_1) \cdots (x + t_1 + \cdots + t_k)} \prod_{i=1}^{k} d\pi(t_i).$$

Proof: Let $\overline{\chi}$ be the quotient space of χ with respect to the set of functions $\{h \in \chi; h = 0 \text{ on } \mathbb{R}'_+\}$, and denote the equivalence class containing h by $\overline{h}$. Define $\mathcal{Y}$ to be the Banach space $\{h \in \chi; \|h\| < \infty\}$, equipped with the supremum norm $\|h\|_{\mathcal{Y}} = \|h\|$. Define the linear spaces

$$\mathcal{Z} = \{\overline{f} \in \overline{\chi}; \sup_{x>0} |f(x)| < \infty\};$$

$$\mathcal{X} = \{\overline{f} \in \overline{\chi}; \sup_{x>0} |xf(x)| < \infty\},$$

equipped with the norms $\|\overline{f}\|_{\mathcal{Z}} = \|f\|_{\mathcal{Y}}$ and $\|\overline{f}\|_{\mathcal{X}} = \sup_{x>0} |xf(x)|$, respectively. It is easy to see that $\mathcal{Z}$ is a Banach space.

Define the linear mappings $M : \mathcal{X} \to \mathcal{Y}$ and $U : \mathcal{X} \to \mathcal{Y}$ through the equations $(M\overline{f})(x) = \int_{\mathbb{R}'_+} tf(x+t) \, d\pi(t)$ and $(U\overline{f})(x) = xf(x)$. Define also the mappings $\overline{M} : \mathcal{X} \to \mathcal{Z}$ and $\overline{U} : \mathcal{X} \to \mathcal{Z}$ by $\overline{Mf} = \overline{Mf}$ and $\overline{Uf} = \overline{Uf}$. It is not difficult to prove that $\overline{U}$ is an isometry and a bijection, making $\mathcal{X}$ a Banach space. Moreover, it is shown in Lemma 3 in Barbour, Chen & Loh (1992) that the linear operator $\overline{M} - \overline{U}$ is a bounded bijection, with an inverse given by

$$(\overline{M} - \overline{U})^{-1}\overline{h} = -\sum_{k=0}^{\infty} (\overline{U}^{-1}\overline{M})^k \overline{U}^{-1}\overline{h}, \qquad \forall \overline{h} \in \mathcal{Z},$$

and such that $\|(\overline{M} - \overline{U})^{-1}\| \leq e^{\|\pi\|}$.

Define the linear mapping $\phi : \mathcal{Z} \to \mathcal{Y}$ by

$$(\phi\overline{h})(x) = \begin{cases} h(x), & \text{if } x > 0; \\ -e^{\|\pi\|} \int_{\mathbb{R}'_+} h \, d\mu_0, & \text{if } x = 0. \end{cases}$$

Clearly, ϕ is 1-1, and maps $\mathcal{Z}$ onto $\{h \in \mathcal{Y}; \int_{\mathbb{R}_+} h \, d\mu_0 = 0\}$. Hence, the operator $\phi \circ (\overline{M} - \overline{U})$ is 1-1 and maps $\mathcal{X}$ onto $\{h \in \mathcal{Y}; \int_{\mathbb{R}_+} h \, d\mu_0 = 0\}$. This proves the first part of the theorem. It is not difficult to verify the explicit expression for the solution f. ∎

The following theorem is stated without proof in Barbour, Chen & Loh (1992).

Theorem 3.2: *Let f_A be the solution of the Stein equation with $h = I_A$, where $A \in \mathcal{B}_{\mathbb{R}_+}$. Then,*

$$\sup_{A \in \mathcal{B}_{\mathbb{R}_+}} \sup_{y \geq x > 0} \left| x(f_A(y) - f_A(x)) \right| \leq e^{\|\pi\|}.$$

From now on we restrict our attention to the discrete case. So let $(S, \mathcal{S}, \mu) = (\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+}, \mu)$, where $(\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+})$ is the set of nonnegative integers equipped with the power σ-algebra, and take $\mu_0 = \mathrm{CP}(\pi)$, where $\sum_{i=1}^{\infty} i\pi_i < \infty$. Let χ be the set of all functions $f : \mathbb{Z}_+ \to \mathbb{R}$, and let $\mathcal{F}_0 = \{f \in \chi;\ \sup_{k \in \mathbb{N}} |kf(k)| < \infty\}$ (a subset of the set of bounded functions). Now define the Stein operator $T_0 : \mathcal{F}_0 \to \chi$ by

$$(T_0 f)(k) = \sum_{i=1}^{\infty} i\pi_i f(k+i) - kf(k) \qquad \forall k \in \mathbb{Z}_+. \tag{3.2}$$

Theorem 3.3: *If $h : \mathbb{Z}_+ \to \mathbb{R}$ is bounded, the Stein equation*

$$T_0 f = h - \int_{\mathbb{Z}_+} h \, d\mu_0$$

has a bounded solution f. The solution f is unique except for $f(0)$, which can be chosen arbitrarily. f is given by

$$f(k) = -\sum_{i=k}^{\infty} a_{i,k} \left(h(i) - \int_{\mathbb{Z}_+} h \, d\mu_0 \right) \qquad \forall k \in \mathbb{N}, \tag{3.3}$$

where $a_{k,k} = 1/k$ and

$$a_{k+i,k} = \sum_{j=1}^{i} \frac{j\pi_j}{k+i} \, a_{k+i-j,k} \qquad \forall i \in \mathbb{N}.$$

Proof: Imitating the proof of Theorem 3.1, we see that there exists a unique solution $f \in \mathcal{F}_0 = \{f \in \chi;\ \sup_{k \in \mathbb{N}} |kf(k)| < \infty\}$. Since we have assumed that $\sum_{i=1}^{\infty} i\pi_i < \infty$, we can prove by contradiction that no other bounded solutions exist. Assume that there exists two bounded solutions f_1 and f_2. Then, $f_1 - f_2$ is a solution of the Stein equation with $h \equiv 0$, and

$$\begin{aligned} \sup_{k \in \mathbb{N}} \left| k(f_1(k) - f_2(k)) \right| &= \sup_{k \in \mathbb{N}} \left| \sum_{i=1}^{\infty} i\pi_i (f_1(k+i) - f_2(k+i)) \right| \\ &\leq \sup_{k \in \mathbb{N}} \sum_{i=1}^{\infty} i\pi_i \left| f_1(k+i) - f_2(k+i) \right| \ < \ \infty, \end{aligned}$$

so $f_1 - f_2 \in \mathcal{F}_0$. Hence, $f_1 - f_2 \equiv 0$ by uniqueness. The proof of (3.3) is omitted, but can be found in Barbour, Chen & Loh (1992). ∎

The following characterization of CP (π) is the analogue of Theorem 2.2, and is proved in the same way.

Theorem 3.4: *A probability measure μ on $(\mathbb{Z}_+, \mathcal{B}_{\mathbb{Z}_+})$ is* CP (π) *if and only if*

$$\int_{\mathbb{Z}_+} (T_0 f) d\mu = 0$$

for all bounded $f : \mathbb{Z}_+ \to \mathbb{R}$.

In Section 2.3, it was seen that "magic factors", or good bounds on the supremum norm of the first difference of the solution of the Stein equation and the supremum norm of the solution itself, are essential for the success of Stein's method for Poisson approximation. This is true also for compound Poisson approximation. However, in this case much more work is required to find such bounds, and the best known bounds are valid only if the compounding distribution satisfies certain conditions. Theorem 3.5 below is due to Barbour, Chen & Loh (1992). Later on, in Theorem 3.17, some other bounds are presented, which are valid under condition (3.16).

Theorem 3.5: *Let f be the unique bounded solution of the Stein equation with h bounded. Let*

$$H_1(\pi) = \sup_{A \subset \mathbb{Z}_+} \|\Delta f_A\|, \qquad H_0(\pi) = \sup_{A \subset \mathbb{Z}_+} \|f_A\|, \tag{3.4}$$

where f_A is the unique bounded solution of the Stein equation with $h = I_A$. Then

$$\max\bigl(H_0(\pi), H_1(\pi)\bigr) \leq \Bigl(1 \wedge \frac{1}{\pi_1}\Bigr) e^{\|\pi\|}. \tag{3.5}$$

Moreover, if

$$i\pi_i - (i+1)\pi_{i+1} \geq 0 \qquad \forall i \in \mathbb{N}, \tag{3.6}$$

then

$$\begin{aligned} H_1(\pi) &\leq \left\{ 1 \wedge \frac{1}{\pi_1 - 2\pi_2} \Bigl(\frac{1}{4(\pi_1 - 2\pi_2)} + \log^+ 2(\pi_1 - 2\pi_2) \Bigr) \right\}; \\ H_0(\pi) &\leq \begin{cases} \frac{1}{\sqrt{\pi_1 - 2\pi_2}} \bigl(2 - \frac{1}{\sqrt{\pi_1 - 2\pi_2}}\bigr), & \text{if } \pi_1 - 2\pi_2 > 1; \\ 1, & \text{if } \pi_1 - 2\pi_2 \leq 1. \end{cases} \end{aligned} \tag{3.7}$$

Proof: The bound (3.5) can be proved using the representation (3.3), as in Barbour, Chen & Loh (1992). However, (3.5) is not very useful unless $\|\pi\|$ is quite small.

The bound (3.7), valid under condition (3.6), is much better. It is proved by means of a probabilistic representation of the solution of the Stein equation, similar to the one given in Theorem 2.4 in the Poisson case. Informally, letting $f(k) = \nabla g(k) = g(k) - g(k-1)$ and assuming that g does not grow too fast, we get

$$\begin{aligned}(T_0 f)(k) &= \sum_{i=1}^{\infty} \big(i\pi_i - (i+1)\pi_{i+1}\big) g(k+i) + kg(k-1) - (\pi_1 + k)g(k) \\ &= (\mathcal{A}g)(k) \qquad \forall k \in \mathbb{Z}_+,\end{aligned}$$

where the operator $\mathcal{A}$ is the generator of a batch immigration-death process $\{Z_t; t \in \mathbb{R}_+\}$; for each $i \in \mathbb{Z}_+$, the immigration rate for batches of size i is $i\pi_i - (i+1)\pi_{i+1}$, while the death rate per individual is 1. It is not difficult to prove that the stationary distribution of $\{Z_t; t \in \mathbb{R}_+\}$ is CP (π). The Poisson's equation corresponding to $\mathcal{A}$ is

$$-\mathcal{A}g = h - \int_{\mathbb{Z}_+} h \, d\mu_0.$$

If h is bounded, then it can be proved, just as in Theorem 2.4, that this equation has the solution

$$g(k) = \int_0^{\infty} \Big(\mathbb{E}(h(Z_t) | Z_0 = k) - \int_{\mathbb{Z}_+} h \, d\mu_0 \Big) dt \qquad \forall k \in \mathbb{Z}_+,$$

and that $f = -\nabla g$ is the unique bounded solution of the Stein equation; note that we have assumed that $\sum_{i=1}^{\infty} i\pi_i < \infty$. To prove (3.7) we define, for each $k \geq 1$, four coupled batch immigration-death processes, where $Z^{(k)}$ starts at k, and

$$\begin{aligned} Z_t^{(k,1)} &= Z_t^{(k)} + I\{\tau_1 > t\}, \\ Z_t^{(k,2)} &= Z_t^{(k)} + I\{\tau_2 > t\}, \\ Z_t^{(k,3)} &= Z_t^{(k,1)} + I\{\tau_2 > t\}. \end{aligned}$$

Here, $\tau_1 \sim \exp(1)$ and $\tau_2 \sim \exp(1)$ are independent of each other and of $Z^{(k)}$. Then

$$
\begin{aligned}
&f_A(k+2) - f_A(k+1) \\
&= -\int_0^\infty \mathbb{E}\big(I_A(Z_t^{(k,3)}) - I_A(Z_t^{(k,2)}) - I_A(Z_t^{(k,1)}) + I_A(Z_t^{(k)})\big)\, dt \\
&= -\int_0^\infty \mathbb{E}\big(I\{\tau_1 \wedge \tau_2 > t\}(I_A(Z_t^{(k)}+2) - 2I_A(Z_t^{(k)}+1) + I_A(Z_t^{(k)}))\big)\, dt \\
&= -\int_0^\infty \mathbb{P}(\tau_1 \wedge \tau_2 > t)\big(\mathbb{P}(Z_t^{(k)} \in A-2) - 2\mathbb{P}(Z_t^{(k)} \in A-1) + \mathbb{P}(Z_t^{(k)} \in A)\big)\, dt \\
&= -\int_0^\infty e^{-2t}\big(\mathbb{P}(Z_t^{(k)} \in A-2) - 2\mathbb{P}(Z_t^{(k)} \in A-1) + \mathbb{P}(Z_t^{(k)} \in A)\big)\, dt.
\end{aligned}
$$

We can write $Z_t^{(k)} = Y_t + W_t$, where Y_t is the number of individuals who have immigrated in batches of size 1 after time 0 and are still alive at time t, and Y_t and W_t are independent. Then, using an inequality derived in Barbour (1988),

$$
\begin{aligned}
&|\mathbb{P}(Z_t^{(k)} \in A-2) - 2\mathbb{P}(Z_t^{(k)} \in A-1) + \mathbb{P}(Z_t^{(k)} \in A)| \\
&\le \sum_{r=0}^\infty \mathbb{P}(W_t = r)\Big|\sum_{s \in A-r}\big(\mathbb{P}(Y_t = s-2) - 2\mathbb{P}(Y_t = s-1) + \mathbb{P}(Y_t = s)\big)\Big| \\
&\le \Big(2 \wedge \frac{1}{(1-e^{-t})(\pi_1 - 2\pi_2)}\Big).
\end{aligned}
$$

An integration completes the proof of the first bound in (3.7). The proof of the second bound, given in Barbour, Chen & Loh (1992), is similar and is therefore omitted. ∎

3.4. *Error estimates in compound Poisson approximation*

For the construction of error estimates in compound Poisson approximation, we shall focus on the case when the distribution to be approximated, μ, is the distribution of a sum of nonnegative integer valued random variables with finite means.

Unless explicitly stated otherwise, we use the following notation, similar to the one in Section 2.3. Γ is the index set, which is finite. In most cases $\Gamma = \{1, \ldots, n\}$. $\{X_i; i \in \Gamma\}$ are nonnegative integer valued random variables with finite means; $W = \sum_{i \in \Gamma} X_i$; $\mu = \mathcal{L}(W)$; and $\mu_0 = \mathrm{CP}(\pi)$, where π is the "canonical" compounding measure, defined by (3.8).

In most of what follows our goal will be to bound the total variation distance between μ and μ_0. As in Section 2.3, using the Stein equation we

obtain the representation

$$d_{TV}(\mu,\mu_0) = \sup_{A\subset\mathbb{Z}_+} |\mu(A)-\mu_0(A)| = \sup_{A\subset\mathbb{Z}_+} \left|\int_{\mathbb{Z}_+} (T_0 f_A)\, d\mu\right|$$
$$= \sup_{A\subset\mathbb{Z}_+} \left|\mathbb{E}\Big(\sum_{i=1}^{\infty} i\pi_i f_A(W+i) - W f_A(W)\Big)\right|,$$

where f_A is the unique bounded solution of the Stein equation with $h = I_A$. Just as for the Poisson distribution, there is a local and a coupling approach to bounding the right hand side.

Theorem 3.6: (The local approach). *Let $W = \sum_{i\in\Gamma} X_i$, where the random variables $\{X_i; i \in \Gamma\}$ are nonnegative and integer valued, and have finite means. For each $i \in \Gamma$, divide $\Gamma \setminus \{i\}$ into three subsets Γ_i^{vs}, Γ_i^{w} and Γ_i^{b}, so that, informally,*

$$\Gamma_i^{vs} = \{j \in \Gamma\setminus\{i\}; X_j \text{ "very strongly" dependent on } X_i\};$$
$$\Gamma_i^{w} = \{j \in \Gamma\setminus\{i\}; X_j \text{ "weakly" dependent on } \{X_k; k \in \{i\}\cup\Gamma_i^{vs}\}\}.$$

Let $Z_i = \sum_{j\in\Gamma_i^{vs}} X_j$, $W_i = \sum_{j\in\Gamma_i^{w}} X_j$ and $U_i = \sum_{j\in\Gamma_i^{b}} X_j$. Define π by

$$\pi_k = \frac{1}{k}\sum_{i\in\Gamma} \mathbb{E}(X_i I\{X_i + Z_i = k\}), \qquad \forall k \in \mathbb{N}. \tag{3.8}$$

Then

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \le H_1(\pi)\sum_{i\in\Gamma}\big(\mathbb{E}(X_i)\mathbb{E}(X_i+Z_i+U_i) + \mathbb{E}(X_iU_i)\big)$$
$$+ H_0(\pi)\sum_{i\in\Gamma}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} j\mathbb{E}\big|\mathbb{P}(X_i=j, X_i+Z_i=k) - \mathbb{P}(X_i=j, X_i+Z_i=k|W_i)\big|,$$

where $H_0(\cdot)$ and $H_1(\cdot)$ are as defined in (3.4).

Proof: The proof parallels that for Poisson approximation in Theorem 2.5.

Direct computation shows that

$$\begin{aligned}
&\mathbb{E}\Big(\sum_{k=1}^{\infty} k\pi_k f_A(W+k) - W f_A(W)\Big) \\
&= \sum_{i\in\Gamma} \mathbb{E}\Big(\sum_{k=1}^{\infty} \mathbb{E}(X_i I\{X_i+Z_i=k\}) f_A(W+k) - X_i f_A(W)\Big) \\
&= \sum_{i\in\Gamma}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} j\mathbb{E}\Big\{\mathbb{P}(X_i=j, X_i+Z_i=k) f_A(W+k) \\
&\qquad - I\{X_i=j, X_i+Z_i=k\} f_A(W_i+U_i+k)\Big\} \\
&= \sum_{i\in\Gamma}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} j\mathbb{E}\Big\{\mathbb{P}(X_i=j, X_i+Z_i=k)(f_A(W+k) - f_A(W_i+k)) \\
&\qquad + (\mathbb{P}(X_i=j, X_i+Z_i=k) - I\{X_i=j, X_i+Z_i=k\}) f_A(W_i+k) \\
&\qquad + I\{X_i=j, X_i+Z_i=k\}(f_A(W_i+k) - f_A(W_i+U_i+k))\Big\}.
\end{aligned}$$

The result follows since, for each $i \in \Gamma$, $j \in \mathbb{N}$ and $k \in \mathbb{N}$,

$$\begin{aligned}
\big|f_A(W+k) - f_A(W_i+k)\big| &\leq \|\Delta f_A\|(X_i+Z_i+U_i); \\
\big|f_A(W_i+k) - f_A(W_i+U_i+k)\big| &\leq \|\Delta f_A\| U_i,
\end{aligned}$$

and

$$\begin{aligned}
&\big|\mathbb{E}\big((\mathbb{P}(X_i=j, X_i+Z_i=k) - I\{X_i=j, X_i+Z_i=k\}) f_A(W_i+k)\big)\big| \\
&\quad \leq \|f_A\| \mathbb{E}\big|\mathbb{P}(X_i=j, X_i+Z_i=k) - \mathbb{P}(X_i=j, X_i+Z_i=k|W_i)\big|. \qquad \blacksquare
\end{aligned}$$

Example 3.7: (Independent random variables). Let $\{X_i; i \in \Gamma\}$ be independent. Choosing $\Gamma_i^{vs} = \Gamma_i^b = \emptyset$ for each $i \in \Gamma$ in Theorem 3.6 gives

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq H_1(\pi) \sum_{i\in\Gamma} \mathbb{E}(X_i)^2, \tag{3.9}$$

where the canonical compounding measure π is

$$\pi_k = \sum_{i\in\Gamma} \mathbb{P}(X_i = k) \qquad \forall k \in \mathbb{N}.$$

This can be compared with bounds obtained using other means than Stein's method. Le Cam (1965) proved that

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq \sum_{i\in\Gamma} \mathbb{P}(X_i > 0)^2,$$

which is sometimes better than (3.9) and sometimes worse, depending on π. In the special case when $\mathcal{L}(X_i|X_i > 0)$ is the same for all $i \in \Gamma$, Michel (1988) proved that

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq \frac{1}{\|\pi\|} \sum_{i \in \Gamma} \mathbb{P}(X_i > 0)^2, \tag{3.10}$$

which is better than both of the preceding bounds. However, the following counter-example shows that the bound (3.10) cannot be valid in general, if the X_i are allowed to have different distributions. For each $i \in \Gamma = \{1, \ldots, n\}$, let $\mathbb{P}(X_i = 3^{i-1}) = 1 - \mathbb{P}(X_i = 0) = p$. The bound (3.10) is then p. However, we know from Section 3.1 that the random variable $Y = \sum_{i=1}^{n} 3^{i-1} Z_i$, where $\{Z_i; i = 1, \ldots, n\}$ are i.i.d. with distribution $\mathrm{Po}(p)$, has the distribution $\mathrm{CP}(\pi)$. Moreover, $\mathcal{L}(W)$ is supported on those nonnegative integers whose ternary representations contain no 2s, so from the definition of total variation distance,

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) &\geq \sum_{i=1}^{n} \mathbb{P}(Z_1 \leq 1, \ldots, Z_{i-1} \leq 1, Z_i = 2) \\ &= \frac{p^2 e^{-p}}{2} \sum_{i=0}^{n-1} (1+p)^i e^{-ip} = \frac{p^2 e^{-p}}{2} \Big(\frac{1 - (1+p)^n e^{-np}}{1 - (1+p)e^{-p}} \Big). \end{aligned}$$

Applying l'Hospital's rule twice shows that

$$\lim_{p \downarrow 0} \frac{(p^2 e^{-p})/2}{1 - (1+p)e^{-p}} = 1,$$

and, for each fixed $p > 0$, $\lim_{n \to \infty} (1+p)^n e^{-np} = 0$. To see that the bound given in (3.10) would be too small here to be true, now choose p very small and n very large.

The next example, where the random variables are dependent, is taken from Roos (1993); it was mentioned previously in Section 3.2. Barbour & Chryssaphinou (2001) consider other examples of similar flavour from reliability theory and the theory of random graphs.

Example 3.8: (Head runs). Let $\{\eta_i; i \in \mathbb{Z}\}$ be an i.i.d. sequence of indicator variables with $\mathbb{E}(\eta_i) = p$. Let $X_i = I\{\eta_i = \eta_{i-1} = \ldots = \eta_{i-r+1} = 1\}$, where for simplicity of computation we identify $i + kn$ with i for each $n \in \mathbb{Z}$. Let $W = \sum_{i=1}^{n} X_i$. Choosing

$$\begin{aligned} \Gamma_i^{vs} &= \{j \in \Gamma; 1 \leq |i-j| \leq r-1\}; \\ \Gamma_i^{b} &= \{j \in \Gamma; r \leq |i-j| \leq 2(r-1)\}; \\ \Gamma_i^{w} &= \Gamma \setminus \{\{i\} \cup \Gamma_i^{vs} \cup \Gamma_i^{b}\}, \end{aligned}$$

the canonical compounding measure π is

$$\pi_k = \begin{cases} np^{r+k-1}(1-p)^2, & \text{if } k = 1, \ldots, r-1; \\ \frac{1}{k} np^{r+k-1}(1-p) & \\ \quad \times (2 + (2r-k-2)(1-p)), & \text{if } k = r, \ldots, 2(r-1); \\ np^{3r-2} \frac{1}{2r-1}, & \text{if } k = 2r-1. \end{cases}$$

Moreover, the last term in the bound of Theorem 3.6 vanishes, so

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) &\leq H_1(\pi) \sum_{i=1}^{n} \big(\mathbb{E}(X_i) \mathbb{E}(X_i + Z_i + U_i) + \mathbb{E}(X_i U_i) \big) \\ &= H_1(\pi)(6r-5) n p^{2r}. \end{aligned}$$

What are the asymptotics of this bound as $n \to \infty$ if $p = p_n$ and $r = r_n$? Noting that $\|\pi\| < \mathbb{E}(W) = np^r$, we can show that: (i) if $np^r \leq C < \infty$, the bound is $O(rp^r)$; (ii) if $np^r \to \infty$, $r \geq 4$ and $p \leq p' < \frac{1}{2}$, the bound is $O(rp^r \log(np^r))$; (iii) if $np^r \to \infty$ and $p \leq p' < \frac{1}{5}$, the bound is $O(rp^r)$. To do this we use Theorem 3.5 for (i) and (ii) and Theorem 3.17 for (iii).

Theorem 3.9: (The coupling approach). *Let $W = \sum_{i \in \Gamma} X_i$, where the $\{X_i; i \in \Gamma\}$ are nonnegative integer valued random variables with finite means. For each $i \in \Gamma$, divide $\Gamma \setminus \{i\}$ into three subsets Γ_i^{vs}, Γ_i^{w}, and Γ_i^{b}. Let $Z_i = \sum_{j \in \Gamma_i^{vs}} X_j$, $W_i = \sum_{j \in \Gamma_i^{w}} X_j$ and $U_i = \sum_{j \in \Gamma_i^{b}} X_j$. Define π by (3.8). For each $i \in \Gamma$, $j \in \mathbb{N}$ and $k \in \mathbb{N}$, let two random variables $\widetilde{W}_i^{j,k}$ and $W_i^{j,k}$ such that*

$$\begin{aligned} \mathcal{L}(\widetilde{W}_i^{j,k}) &= \mathcal{L}(W_i | X_i = j, X_i + Z_i = k); \\ \mathcal{L}(W_i^{j,k}) &= \mathcal{L}(W_i), \end{aligned}$$

be defined on the same probability space. Then

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \;\leq\; & H_1(\pi) \sum_{i \in \Gamma} \big(\mathbb{E}(X_i) \mathbb{E}(X_i + Z_i + U_i) + \mathbb{E}(X_i U_i) \big) \\ & + H_1(\pi) \sum_{i \in \Gamma} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} j \mathbb{P}(X_i = j, X_i + Z_i = k) \mathbb{E} \big| W_i^{j,k} - \widetilde{W}_i^{j,k} \big|, \end{aligned}$$

where $H_1(\cdot)$ is as defined in (3.4).

Proof: The same as for Theorem 3.6, except that instead of the last inequality in that proof we use the couplings to obtain, for each $i \in \Gamma$, $j \in \mathbb{N}$

and $k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}\big((\mathbb{P}(X_i = j, X_i + Z_i = k) - I\{X_i = j, X_i + Z_i = k\})f_A(W_i + k)\big) \\
&= \mathbb{P}(X_i = j, X_i + Z_i = k)\big\{\mathbb{E}(f_A(W_i + k)) \\
&\qquad -\mathbb{E}(f_A(W_i + k)|X_i = j, X_i + Z_i = k)\big\} \\
&= \mathbb{P}(X_i = j, X_i + Z_i = k)\mathbb{E}\big(f_A(W_i^{j,k} + k) - f_A(\widetilde{W}_i^{j,k} + k)\big),
\end{aligned}$$

implying that

$$\begin{aligned}
&\big|\mathbb{E}\big((\mathbb{P}(X_i = j, X_i + Z_i = k) - I\{X_i = j, X_i + Z_i = k\})f_A(W_i + k)\big)\big| \\
&\leq \|\Delta f_A\|\mathbb{P}(X_i = j, X_i + Z_i = k)\mathbb{E}\big|W_i^{j,k} - \widetilde{W}_i^{j,k}\big|. \qquad \blacksquare
\end{aligned}$$

In the case when the random variables $\{X_i; i \in \Gamma\}$ are independent, by choosing $\Gamma_i^{vs} = \Gamma_i^b = \emptyset$ and $\widetilde{W}_i^{j,k} = W_i^{j,k} = W_i$ in Theorem 3.9, for each $i \in \Gamma$, we again get (3.9). We now turn to an example where the random variables are dependent, taken from Erhardsson (1999).

Example 3.10: Let $\{\eta_i; i \in \mathbb{Z}\}$ be a stationary irreducible discrete time Markov chain on the finite state space S, with stationary distribution ν. Let $W = \sum_{i=1}^n I\{\eta_i \in B\}$, where $B \subset S$. If B is a rare subset of S, meaning that $\nu(B)$ is small, we expect W to be approximately Poisson distributed. However, conditional on $\{\eta_0 \in B\}$, the probability of returning to B after a short time might be large, in which case visits by η to B would tend to occur in clumps. As explained in Section 3.2, compound Poisson approximation should then be preferable to Poisson approximation.

We first introduce suitable random variables $\{X_i; i \in \Gamma\}$, as follows. Choose $a \in B^c$. For each $i \in \mathbb{Z}$, let $\tau_i^a = \min\{t \geq i; \eta_t = a\}$, and let

$$X_i = I\{\eta_i = a\} \sum_{j=i+1}^{\tau_{i+1}^a} I\{\eta_j \in B\}.$$

In words, we consider η as a regenerative random sequence, the regenerations being the visits by η to the state a. X_i is the number of visits to by η to B occurring during the regeneration cycle starting at time i, if a regeneration cycle starts at time i; otherwise $X_i = 0$. Let $W' = \sum_{i=1}^n X_i$. The basic coupling inequality gives

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(W')) \leq \mathbb{P}(W \neq W'),$$

where the right hand side is small if B is a rare set. We next apply the coupling approach to W'. For each $i \in \Gamma$, choose $\Gamma_i^{vs} = \Gamma_i^b = \emptyset$, so that

$$\pi_k = n\mathbb{P}(X_i = k), \qquad \forall k \geq 1.$$

For each $i \in \Gamma$ and $j \in \mathbb{N}$, we shall now construct a random sequence $\{\widetilde{\eta}_t^{i,j}; t \in \mathbb{Z}\}$ on the same probability space as η, such that

$$\mathcal{L}(\widetilde{\eta}_t^{i,j}) = \mathcal{L}(\eta | X_i = j),$$

and such that the two sequences differ only in one regeneration cycle. Formally, let the random sequence $\{\widehat{\eta}_t^{0,j}; t \in \mathbb{Z}\}$ be independent of η and distributed as $\mathcal{L}(\eta | X_0 = j)$. For each $i \in \mathbb{Z}$ and each $j \in \mathbb{N}$, let $\tau_i^a(\widehat{\eta}^{0,j}) = \min\{t \geq i; \widehat{\eta}_t^{0,j} = a\}$ and $\overline{\tau}_i^a = \max\{t \leq i; \eta_t = a\}$. Define $\{\widetilde{\eta}_t^{0,j}; t \in \mathbb{Z}\}$ by

$$\widetilde{\eta}_t^{0,j} = \begin{cases} \widehat{\eta}_t^{0,j}, & 0 \leq t \leq \tau_1^a(\widehat{\eta}^{0,j}); \\ \eta_{t+\overline{\tau}_0^a}, & t < 0; \\ \eta_{t-\tau_1^a(\widehat{\eta}^{0,j})+\tau_0^a}, & t > \tau_1^a(\widehat{\eta}^{0,j}). \end{cases}$$

We then define, for each $t \in \mathbb{Z}$, $\widetilde{\eta}_t^{i,j} = \widetilde{\eta}_{t-i}^{0,j}$. It follows from the strong Markov property that $\widetilde{\eta}^{i,j}$ has the desired distribution. Let

$$\widetilde{X}_k^{i,j} = I\{\widetilde{\eta}_k^{i,j} = a\} \sum_{r=k+1}^{\tau_{k+1}^a(\widetilde{\eta}^{i,j})} I\{\widetilde{\eta}_r^{i,j} \in B\}, \qquad \forall k \in \mathbb{Z},$$

where $\tau_i^a(\widetilde{\eta}^{i,j}) = \min\{t \geq i; \widetilde{\eta}_t^{i,j} = a\}$, and define $\widetilde{W}_i^j = \sum_{k \in \Gamma_i^w} \widetilde{X}_k^{i,j}$. Then

$$\mathcal{L}(\widetilde{W}_i^j) = \mathcal{L}(W_i | X_i = j).$$

Similarly, let $X_k^{i,j} = X_{k-i}$, and then define $W_i^j = \sum_{k \in \Gamma_i^w} X_k^{i,j}$. Clearly, $\mathcal{L}(W_i^j) = \mathcal{L}(W_i)$. If B is a rare set, $|W_i^j - \widetilde{W}_i^j|$ should be small with high probability. From Theorem 3.9, the triangle inequality for the total variation distance, and calculations in Erhardsson (1999) not repeated here, we get

$$\begin{aligned} &d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \\ &\leq H_1(\pi) \sum_{i=1}^n \mathbb{E}(X_i)^2 + H_1(\pi) \sum_{i=1}^n \sum_{j=1}^\infty j \mathbb{P}(X_i = j) \mathbb{E}\big|W_i^j - \widetilde{W}_i^j\big| + \mathbb{P}(W \neq W') \\ &\leq 2H_1(\pi)\Big(\mathbb{E}(\tau_0^a | \eta_0 \in B) - \mathbb{E}(\overline{\tau}_0^a | \eta_0 \in B) + \frac{\mathbb{E}(\tau_0^a)}{\nu(a)}\Big) n\nu(B)^2 + 2\mathbb{P}(\tau_0^B < \tau_0^a). \end{aligned}$$

As an application, consider again Example 3.8 (Head runs). We define a stationary Markov chain ζ by $\zeta_i = \min\{j \geq 0; \eta_{i-j} = 0\} \wedge r$ for each $i \in \mathbb{Z}$, so that $\mathcal{L}(W) = \mathcal{L}(\sum_{i=1}^n I\{\zeta_i = r\})$. With $B = \{r\}$ and $a = 0$, simple calculations show that $\pi_k = (n - r + 1)p^{r+k-1}(1-p)^2$ for each $k \in \mathbb{N}$ (see

Erhardsson (2000)), and

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi))$$
$$\leq H_1(\pi)\Big(2r+2+\frac{4p}{1-p}+\frac{2p}{(1-p)^2}\Big)(n-r+1)p^{2r}+\Big(2(1-p)r+2p\Big)p^r.$$

The asymptotics of the bound as $n \to \infty$ are similar to those given in Example 3.8.

Just as for Poisson approximation, the following slight extension of the coupling approach is sometimes useful.

Theorem 3.11: (The detailed coupling approach). *Let $W = \sum_{i\in\Gamma} X_i$, where $\{X_i; i \in \Gamma\}$ are nonnegative integer valued random variables with finite means. For each $i \in \Gamma$, divide $\Gamma \setminus \{i\}$ into three subsets Γ_i^{vs}, Γ_i^{w}, and Γ_i^{b}. Let $Z_i = \sum_{j\in\Gamma_i^{vs}} X_j$, $W_i = \sum_{j\in\Gamma_i^{w}} X_j$ and $U_i = \sum_{j\in\Gamma_i^{b}} X_j$. Define π by (3.8). Let a random variable σ_i be defined on the same probability space as X_i, Z_i and W_i, and let, for each $i \in \Gamma$, $j \in \mathbb{N}$, $k \in \mathbb{N}$ and $x \in \mathbb{R}$, two random variables $\widetilde{W}_i^{j,k,x}$ and $W_i^{j,k,x}$ such that*

$$\mathcal{L}(\widetilde{W}_i^{j,k,x}) = \mathcal{L}(W_i | X_i = j, X_i + Z_i = k, \sigma_i = x);$$
$$\mathcal{L}(W_i^{j,k,x}) = \mathcal{L}(W_i),$$

be defined on the same probability space. Then

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq H_1(\pi)\sum_{i=1}^{n}\big(\mathbb{E}(X_i)\mathbb{E}(X_i+Z_i+U_i)+\mathbb{E}(X_iU_i)\big)$$
$$+ H_1(\pi)\sum_{i=1}^{n}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} j\mathbb{E}\left(I\{X_i=j, X_i+Z_i=k\}\mathbb{E}\big|W_i^{j,k,x}-\widetilde{W}_i^{j,k,x}\big|_{x=\sigma_i}\right),$$

where $H_1(\cdot)$ is as defined in (3.4).

Our third example, taken from Barbour & Månsson (2000), concerns the number of k-sets of random points in the unit square that can be covered by a translate of a square with side length c. This random variable is closely related to the scan statistic in two dimensions, which, by definition, is the maximal number of random points that can be covered by such a translate.

Example 3.12: Let n numbered random points be uniformly and independently distributed in the unit square A. Identify each point with its number. Let Γ be the set of k-subsets of $\{1, \ldots, n\}$, i.e., $\Gamma = \{i \subset \{1, \ldots, n\}; |i| = k\}$. For each $i \in \Gamma$, let X_i be the indicator that all points in i can be covered by a translate of a square C with side length c. Let $W = \sum_{i\in\Gamma} X_i$.

We use Theorem 3.11. For each $i \in \Gamma$, let R_i be the square centred at a randomly chosen point in i, with side length $4c$. Choose

$$\Gamma_i^{vs} = \{j \in \Gamma \setminus \{i\}; \text{all points in } j \text{ lie in } R_i\};$$
$$\Gamma_i^{w} = \{j \in \Gamma; j \cap i = \emptyset; \text{no points in } j \text{ lie in } R_i\};$$
$$\Gamma_i^{b} = \Gamma \setminus (\{i\} \cup \Gamma_i^{vs} \cup \Gamma_i^{w}).$$

This means that we are using random neighborhoods instead of fixed ones, but it is easy to see that the error estimates are still valid. For each $i \in \Gamma$, let the random variable σ_i be the number of points not in i that lie in R_i. For each $i \in \Gamma$, $r \in \mathbb{N}$, and $s \in \mathbb{Z}_+$, we construct two random variables $\widetilde{W}_i^{1,r,s}$ and $W_i^{1,r,s}$ on the same probability space, such that

$$\mathcal{L}(\widetilde{W}_i^{1,r,s}) = \mathcal{L}(W_i | X_i = 1, X_i + Z_i = r, \sigma_i = s),$$

and $\mathcal{L}(\widetilde{W}_i^{1,r,s}) = \mathcal{L}(W_i)$. This is achieved as follows. Let $n-k$ points be independently and uniformly distributed on $A \setminus R_i$, and then sample $N_i \sim \mathrm{Bin}(n-k, 16|C|)$ independently of the $n-k$ points. Now define $M_{i,s} = \max\{n-k-s, n-k-N_i\}$ and $m_{i,s} = \min\{n-k-s, n-k-N_i\}$. Number the k-subsets of $\{1, \ldots, M_{i,s}\}$ in such a way that the k-subsets of $\{1, \ldots, m_{i,s}\}$ are given numbers from 1 to $\binom{m_{i,s}}{k}$. Let $\widetilde{X}_l$ be the indicator that all points in the lth k-subset can be covered by a translate of C. We then define

$$\widetilde{W}_i^{1,r,s} = \sum_{l=1}^{\binom{n-k-s}{k}} \widetilde{X}_l; \qquad W_i^{1,r,s} = \sum_{l=1}^{\binom{n-k-N_i}{k}} \widetilde{X}_l.$$

These random variables have the desired distributions. Hence, Theorem 3.11 gives

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq H_1(\pi) \sum_{i=1}^{n} (\mathbb{E}(X_i)\mathbb{E}(X_i + Z_i + U_i) + \mathbb{E}(X_i U_i))$$
$$+ H_1(\pi) \sum_{i=1}^{n} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \mathbb{P}(X_i = 1, X_i + Z_i = r, \sigma_i = s) \mathbb{E}\left|W_i^{1,r,s} - \widetilde{W}_i^{1,r,s}\right|.$$

After some computations, which are not repeated here, the following bound is finally obtained:

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) \leq H_1(\pi) \binom{n}{k} \left[16 \binom{n-k}{k} k^4 |C|^{2k-1} \right.$$
$$+ k^4 |C|^{2(k-1)} \left\{ \sum_{l=1}^{k-1} \binom{k}{l} \binom{n-k}{k-l} + 25 \binom{n-k}{k} |C| + 1 \right\}$$
$$\left. + 32 k^4 |C|^k \left(\frac{|C|}{1 - 16|C|}\right)^{k-1} (n-k) \binom{n-k}{k-1} \right].$$

Asymptotically, as $n \to \infty$, if $C = C_n$ in such a way that

$$\mathbb{E}(W) = \binom{n}{k} k^2 |C|^{k-1} \leq K < \infty,$$

the bound is $O(n^{2k-1}|C|^{2k-2})$. Since π does not satisfy (3.6), the bound does not converge to 0 when $\mathbb{E}(W) \to \infty$. However, one can use results due to Barbour & Utev (1998, 1999), to be presented in the next section, to prove a slightly different result: there exists a constant $c > 0$ so that, as $n \to \infty$, if $n|C_n| \to 0$ and $\mathbb{E}(W) \to \infty$,

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\pi)) = O\big((n|C_n|)^{k-1} + \exp(-c\mathbb{E}(W))\big).$$

3.5. *The Barbour-Utev version of Stein's method for compound Poisson approximation*

We here describe a version of Stein's method for compound Poisson approximation due to Barbour & Utev (1998, 1999). Using their approach, it is often possible to obtain much better bounds on the *rate of convergence* to an approximating compound Poisson distribution than using the original Stein's method, but the corresponding error estimates are not so well suited for numerical evaluation.

The work of Barbour & Utev was motivated by the desire to improve the bounds given in Theorem 3.5 (the magic factors). Unfortunately, (3.5) cannot be much improved in general: it is not difficult to find a π for which the condition (3.6) fails, and for which there exists $\beta > 0$ and $C(\overline{\pi})$ so that $H_1(\pi) \geq C(\overline{\pi})e^{\beta\|\pi\|}$. To get around this difficulty we define, for each $a \geq 1$,

$$H_1^a(\pi) = \sup_{A \subset \mathbb{Z}_+} \sup_{k>a} |f_A(k+1) - f_A(k)|,$$

$$H_0^a(\pi) = \sup_{A \subset \mathbb{Z}_+} \sup_{k>a} |f_A(k)|.$$

Theorem 3.13: *Let W be a nonnegative integer valued random variable, let $\mu = \mathcal{L}(W)$, and let $\mu_0 = \mathrm{CP}(\pi)$, where $m_2 = \sum_{i=1}^{\infty} i^2\overline{\pi}_i < \infty$. For each $0 < a < b < \infty$, let*

$$u_{a,b}(x) = \Big(\frac{x-a}{b-a}\Big) I_{(a,b]}(x) + I_{(b,\infty)}(x).$$

Then,

$$d_{TV}(\mu, \mu_0) \leq \sup_{A \subset \mathbb{Z}_+} \left| \mathbb{E}\Big(\sum_{i=1}^{\infty} i\pi_i f_A(W+i) u_{a,b}(W+i) - W f_A(W) u_{a,b}(W) \Big) \right|$$

$$+ \mathbb{P}(W \leq b)\Big(1 + \frac{\|\pi\| m_2 H_0^a(\pi)}{b-a}\Big). \tag{3.11}$$

In particular, we may choose $a = c\|\pi\|m_1$ *and* $b = \frac{1}{2}(1+c)\|\pi\|m_1$, *where* $m_1 = \sum_{i=1}^{\infty} i\overline{\pi}_i$, *for* $0 < c < 1$.

Proof: For each $A \subset \mathbb{Z}_+$ and each $k \in \mathbb{Z}_+$,

$$
\begin{aligned}
&I_A(k) - \mu_0(A) \\
&= u_{a,b}(k)\Big(\sum_{i=1}^{\infty} i\pi_i f_A(k+i) - k f_A(k)\Big) + (1 - u_{a,b}(k))(I_A(k) - \mu_0(A)) \\
&= \sum_{i=1}^{\infty} i\pi_i f_A(k+i) u_{a,b}(k+i) - k f_A(k) u_{a,b}(k) \\
&\quad + (1 - u_{a,b}(k))(I_A(k) - \mu_0(A)) - \sum_{i=1}^{\infty} i\pi_i f_A(k+i)(u_{a,b}(k+i) - u_{a,b}(k)).
\end{aligned}
$$

Moreover,

$$
|u_{a,b}(k+i) - u_{a,b}(k)| \leq \min\Big\{1, \frac{i}{b-a}\Big\} I_{[0,b]}(k) I_{(a,\infty)}(k+i). \qquad \blacksquare
$$

To bound the first term on the right hand side in (3.11) we can use the local, coupling or detailed coupling approaches together with explicit bounds on $H_0^a(\pi)$ and $H_1^a(\pi)$, since it easily seen that

$$
\begin{aligned}
\sup_{A \subset \mathbb{Z}_+} \|\Delta(f_A u_{a,b})\| &\leq H_1^a(\pi) + \frac{H_0^a(\pi)}{b-a}; \\
\sup_{A \subset \mathbb{Z}_+} \|f_A u_{a,b}\| &\leq H_0^a(\pi).
\end{aligned}
$$

For the second term we also need a bound on $\mathbb{P}(W \leq b)$. This can be obtained in various ways, for example using Chebyshev's inequality, Janson's inequality, included as Theorem 2.S in Barbour, Holst & Janson (1992), or Janson's extension of Suen's inequality; see Janson (1998).

Explicit bounds on $H_0^a(\pi)$ and $H_1^a(\pi)$ are given in the following theorem due to Barbour & Utev (1999). The very long and complicated proof is omitted.

Theorem 3.14: *Assume that the generating function* φ *of* $\overline{\pi}$ *has radius of convergence* $R > 1$, *that* $\|\pi\| \geq 2$, *and that*

$$
\min\big(\rho_1^*(\zeta), \frac{1}{2}\rho_2^*(\zeta)\big) > 0 \qquad \forall 0 < \zeta \leq \pi, \tag{3.12}
$$

where

$$\rho_1^*(\zeta) = \inf_{\zeta \le \theta \le \pi} \bigl(1 - \sum_{k=1}^{\infty} \overline{\pi}_k \cos(k\theta)\bigr),$$

$$\rho_2^*(\zeta) = \inf_{\zeta \le \theta \le \pi} \bigl(1 - \frac{1}{m_1} \sum_{k=1}^{\infty} k\overline{\pi}_k \cos(k\theta)\bigr),$$

and $m_1 = \sum_{k=1}^{\infty} k\overline{\pi}_k$. *Then there exist explicit constants* $C_0(\overline{\pi})$, $C_1(\overline{\pi})$ *and* $C_2(\overline{\pi}) < 1$, *such that, for each* $a \ge C_2(\overline{\pi})\|\pi\|m_1 + 1$,

$$\begin{aligned} H_1^a(\pi) &\le C_1(\overline{\pi})\|\pi\|^{-1}, \\ H_0^a(\pi) &\le C_0(\overline{\pi})\|\pi\|^{-1/2}. \end{aligned} \tag{3.13}$$

Proof: [Starting points of proof.] The unique bounded solution of the Stein equation with $h(k) = z^k$, where z is a complex number such that $|z| \le 1$, is

$$f^z(k) = e^{\|\pi\|\varphi(z)} \int_{\Gamma_{z,1}} w^{k-1} e^{-\|\pi\|\varphi(w)} dw \qquad \forall k \in \mathbb{N},$$

where φ is the generating function of $\overline{\pi}$, and $\Gamma_{z,1}$ is a path in the unit disc from z to 1. Moreover, for each $A \subset \mathbb{Z}_+$,

$$\begin{aligned} f_A(k) &= \sum_{l \in A} \frac{1}{2\pi i} \int_{|z|=1} z^{-l-1} f^z(k)\, dz \\ &= \sum_{l \in A} \frac{1}{2\pi i} \int_0^1 t^{k-1} e^{\|\pi\|(1-\varphi(t))} (\mu_t(l-k) - \mu_0(l)) dt, \qquad \forall k \in \mathbb{N}, \end{aligned}$$

where $\mu_t = \mathrm{CP}(\pi^t)$, and $\pi_k^t = \pi_k(1 - t^k)$ for each $k \in \mathbb{N}$. Using these two representations of f_A, the bounds can be derived. ∎

Condition (3.12) is easily seen to be satisfied for all aperiodic π.

The explicit expressions for $C_0(\overline{\pi})$, $C_1(\overline{\pi})$ and $C_2(\overline{\pi})$ are very complicated, and no effort was made to optimize them, so we do not reproduce them here. However, by examining these expressions we can find sufficient conditions for $C_0(\mu)$, $C_1(\mu)$ and $C_2(\mu)$ to be uniformly bounded over a set of probability measures. For example, if $\overline{\pi}$ has radius of convergence $R > 1$ and satisfies condition (3.12), then there exists an $\varepsilon > 0$ such that $C_0(\mu)$ and $C_1(\mu)$ are uniformly bounded above, and $C_2(\mu)$ is uniformly bounded away from 1, over the set of probability measures $\{\mu; \sup_{i \ge 1} R^i |\mu_i - \overline{\pi}_i| \le \varepsilon\}$. In this fashion, one can obtain results like the following, due to Barbour & Månsson (2000).

Theorem 3.15: *Let $\{W_n; n \geq 1\}$ be nonnegative integer valued random variables. Assume that the sequence $\{\pi^n; n \geq 1\}$ of compounding measures satisfies the following conditions: (i) $\lim_{n\to\infty} \pi_i^n = \pi_i$ for each $i \geq 1$; (ii) $\sup_{n\geq 1} \sum_{i=1}^{\infty} \pi_i^n R^i < \infty$ for some $R > 1$; (iii) $\inf_{n\geq 1} \|\pi^n\| > 2$; (iv) $\inf_{n\geq 1} \pi_1^n > 0$. Assume also that, for each $n \geq 1$ and each bounded $f : \mathbb{Z}_+ \to \mathbb{R}$,*

$$\left| \mathbb{E}\Big(\sum_{i=1}^{\infty} i\pi_i^n f(W_n + i) - W_n f(W_n) \Big) \right| \leq \|\Delta f\| B_n.$$

Then there exist positive constants $K < \infty$ and $c_2 < 1$ such that, for any $c_2 \leq c < 1$ and any n such that $\mathbb{E}(W_n) \geq (c - c_2)^{-1}$,

$$d_{TV}(\mathcal{L}(W_n), \mathrm{CP}(\pi^n)) \leq \frac{K}{1-c} \Big(\frac{B_n}{\|\pi^n\|} + \mathbb{P}(W_n \leq \tfrac{1}{2}(1+c)\mathbb{E}(W_n)) \Big).$$

3.6. *Stein's method and Kolmogorov distance*

So far, we have constructed bounds for the total variation distance between a distribution μ and a compound Poisson distribution μ_0, using the chain of equalities

$$\begin{aligned} d_{TV}(\mu, \mu_0) &= \sup_{A \subset \mathbb{Z}_+} |\mu(A) - \mu_0(A)| = \sup_{A \subset \mathbb{Z}_+} \left| \int_{\mathbb{Z}_+} (T_0 f_A)\, d\mu \right| \\ &= \sup_{A \subset \mathbb{Z}_+} \left| \mathbb{E}\Big(\sum_{i=1}^{\infty} i\pi_i f_A(W + i) - W f_A(W) \Big) \right|, \end{aligned}$$

where f_A is the unique bounded solution of the Stein equation with $h = I_A$. However, it should be clear that we could just as well try to bound

$$\sup_{h \in \chi_0} \left| \int_{\mathbb{Z}_+} h\, d\mu - \int_{\mathbb{Z}_+} h\, d\mu_0 \right| = \sup_{h \in \chi_0} \left| \int_{\mathbb{Z}_+} (T_0 f_h)\, d\mu \right|,$$

where f_h is the unique bounded solution of the Stein equation with $h \in \chi_0$, for any collection of bounded functions $\chi_0 \subset \chi$. For example, we could try to bound the Kolmogorov distance, defined as

$$d_K(\mu, \mu_0) = \sup_{k \in \mathbb{Z}_+} |\mu([k, \infty)) - \mu_0([k, \infty))|.$$

The following theorem is due to Barbour & Xia (2000).

Theorem 3.16: *Let $f_{[k,\infty)}$ be the unique bounded solution of the Stein equation with $h = I_{[k,\infty)}$, where $k \in \mathbb{Z}_+$. Define*

$$J_1(\pi) = \sup_{k \in \mathbb{Z}_+} \|\Delta f_{[k,\infty)}\|, \qquad J_0(\pi) = \sup_{k \in \mathbb{Z}_+} \|f_{[k,\infty)}\|.$$

If condition (3.6) *holds, then*

$$J_1(\pi) \le \frac{1}{2} \wedge \left(\frac{1}{\pi_1 + 1}\right) \quad \text{and} \quad J_0(\pi) \le 1 \wedge \sqrt{\frac{2}{e\pi_1}}. \tag{3.14}$$

Proof: [Sketch of proof.] As in the proof of (3.7), we use the probabilistic representation of the solution of the Stein equation, and get

$$\begin{aligned} & f_{[k,\infty)}(i+2) - f_{[k,\infty)}(i+1) \\ &= \int_0^\infty e^{-2t}\big(\mathbb{P}(Z_t^{(i)} \ge k-2) - 2\mathbb{P}(Z_t^{(i)} \ge k-1) + \mathbb{P}(Z_t^{(i)} \ge k)\big)\, dt \\ &= \mathbb{E}\Big(\int_0^\infty e^{-2t}(I\{Z_t^{(i)} = k-2\} - I\{Z_t^{(i)} = k-1\})\, dt\Big), \qquad \forall i \in \mathbb{Z}_+, \end{aligned}$$

where $\{Z_t^{(i)}; t \in \mathbb{R}_+\}$ is a batch immigration-death process with generator $\mathcal{A}$, starting at i. Since, unless $Z_t^{(i)} \in \{k-1, k-2\}$, the integrand assumes the value 0, and since $\{Z_t^{(i)}; t \in \mathbb{R}_+\}$ is a strong Markov process which makes only unit downward jumps, we can prove that

$$\begin{aligned} f_{[k,\infty)}(k+1) - f_{[k,\infty)}(k) &\le f_{[k,\infty)}(i+2) - f_{[k,\infty)}(i+1) \\ &\le f_{[k,\infty)}(k) - f_{[k,\infty)}(k-1), \end{aligned}$$

and, furthermore,

$$0 \le f_{[k,\infty)}(k) - f_{[k,\infty)}(k-1) \le \frac{1}{\pi_1 + 1}, \qquad \forall k > 1;$$

$$0 \le -\big(f_{[k,\infty)}(k+1) - f_{[k,\infty)}(k)\big) \le \frac{1}{\pi_1 + 1}, \qquad \forall k \ge 1. \qquad \blacksquare$$

3.7. *Stein's method for translated signed discrete compound Poisson measure approximation*

We here describe Stein's method for approximation of distributions of integer valued random variables with translated signed discrete compound Poisson measures, a class of signed measures which extends the discrete compound Poisson distributions. The results presented are due to Barbour & Xia (1999) and Barbour & Čekanavičius (2002).

This class of signed measures is defined as follows. Let π be a signed measure on $(\mathbb{Z}, \mathcal{B}_\mathbb{Z})$ such that $\sum_{i\in\mathbb{Z}} |i\pi_i| < \infty$, and such that $\pi_0 = 0$. For each $\gamma \in \mathbb{Z}$, the translated signed discrete compound Poisson measure $\mu_{\pi,\gamma}$ is the signed measure on $(\mathbb{Z}, \mathcal{B}_\mathbb{Z})$ with generating function

$$\varphi_{\pi,\gamma}(z) = z^\gamma \exp\Big(-\sum_{k\in\mathbb{Z}} (1 - z^k)\pi_k\Big). \tag{3.15}$$

Note that $\mu_{\pi,\gamma}(\mathbb{Z}) = 1$. We shall use these signed measures as approximations of discrete probability distributions which are close to the normal distribution. Compared to the normal distribution and to Edgeworth signed measures, they have the advantages of being themselves discrete and simpler in structure. Also, Stein's method can be used to produce explicit error estimates.

Let $(S, \mathcal{S}, \mu) = (\mathbb{Z}, \mathcal{B}_{\mathbb{Z}}, \mu)$, and let χ be all functions $f : \mathbb{Z} \to \mathbb{R}$. Let $\mathcal{F}_1 = \{f \in \chi; \|f\| < \infty\}$, and define the Stein operator $T_{\pi,\gamma} : \mathcal{F}_1 \to \chi$ by

$$(T_{\pi,\gamma} f)(k) = \sum_{i \in \mathbb{Z}} i\pi_i f(i+k) - (k-\gamma) f(k) \qquad \forall k \in \mathbb{Z}.$$

Theorem 3.17: *Assume that*

$$\lambda = \sum_{i \in \mathbb{Z}} i\pi_i > 0; \quad \theta = \frac{1}{\lambda} \sum_{i \in \mathbb{Z}} i(i-1)|\pi_i| < \frac{1}{2}. \tag{3.16}$$

Then, for each bounded $h \in \chi$ and each $\gamma \in \mathbb{Z}$, there exists an $f \in \mathcal{F}_1$ such that

1. $f(k) = 0 \qquad \forall\, k \leq \gamma;$
2. $\left|(T_{\pi,\gamma} f)(k) - (h(k) - \int_{\mathbb{Z}} h \, d\mu_{\pi,\gamma})\right| \leq \frac{2}{1-2\theta} \sum_{j<0} |\mu_{\pi,0}(j)| \|h\| \quad \forall\, k \geq \gamma;$
3. $\|f\| \leq \frac{2}{1-2\theta} \left(1 \wedge \frac{1}{\sqrt{\lambda}}\right) \|h\|;$
4. $\|\Delta f\| \leq \frac{2}{1-2\theta} \left(\frac{1-e^{-\lambda}}{\lambda}\right) \|h\|.$

Proof: We first consider the case $\gamma = 0$. Define the operator $U : \mathcal{F}_1 \to \chi$ by

$$(Uf)(k) = (T_{\pi,0} f)(k) - \lambda f(k+1) + k f(k).$$

Direct calculation shows that

$$\|Uf\| \leq \sum_{i \in \mathbb{Z}} i(i-1)|\pi_i| \|\Delta f\| = \lambda\theta \|\Delta f\|. \tag{3.17}$$

In particular, U is a bounded operator. For each bounded f, let Sf denote the unique bounded solution $\widetilde{f}$ of the equation

$$\lambda \widetilde{f}(k+1) - k\widetilde{f}(k) = h(k) - (Uf)(k) - \int_{\mathbb{Z}_+} (h - Uf) \, d\widetilde{\mu}, \quad k \geq 0,$$

for which $\widetilde{f}(k) = 0$ for all $k \leq 0$, where $\widetilde{\mu} = \mathrm{Po}(\lambda)$. Now define $f_0 \equiv 0$, $f_n = Sf_{n-1}$, and $g_n = f_n - f_{n-1}$. It follows that

$$\lambda g_n(k+1) - kg_n(k) = -(Ug_{n-1})(k) + \int_{\mathbb{Z}_+} (Ug_{n-1})\, d\widetilde{\mu} \quad \forall\, k \geq 0.$$

Theorem 2.3 and (3.17) give

$$\begin{aligned}\|g_n\| &\leq 2\Big(1 \wedge \frac{1}{\sqrt{\lambda}}\Big)\|Ug_{n-1}\| \leq 2\lambda\theta\Big(1 \wedge \frac{1}{\sqrt{\lambda}}\Big)\|\Delta g_{n-1}\| \\ &\leq 2(2\theta)^{n-1}\Big(1 \wedge \frac{1}{\sqrt{\lambda}}\Big)\|h\|.\end{aligned}$$

Hence, the f_n converge uniformly as $n \to \infty$ to a bounded function f. Moreover,

$$\begin{aligned}\|f\| &\leq \sum_{n=1}^{\infty} \|g_n\| = \frac{2}{1-2\theta}\Big(1 \wedge \frac{1}{\sqrt{\lambda}}\Big)\|h\|; \\ \|\Delta f\| &\leq \sum_{n=1}^{\infty} \|\Delta g_n\| = \frac{2}{1-2\theta}\Big(\frac{1-e^{-\lambda}}{\lambda}\Big)\|h\|.\end{aligned}$$

Finally, from the fact that

$$\begin{aligned}(Uf)(k) - \lambda f(k+1) + kf(k) \;&=\; (T_{\pi,0}f)(k) \\ &=\; h(k) - \int_{\mathbb{Z}_+} (h - Uf)\, d\widetilde{\mu}, \qquad \forall\, k \geq 0,\end{aligned}$$

and since $\int_{\mathbb{Z}} (T_{\pi,0}f)\, d\mu_{\pi,0} = 0$, which can be seen by differentiating (3.15) with respect to z and equating coefficients, we get

$$\begin{aligned}&\Big|(T_{\pi,0}f)(k) - \Big(h(k) - \int_{\mathbb{Z}} h\, d\mu_{\pi,0}\Big)\Big| = \Big|\int_{\mathbb{Z}} h\, d\mu_{\pi,0} - \int_{\mathbb{Z}_+} (h - Uf) d\widetilde{\mu}\Big| \\ &\leq \sum_{j<0} |\mu_{\pi,0}(j)| \left(\|h\| + \int_{\mathbb{Z}_+} |h - Uf|\, d\widetilde{\mu} + \|Uf\|\right) \\ &\leq 2\sum_{j<0} |\mu_{\pi,0}(j)|(\|h\| + \|Uf\|) \leq \frac{2}{1-2\theta} \sum_{j<0} |\mu_{\pi,0}(j)|\|h\|.\end{aligned}$$

The case $\gamma \neq 0$ can be reduced to the first case by defining the function $\hat{h} = h(\cdot + \gamma)$, denoting by $\hat{f}$ the function corresponding to $\hat{h}$ defined above, and letting $f = \hat{f}(\cdot - \gamma)$. ∎

We remark that if π is a nonnegative measure supported on the positive integers, so that $\mu_{\pi,0}$ is a compound Poisson distribution, then, for

$\gamma = 0$, the function f in Theorem 3.17 is the unique bounded solution of the Stein equation in Theorem 3.3, and the bounds on $\|\Delta f\|$ and $\|f\|$ given in Theorem 3.17 can be used as an alternative to the bounds in Theorem 3.5, provided that condition (3.16) is satisfied. This can sometimes be a considerable improvement.

Theorem 3.18: *Let W be an integer valued random variable, let π be a signed measure satisfying the conditions of Theorem 3.17, let $h \in \chi$ be bounded, and let f be the function corresponding to h defined in Theorem 3.17. Then*

$$\left|\mathbb{E}(h(W)) - \int_{\mathbb{Z}} h\,d\mu_{\pi,\gamma}\right| \leq |\mathbb{E}(T_{\pi,\gamma}f(W))| + \frac{2}{1-2\theta}\Big\{\sum_{j<0}|\mu_{\pi,0}(j)|\|h\| + \Big(1+\sum_{j<0}|\mu_{\pi,0}(j)|\Big)\|h\|\,\mathbb{P}(W<\gamma)\Big\},$$

where θ is as defined in (3.16).

Proof: Theorem 3.17 gives

$$\left|\mathbb{E}\Big\{\Big(h(W) - \int_{\mathbb{Z}} h\,d\mu_{\pi,\gamma}\Big)\big(I\{W \geq \gamma\} + I\{W < \gamma\}\big)\Big\}\right| \leq |\mathbb{E}(T_{\pi,\gamma}f(W))| + \frac{2}{1-2\theta}\sum_{j<0}|\mu_{\pi,0}(j)|\|h\| + (1+\|\mu_{\pi,0}\|)\|h\|\mathbb{P}(W<\gamma) + |\mathbb{E}(T_{\pi,\gamma}f(W)I\{W<\gamma\})|.$$

To complete the proof, we use the fact that $T_{\pi,\gamma}f(k) = Uf(k)$ for each $k < \gamma$, and also that

$$\|\mu_{\pi,0}\| \leq \frac{2}{1-2\theta}\sum_{j<0}|\mu_{\pi,0}(j)| + 1 + \frac{2\theta}{1-2\theta},$$

which follows from the last part of the proof of Theorem 3.17. ∎

The next theorem gives a bound for the total variation norm of the difference between the distribution $\mathcal{L}(W)$ of a sum of independent integer valued random variables and a centered Poisson distribution, by which we mean an integrally translated Poisson distribution with mean $\mathbb{E}(W)$ and variance as close as possible to $\mathrm{Var}\,(W)$. Recall that the total variation norm is defined for all finite signed measures on $\mathbb{Z}$, and that for the difference $\nu_1 - \nu_2$ between two probability measures ν_1 and ν_2, it is given by

$$\|\nu_1 - \nu_2\| = \sup_{\|h\|\leq 1}\left|\int_{\mathbb{Z}} h\,d\nu_1 - \int_{\mathbb{Z}} h\,d\nu_2\right| = 2d_{TV}(\nu_1,\nu_2).$$

Theorem 3.19: *Let* $W = \sum_{i\in\Gamma} X_i$, *where* $\{X_i; i \in \Gamma\}$ *are independent integer valued random variables such that* $\mathbb{E}(X_i) = \beta_i$, $\mathrm{Var}\,(X_i) = \sigma_i^2$ *and* $\mathbb{E}|X_i^3| < \infty$, *for each* $i \in \Gamma$. *Define* $\pi_i = 0$ *for* $i \neq 1$, *and*

$$\pi_1 = \sigma^2 + \delta; \quad \gamma = \lfloor \beta - \sigma^2 \rfloor,$$

where $\beta = \mathbb{E}(W)$, $\sigma^2 = \mathrm{Var}\,(W)$, *and* $\delta = (\beta - \sigma^2) - \lfloor \beta - \sigma^2 \rfloor$. *For each* $i \in \Gamma$, *let* $W_i = W - X_i$,

$$\begin{aligned} v_i &= \min\left\{\tfrac{1}{2}, 1 - d_{TV}(\mathcal{L}(X_i), \mathcal{L}(X_i + 1))\right\} \quad \text{and} \\ \psi_i &= \sigma_i^2 \mathbb{E}(X_i(X_i - 1)) + |\beta_i - \sigma_i^2| \mathbb{E}((X_i - 1)(X_i - 2)) \\ &\qquad + \mathbb{E}|X_i(X_i - 1)(X_i - 2)|. \end{aligned}$$

Then,

$$\|\mathcal{L}(W) - \mu_{\pi,\gamma}\| \leq \frac{1}{\sigma^2}\Big(d \sum_{i\in\Gamma} \psi_i + 2\delta\Big) + 2\mathbb{P}(W < \gamma),$$

where

$$d = 2 \max_{i\in\Gamma} d_{TV}(\mathcal{L}(W_i), \mathcal{L}(W_i + 1)) \leq \frac{2}{(\sum_{i\in\Gamma} v_i - \max_{i\in\Gamma} v_i)^{1/2}}.$$

Proof: Let $h \in \chi$ be bounded, and let f be the function corresponding to h defined in Theorem 3.17. Then

$$\begin{aligned} &\mathbb{E}(T_{\pi,\gamma} f(W)) = \mathbb{E}\big(\pi_1 f(W + 1) - (W - \gamma) f(W)\big) \qquad (3.18) \\ &= \sum_{i=1}^{n} \big\{\sigma_i^2 \mathbb{E}(f(W + 1)) + (\beta_i - \sigma_i^2)\mathbb{E}(f(W)) - \mathbb{E}(X_i f(W))\big\} + \delta \mathbb{E}(\Delta f(W)). \end{aligned}$$

Newton's expansion gives

$$\begin{aligned} f(W_i + l) &= f(W_i + 1) + (l - 1)\Delta f(W_i + 1) \\ &\quad + \begin{cases} \sum_{s=1}^{l-2} (l - 1 - s)\Delta^2 f(W_i + s), & l \geq 3; \\ 0, & 1 \leq l \leq 2; \\ \sum_{s=0}^{-l} (-l - s + 1)\Delta^2 f(W_i - s), & l \leq 0, \end{cases} \end{aligned}$$

from which we get

$$\begin{aligned} &|\mathbb{E}(f(W_i + l)) - \mathbb{E}(f(W_i + 1)) - (l - 1)\mathbb{E}(\Delta f(W_i + 1))| \\ &\quad \leq \tfrac{1}{2}(l - 1)(l - 2) d \|\Delta f\|, \qquad \forall\, i \in \Gamma, \end{aligned}$$

since, for each $l \in \mathbb{Z}$,

$$|\mathbb{E}(\Delta^2 f(W_i + l))| \leq 2\|\Delta f\| d_{TV}(\mathcal{L}(W_i), \mathcal{L}(W_i + 1)).$$

Hence,

$$\begin{aligned}\mathbb{E}(f(W+1)) &= \sum_{j\in\mathbb{Z}} \mathbb{P}(X_i = j)\mathbb{E}(f(W_i + j + 1)) \\ &= \mathbb{E}(f(W_i+1)) + \beta_i \mathbb{E}(\Delta f(W_i+1)) + r_{i,1}, \qquad \forall\, i \in \Gamma,\end{aligned}$$

where $|r_{i,1}| \leq \frac{1}{2}\mathbb{E}(X_i(X_i-1))d\|\Delta f\|$. The other terms in (3.18) can be treated similarly. Application of Theorems 3.17 and 3.18 completes the proof, except for the bound on d, which is Proposition 4.6 in Barbour & Xia (1999). ∎

Example 3.20: Let $\{X_i; i = 1, \ldots, n\}$ be independent integer valued random variables satisfying the conditions of Theorem 3.19. If, for $n \geq 2$ and $i = 1, \ldots, n$, $\sigma_i^2 \geq a > 0$, $v_i \geq b > 0$, and $\psi_i/\sigma_i^2 \leq c < \infty$, then

$$\|\mathcal{L}(W) - \mu_{\pi,\gamma}\| \leq \frac{2c}{\sqrt{(n-1)b}} + \frac{2(1+\delta)}{na}. \tag{3.19}$$

If $X_i \sim \mathrm{Be}(p)$ with $p \leq \frac{1}{2}$, then we may take $a = p(1-p)$, $b = p$, and $c = 2p$, which gives

$$\|\mathcal{L}(W) - \mu_{\pi,\gamma}\| \leq 4\left(\frac{p}{\sqrt{(n-1)p}} + \frac{1}{np(1-p)}\right). \tag{3.20}$$

Recall that Stein's method for Poisson approximation in this case yields the total variation distance bound p, which is of the right order. If $np^2 \to \infty$ as $n \to \infty$, the rate of convergence to 0 for the bound in (3.20) is faster than p. On the other hand, if $np^2 < 1$, since $\mathbb{P}(W < \gamma) = 0$, the second term in (3.19) can be replaced by

$$\frac{2\delta}{na} = \frac{2np^2}{na} = \frac{2p}{1-p};$$

hence, if also np is bounded away from 0, the rate of convergence to 0 for the error bound in (3.20) is the same as for the Poisson approximation.

In Barbour & Čekanavičius (2002), Theorems 3.17–3.18 are also used to obtain error estimates in approximations of $\mathcal{L}(W)$ with more complex translated signed discrete compound Poisson measures, as well as with the signed measures $\widetilde{\nu}_r$, $r \geq 1$, defined by

$$\widetilde{\nu}_r(i) = \left(\frac{e^{-\beta}\beta^i}{i!}\right)\left(1 + \sum_{s=1}^{3(r-1)\wedge 1} (-1)^s b_s C_s(\beta, i)\right), \qquad \forall\, i \in \mathbb{Z}_+,$$

where $\beta = \mathbb{E}(W)$,

$$1 + \sum_{s=1}^{\infty} b_s z^s = \exp\Big(\sum_{j=2}^{r+1} \frac{\kappa_j z^j}{j!}\Big),$$

κ_j is the jth factorial cumulant of W, and $C_s(\beta, i)$ is the sth order Charlier polynomial; see (2.7).

3.8. *Compound Poisson approximation via Poisson process approximation*

We here describe a completely different way in which Stein's method can be used to obtain error estimates in compound Poisson approximation. The Stein operator (3.2) is not used; instead an error estimate is obtained as a by-product from Stein's method for *Poisson process* approximation, which is described in detail in Chapter 3. This "indirect" approach, which is older than the Stein operator for the compound Poisson distribution, was developed in Barbour (1988), Arratia, Goldstein & Gordon (1989, 1990) and Barbour & Brown (1992a).

In order to briefly describe the approach, let $\{X_i; i \in \Gamma\}$ be nonnegative integer valued random variables with finite means on a finite set Γ, and let $W = \sum_{i\in\Gamma} X_i$. Now define the indicator random variables $Y_{ij} := I[X_i = j]$, $(i,j) \in \Gamma \times \mathbb{N}$, noting that

$$\mathbb{E}W = \sum_{i\in\Gamma} \mathbb{E}X_i = \sum_{i\in\Gamma}\sum_{j=1}^{\infty} j\mu_{ij},$$

where $\mu_{ij} = \mathbb{E}(Y_{ij}) = \mathbb{P}[X_i = j]$. Then W can formally be expressed as an integral with respect to a point process $\widetilde{\Xi}$ on $\Gamma \times \mathbb{R}_+$:

$$W = \int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}(dx, dy),$$

where the expectation measure μ of $\widetilde{\Xi}$ is concentrated on $\Gamma \times \mathbb{N}$, with $\mu\{(i,j)\} = \mu_{ij}$, $(i,j) \in \Gamma \times \mathbb{N}$. A natural approximating distribution for $\mathcal{L}(W)$ is thus $\mathcal{L}(\int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}^0(dx,dy))$, where $\widetilde{\Xi}^0$ is a Poisson point process on $\Gamma \times \mathbb{R}_+$ with the same expectation measure μ; it is easy to see that $\mathcal{L}(\int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}^0(dx,dy)) = \mathrm{CP}(\pi)$, where $\pi\{j\} = \sum_{i\in\Gamma}\mu\{(i,j)\}$. By definition, since $\int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}(dx,dy)$ is a measurable function on the space of locally finite counting measures on $\Gamma \times \mathbb{R}_+$,

$$d_{TV}\Big(\mathcal{L}\Big(\int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}(dx,dy)\Big), \mathcal{L}\Big(\int_{\Gamma\times\mathbb{R}_+} y\,\widetilde{\Xi}^0(dx,dy)\Big)\Big) \le d_{TV}\big(\mathcal{L}(\widetilde{\Xi}), \mathcal{L}(\widetilde{\Xi}^0)\big). \tag{3.21}$$

Using Stein's method for Poisson process approximation, the right hand side can be bounded, if the random variables X_i have suitable structure: see Arratia, Goldstein & Gordon (1989, 1990) and Chapter 3, Section 5.3. However, no "magic factors" as good as those in Theorem 2.3 are known. If the total variation distance is replaced by the Wasserstein d_2–metric on the right hand side, good "magic factors" can be reintroduced, but then the inequality in (3.21) no longer holds, so this is of no use for compound Poisson approximation.

A recently proposed way to circumvent these difficulties is to use Stein's method for compound Poisson process approximation. One observes that $\mathcal{L}(W) = \mathcal{L}(\Xi(\Gamma))$, where Ξ is a point process on Γ which is no longer simple: $\Xi\{i\} = X_i$. A natural approximating distribution for $\mathcal{L}(W)$ is $\mathcal{L}(\Xi^0(\Gamma))$, where Ξ^0 is a suitably chosen compound Poisson point process on Γ; clearly, $\mathcal{L}(\Xi^0(\Gamma))$ is a compound Poisson distribution. Moreover,

$$d_{TV}(\mathcal{L}(\Xi(\Gamma)), \mathcal{L}(\Xi^0(\Gamma))) \leq d_2(\mathcal{L}(\Xi), \mathcal{L}(\Xi^0)).$$

It is shown in Barbour & Månsson (2002) how Stein's method can be used to bound the right hand side.

3.9. *Compound Poisson approximation on groups*

In the last section we describe an alternative to the Stein operator (3.2), proposed in Chen (1998).

Once again recall the general setting in Section 1. Let $(S, \mathcal{S})$ be a measurable abelian group, i.e., an abelian group such that the group operation is a measurable mapping from S^2 to S. Let $\mu_0 = \mathrm{CP}(\pi)$, where π is a finite measure on $(S, \mathcal{S})$ with no mass on the identity element. This means that $\mu_0 = \mathcal{L}(\sum_{i=1}^{U} T_i)$, where all random variables are independent, $\mathcal{L}(T_i) = \overline{\pi}$ for each $i \geq 1$, and $\mathcal{L}(U) = \mathrm{Po}(\|\pi\|)$. For some suitable $\mathcal{F}_0 \subset L^2(S, \mu_0)$, define a Stein operator $T_0 : \mathcal{F}_0 \to L^2(S, \mu_0)$ by

$$(T_0 f)(x) = \int_S f(x+t) d\pi(t) - \mathbb{E}\left(U \,\middle|\, \sum_{i=1}^{U} T_i = x \right) f(x), \qquad \forall\, x \in S.$$

This Stein operator T_0 can be obtained through Chen's L^2 approach, described in Section 1. If the linear operator $A : \mathcal{F}_0 \to L^2(S, \mu_0)$ is defined by $(Af)(x) = \int_S f(x+t)\, d\pi(t)$, its adjoint A^* satisfies

$$(A^* f)(x) = \mathbb{E}\left\{ Uf\left(\sum_{i=1}^{U-1} T_i \right) \,\middle|\, \sum_{i=1}^{U} T_i = x \right\}, \qquad \forall x \in S,$$

so we can define $T_0 = A - (A^*1)I$. As Chen points out, it is most natural to use T_0 when there exists a subgroup $\mathcal{K}_0$ of S and a coset $\mathcal{K}_1$ of $\mathcal{K}_0$, such that $\mathcal{K}_1$ is of infinite order in the quotient group $S/\mathcal{K}_0$, and such that π is supported on $\mathcal{K}_1$. If this is the case, then $U = \psi\left(\sum_{i=1}^{U} T_i\right)$ for some function $\psi : S \to \mathbb{Z}_+$. For example, if S is the additive group $\mathbb{Z}^d$, we might take $\mathcal{K}_1 = \{(x_1, \ldots, x_d) \in \mathbb{Z}^d; \sum_{i=1}^{d} x_i = 1\}$.

The properties of the solutions of the corresponding Stein equation still need to be investigated.

References

1. D. J. Aldous (1989) *Probability approximations via the Poisson clumping heuristic.* Springer, New York.
2. R. Arratia, L. Goldstein & L. Gordon (1989) Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* 17, 9–25.
3. R. Arratia, L. Goldstein & L. Gordon (1990) Poisson approximation and the Chen-Stein method. *Statist. Sci.* 5, 403–434.
4. A. D. Barbour (1987) Asymptotic expansions in the Poisson limit theorem. *Ann. Probab.* 15, 748–766.
5. A. D. Barbour (1988) Stein's method and Poisson process convergence. *J. Appl. Probab.* 15A, 175–184.
6. A. D. Barbour (1997) Stein's method. In: *Encyclopedia of statistical science,* 2nd ed. Wiley, New York.
7. A. D. Barbour & T. C. Brown (1992a) Stein's method and point process approximation. *Stoch. Proc. Appl.* 43, 9–31.
8. A. D. Barbour & T. C. Brown (1992b) The Stein-Chen method, point processes and compensators. *Ann. Probab.* 20, 1504–1527.
9. A. D. Barbour & V. Čekanavičius (2002) Total variation asymptotics for sums of independent integer random variables. *Ann. Probab.* 30, 509–545.
10. A. D. Barbour, L. H. Y. Chen & K. P. Choi (1995) Poisson approximation for unbounded functions, I: independent summands. *Statist. Sinica* 5, 749–766.
11. A. D. Barbour, L. H. Y. Chen & W.-L. Loh (1992) Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.* 20, 1843–1866.
12. A. D. Barbour & O. Chryssaphinou (2001) Compound Poisson approximation: a user's guide. *Ann. Appl. Probab.* 11, 964–1002.
13. A. D. Barbour & G. K. Eagleson (1983) Poisson approximation for some statistics based on exchangeable trials. *Adv. Appl. Probab.* 15, 585–600.
14. A. D. Barbour & P. Hall (1984) On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* 95, 473–480.
15. A. D. Barbour & L. Holst (1989) Some applications of the Stein-Chen method for proving Poisson convergence. *Adv. Appl. Probab.* 21, 74–90.

16. A. D. BARBOUR, L. HOLST & S. JANSON (1992) *Poisson approximation*. Oxford University Press.
17. A. D. BARBOUR & J. L. JENSEN (1989) Local and tail approximations near the Poisson limit. *Scand. J. Statist.* 16, 75–87.
18. A. D. BARBOUR & M. MÅNSSON (2000) Compound Poisson approximation and the clustering of random points. *Adv. in Appl. Probab.* 32, 19–38.
19. A. D. BARBOUR & M. MÅNSSON (2002) Compound Poisson process approximation. *Ann. Probab.* 30, 1492–1537.
20. A. D. BARBOUR & S. UTEV (1998) Solving the Stein equation in compound Poisson approximation. *Adv. Appl. Probab.* 30, 449–475.
21. A. D. BARBOUR & S. UTEV (1999) Compound Poisson approximation in total variation. *Stoch. Proc. Appl.* 82, 89–125.
22. A. D. BARBOUR & A. XIA (1999) Poisson perturbations. *ESAIM Probab. Statist.* 3, 131–150.
23. A. D. BARBOUR & A. XIA (2000) Estimating Stein's constants for compound Poisson approximation. *Bernoulli* 6, 581–590.
24. I. S. BORISOV (2002) Poisson approximation for expectations of unbounded functions of independent random variables. *Ann. Probab.* 30, 1657–1680.
25. L. LE CAM (1960) An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* 10, 1181–1197.
26. L. LE CAM (1965) On the distribution of sums of independent random variables. In: J. Neyman & L. Le Cam (eds). *Bernoulli, Bayes, Laplace*. Springer, New York.
27. L. H. Y. CHEN (1975) Poisson approximation for dependent trials. *Ann. Probab.* 3, 534–545.
28. L. H. Y. CHEN (1998) Stein's method: some perspectives with applications. In: L. Accardi & C. C. Heyde (eds). *Probability towards 2000*. Lecture notes in statistics, Vol. 128, 97–122. Springer, New York.
29. L. H. Y. CHEN & K. P. CHOI (1990) Some asymptotic and large deviation results in Poisson approximation. *Ann. Probab.* 20, 1867–1876.
30. D. J. DALEY & D. VERE-JONES (1988) *An introduction to the theory of point processes*. Springer, New York.
31. P. DEHEUVELS & D. PFEIFER (1988) On a relationship between Uspensky's theorem and Poisson approximations. *Ann. Inst. Statist. Math.* 40, 671–681.
32. T. ERHARDSSON (1999) Compound Poisson approximation for Markov chains using Stein's method. *Ann. Probab.* 27, 565–596.
33. T. ERHARDSSON (2000) Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. *Ann. Appl. Probab.* 10, 573–591.
34. J. D. ESARY, F. PROSCHAN & D. W. WALKUP (1967) Association of random variables, with applications. *Ann. Math. Statist.* 38, 1466–1474.
35. S. JANSON (1998) New versions of Suen's correlation inequality. *Random Structures Algorithms* 13, 467–483.
36. K. JOAG-DEV & F. PROSCHAN (1983) Negative association of random variables, with applications. *Ann. Statist.* 11, 451–455.

37. J. KERSTAN (1964) Verallgemeinerung eines Satzes von Prochorow und Le Cam. *Z. Wahrsch. Verw. Gebiete* 2, 173–179.
38. T. M. LIGGETT (1985) *Interacting particle systems*. Springer, New York.
39. R. MICHEL (1988) An improved error bound for the compound Poisson approximation of a nearly homogenous portfolio. *Astin Bull.* 17, 165–169.
40. M. ROOS (1993) Stein-Chen method for compound Poisson approximation. PhD thesis, University of Zürich.
41. C. STEIN (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. II: Probability theory, 583–602. Univ. California Press.
42. C. STEIN (1986) *Approximate computation of expectations.* IMS Lecture Notes Vol. 7. Hayward, California.

Stein's method and Poisson process approximation

Aihua Xia
Department of Mathematics and Statistics
University of Melbourne, VIC 3010
Australia
and
Department of Statistics and Applied Probability
National University of Singapore
6 Science Drive 2, Singapore 117546
E-mail: xia@ms.unimelb.edu.au

The chapter begins with an introduction to Poisson processes on the real line and then to Poisson point processes on a locally compact complete separable metric space. The focus is on the characterization of Poisson point processes. The next section reviews the basics of Markov immigration-death processes, and then of Markov immigration-death point processes. We explain how a Markov immigration-death point process evolves, establish its generator and find its equilibrium distribution. We then discuss our choice of metrics for Poisson process approximation, and illustrate their use in the context of stochastic calculus bounds for the accuracy of Poisson process approximation to a point process on the real line. These considerations are combined in constructing Stein's method for Poisson process approximation. Some of the key estimates given here are sharper than those found elsewhere in the literature, and have simpler proofs. In the final part, we show how to apply the bounds in various examples, from the easiest Bernoulli process to more complicated networks of queues.

Contents

1. Introduction

The law of rare events asserts that if certain events may occur in time or space with small probability and short range of dependence, then the total number of events that take place in each subinterval or subspace should approximately follow the Poisson distribution. This universal law makes the Poisson behaviour very pervasive in natural phenomena and the Poisson process is often regarded as a cornerstone of stochastic modelling.

A large number of intractable problems in probability can be considered as point processes of rare events, such as incoming calls in telecommunications, or clusters of palindromes in a family of herpes virus genomes in DNA sequence analysis, with each event being associated with a random set, in the line, on the plane, or in a Euclidean space; for many examples, see Aldous (1989). These problems appear in a large number of fields of science and engineering as well as in many of the major areas of probability theory, including Markov processes and diffusion, stationary processes,

combinatorics, extreme value theory, stochastic geometry, random fields and queueing. The resulting point processes often permit Poisson process approximate solutions which may be sufficiently good for practical purposes. For example, in queueing or telecommunication networks, particularly in large systems with diverse routing, flows of customers or calls along links in the network may approximately follow a Poisson process [see Barbour & Brown (1996)]. If these approximations are to be used in any particular case, it is essential to have some idea of the errors involved; in intractable problems, the best that can be done is to estimate the errors in these approximations. If the approximation is precise enough, it can be used in place of the original process. Its behaviour can be analyzed in a simple fashion using standard and well-developed methods, and it can then be applied in statistics, as in Andersen, Borgan, Gill & Keiding (1993).

We begin with an introduction to Poisson processes on the real line and then to Poisson point processes on a locally compact complete separable metric space. The focus is on the characterization of Poisson point processes. The next section reviews the basics of Markov immigration-death processes, and then of Markov immigration-death point processes. We explain how a Markov immigration-death point process evolves, establish its generator and find its equilibrium distribution. We then discuss our choice of metrics for Poisson process approximation, and illustrate their use in the context of stochastic calculus bounds for the accuracy of Poisson process approximation to a point process on the real line. These considerations are combined in constructing Stein's method for Poisson process approximation. Some of the key estimates given here are sharper than those found elsewhere in the literature, and have simpler proofs. In the final part, we show how to apply the bounds in various examples, from the easiest Bernoulli process to more complicated networks of queues.

2. Poisson point processes

In this section, we briefly review the basic knowledge of Poisson processes on the real line and on a general metric space. For more details, interested readers could refer to [Kingman (1993), Daley & Vere-Jones (1988) and Kallenberg (1976)].

2.1. *Poisson processes on the real line*

Suppose that, at time $t = 0$, we start recording the random times of occurrence $0 \leq T_1 \leq T_2 \leq \cdots \leq T_n \leq \cdots$ of a sequence of events, e.g. customers

who join a queue or the sales of an item. Let $N_t = \#\{i : T_i \leq t\}$ be the number of events that have occurred in time interval $[0,t]$; then $\{N_t\}_{t\geq 0}$ is called a *counting process*. The number of events that have occurred in a time interval $(s,t]$ can be written as $N_t - N_s$, and the latter is called the increment of the counting process over the interval $(s,t]$.

Definition 2.1: The *stationary Poisson process* $\{N_t\}_{t\geq 0}$ is defined as a counting process which has independent increments on disjoint time intervals and such that, for each $t \geq 0$, N_t follows a Poisson distribution with parameter λt, denoted as $\mathrm{Po}(\lambda t)$.

There are a few equivalent ways to view a stationary Poisson process.

Proposition 2.2: *A counting process $\{N_t : t \geq 0\}$ is a stationary Poisson process if and only if*

(a) $N_0 = 0$,
(b) the process has stationary, independent increments,
(c) $\mathbb{P}(N_t \geq 2) = o(t)$ *as* $t \to 0$,
(d) $\mathbb{P}(N_t = 1) = \lambda t + o(t)$ *as* $t \to 0$, *for some constant* λ.

Proposition 2.3: *A counting process $\{N_t\}_{t\geq 0}$ is a stationary Poisson process with rate λ if and only if*

(a) for each fixed t, N_t follows the Poisson distribution with parameter λt;
(b) given that $N_t = n$, the n arrival times $T_1, T_2, \ldots, T_n$ have the same joint distribution as the order statistics of n independent uniformly distributed random variables on $[0,t]$.

Remark 2.4: If the arrival times are ordered as $T_1 < T_2 < \cdots < T_n$, then, given that n arrivals occurred in $[0,t]$, the unordered arrival times (a random permutation of $T_1, T_2, \ldots, T_n$) look like n independent uniformly distributed random variables on $[0,t]$.

2.2. *Poisson point processes*

Let Γ be a locally compact complete separable metric space with the Borel algebra $\mathcal{B}$ and the ring $\mathcal{B}_b$ consisting of all *bounded* (note that a set is bounded if its closure is compact) Borel sets. Such a space is necessarily $\sigma-$compact. A measure μ on $(\Gamma, \mathcal{B})$ is called *locally finite* if $\mu(B) < \infty$ for all $B \in \mathcal{B}_b$. A locally finite measure ξ is then called a *point measure* if $\xi(B) \in \mathbb{Z}_+ := \{0, 1, 2, \cdots\}$ for all $B \in \mathcal{B}_b$.

For each $x \in \Gamma$, let δ_x be the Dirac measure at x, namely

$$\delta_x(B) = \begin{cases} 1, \text{ if } x \in B; \\ 0, \text{ if } x \notin B. \end{cases}$$

Since Γ is $\sigma-$compact and a point measure ξ must be locally finite, it is possible to write $\xi = \sum_{i=1}^{n} \delta_{x_i}$, where $n < \infty$ or $n = \infty$. We use $\mathcal{H}$ to stand for the space of all point measures on $(\Gamma, \mathcal{B}_b)$:

$$\mathcal{H} = \left\{ \sum_{i=1}^{n} \delta_{x_i} : \ n < \infty \text{ or } n = \infty \text{ and } \#\{x_i : \ x_i \in B\} < \infty \text{ for all } B \in \mathcal{B}_b \right\}.$$

For each measure μ on Γ and set $B \in \mathcal{B}$, we use μ, $|\mu|$ or $\mu(\Gamma)$ to stand for the total measure of μ and $\mu|_B$ to represent the restriction of μ to B: $\mu|_B(C) = \mu(B \cap C)$ for all $C \in \mathcal{B}_b$.

For a sequence $\{\xi_n\}_{n \geq 1} \subset \mathcal{H}$, we say that ξ_n converges to $\xi \in \mathcal{H}$ *vaguely* if $\int_\Gamma f(x)\xi_n(dx) \to \int_\Gamma f(x)\xi(dx)$ for all continuous functions on Γ with compact support. The following two Propositions can be found in Kallenberg [1976, pp. 94–95].

Proposition 2.5: *The following statements are equivalent:*

(i) ξ_n *converges to* ξ *vaguely;*
(ii) $\xi_n(B) \to \xi(B)$ *for all* $B \in \mathcal{B}_b$ *whose boundary* ∂B *satisfies* $\xi(\partial B) = 0$*;*
(iii) $\limsup_{n\to\infty} \xi_n(F) \leq \xi(F)$ *and* $\liminf_{n\to\infty} \xi_n(G) \geq \xi(G)$ *for all closed* $F \in \mathcal{B}_b$ *and open* $G \in \mathcal{B}_b$.

Proposition 2.6: *The space* $\mathcal{H}$ *is Polish in the vague topology.*

In other words, there exists some separable and complete metric in $\mathcal{H}$ generating the vague topology. We use $\mathcal{B}(\mathcal{H})$ to stand for the Borel σ-algebra generated by the vague topology.

Definition 2.7: [Kallenberg (1976), p. 5] A point process Ξ is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. The measure $\boldsymbol{\lambda}$ defined by $\boldsymbol{\lambda}(B) = \mathbb{E}\Xi(B)$, $B \in \mathcal{B}_b$, is called the mean measure of Ξ.

We now can generalize the Poisson process concept from the real line to the carrier space Γ.

Definition 2.8: A point process Ξ with locally finite mean measure $\boldsymbol{\lambda}$ is a Poisson process, denoted as $\mathrm{Po}(\boldsymbol{\lambda})$, if

(P1) for any bounded Borel set B, the random variable $\Xi(B)$ is Poisson distributed with mean $\boldsymbol{\lambda}(B)$;

(P2) for any $k \in \mathbb{N} := \{1, 2, \cdots\}$, and disjoint sets $B_1, \cdots, B_k$, the random variables $\Xi(B_1), \cdots, \Xi(B_k)$ are independent.

The following Proposition is a generalization of Proposition 2.3.

Proposition 2.9: *A point process Ξ is a Poisson process with mean measure λ if and only if, for each bounded Borel set B, $\Xi(B) \sim \text{Po}(\lambda(B))$ and*

(P3) given $\Xi(B) = k$, $\Xi|_B$ has the same distribution as $X_k = \sum_{i=1}^{k} \delta_{\theta_i}$, where θ_i, $1 \le i \le k$, are independent and identically distributed B-valued random elements with the common distribution $\lambda|_B/\lambda(B)$.

Proof: Assume that Ξ is a Poisson point process; then, for every bounded Borel set B and every partition of B consisting (bounded) Borel sets $B_1, \cdots, B_m$, and $i_1 + \cdots + i_m = k$,

$$\mathbb{P}\left(\cap_{j=1}^{m}\{\Xi(B_j) = i_j\} \middle| \Xi(B) = k\right) = \frac{\mathbb{P}\left(\bigcap_{j=1}^{m}\{\Xi(B_j) = i_j\}\right)}{\mathbb{P}(\Xi(B) = k)}$$

$$= \frac{k!}{\prod_{j=1}^{m} i_j!} \cdot \frac{\prod_{j=1}^{m} \lambda(B_j)^{i_j}}{\lambda(B)^k} = \mathbb{P}\left(\cap_{j=1}^{m}\{X(B_j) = i_j\}\right).$$

Conversely, assume that (P3) holds. Then, for any $m \in \mathbb{N}$ and bounded Borel sets $B_1, \cdots, B_m$, the set $B := \cup_{i=1}^{m} B_i$ is also bounded, and, writing $k = i_1 + \cdots + i_m$, we have

$$\mathbb{P}\left(\cap_{j=1}^{m}\{\Xi(B_j) = i_j\}\right) = \mathbb{P}\left(\cap_{j=1}^{m}\{\Xi(B_j) = i_j\} \middle| \Xi(B) = k\right)\mathbb{P}(\Xi(B) = k)$$

$$= \frac{k!}{i_1! \cdots i_m!} \cdot \frac{\lambda(B_1)^{i_1} \cdots \lambda(B_m)^{i_m}}{\lambda(B)^k} \cdot \frac{e^{-\lambda(B)}\lambda(B)^k}{k!}$$

$$= \prod_{j=1}^{m} \frac{e^{-\lambda(B_j)}\lambda(B_j)^{i_j}}{i_j!}. \qquad \blacksquare$$

Thus, to construct a Poisson point process on Γ, since Γ is $\sigma-$compact, one can partition the space Γ into at most countably many bounded subsets Γ_i satisfying $0 < \lambda(\Gamma_i) < \infty$. For each i, one can then independently define a Poisson process using Proposition 2.9, and the union of these independent Poisson processes is the desired Poisson process [see Reiss (1993), p. 46].

2.3. *Characterization of Poisson point processes*

From the structure and properties of Poisson processes, it is possible to find many ways to characterize a Poisson process on a Polish space. If the

process is simple, i.e., $\mathbb{P}(\Xi\{x\} = 0 \text{ or } 1, \text{ for all } x) = 1$, we can even use the avoidance function or one-dimensional distribution only [Renyi (1967), Brown & Xia (2002)]. However, we will concentrate on the following three tools: Palm distributions, Janossy densities and compensators.

2.3.1. *Palm theory*

For a non-negative integer-valued random variable X, $X \sim \text{Po}(\lambda)$ if and only if

$$\frac{\mathbb{E}g(X)X}{\mathbb{E}X} = \frac{\mathbb{E}g(X)X}{\lambda} = \mathbb{E}g(X+1) \text{ for all bounded function } g \text{ on } \mathbb{Z}_+. \tag{2.1}$$

Equation (2.1) completely characterizes the Poisson distribution and is the starting point for Stein's identity. Its counterpart for the Poisson point process is based on Palm theory, which was developed from the work of Palm (1943), see Daley & Vere-Jones (1988), p. 13, for a brief history.

Heuristically, since a Poisson process is a point process with independent increments and its one dimensional distributions are Poisson, we may imagine that a Poisson process is pieced together by lots of independent "Poisson components" (if the location is an atom, the "component" will be a Poisson random variable, but if the location is diffuse, then the "component" is either 0 or 1). Hence, to specify a Poisson process, one needs to check that "each component" is Poisson, by virtue of (2.1), and also independent of the rest. Intuitively, this can be done by checking that

$$\frac{\mathbb{E}g(\Xi)\Xi(d\alpha)}{\mathbb{E}\Xi(d\alpha)} = \mathbb{E}g(\Xi + \delta_\alpha), \tag{2.2}$$

for all bounded function g on $\mathcal{H}$ and all $\alpha \in \Gamma$. For example, Ξ is a stationary Poisson process on $\mathbb{R}_+ := [0, \infty)$ if and only if, for each $s \in \mathbb{R}_+$ and bounded function g on $\mathcal{H}$,

$$\lim_{\Delta s \to 0} \mathbb{E}\{g(\Xi) | \Xi[s, s+\Delta s) > 0\} = \mathbb{E}g(\Xi + \delta_s).$$

To make the argument rigorous, one needs the tools of Campbell measures and Radon-Nikodym derivatives [Kallenberg (1976), p. 69].

In general, for each point process Ξ with locally finite mean measure $\boldsymbol{\lambda}$, we may define probability measures $\{P_\alpha, \ \alpha \in \Gamma\}$, called *Palm distributions*, on $\mathcal{B}(\mathcal{H})$ [Kallenberg (1976), p. 69] satisfying

$$P_\alpha(\mathbf{B}) = \frac{\mathbb{E}\left[1_{[\Xi \in \mathbf{B}]}\Xi(d\alpha)\right]}{\boldsymbol{\lambda}(d\alpha)}, \ \alpha \in \Gamma \ \boldsymbol{\lambda} - \text{almost surely}, \ \mathbf{B} \in \mathcal{B}(\mathcal{H}).$$

A point process Ξ_α on Γ is called a *Palm process* of Ξ at location α if it has the Palm distribution P_α of Ξ at α. Since the Palm process of a Poisson process has the same distribution as the original process, except for the one additional point added at α, it is more convenient to work with $\Xi_\alpha - \delta_\alpha$, which is called the *reduced Palm process* [Kallenberg (1976), p. 70]. Moreover, for any measurable function $f : \Gamma \times \mathcal{H} \to \mathbb{R}_+$,

$$\mathbb{E}\left(\int_B f(\alpha, \Xi)\Xi(d\alpha)\right) = \mathbb{E}\left(\int_B f(\alpha, \Xi_\alpha)\boldsymbol{\lambda}(d\alpha)\right) \tag{2.3}$$

$$\mathbb{E}\left(\int_B f(\alpha, \Xi - \delta_\alpha)\Xi(d\alpha)\right) = \mathbb{E}\left(\int_B f(\alpha, \Xi_\alpha - \delta_\alpha)\boldsymbol{\lambda}(d\alpha)\right) \tag{2.4}$$

for all Borel set $B \subset \Gamma$.

Palm distributions are a powerful tool in point process theory [Franken, König, Arndt & Schmidt (1982), Kallenberg (1976), Karr (1986) and Daley & Vere-Jones (1988)]. One can even use the equation (2.3) to establish the Stein identity for Poisson process approximation [Chen & Xia (2004)]. For a characterization of Poisson processes, we summarize the heuristic analysis in the following theorem.

Theorem 2.10: [cf Kallenberg (1976), Theorem 11.5, p. 82] *A point process Ξ with locally finite mean measure $\boldsymbol{\lambda}$ is a Poisson process if and only if its reduced Palm processes have the same distribution as that of the original process.*

2.3.2. *Janossy densities*

For Ξ a Poisson process on Γ with a finite mean measure $\boldsymbol{\lambda}$, let $\boldsymbol{\nu}(d\alpha) = \boldsymbol{\lambda}(d\alpha)/\lambda$. Then, by Proposition 2.9, for each bounded measurable function f on $\mathcal{H}$,

$$\mathbb{E}f(\Xi) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} \int_{\Gamma^n} f\left(\sum_{j=1}^{n} \delta_{\alpha_j}\right) \boldsymbol{\nu}^n(d\alpha_1, \cdots, d\alpha_n), \tag{2.5}$$

where $\Gamma^n := \overbrace{\Gamma \times \cdots \times \Gamma}^{n \text{ times}}$. There is a similar expression for general point processes, and the expression is in terms of the so-called Janossy densities, named after their first introduction in Janossy (1950) in the context of particle showers [see Daley & Vere-Jones (1988), p. 122].

To introduce the Janossy density, it is necessary to make a few assumptions:

(JA1) the point process Ξ is finite with distribution $R_n = \mathbb{P}(|\Xi| = n)$, $\sum_{n=0}^{\infty} R_n = 1$.

(JA2) Given the total number of points equals $n \geq 1$, there is a probability distribution $\Pi_n(\cdot)$ on Γ^n which determines the joint distribution of the positions of the points of the point process Ξ.

From these assumptions, we can see that

$$\begin{aligned}\mathbb{E}f(\Xi) &= \sum_{n=0}^{\infty} R_n \mathbb{E}[f(\Xi)||\Xi| = n] \\ &= \sum_{n=0}^{\infty} R_n \int_{\Gamma^n} f\left(\sum_{j=1}^{n} \delta_{\alpha_j}\right) \Pi_n(d\alpha_1, \cdots, d\alpha_n).\end{aligned}$$

As we are working with general point processes, there is no preferred order of the points, so we may replace Π_n by its symmetrized version

$$\Pi_n^s(d\alpha_1, \cdots, d\alpha_n) = \frac{1}{n!} \sum_{\pi} \Pi_n(d\alpha_{\pi(1)}, \cdots, d\alpha_{\pi(n)}),$$

and then write

$$\begin{aligned}J_n(d\alpha_1, \cdots, d\alpha_n) &= R_n \sum_{\pi} \Pi_n(d\alpha_{\pi(1)}, \cdots, d\alpha_{\pi(n)}) \\ &= n! R_n \Pi_n^s,\end{aligned}$$

where the sums range over all permutations π of $(1, \cdots, n)$. The measures $\{J_n\}$ are called *Janossy measures*. If further we can find a reference measure μ such that J_n is absolutely continuous with respect to the product measure μ^n on Γ^n, we denote its Radon-Nikodym derivative by j_n; then, for any measurable function $f : \mathcal{H} \to \mathbb{R}_+$,

$$\mathbb{E}(f(\Xi)) = \sum_{n \geq 0} \frac{1}{n!} \int_{\Gamma^n} f\left(\sum_{i=1}^{n} \delta_{\alpha_i}\right) j_n(\alpha_1, \cdots, \alpha_n) \mu^n(d\alpha_1, \cdots, d\alpha_n). \tag{2.6}$$

The derivatives $\{j_n\}$ are called *Janossy densities*. One may compare (2.5) and (2.6) to reach the following conclusion.

Theorem 2.11: *Ξ is a Poisson point process with finite mean measure λ if and only if, with respect to λ, its Janossy densities j_n are constant and equal to $e^{-\lambda}$ for all $n \in \mathbb{Z}_+$.*

Using the Janossy densities, we can express the density of the mean measure $\boldsymbol{\lambda}$ of the point process Ξ with respect to $\boldsymbol{\mu}$ as

$$\phi(\alpha) = \sum_{n\geq 0} \int_{\Gamma^n} (n!)^{-1} j_{n+1}(\alpha, \alpha_1, \cdots, \alpha_n) \boldsymbol{\mu}^n(d\alpha_1, \cdots, d\alpha_n). \tag{2.7}$$

In fact, by (2.6), for each $B \in \mathcal{B}_b$,

$$\begin{aligned}
\boldsymbol{\lambda}(B) = \mathbb{E}\Xi(B) &= \sum_{n\geq 0} \int_{\Gamma^n} \frac{1}{n!} \sum_{i=1}^{n} \delta_{\alpha_i}(B) j_n(\alpha_1, \cdots, \alpha_n) \boldsymbol{\mu}^n(d\alpha_1, \cdots, d\alpha_n) \\
&= \sum_{n\geq 1} \int_B \int_{\Gamma^{n-1}} \frac{1}{(n-1)!} j_n(\alpha_1, \cdots, \alpha_n) \boldsymbol{\mu}^n(d\alpha_1, \cdots, d\alpha_n) \\
&= \int_B \sum_{m\geq 0} \int_{\Gamma^m} \frac{1}{m!} j_{m+1}(\alpha, \alpha_1, \cdots, \alpha_m) \boldsymbol{\mu}^m(d\alpha_1, \cdots, d\alpha_m) \boldsymbol{\mu}(d\alpha),
\end{aligned}$$

which implies (2.7).

The Janossy densities can also be used to describe the density of a point being at α, given the configuration Ξ^α of Ξ outside $N_\alpha \in \mathcal{B}$ with $\alpha \in N_\alpha$. To achieve this, let $m \in \mathbb{N}$ be fixed and $B_1, \cdots, B_m$ be bounded Borel subsets of N_α^c, the complement of N_α. Write $\boldsymbol{x}_n = (x_1, \cdots, x_n)$ and $\boldsymbol{y}_m = (y_1, \cdots, y_m)$. Set

$$\mathbf{B} = \left\{ \boldsymbol{\delta}_{\boldsymbol{y}_m} : \boldsymbol{y}_m \in B_1 \times \cdots \times B_m \right\}$$

and

$$\mathbf{B}_e = \left\{ \boldsymbol{\delta}_{\boldsymbol{x}_n} + \boldsymbol{\delta}_{\boldsymbol{y}_m} : \boldsymbol{x}_n \in N_\alpha^n, n \geq 0;\ \boldsymbol{y}_m \in B_1 \times \cdots \times B_m \right\},$$

where $\boldsymbol{\delta}_{\boldsymbol{x}_n} := \sum_{i=1}^n \delta_{x_i}$ and $\boldsymbol{\delta}_{\boldsymbol{y}_m} := \sum_{i=1}^m \delta_{y_j}$. Then

$$\begin{aligned}
\mathbb{P}\left(\Xi|_{(N_\alpha)^c} \in \mathbf{B}\right) = \mathbb{E}1_{\mathbf{B}_e}(\Xi) &= \sum_{k\geq 0} \int_{\Gamma^k} \frac{1}{k!} 1_{\mathbf{B}_e}\left(\boldsymbol{\delta}_{\boldsymbol{x}_k}\right) j_k(\boldsymbol{x}_k) \boldsymbol{\mu}^k(d\boldsymbol{x}_k) \\
&= \int_{B_1 \times \cdots \times B_m} \sum_{n\geq 0} \int_{N_\alpha^n} \frac{1}{n!} j_{m+n}(\boldsymbol{y}_m, \boldsymbol{x}_n) \boldsymbol{\mu}^n(d\boldsymbol{x}_n) \boldsymbol{\mu}^m(d\boldsymbol{y}_m),
\end{aligned}$$

so for $\boldsymbol{y}_m \in (N_\alpha^c)^m$, the density of $\Xi|_{(N_\alpha)^c}$ at $\boldsymbol{\delta}_{\boldsymbol{y}_m}$ is

$$\sum_{s\geq 0} \int_{N_\alpha^s} j_{m+s}(\boldsymbol{y}_m, \boldsymbol{x}_s)(s!)^{-1} \boldsymbol{\mu}^s(d\boldsymbol{x}_s).$$

Similarly, the density of $\Xi|_{(N_\alpha)^c \cup \{\alpha\}}$ at $\delta_\alpha + \boldsymbol{\delta}_{\boldsymbol{y}_m}$ is

$$\sum_{r\geq 0} \int_{N_\alpha^r} j_{1+m+r}(\alpha, \boldsymbol{y}_m, \boldsymbol{x}_r)(r!)^{-1} \boldsymbol{\mu}^r(d\boldsymbol{x}_r),$$

so the conditional density of a point being at α given $\Xi|_{(N_\alpha)^c} = \boldsymbol{\delta}_{\boldsymbol{y}_m}$ is

$$g(\alpha, \boldsymbol{y}_m) := \frac{\sum_{r\geq 0} \int_{N_\alpha^r} j_{1+m+r}(\alpha, \boldsymbol{y}_m, \boldsymbol{x}_r)(r!)^{-1}\boldsymbol{\mu}^r(d\boldsymbol{x}_r)}{\sum_{s\geq 0} \int_{N_\alpha^s} j_{m+s}(\boldsymbol{y}_m, \boldsymbol{x}_s)(s!)^{-1}\boldsymbol{\mu}^s(d\boldsymbol{x}_s)}, \tag{2.8}$$

where the term with $r = 0$ is interpreted as $j_{m+1}(\alpha, \boldsymbol{y}_m)$ and the term with $s = 0$ similarly.

2.3.3. *Compensator*

When a point process is defined on the time axis, $\mathbb{R}_+$, its properties are most often expressed in terms of its evolution over time with respect to a right-continuous filtration $\mathcal{F} = (\mathcal{F}_s)_{s\geq 0}$, its counting process $\{N_s\}_{s\geq 0}$ being taken to be $\mathcal{F}$–adapted and right continuous. The point process $(N, \mathcal{F})$ is then called simple if the jumps $\Delta N_s := N_s - N_{s-}$ can take only the values 0 or 1, $\forall s \geq 0$ almost surely. The compensator A of $(N, \mathcal{F})$ is the unique previsible right-continuous increasing process such that $N - A$ is an $(\mathcal{F}_s)_{s\geq 0}$ local martingale [Jacod & Shiryaev (1987), p. 32 or Dellacherie & Meyer (1982)].

Note that the Janossy densities and the Palm distributions involve the joint distribution of the process outside and inside an interval, and hence require knowledge of the conditional evolution backwards in time as well as forwards, something not easy to deduce from martingale characteristics, while the compensator needs only the forward conditional evolution in time. A well-known fact is that the compensator of a Poisson process with respect to its intrinsic filtration is a deterministic function, and in this case the Poisson process is simple if and only if its compensator is continuous. It is also well-known that any simple point process with continuous compensator is locally Poisson in character, in the sense that there exists a time transformation that converts the process into a Poisson process. More precisely, the transformation is given by the inverse of the compensator A of the simple point process $\{N_t\}_{t\geq 0}$:

$$\sigma(t) := \inf\{s : A_s > t\}.$$

The continuity of A ensures that $\sigma(t)$ is an $(\mathcal{F}_t)_{t\geq 0}$ stopping time, and in the case where $\lim_{t\uparrow\infty} A_t = \infty$ almost surely, the transformed process $\bar{N}_t := N_{\sigma(t)}$ is a Poisson process with unit rate with respect to filtration $(\bar{\mathcal{F}}_t)_{t\geq 0}$, where $\bar{\mathcal{F}}_t = \mathcal{F}_{\sigma(t)}$ [see Liptser & Shiryaev (1978), Theorem 18.10]. Thus, if M is a Poisson process with compensator $B_t = t$, $0 \leq t < \infty$, then $(N, \bar{N})$ is a natural coupling of the distributions of M and N.

3. Immigration-death point processes

3.1. *Immigration-death processes*

Consider a process whose state X_t at time t is the number of objects in a system at that time. Suppose that arrivals and departures occur independently of one another, but at rates depending on the state of the process. Thus, when there are n objects in the system, new arrivals enter according to a Poisson process of rate $\lambda > 0$, and each of the n objects in the system departs after spending a lifetime having negative exponential distribution NE(1) with mean 1 and independent of the others: we call this departure behaviour *unit per capita death rate.* Hence, when there are n objects in the system, the time to the next arrival or departure is negative exponentially distributed with mean $1/(\lambda + n)$, the probability of its being an arrival being $\lambda/(\lambda + n)$, and of a departure $n/(\lambda + n)$. The process $\{X_t; t \geq 0\}$ is called an immigration-death process, and the parameters λ and $\{n\}$ are the arrival (or immigration) rates and departure (or death) rates. It is a continuous time Markov chain with state space $\mathbb{Z}_+$, and in which a transition from state n can only be to one of its adjoining states $n+1$ or $n-1$. It is well-known that this Markov chain is ergodic, and that its equilibrium distribution is a Poisson distribution with mean λ:

$$\pi_j := \mathrm{Po}(\lambda)\{j\} = \frac{e^{-\lambda}\lambda^j}{j!},\ j \in \mathbb{Z}_+.$$

The idea of using Markov immigration-death processes to interpret the Stein-Chen method was initiated by Barbour (1988). Brown & Xia (2001) extended the idea to start with an approximating distribution and then search for an immigration-death process with the approximating distribution as equilibrium distribution to establish a Stein identity.

For each $i \in \mathbb{Z}_+$, let Z_i be a copy of $\{X_t; t \geq 0\}$ with initial state i. There are two methods to couple Z_i and Z_j with $i \neq j$. The first is based on *coupling by the states*: let $(Z_i(t), Z_j(t))$ run in certain way, for instance independently, until they meet, at which point they are *coupled*, and have identical paths thereafter. The other method is to *couple by the time*: let Z_i run until the first time τ_{ij} that it hits j, then make it follow the Z_j-path thereafter, so that $Z_i(\tau_{ij} + t) = Z_j(t)$ for all $t > 0$. The latter approach was first used by Xia (1999) to estimate Stein's factors for Poisson random variable approximation, and was fully exploited for a large class of approximating distributions, named polynomial birth-death distributions,

in Brown & Xia (2001). We use the notation of Brown & Xia (2001), setting

$$\tau_{ij} = \inf\{t : Z_i(t) = j\},\ \tau_j^+ = \tau_{j,j+1},\ \tau_j^- = \tau_{j,j-1}, \tag{3.1}$$

for $i, j \in \mathbb{Z}_+$, and

$$\overline{\tau_j^+} = \mathbb{E}\left(\tau_j^+\right);\ \overline{\tau_j^-} = \mathbb{E}\left(\tau_j^-\right).$$

Lemma 3.1: *For* $j \in \mathbb{Z}_+$,

$$\overline{\tau_j^+} = \frac{F(j)}{\lambda\pi_j} \quad \text{and} \quad \overline{\tau_j^-} = \frac{\bar{F}(j)}{\lambda\pi_{j-1}}, \tag{3.2}$$

where

$$F(j) = \sum_{i=0}^{j} \pi_i;\ \bar{F}(j) = \sum_{i=j}^{\infty} \pi_i.$$

Proof: [cf Keilson (1979), p. 61 and Wang & Yang (1992)] We prove (3.2) by conditioning on the time of the first jump after leaving the initial state, and producing a recurrence relation using the fact that the time of transition between states which are two apart is the sum of the times of transition between two pairs of adjoining states. More precisely, defining the first jump time $L_i := \inf\{t : Z_i(t) \neq i\}$, we have

$$\begin{aligned}
\overline{\tau_i^+} &= \mathbb{E}(L_i) + \mathbb{E}(\tau_i^+ - L_i | Z_i(L_i) = i-1)\mathbb{P}(Z_i(L_i) = i-1) \\
&\quad + \mathbb{E}(\tau_i^+ - L_i | Z_i(L_i) = i+1)\mathbb{P}(Z_i(L_i) = i+1) \\
&= \frac{1}{\lambda + i} + \mathbb{E}(\tau_{i-1,i+1})\frac{i}{\lambda + i} \\
&= \frac{1}{\lambda + i} + (\overline{\tau_{i-1}^+} + \overline{\tau_i^+})\frac{i}{\lambda + i},
\end{aligned}$$

where the second equation is due to the strong Markov property and the last equation is due to the fact that $\tau_{i-1,i+1} = \tau_{i-1}^+ + \tau_i^+$. This implies that

$$\lambda\overline{\tau_i^+} = 1 + i\overline{\tau_{i-1}^+}. \tag{3.3}$$

Multiplying both sides of (3.3) by π_i and using the fact that $i\pi_i = \lambda\pi_{i-1}$, it follows that

$$\lambda\pi_i\overline{\tau_i^+} - \lambda\pi_{i-1}\overline{\tau_{i-1}^+} = \pi_i,$$

which yields $\lambda\pi_j\overline{\tau_j^+} = \sum_{i=0}^{j}\pi_i$ and so $\overline{\tau_j^+} = \frac{F(j)}{\lambda\pi_j}$.

Similarly, using the fact that $\tau_{i+1,i-1} = \tau_{i+1}^- + \tau_i^-$, we have

$$\begin{aligned}\overline{\tau_i^-} &= \mathbb{E}(L_i) + \mathbb{E}(\tau_i^- - L_i | Z_i(L_i) = i-1)\mathbb{P}(Z_i(L_i) = i-1) \\ &\quad + \mathbb{E}(\tau_i^- - L_i | Z_i(L_i) = i+1)\mathbb{P}(Z_i(L_i) = i+1) \\ &= \frac{1}{\lambda+i} + \mathbb{E}(\tau_{i+1,i-1})\frac{\lambda}{\lambda+i} \\ &= \frac{1}{\lambda+i} + (\overline{\tau_{i+1}^-} + \overline{\tau_i^-})\frac{\lambda}{\lambda+i},\end{aligned}$$

hence

$$i\overline{\tau_i^-} = 1 + \lambda\overline{\tau_{i+1}^-}. \tag{3.4}$$

We now multiply both sides of (3.4) by π_i and replace $\lambda\pi_i$ by $(i+1)\pi_{i+1}$, then sum up for i from j to ∞, the claim for $\overline{\tau_j^-}$ follows. ■

Lemma 3.2: *$\overline{\tau_j^+}$ is increasing in j and $\overline{\tau_j^-}$ is decreasing in j.*

Proof: The claims are intuitively obvious, as the immigration-death process has constant birth rate and unit per capita death rate, and so the time Z needs to move from state $i+1$ to state i should be shorter than the time it takes to move from state i to state $i-1$ and the time for Z to move from $i-1$ to i should be less than the time from i to $i+1$. To make the argument mathematically rigorous, noting that

$$\frac{\pi_i}{\pi_j} \le \frac{\pi_{i+1}}{\pi_{j+1}} \text{ for all } i \le j, \text{ and that } \frac{\pi_i}{\pi_{j-1}} > \frac{\pi_{i+1}}{\pi_j} \text{ for all } i \ge j,$$

we have

$$\frac{F(j)}{\pi_j} \le \frac{\pi_1 + \cdots + \pi_{j+1}}{\pi_{j+1}} < \frac{F(j+1)}{\pi_{j+1}} \tag{3.5}$$

and

$$\frac{\bar{F}(j)}{\pi_{j-1}} = \frac{\pi_j + \pi_{j+1} + \cdots}{\pi_{j-1}} > \frac{\pi_{j+1} + \pi_{j+2} + \cdots}{\pi_j} = \frac{\bar{F}(j+1)}{\pi_j},$$

hence the claims follow from Lemma 3.1. ■

Lemma 3.3: *With the above setup, we have*

$$Z_n(t) = D_n(t) + Z_0(t),$$

where $D_n(t)$ is a pure death process with unit per capita death rate and independent of $Z_0(t)$. Moreover, for each $t \ge 0$, $D_n(t) \sim B(n, e^{-t})$ and $Z_0(t) \sim \text{Po}(\lambda_t)$ with $\lambda_t := \lambda(1-e^{-t})$.

The lemma is quoted from [Theorem 2, Wang & Yang (1992), page 190]. The proof utilizes the Kolmogrov's forward equations and works on the probability generating function of $Z_n(t)$ to conclude that it is the same as the probability generating function of the sum of independent binomial random variable having distribution $B(n, e^{-t})$ and Poisson random variable having distribution $\mathrm{Po}(\lambda_t)$. It is a special case of Proposition 3.5 when the carrier space Γ in the spatial immigration-death process consists of only one point, so we omit the proof.

3.2. *Spatial immigration-death processes*

The structure of the immigration-death process enables us to represent the Stein's equation in probabilistic terms. This idea has been used in studying Poisson random variable approximation since Barbour (1988). It enables probabilistic formulae for the solutions to the Stein equation to be derived, and these in turn can be used to determine Stein's factors. The immigration-death process is chosen to have the distribution being used as approximation, here $\mathrm{Po}(\lambda)$, as the equilibrium distribution, and solutions to the Stein equation can then be written in terms of differences of expectations under different initial conditions. This enables techniques of stochastic processes, such as coupling, to be exploited.

To study Poisson process approximation, we need not only the total number of events occurring, but also the locations at which they occur. This leads to an immigration-death process with birth and death rates depending on the positions of the individuals in the carrier space. Such a process is necessary $\mathcal{H}$-valued instead of integer-valued and we call it a *spatial immigration-death process* [Preston (1975)]. If we consider the total number of individuals only, a spatial immigration-death process is reduced to an integer-valued immigration-death process.

There is also some connection between the Markov chain Monte Carlo approach and Stein's method with the Markov process interpretation. Monte Carlo methods introduce random samples as a computational tool when certain characteristics of the problem under study are intractable. Such procedures are used in a wide range of problems, such as the evaluation of intractable integrals, optimization problems, statistical inference and statistical mechanics [see Fishman (1996), Sokal (1989) and Tierney (1994) and the references therein]. In the study of statistical inference problems in Markov point processes, including Poisson point processes, binomial point processes and hard core processes, one of the popular techniques

for simulating the process is to run a spatial immigration-death process [see Lieshout (2000), pp. 85-89 for more details]. Stein's method with a Markov process interpretation, on the other hand, utilizes coupling properties of the immigration-death process as a tool in bounding the errors of approximation.

To avoid technical difficulties, taking a subspace if necessary, we assume from now on that Γ is a compact metric space. We assume $\boldsymbol{\lambda}$ to be a finite measure on Γ, the mean measure of the approximating Poisson process. There are infinitely many spatial immigration-death processes with equilibrium distribution $\mathrm{Po}(\boldsymbol{\lambda})$. The one we use is as follows. Given that the process takes a configuration $\xi \in \mathcal{H}$, the process stays in state ξ for a negative exponentially distributed random sojourn time with mean $1/(|\xi|+\lambda)$ with $\lambda = \boldsymbol{\lambda}(\Gamma)$; then, with probability $\lambda/(|\xi|+\lambda)$, a new point is added to the existing configuration, its position in Γ chosen from the distribution $\boldsymbol{\lambda}/\lambda$, independently of the existing configuration; and with probability $|\xi|/(|\xi|+\lambda)$, a point is chosen uniformly at random from the existing configuration ξ, and is deleted from the system. This is equivalent to say that each point in ξ has an independent, negative exponentially distributed lifetime with mean 1, again referred to as *unit per capita death rate*. Such a spatial immigration-death process, denoted by $\mathbf{Z}_\xi(t)$, can also be constructed as follows. First, define the transition kernel $\mu : \mathcal{H} \times \mathcal{B}(\mathcal{H}) \to \mathbb{R}_+$ as

$$\mu(\eta, B) = \frac{1}{\lambda + |\eta|}\left(\int_\Gamma 1_B(\eta + \delta_x)\boldsymbol{\lambda}(dx) + \int_\Gamma 1_B(\eta - \delta_x)\eta(dx)\right),$$

and construct an $\mathcal{H}$-valued Markov chain $\{Y_k\}_{k\geq 0}$ with initial distribution ξ and transition kernel $\mu(\eta, B)$. Next, construct ζ_i, $i \geq 0$, as independent and identically distributed negative exponential random variables with mean 1 such that $\{\zeta_i\}_{i\geq 0}$ is independent of $\{Y_k\}_{k\geq 0}$. Set

$$\mathbf{Z}_\xi(t) = \begin{cases} Y_0, \ 0 \leq t < \frac{\zeta_0}{\lambda + |Y_0|}, \\ Y_k, \ \sum_{j=0}^{k-1} \frac{\zeta_j}{\lambda+|Y_j|} \leq t < \sum_{j=0}^{k} \frac{\zeta_j}{\lambda+|Y_j|}, \quad k \in \mathbb{N}, \end{cases}$$

[Ethier & Kurtz (1986), pp. 162-163 and Problem 5, page 262]. The spatial immigration-death process has generator $\mathcal{A}$ specified by

$$\mathcal{A}h(\xi) = \int_\Gamma [h(\xi + \delta_\alpha) - h(\xi)]\boldsymbol{\lambda}(d\alpha) + \int_\Gamma [h(\xi - \delta_\alpha) - h(\xi)]\xi(d\alpha), \quad \forall \xi \in \mathcal{H} \tag{3.6}$$

for suitable functions h on $\mathcal{H}$. To prove that the spatial immigration-death process constructed is the unique spatial Markov process determined by the generator, let $Q_t(\xi, B) = \mathbb{P}(\mathbf{Z}_\xi(t) \in B)$. Then it is possible to check

that the transition semigroup Q_t is the unique solution of the Kolmogorov backward equations [Proposition 5.1, Preston (1975)]:

$$\frac{\partial}{\partial t}Q_t(\xi, B) = -(\lambda + |\xi|)Q_t(\xi, B) + (\lambda + |\xi|)\int_{\mathcal{H}} Q_t(\eta, B)\mu(\xi, d\eta) \quad (3.7)$$

with initial conditions

$$Q_0(\xi, B) = 1_B(\xi). \quad (3.8)$$

The corresponding Kolmogorov forward equations are

$$\frac{\partial}{\partial t}Q_t(\xi, B) = -\int_B (\lambda + |\eta|)Q_t(\xi, d\eta) + \int_{\mathcal{H}} (\lambda + |\eta|)\mu(\eta, B)Q_t(\xi, d\eta) \quad (3.9)$$

with the same initial conditions (3.8) [Preston (1975), p. 374]. The right hand side of (3.9) is seen, after some reorganization, to be $\mathbb{E}\mathcal{A}1_B(\mathbf{Z}_\xi(t))$.

If Ξ is a Poisson process with mean measure $\boldsymbol{\lambda}$, then it follows from (2.4) and Theorem 2.10 that for every bounded function h on $\mathcal{H}$,

$$\mathbb{E}\int_\Gamma [h(\Xi - \delta_\alpha) - h(\Xi)]\Xi(d\alpha) = \mathbb{E}\int_\Gamma [h(\Xi) - h(\Xi + \delta_\alpha)]\boldsymbol{\lambda}(d\alpha),$$

which in turn implies that

$$\mathbb{E}\mathcal{A}h(\Xi) = 0. \quad (3.10)$$

This fact, together with [Theorem 7.1, Preston (1975)] gives the following Proposition.

Proposition 3.4: *The spatial immigration-death process has a unique equilibrium distribution* $\mathrm{Po}(\boldsymbol{\lambda})$ *to which it converges in distribution from any initial state.*

Moreover, one may argue that Ξ is a Poisson process with mean measure $\boldsymbol{\lambda}$ if and only if (3.10) holds for a sufficiently rich class of bounded functions h.

We now state another important property of the spatial immigration-death process which will be needed when we estimate Stein's factors for Poisson process approximation.

Proposition 3.5: *Let* $\xi = \sum_{i=1}^{|\xi|} \delta_{z_i}$, *and let* $\zeta_1, \zeta_2, \cdots, \zeta_{|\xi|}$ *be independent negative exponentially distributed random variables with mean 1, also independent of* $\mathbf{Z}_0$. *Then*

$$\mathbf{Z}_\xi(t) =_d \mathbf{D}_\xi(t) + \mathbf{Z}_0(t), \quad (3.11)$$

where $\mathbf{D}_\xi(t) = \sum_{i=1}^{|\xi|} \delta_{z_i} 1_{\zeta_i > t}$ *is a pure death process. Note that* $\mathbf{Z}_0(t)$ *is a Poisson process with mean measure* $(1 - e^{-t})\boldsymbol{\lambda}$.

Proof: Define $\mathcal{Q}_t(\xi, h) := \mathbb{E}h(\mathbf{D}_\xi(t) + \mathbf{Z}_0(t))$. As the solution to the Kolmogorov backward equations (3.7) is unique, it suffices to show that, for every bounded function h on $\mathcal{H}$,

$$\frac{\partial}{\partial t}\mathcal{Q}_t(\xi, h) = -(\lambda + |\xi|)\mathcal{Q}_t(\xi, h) + (\lambda + |\xi|)\int_{\mathcal{H}} \mathcal{Q}_t(\eta, h)\mu(\xi, d\eta). \quad (3.12)$$

Let $\boldsymbol{\tau} := \inf\{t : \mathbf{Z}_\xi(t) \neq \xi\}$ be the time of the first jump of $\mathbf{Z}_\xi$. Then it follows that $\boldsymbol{\tau} \sim \mathrm{NE}\,(\lambda+|\xi|)$, where $\mathrm{NE}\,(\mu)$ denotes the negative exponential distribution with mean $1/\mu$, and, conditioning on the time and outcome of the first jump, we obtain

$$\begin{aligned}
&e^{(\lambda+|\xi|)t}\mathcal{Q}_t(\xi, h)\\
&= e^{(\lambda+|\xi|)t}\int_0^\infty (\lambda + |\xi|)\mathbb{E}\left(h(\mathbf{D}_\xi(t) + \mathbf{Z}_0(t)) \,|\, \boldsymbol{\tau} = s\right) e^{-(\lambda+|\xi|)s} ds\\
&= h(\xi) + e^{(\lambda+|\xi|)t}\int_0^t (\lambda + |\xi|)\mathbb{E}\left(\mathcal{Q}_{t-s}(\mathbf{Z}_\xi(\boldsymbol{\tau}), h) \,|\, \boldsymbol{\tau} = s\right) e^{-(\lambda+|\xi|)s} ds\\
&= h(\xi) + \int_0^t (\lambda + |\xi|)\mathbb{E}\left(\mathcal{Q}_v(\mathbf{Z}_\xi(\boldsymbol{\tau}), h) \,|\, \boldsymbol{\tau} = s\right) e^{(\lambda+|\xi|)v} dv\\
&= h(\xi) + \int_0^t e^{(\lambda+|\xi|)v} dv \left\{\int_\Gamma \mathcal{Q}_v(\xi + \delta_x, h)\boldsymbol{\lambda}(dx) + \sum_{i=1}^{|\xi|} \mathcal{Q}_v(\xi - \delta_{z_i}, h)\right\}.
\end{aligned}$$

Hence $\mathcal{Q}_t(\xi, h)$ is differentiable in terms of t, and we differentiate on both sides and cancel the exponential factor, giving

$$(\lambda + |\xi|)\mathcal{Q}_t(\xi, h) + \frac{\partial}{\partial t}\mathcal{Q}_t(\xi, h) = \int_\Gamma \mathcal{Q}_t(\xi + \delta_x, h)\boldsymbol{\lambda}(dx) + \sum_{i=1}^{|\xi|} \mathcal{Q}_t(\xi - \delta_{z_i}, h),$$

which is the same as (3.12). ∎

4. Poisson process approximation by coupling

4.1. *Metrics for quantifying the weak topology in $\mathcal{H}$*

Since we assume Γ is compact, the vague topology is the same as the weak topology defined as follows. We say a sequence $\{\xi_n\} \subset \mathcal{H}$ converges to $\xi \in \mathcal{H}$ *weakly* if $\int_\Gamma f(x)\xi_n(dx) \to \int_\Gamma f(x)\xi(dx)$ for all bounded continuous functions on Γ.

There are a few metrics to quantify the weak topology. The one we shall use is a type of Wasserstein metric on $\mathcal{H}$ introduced in Barbour & Brown (1992a). Assume Γ is equipped with a metric $d_0 \leq 1$. Let $\mathcal{K}$ stand for the

set of functions $k : \Gamma \to [0,1]$ such that

$$s_1(k) = \sup_{y_1 \neq y_2 \in \Gamma} \frac{|\, k(y_1) - k(y_2)\,|}{d_0(y_1, y_2)} \leq 1$$

and define

$$d_1(\xi_1, \xi_2) = \begin{cases} 1, & \text{if } \xi_1(\Gamma) \neq \xi_2(\Gamma), \\ \frac{1}{m} \sup_{k \in \mathcal{K}} \left| \int_\Gamma k(x)\xi_1(dx) - \int_\Gamma k(x)\xi_2(dx) \right|, & \text{if } \xi_1(\Gamma) = \xi_2(\Gamma) = m > 0. \end{cases}$$

By the Kantorovich-Rubinstein duality theorem [Rachev (1991), Theorem 8.1.1, p. 168], if $\xi_1 = \sum_{i=1}^m \delta_{y_i}$, $\xi_2 = \sum_{i=1}^m \delta_{z_i}$, then $d_1(\xi_1, \xi_2)$ can be interpreted as the average distance between a best coupling of the points of ξ_1 and ξ_2:

$$d_1(\xi_1, \xi_2) = m^{-1} \min_\pi \sum_{i=1}^m d_0(y_i, z_{\pi(i)}),$$

where π ranges over all permutations of $(1, \cdots, m)$.

To state the errors of Poisson process approximation, we need a metric corresponding to d_1 but without the average feature. For $\xi_1 = \sum_{i=1}^m \delta_{x_i}$ and $\xi = \sum_{i=1}^n \delta_{y_i}$ with $m \geq n$, the metric is defined by

$$d_1'(\xi_1, \xi_2) = (m - n) + \min_\pi \sum_{i=1}^n d_0(x_{\pi(i)}, y_i)$$

for π ranging over all permutations of $(1, \cdots, m)$ [Brown & Xia (1995a)].

As the elements of $\mathcal{H}$ are (point) measures on Γ, one may immediately think of the well-known Prohorov (1956) metric to quantify the weak topology, as conventionally used in probability theory [see Billingsley (1968)]. In our context, with the above setup, the Prohorov metric is defined by

$$\rho_1(\xi_1, \xi_2) = \begin{cases} 1, & \text{if } n \neq m, \\ \min_\pi \max_{1 \leq i \leq n} d_0(y_i, z_{\pi(i)}), & \text{if } n = m > 0. \end{cases}$$

The fact that the maximum appears here makes the Prohorov metric behave in a similar fashion to the total variation metric $\rho_{TV}(\xi_1, \xi_2) = 1_{\xi_1 \neq \xi_2}$, as studied in Xia (1994), and is thus also typically too strong to use [see Section 4.2 and Remark 5.13].

Example 4.1: Take Γ to be the interval $[0,1]$ with metric $d_0(x,y) = |x-y|$. If $\xi_1 = \sum_{i=1}^n \delta_{t_i}$ with $0 \leq t_1 \leq \cdots \leq t_n \leq 1$ and $\xi_2 = \sum_{i=1}^n \delta_{s_i}$ with

$0 \leq s_1 \leq \cdots \leq s_n \leq 1$, then

$$\sum_{i=1}^{n} |t_i - s_i| \leq \sum_{i=1}^{n} |t_i - s_{\pi(i)}| \tag{4.1}$$

for all permutations π of $(1, \cdots, n)$, and hence

$$d_1(\xi_1, \xi_2) = \frac{1}{n} \sum_{i=1}^{n} |t_i - s_i|.$$

The following proof is taken from Xia (1994).

Proof: We use mathematical induction to prove the claim. Take first the case where $n = 2$. Without loss of generality, we may assume that we have $t_1 = \min\{t_1, t_2, s_1, s_2\}$. Then it suffices to consider the following three cases.

(i) If $t_2 \geq s_2$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + t_2 - s_2 \leq |t_1 - s_2| + |t_2 - s_1|.$$

(ii) If $s_1 \leq t_2 < s_2$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + s_2 - t_2 \leq |t_1 - s_2| + |t_2 - s_1|.$$

(iii) If $t_2 < s_1$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + s_2 - t_2 = |t_1 - s_2| + |t_2 - s_1|.$$

Now, suppose (4.1) holds for $n \leq k$ with $k \geq 2$, we shall prove it holds for $n = k + 1$ and all permutations π of $(1, \cdots, k + 1)$. As a matter of fact, the claim is obvious if $\pi(k + 1) = k + 1$. Assume $\pi(k + 1) \neq k + 1$, then it follows that

$$\begin{aligned}
&\sum_{i=1}^{k+1} |t_i - s_{\pi(i)}| \\
&= \sum_{i \neq k+1, i \neq \pi^{-1}(k+1)} |t_i - s_{\pi(i)}| + [|t_{k+1} - s_{\pi(k+1)}| + |t_{\pi^{-1}(k+1)} - s_{k+1}|] \\
&\geq \sum_{i \neq k+1, i \neq \pi^{-1}(k+1)} |t_i - s_{\pi(i)}| + |t_{\pi^{-1}(k+1)} - s_{\pi(k+1)}| + |t_{k+1} - s_{k+1}| \\
&\geq \sum_{i=1}^{k} |t_i - s_i| + |t_{k+1} - s_{k+1}|,
\end{aligned}$$

where the inequalities are due to the induction assumptions. ∎

Now we show that both ρ_1 and d_1 generate the weak topology.

Proposition 4.2: *If Γ is compact, the following statements are equivalent:*

(i) ξ_n converges to ξ vaguely;
(ii) $\xi_n(B) \to \xi(B)$ for all sets $B \in \mathcal{B}_b$ whose boundary ∂B satisfies $\xi(\partial B) = 0$;
(iii) $\limsup_{n\to\infty} \xi_n(F) \le \xi(F)$ and $\liminf_{n\to\infty} \xi_n(G) \ge \xi(G)$ for all closed $F \in \mathcal{B}_b$ and open $G \in \mathcal{B}_b$;
(iv) ξ_n converges to ξ weakly;
(v) $d_1(\xi_n, \xi) \to 0$;
(vi) $\rho_1(\xi_n, \xi) \to 0$.

Proof: The equivalence of (i)-(iv) is well-known and can be found in Kallenberg (1976), pp. 94-95, and (vi) is equivalent to (v) due to the fact that if $\xi(\Gamma) \neq 0$,

$$d_1(\xi_n, \xi) \le \rho_1(\xi_n, \xi) \le \xi(\Gamma) d_1(\xi_n, \xi).$$

Hence it suffices to show that (iv) is equivalent to (v). Assume first that $d_1(\xi_n, \xi) \to 0$, then there exists an n_0 such that $\xi_n(\Gamma) = \xi(\Gamma)$ for all $n \ge n_0$. For any bounded continuous function f, as Γ is compact, f is uniformly continuous and for $n \ge n_0$,

$$\left| \int_\Gamma f(x)\xi_n(dx) - \int_\Gamma f(x)\xi(dx) \right| \le |\xi| \sup_{d_0(x,y) \le |\xi| d_1(\xi_n, \xi)} |f(x) - f(y)| \to 0.$$

Conversely, choose $f \equiv 1$, we have $\xi_n(\Gamma) \to \xi(\Gamma) := m$, so when n is sufficiently large, say, $n \ge m_0$, $\xi_n(\Gamma) = \xi(\Gamma)$. We use the basic fact that $d_1(\xi_n, \xi) \to 0$ if and only if for every sequence $\{n_k\} \subset \mathbb{N}$, there exists a subsequence $\{n_{k_j}\} \subset \{n_k\}$ such that $d_1(\xi_{n_{k_j}}, \xi) \to 0$ to prove the claim. Let $\xi = \sum_{i=1}^m \delta_{z_i}$, and relabel if necessary, for $n \ge m_0$, we may write $\xi_n = \sum_{i=1}^m \delta_{y_i^n}$ so that $d_1(\xi_n, \xi) = [\sum_{i=1}^m d_0(y_i^n, z_i)]/m$. Since Γ is compact, for any sequence $\{n_k\} \subset \mathbb{N}$, there exists a subsequence $\{n_{k_j}\}$ and $\tilde{z}_i$ such that $d_0(y_i^{n_{k_j}}, \tilde{z}_i) \to 0$ for all $i = 1, \cdots, m$. Set $\tilde{\xi} = \sum_{i=1}^m \delta_{\tilde{z}_i}$, then $\int_\Gamma f(x)\xi_{n_{k_j}}(dx) \to \int_\Gamma f(x)\tilde{\xi}(dx)$ and by the assumption, $\int_\Gamma f(x)\xi_{n_{k_j}}(dx) \to \int_\Gamma f(x)\xi(dx)$, so $\int_\Gamma f(x)\tilde{\xi}(dx) = \int_\Gamma f(x)\xi(dx)$ for all bounded functions f, which implies $\tilde{\xi} = \xi$ and

$$\sum_{i=1}^m d_0(y_i^{n_{k_j}}, z_i) \le \sum_{i=1}^m d_0(y_i^{n_{k_j}}, \tilde{z}_i) \to 0,$$

since $(y_i^{n_{k_j}}, z_i)$, $i = 1, \cdots, m$, is a best matching. Hence $d_1(\xi_n, \xi) \to 0$. ∎

Proposition 4.3: *$(\mathcal{H}, d_1)$ is a locally compact separable metric space.*

Proof: By Proposition 2.6, it suffices to show $\mathcal{H}$ is locally compact. In fact, define $\mathcal{H}_k = \{\xi \in \mathcal{H} : \xi(\Gamma) = k\}$, $\mathcal{H} = \cup_{k=0}^{\infty}\mathcal{H}_k$ and for each k, $\mathcal{H}_k$ is a compact, open and as well as closed set. ∎

4.2. *The metrics for point process approximation*

For any two probability measures $\mathbf{Q}_1$ and $\mathbf{Q}_2$ on the same domain, the total variation distance, d_{TV}, between $\mathbf{Q}_1$ and $\mathbf{Q}_2$ is defined as

$$d_{TV}(\mathbf{Q}_1, \mathbf{Q}_2) = \sup |\mathbf{Q}_1(D) - \mathbf{Q}_2(D)|,$$

where the supremum is taken over all sets D in the common domain. Thus, if X_1 and X_2 are two non-negative integer-valued random variables, then they induce two probability measures, $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$, on the domain consisting of all subsets of $\mathbb{Z}_+$ and

$$\begin{aligned} d_{TV}(\mathcal{L}(X_1), \mathcal{L}(X_2)) &= \sup_{A \subset \mathbb{Z}_+} |\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)| \\ &= \frac{1}{2}\sum_{i=0}^{\infty} |\mathbb{P}(X_1 = i) - \mathbb{P}(X_2 = i)| \\ &= \inf \mathbb{P}(X_1' \neq X_2'), \end{aligned}$$

where the last equality is due to the duality theorem [see Rachev (1991), p. 168] and the infimum is taken over all *couplings* of (X_1', X_2') such that $\mathcal{L}(X_i') = \mathcal{L}(X_i)$, $i = 1, 2$. Likewise, two point processes Ξ_1 and Ξ_2 defined on Γ generate two probability measures, denoted by $\mathcal{L}(\Xi_1)$ and $\mathcal{L}(\Xi_2)$, on the domain $\mathcal{B}(\mathcal{H})$, and the total variation distance between the distributions of the two point processes is

$$\begin{aligned} d_{TV}(\mathcal{L}(\Xi_1), \mathcal{L}(\Xi_2)) &= \sup_{D \in \mathcal{B}(\mathcal{H})} |\mathbb{P}(\Xi_1 \in D) - \mathbb{P}(\Xi_2 \in D)| \\ &= \inf \mathbb{P}(\Xi_1' \neq \Xi_2'), \end{aligned}$$

where, again, the last equality is from the duality theorem and the infimum ranges over all couplings of (Ξ_1', Ξ_2') such that $\mathcal{L}(\Xi_i') = \mathcal{L}(\Xi_i)$, $i = 1, 2$.

The major advantage of using the total variation metric for the distributions of point processes is that, if the distributions of two processes are sufficiently close with respect to the total variation distance, the same accuracy is also valid for the total variation distance between the distributions of any functional of the two processes. However, although the total variation metric for random variable approximation is very successful, the total

variation metric for point processes turns out to be too strong to use in practical problems. For example, if the two point processes have the same numbers of points, but distributed in slightly different locations, they would be at total variation distance 1 from one another. This inspired people to look for weaker metrics which would be small in such situations. The metric d_1 defined in Section 4.1 and the metric d_2 derived from d_1 serve the purpose very well.

To define d_2, let Ψ denote the sets of functions $f : \mathcal{H} \mapsto [0,1]$ such that

$$\sup_{\xi_1 \neq \xi_2 \in \mathcal{H}} \frac{|f(\xi_1) - f(\xi_2)|}{d_1(\xi_1, \xi_2)} \leq 1. \tag{4.2}$$

Then the second Wasserstein metric d_2 between probability measures $\mathbf{Q}_1$ and $\mathbf{Q}_2$ over $\mathcal{B}(\mathcal{H})$ with respect to d_1 is defined as [Barbour & Brown (1992a)]

$$\begin{aligned} d_2(\mathbf{Q}_1, \mathbf{Q}_2) &= \sup_{f \in \Psi} \left| \int f d\mathbf{Q}_1 - \int f d\mathbf{Q}_2 \right| \\ &= \inf \mathbb{E} d_1(\Xi_1, \Xi_2) \\ &= \inf \left\{ \mathbb{P}(|\Xi_1| \neq |\Xi_2|) + \mathbb{E} d_1(\Xi_1, \Xi_2) 1_{[|\Xi_1| = |\Xi_2|]} \right\} \end{aligned} \tag{4.3}$$

where, again, the infimum is taken over all couplings of (Ξ_1, Ξ_2) with $\mathcal{L}(\Xi_1) = \mathbf{Q}_1$ and $\mathcal{L}(\Xi_2) = \mathbf{Q}_2$ and the second equality of (4.3) is due to the duality theorem [Rachev (1991), p. 168]. The last equality of (4.3) offers a nice interpretation of d_2 metric. That is, it is the total variation distance between the distributions of the total numbers of points of the two point processes, and given the two distributions have the same number of points, it then measures the average distance of the matched pairs under the best matching. It is also clear that for two point processes on Γ,

$$d_{TV}(\mathcal{L}(\Xi_1), \mathcal{L}(\Xi_2)) \geq d_2(\mathcal{L}(\Xi_1), \mathcal{L}(\Xi_2)) \geq d_{TV}(\mathcal{L}(|\Xi_1|), \mathcal{L}(|\Xi_2|)),$$

hence, when we bound the errors of point process approximation in terms of the metric d_2, the best possible order we can hope for is the same as that for one dimensional total variation metric.

4.3. *Poisson process approximation using stochastic calculus*

Before Stein's method was introduced into the area of Poisson process approximation, the errors of process approximation were mainly estimated using coupling methods [e.g., Freedman (1974), Serfling (1975) and

Brown (1983)]. From the duality theorem, any coupling will give an upper bound for the distance under study and the effort is in finding those which give upper bounds as close to the true distance as possible. A good coupling often requires very sophisticated skills and knowledge of probability theory and, unlike the universality of Stein's method, it is usually found in an ad hoc way. In some special cases where a near perfect coupling can be constructed, an estimate using coupling may be better than one derived by Stein's method; however, when the complexity of the structure of the point processes is beyond our comprehension, Stein's method usually performs better than our manipulation of coupling skills.

In this section, we quote two results of Poisson process approximation using coupling [see Section 2.3.3] and stochastic calculus, one in terms of the total variation metric and the other for the d_2 distance. We omit the proofs, as the proofs require advanced knowledge in stochastic calculus, which is not a necessity for understanding Stein's method.

Theorem 4.4: [Brown (1983)] *If $(M, \mathcal{F})$ is a simple point process with compensator A, $\boldsymbol{\mu}$ is a measure on $[0, \infty)$ and $t \geq 0$ is non-random, then*

$$d_{TV}(\mathcal{L}(M^t), \mathrm{Po}\,(\boldsymbol{\mu}^t)) \leq \mathbb{E}|A - \boldsymbol{\mu}|_t + \mathbb{E}\left\{\sum_{s \leq t}(\Delta A(s))^2\right\}, \tag{4.4}$$

where $\mathcal{L}(M^t)$ is the distribution of M confined to $[0,t]$, $\mathrm{Po}\,(\boldsymbol{\mu}^t)$ is the distribution of a Poisson process on $[0,t]$ with mean measure $\boldsymbol{\mu}|_{[0,t]}$ and $|A - \boldsymbol{\mu}|_t$ is the pathwise variation of the signed measure of $A - \boldsymbol{\mu}$ on $[0,t]$, i.e. $|A - \boldsymbol{\mu}|_t = \int_0^t |A(ds) - \boldsymbol{\mu}(ds)|$.

This result is optimal in the sense that we can find examples with total variation distance having the same order as the bounds derived from the result.

Example 4.5: (Poisson process approximation to a Bernoulli process). Let $I_1, \cdots, I_n$ be independent indicators with

$$\mathbb{P}(I_i = 1) = 1 - \mathbb{P}(I_i = 0) = p_i, \ i = 1, \cdots, n.$$

Let $\Gamma = [0,1]$, $\Xi = \sum_{i=1}^n I_i \delta_{i/n}$ and $\boldsymbol{\lambda} = \sum_{i=1}^n p_i \delta_{i/n}$ be the mean measure of Ξ. With respect to its intrinsic filtration $\mathcal{F} = (\mathcal{F}_s)_{s \geq 0}$, where $\mathcal{F}_s$ is just $\sigma\{\Xi[0,r] : r \leq s\}$, Ξ has compensator $A(s) = \sum_{i \leq ns} p_i = \boldsymbol{\lambda}[0,s]$. Using (4.4), we have

$$d_{TV}(\mathcal{L}(\Xi), \mathrm{Po}\,(\boldsymbol{\lambda})) \leq \sum_{i=1}^n p_i^2.$$

The bound is in the right order. To see this, let

$$B = \{\xi : \text{ there is an } i \text{ such that } \xi\{i/n\} \geq 2\};$$

then $\mathbb{P}(\Xi \in B) = 0$, while $\mathrm{Po}\,(\boldsymbol{\lambda})(B) = O\left(\sum_{i=1}^{n} p_i^2\right)$.

We now state the estimate in terms of the d_2–metric.

Theorem 4.6: [Xia (2000)] *Suppose $(M, \mathcal{F})$ is a simple point process with compensator A, and B is a non-random continuous increasing function with $B(0) = 0$. Let $p, q \in [1, \infty]$ satisfying $1/p + 1/q = 1$. Then, for any fixed $\delta > 0$ and $T > 0$, we have*

$$d_2(\mathcal{L}(M^T), \mathrm{Po}\,(B^T)) \leq C_q(\lambda)\|(B, A)\|_p + \mathbb{E}\{\delta \wedge |A(T^-) - B(T)|\}$$
$$+ \mathbb{P}(|A(T^-) - B(T)| \geq \delta) + \mathbb{E}\Delta A(T) + \mathbb{E}\left\{\sum_{s<T} (\Delta A(s))^2\right\}, \quad (4.5)$$

where $\lambda = B(T)$, $C_q(\lambda) = \left[\sum_{i=0}^{\infty} \frac{1}{(i+1)^q} \frac{e^{-\lambda}\lambda^i}{i!}\right]^{1/q}$,

$$\|(B, A)\|_p = \left\{\mathbb{E}\left[\int_{[0,T)} 1 \wedge (|B^{-1} \circ A(t) - t| \vee |B^{-1} \circ A(t^-) - t|) dA(t)\right]^p\right\}^{1/p},$$

and the quantity $\|(B, A)\|_\infty$ is understood to be an essential supremum [see Shiryaev (1984), p. 259].

Remark 4.7: In calculating the upper bound of (4.5), the values of p and q are chosen according to the randomness of A. Usually, we choose $p = q = 2$, while sometimes other choices are also possible. In particular, if A is non-random, then we can choose $p = \infty$ and $q = 1$ [see example 4.11].

The following proposition is useful in applications.

Proposition 4.8: *The values of C_1 and C_2 are given by*

$$(i)\ \ C_1(\lambda) = \frac{1 - e^{-\lambda}}{\lambda};$$

$$(ii)\ \ C_2(\lambda) = \sqrt{\frac{e^{-\lambda}}{\lambda} \int_0^\lambda \frac{e^x - 1}{x} dx} \leq \left(1 \wedge \frac{\sqrt{2}}{\lambda}\right),$$

and $C_2(\lambda) \sim 1/\lambda$ as $\lambda \to \infty$.

Proof: The proof of (i) is straightforward. Apropos of (ii), we have

$$\sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{(i+1)^2 i!} = \frac{e^{-\lambda}}{\lambda} \sum_{i=0}^{\infty} \frac{1}{i+1} \frac{\lambda^{i+1}}{(i+1)!} = \frac{e^{-\lambda}}{\lambda} \int_0^\lambda \frac{e^x - 1}{x} dx.$$

Since $\frac{e^x-1}{x}$ is an increasing function of x, it follows that

$$\frac{e^{-\lambda}}{\lambda}\int_0^\lambda \frac{e^x-1}{x}dx \le \frac{e^{-\lambda}}{\lambda}\frac{e^\lambda-1}{\lambda}\lambda \le 1.$$

On the other hand,

$$\frac{e^{-\lambda}}{\lambda}\sum_{i=0}^{\infty}\frac{1}{i+1}\frac{\lambda^{i+1}}{(i+1)!} \le \frac{e^{-\lambda}}{\lambda}\sum_{i=0}^{\infty}\frac{2}{i+2}\frac{\lambda^{i+1}}{(i+1)!} \le \frac{2}{\lambda^2}.$$ ∎

Remark 4.9: Although we omitted the proofs, a simple comparison of the bounds (4.4) and (4.5) would reveal that the coupling idea works well for the continuous part of the compensator but does not handle the jumps of the compensator efficiently [see Kabanov, Liptser & Shiryaev (1983), Lemma 3.2].

Example 4.10: If M is a simple point process with continuous compensator A satisfying $A(T) = B(T), a.s.$, then applying (4.5) gives

$$d_2(\mathcal{L}(M^T), \mathrm{Po}\,(B^T)) \le C_q(\lambda)\|(B,A)\|_p.$$

This upper bound is of a kind similar to the upper bound on the total variation distance between a mixed Poisson random variable and a Poisson random variable established in Pfeifer (1987) [see also Barbour, Holst & Janson (1992), p. 12].

Note that, if $A(T) = B(T)$ $a.s.$, $M(T)$ has distribution $\mathrm{Po}(B(T))$, but it is not necessarily true that M is a Poisson process.

Example 4.11: If $T = 1$, $B(t) = at, A(t) = bt$ with $a \ge b > 0$, and $p = \infty, q = 1$, then (4.5) implies that

$$d_2(\mathrm{Po}\,(A^T), \mathrm{Po}\,(B^T)) \le C_1(a)\frac{b|b-a|}{2a} + |b-a| \asymp |b-a|. \tag{4.6}$$

On the other hand, the result in Barbour, Brown & Xia (1998), who combined stochastic calculus with Stein's method, gives

$$d_2(\mathrm{Po}\,(A^T), \mathrm{Po}\,(B^T)) \le (1 \wedge 1.65b^{-\frac{1}{2}})|b-a|,$$

which is significantly better than (4.6) if b is large.

5. Poisson process approximation via Stein's method

5.1. *Stein's equation and Stein's factors*

As shown in the previous section, the coupling idea applied directly can solve the problem of Poisson process approximation with a certain degree of success, but with little hope of further improvement. We need a mechanism for quantifying Poisson process approximation which is more routinely applicable, and Stein's method provides the right answer. The foundation of Stein's method for Poisson process approximation was established by Barbour (1988), adapted from the Stein–Chen method in Chen (1975) for Poisson random variable approximation, and was further refined in Barbour & Brown (1992a). There it is shown, in particular, how Stein's method can be combined with standard tools of point process theory such as Palm distributions and Janossy densities. They used an auxiliary spatial immigration-death process, as in Section 3.2, to investigate the approximation of a point process by a Poisson process. Recalling that the generator $\mathcal{A}$ of the spatial immigration-death process is given in (3.6) and a point process Ξ is a Poisson process with mean measure λ if and only if $\mathbb{E}\mathcal{A}f(\Xi) = 0$ for a sufficiently rich class of functions f on $\mathcal{H}$, we conclude that $\mathcal{L}(\Xi)$ is close to $\mathrm{Po}(\lambda)$ if and only if $\mathbb{E}\mathcal{A}f(\Xi) \approx 0$ for this class of functions f. On the other hand, to evaluate the distance between $\mathcal{L}(\Xi)$ and $\mathrm{Po}(\lambda)$, we need to work on the functional form $\mathbb{E}(h(\Xi)) - \mathrm{Po}(\lambda)(h)$ for the test functions h which define the metric of our interest; hence we look for a function f satisfying

$$\begin{aligned}\mathcal{A}f(\xi) &= \int_{\Gamma}[f(\xi+\delta_\alpha) - f(\xi)]\lambda(d\alpha) + \int_{\Gamma}[f(\xi-\delta_\alpha) - f(\xi)]\xi(d\alpha)\\ &= h(\xi) - \mathrm{Po}(\lambda)(h),\end{aligned} \tag{5.1}$$

called the *Stein equation*, and if such an f can be found, we can estimate the quantity $|\mathbb{E}(h(\Xi)) - \mathrm{Po}(\lambda)(h)|$ by investigating $|\mathbb{E}\mathcal{A}f(\Xi)|$. The estimate of the approximation error is then just the supremum of $|\mathbb{E}\mathcal{A}f(\Xi)|$, taken over all the solutions f corresponding to the test functions h of interest [see Theorem 5.3].

Example 5.1: (Poisson process approximation to a Bernoulli process). With the setup of Example 4.5, we have

$$\begin{aligned}
&\mathbb{E}h(\Xi) - \mathrm{Po}\,(\boldsymbol{\lambda})(h) = \mathbb{E}\mathcal{A}f(\Xi) \\
&= \mathbb{E}\int_0^1 [f(\Xi + \delta_\alpha) - f(\Xi)]\boldsymbol{\lambda}(d\alpha) + \mathbb{E}\int_0^1 [f(\Xi - \delta_\alpha) - f(\Xi)]\Xi(d\alpha) \\
&= \sum_{i=1}^n \mathbb{E}\{[f(\Xi + \delta_{i/n}) - f(\Xi)] + [f(\Xi^i) - f(\Xi^i + \delta_{i/n})]\}p_i \\
&= \sum_{i=1}^n \mathbb{E}[f(\Xi^i + 2\delta_{i/n}) - 2f(\Xi^i + \delta_{i/n}) + f(\Xi^i)]p_i^2,
\end{aligned}$$

where $\Xi^i = \Xi - I_i\delta_{i/n}$.

To find the solution to (5.1), recall that the resolvent of $\mathcal{A}$ at $\rho > 0$ is given by

$$(\rho - \mathcal{A})^{-1}g(\xi) = \int_0^\infty e^{-\rho t}\mathbb{E}g(\mathbf{Z}_\xi(t))\,dt \tag{5.2}$$

[Ethier & Kurtz (1986), p. 10]. What we need to do is to argue that when $\rho = 0$ the equation (5.2) still holds for suitable functions g and we summarize the results in the following lemma [cf Barbour & Brown (1992a)].

Lemma 5.2: *For any bounded function h, the integral*

$$\int_0^\infty [\mathbb{E}h(\mathbf{Z}_\xi(t)) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)]\,dt \tag{5.3}$$

is well-defined and the solution to (5.1) is

$$f(\xi) = -\int_0^\infty [\mathbb{E}h(\mathbf{Z}_\xi(t)) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)]\,dt. \tag{5.4}$$

Proof: Set $|\xi| = n$,

$$\tau_{n0} = \inf\{t:\ \mathbf{D}_\xi(t) = 0\} = \inf\{t:\ |\mathbf{D}_\xi(t)| = 0\},$$

and

$$\tau_{P0} = \inf\{t:\ \mathbf{D}_P(t) = 0\} = \inf\{t:\ |\mathbf{D}_P(t)| = 0\},$$

where $\mathbf{D}_P$ is a pure death process with initial distribution the same as $\mathrm{Po}\,(\boldsymbol{\lambda})$. Then, in view of Proposition 3.5, we can construct a pair of processes $\mathbf{Z}_P(t) = \mathbf{D}_P(t) + \mathbf{Z}_0(t)$ and $\mathbf{Z}_\xi(t) = \mathbf{D}_\xi(t) + \mathbf{Z}_0(t)$ such that $\mathbf{Z}_P$ is in

equilibrium and $\mathbf{Z}_\xi$ is an immigration-death process starting in ξ, which satisfy $\mathbf{Z}_P(t) = \mathbf{Z}_\xi(t)$ for $t > \max\{\tau_{n0}, \tau_{P0}\}$. This implies that

$$\begin{aligned}\int_0^\infty |\mathbb{E}h(\mathbf{Z}_\xi(t)) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)|dt &= \int_0^\infty |\mathbb{E}h(\mathbf{Z}_\xi(t)) - \mathbb{E}h(\mathbf{Z}_P(t))|dt \\ &\le \mathbb{E}(\tau_{n0} + \tau_{P0})\|f\| \le (n+\lambda)\|f\|.\end{aligned}$$

Since the integral (5.3) is absolutely convergent, we may split it at the first time of leaving the initial state $\tau = \inf\{t : \mathbf{Z}_\xi(t) \neq \xi\}$:

$$\begin{aligned}&f(\xi) \\ &= \frac{-1}{\lambda+n}\left\{[h(\xi) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)] + (\lambda+n)\mathbb{E}\int_\tau^\infty [h(\mathbf{Z}_\xi(t)) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)]dt\right\} \\ &= \frac{-1}{\lambda+n}\left\{[h(\xi) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)] - \int_\Gamma f(\xi+\delta_\alpha)\boldsymbol{\lambda}(d\alpha) - \int_\Gamma f(\xi-\delta_\alpha)\xi(d\alpha)\right\},\end{aligned}$$

where the last equation is because of the strong Markov property. Now, some reorganization leads to equation (5.1). ∎

Assume that, for each α, there is a Borel set $A_\alpha \subset \Gamma$ such that $\alpha \in A_\alpha$ and that the mapping

$$\Gamma \times \mathcal{H} \to \Gamma \times \mathcal{H} : (\alpha, \xi) \mapsto (\alpha, \xi|_{A_\alpha^c})$$

is product measurable. This is true so long as $A = \{(x,y) : y \in A_x, x \in \Gamma\}$ is measurable in Γ^2 [see Chen & Xia (2004)].

Now,

$$\begin{aligned}&\mathbb{E}\int_\Gamma [f(\Xi - \delta_\alpha) - f(\Xi)]\Xi(d\alpha) \\ &= \mathbb{E}\int_\Gamma \{[f(\Xi-\delta_\alpha) - f(\Xi)] - [f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)]\}\Xi(d\alpha) \\ &\quad + \mathbb{E}\int_\Gamma [f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)][\Xi(d\alpha) - \boldsymbol{\lambda}(d\alpha)] \\ &\quad + \mathbb{E}\int_\Gamma \{[f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)] - [f(\Xi) - f(\Xi + \delta_\alpha)]\}\boldsymbol{\lambda}(d\alpha) \\ &\quad + \mathbb{E}\int_\Gamma [f(\Xi) - f(\Xi + \delta_\alpha)]\boldsymbol{\lambda}(d\alpha),\end{aligned}$$

which gives

$$\begin{aligned}
&\mathbb{E}h(\Xi) - \mathrm{Po}(\boldsymbol{\lambda})(h) = \mathbb{E}\mathcal{A}f(\Xi)\\
&= \mathbb{E}\int_\Gamma \{[f(\Xi-\delta_\alpha)-f(\Xi)]-[f(\Xi|_{A_\alpha^c})-f(\Xi|_{A_\alpha^c}+\delta_\alpha)]\}\Xi(d\alpha)\\
&\quad +\mathbb{E}\int_\Gamma [f(\Xi|_{A_\alpha^c})-f(\Xi|_{A_\alpha^c}+\delta_\alpha)][\Xi(d\alpha)-\boldsymbol{\lambda}(d\alpha)]\\
&\quad +\mathbb{E}\int_\Gamma \{[f(\Xi|_{A_\alpha^c})-f(\Xi|_{A_\alpha^c}+\delta_\alpha)]-[f(\Xi)-f(\Xi+\delta_\alpha)]\}\boldsymbol{\lambda}(d\alpha). \quad (5.5)
\end{aligned}$$

Example 5.1 and (5.5) tell us that, in order to implement Stein's method, it is essential to have good estimates of

$$\begin{aligned}
&\Delta^2 f(\xi;\alpha,\beta) = f(\xi+\delta_\alpha+\delta_\beta)-f(\xi+\delta_\alpha)-f(\xi+\delta_\beta)+f(\xi),\\
&\Delta f(\xi) = \sup_{x\in\Gamma}|f(\xi+\delta_x)-f(\xi)|,\\
&\Delta^2 f(\xi) = \sup_{\eta-\xi\in\mathcal{H},x,y\in\Gamma}\left|\Delta^2 f(\eta;x,y)\right|.
\end{aligned}$$

With these quantities, if $\xi_1 \in \mathcal{H}$ and $\xi_2 = \xi_1 + \sum_{i=1}^k \delta_{x_i}$, we can define $\eta_j = \xi_1 + \sum_{i=1}^j \delta_{x_i}$, so that then $\eta_0 = \xi_1$, $\eta_k = \xi_2$ and

$$\begin{aligned}
&|[f(\xi_2)-f(\xi_2+\delta_\alpha)]-[f(\xi_1)-f(\xi_1+\delta_\alpha)]| \quad (5.6)\\
&= \left|\sum_{j=1}^k \{[f(\eta_j)-f(\eta_j+\delta_\alpha)]-[f(\eta_{j-1})-f(\eta_{j-1}+\delta_\alpha)]\}\right| \le k\Delta^2 f(\xi_1),
\end{aligned}$$

giving

$$\begin{aligned}
&\left|\mathbb{E}\int_\Gamma \{[f(\Xi-\delta_\alpha)-f(\Xi)]-[f(\Xi|_{A_\alpha^c})-f(\Xi|_{A_\alpha^c}+\delta_\alpha)]\}\Xi(d\alpha)\right|\\
&\quad \le \mathbb{E}\int_{\alpha\in\Gamma}\Delta^2 f(\Xi|_{A_\alpha^c})(\Xi(A_\alpha)-1)\Xi(d\alpha), \quad (5.7)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\mathbb{E}\int_\Gamma \{[f(\Xi|_{A_\alpha^c})-f(\Xi|_{A_\alpha^c}+\delta_\alpha)]-[f(\Xi)-f(\Xi+\delta_\alpha)]\}\boldsymbol{\lambda}(d\alpha)\right|\\
&\quad \le \mathbb{E}\int_{\alpha\in\Gamma}\Delta^2 f(\Xi|_{A_\alpha^c})\boldsymbol{\lambda}(d\alpha)\Xi(A_\alpha). \quad (5.8)
\end{aligned}$$

For the second term of (5.5), we have

$$\begin{aligned}
&\mathbb{E}\int_{\Gamma}[f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)][\Xi(d\alpha) - \boldsymbol{\lambda}(d\alpha)] \\
&\quad = \mathbb{E}\int_{\Gamma}[f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)][g(\alpha, \Xi|_{A_\alpha^c}) - \phi(\alpha)]\boldsymbol{\mu}(d\alpha) \quad (5.9) \\
&\quad = \mathbb{E}\int_{\Gamma}\{[f(\Xi_\alpha|_{A_\alpha^c}) - f(\Xi_\alpha|_{A_\alpha^c} + \delta_\alpha)] \\
&\qquad - [f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)]\}\boldsymbol{\lambda}(d\alpha), \quad (5.10)
\end{aligned}$$

where $\phi(\alpha)$ and g are defined in (2.7) and (2.8). The estimates (5.7)-(5.10) are summarized in the following theorem.

Theorem 5.3: [Chen & Xia (2004)] *For each bounded measurable function h on $\mathcal{H}$,*

$$\begin{aligned}
&|\mathbb{E}h(\Xi) - \mathrm{Po}\,(\boldsymbol{\lambda})(h)| \\
&\quad \le \mathbb{E}\int_{\alpha\in\Gamma}\Delta^2 f(\Xi|_{A_\alpha^c})(\Xi(A_\alpha) - 1)\Xi(d\alpha) + \min\{\epsilon_1(h,\Xi), \epsilon_2(h,\Xi)\} \\
&\qquad + \mathbb{E}\int_{\alpha\in\Gamma}\Delta^2 f(\Xi|_{A_\alpha^c})\boldsymbol{\lambda}(d\alpha)\Xi(A_\alpha),
\end{aligned}$$

where

$$\epsilon_1(h,\Xi) = \mathbb{E}\int_{\alpha\in\Gamma}\Delta f(\Xi|_{A_\alpha^c})|g(\alpha, \Xi|_{A_\alpha^c}) - \phi(\alpha)|\boldsymbol{\mu}(d\alpha)$$

which is valid if Ξ is a simple point process, and

$$\epsilon_2(h,\Xi) = \mathbb{E}\int_{\alpha\in\Gamma}|[f(\Xi|_{A_\alpha^c}) - f(\Xi|_{A_\alpha^c} + \delta_\alpha)] - [f(\Xi_\alpha|_{A_\alpha^c}) - f(\Xi_\alpha|_{A_\alpha^c} + \delta_\alpha)]|\boldsymbol{\lambda}(d\alpha).$$

Remark 5.4: How judiciously the $(A_\alpha; \alpha \in \Gamma)$ are chosen is reflected in the upper bound.

The bounds on $\Delta f(\xi)$ and $\Delta^2 f(\xi)$ depend on the metrics we use for measuring the accuracy of Poisson process approximation and will be dealt with in terms of three metrics: one dimensional total variation metric, the total variation metric for the distributions of point processes and the d_2 metric. The bounds can be summarized in the following table:

	$\Delta f(\xi)$	$\Delta^2 f(\xi)$
d_{TV}: one dimensional	$1 \wedge \sqrt{\frac{2}{e\lambda}}$ (Prop. 5.7)	$\frac{1-e^{-\lambda}}{\lambda}$ (Prop. 5.6)
d_{TV}: process distributions	1 (Prop. 5.12)	1 (Prop. 5.12)
d_2 (uniform)	$1 \wedge \frac{1.647}{\sqrt{\lambda}}$ (Prop. 5.16)	$1 \wedge \left[\frac{11}{6\lambda}\left(1+2\ln^+\left(\frac{6\lambda}{11}\right)\right)\right]$ (Prop. 5.17)
d_2 (non-uniform)		$\frac{3.5}{\lambda} + \frac{2.5}{\lvert\xi\rvert+1}$ (Prop. 5.21)

5.2. *One dimensional Poisson process approximation*

In this section, we focus on the distribution of $\Xi(B)$, the total number of points in a Borel set B of a point process Ξ, approximated by a Poisson distribution: see also the discussion in Chapter 2, Section 2.5. For simplicity, we only consider the test functions $h(\xi) = 1_A(|\xi|)$ with $A \subset \mathbb{Z}_+$; one can adapt to functions of the form $1_A(\xi(B))$ by changing λ to $\boldsymbol{\lambda}(B)$ appropriately. The corresponding solution to the Stein equation (5.1) is denoted by $f_A(\cdot)$ and we write f_j for $f_{\{j\}}$. Noting that, when we consider only the total number of points, the spatial immigration-death process is reduced to an immigration-death process as in Section 3.1, we simply write $|\mathbf{Z}_\xi(t)|$ as $Z_{|\xi|}(t)$. For each fixed j, if $i \le j$, it follows from (5.4) and the strong Markov property that

$$\begin{aligned} f_j(i-1) &= -\mathbb{E}\int_0^{\tau_{i-1}^+}[1_{\{j\}}(Z_{i-1}(t)) - \pi_j]dt - \mathbb{E}\int_{\tau_{i-1}^+}^{\infty}[1_{\{j\}}(Z_{i-1}(t)) - \pi_j]dt \\ &= \pi_j\overline{\tau_{i-1}^+} + f_j(i), \end{aligned}$$

giving

$$f_j(i) - f_j(i-1) = -\pi_j\overline{\tau_{i-1}^+}. \tag{5.11}$$

Similarly, for $i \ge j+1$,

$$\begin{aligned} f_j(i) &= -\mathbb{E}\int_0^{\tau_i^-}[1_{\{j\}}(Z_i(t)) - \pi_j]dt - \mathbb{E}\int_{\tau_i^-}^{\infty}[1_{\{j\}}(Z_i(t)) - \pi_j]dt \\ &= \pi_j\overline{\tau_i^-} + f_j(i-1), \end{aligned}$$

which implies that

$$f_j(i) - f_j(i-1) = \pi_j\overline{\tau_i^-}. \tag{5.12}$$

Combining (5.11), (5.12) and Lemma 3.2, we reach the following lemma.

Lemma 5.5: *For each $j \in \mathbb{Z}_+$, $f_j(i+1) - 2f_j(i) + f_j(i-1)$ is negative for all i save $i = j$.*

As a matter of fact, the claim in Lemma 5.5 is valid for a large class of approximating distributions including Poisson [Chen (1975), Barbour & Eagleson (1983)], binomial [Ehm (1991)], negative binomial [Brown & Phillips (1999)] and many other polynomial birth-death distributions [Brown & Xia (2001)].

The following estimates of the Stein factors were first achieved by Barbour & Eagleson (1983) using an analytical method, and were revisited by Xia (1999) using a probabilistic method; see also Chapter 2, Theorem 2.3.

Proposition 5.6: *We have*

$$\Delta^2 f_A(\xi) \le k_2(\lambda) := \left\{ \frac{1-e^{-\lambda}}{\lambda} \right\}.$$

Proof: [Barbour & Eagleson (1983), Xia (1999).] It is immediate that $f_{\mathbb{Z}_+}(i+1) - 2f_{\mathbb{Z}_+}(i) + f_{\mathbb{Z}_+}(i-1) = 0$, so it suffices to show that

$$f_A(i+1) - 2f_A(i) + f_A(i-1) \le \frac{1-e^{-\lambda}}{\lambda},$$

for all $A \subset \mathbb{Z}_+$ and all $i \in \mathbb{Z}_+$. To this end, using Lemma 5.5 and (3.5), we have

$$\begin{aligned}
&f_A(i+1) - 2f_A(i) + f_A(i-1) \\
&\le f_i(i+1) - 2f_i(i) + f_i(i-1) = \pi_i \left(\overline{\tau^+_{i-1}} + \overline{\tau^-_{i+1}} \right) \\
&= \frac{\pi_i}{\lambda} \left(\frac{F(i-1)}{\pi_{i-1}} + \frac{\bar{F}(i+1)}{\pi_i} \right) \le \frac{\pi_i}{\lambda} \left(\frac{\pi_1 + \cdots + \pi_i}{\pi_i} + \frac{\bar{F}(i+1)}{\pi_i} \right) \\
&= \frac{1-\pi_0}{\lambda},
\end{aligned}$$

completing the proof. ∎

Proposition 5.7: *We have*

$$\Delta f_A(\xi) \le k_1(\lambda) := \left\{ 1 \wedge \sqrt{\frac{2}{e\lambda}} \right\}.$$

Proof: It follows from (5.11) and (5.12) that

$$\begin{aligned}
f_A(i+1) - f_A(i) \;&\le\; f_{[0,i]}(i+1) - f_{[0,i]}(i) \\
= -\mathbb{E} \int_0^\infty \left[1_{[0,i]}(Z_{i+1}(t)) - 1_{[0,i]}(Z_i(t)) \right] dt \;&=\; \int_0^\infty e^{-t} \mathbb{P}(Z_i(t) = i) dt \;\le\; 1.
\end{aligned}$$

Also,

$$\begin{aligned}\int_0^\infty e^{-t}\max_j \mathbb{P}(Z_0(t)=j)dt &\le \int_0^\infty e^{-t}\left(1\wedge\frac{1}{\sqrt{2e\lambda_t}}\right)dt\\ &= \int_0^1\left(1\wedge\frac{1}{\sqrt{2e\lambda s}}\right)ds = \int_0^{\frac{1}{2e\lambda}} ds + \int_{\frac{1}{2e\lambda}}^1 \frac{1}{\sqrt{2e\lambda s}}ds\\ &\le \sqrt{\frac{2}{e\lambda}} - \frac{1}{2e\lambda}, \end{aligned} \tag{5.13}$$

where the first inequality is taken from Barbour, Holst & Janson (1992), p. 262. ∎

In view of Proposition 5.6 and Proposition 5.7, the following theorem is an obvious corollary of Theorem 5.3.

Theorem 5.8: *If Ξ is a point process on Γ with mean measure $\boldsymbol{\lambda}$, then for every bounded Borel set B,*

$$\begin{aligned} d_{TV}(\mathcal{L}(\Xi(B)), \mathrm{Po}\,(\boldsymbol{\lambda}(B))) \le k_2(\boldsymbol{\lambda}(B))\mathbb{E}\int_{\alpha\in B}(\Xi(B\cap A_\alpha)-1)\Xi(d\alpha)\\ + \min\{\epsilon_1,\epsilon_2\} + k_2(\boldsymbol{\lambda}(B))\int_{\alpha\in B}\boldsymbol{\lambda}(d\alpha)\boldsymbol{\lambda}(B\cap A_\alpha), \end{aligned} \tag{5.14}$$

where

$$\epsilon_1 = k_1(\boldsymbol{\lambda}(B))\mathbb{E}\int_{\alpha\in B}|g(\alpha,\Xi|_{A_\alpha^c}) - \phi(\alpha)|\boldsymbol{\mu}(d\alpha)$$

which is valid if Ξ is a simple point process, and

$$\epsilon_2 = k_2(\boldsymbol{\lambda}(B))\mathbb{E}\int_{\alpha\in B}|\Xi(B\cap A_\alpha^c) - \Xi_\alpha(B\cap A_\alpha^c)|\,\boldsymbol{\lambda}(d\alpha).$$

Remark 5.9: Consider $A_\alpha = \{\alpha\}$ with ϵ_2, then Theorem 5.8 is reduced to

$$\begin{aligned} &d_{TV}(\mathcal{L}(\Xi(B)), \mathrm{Po}\,(\boldsymbol{\lambda}(B)))\\ &\le k_2(\boldsymbol{\lambda}(B))\left\{\mathbb{E}\int_{\alpha\in B}|\Xi(B\backslash\{\alpha\}) - \Xi_\alpha(B\backslash\{\alpha\})|\,\boldsymbol{\lambda}(d\alpha) + \sum_{\alpha\in B}(\boldsymbol{\lambda}\{\alpha\})^2\right\}, \end{aligned}$$

which is slightly different from Theorem 3.1 of Barbour & Brown (1992b).

If the approximating Poisson random variable has a different mean, then the following theorem together with the triangle inequality would be sufficient [Pfeifer (1987) or Barbour, Holst & Janson (1992), p. 12].

Theorem 5.10: *If ν_1 and ν_2 are two positive real numbers with $\nu_1 > \nu_2$, then*

$$d_{TV}(\mathrm{Po}\,(\nu_1), \mathrm{Po}\,(\nu_2)) \leq k_1(\nu_1 \vee \nu_2)(\nu_1 - \nu_2).$$

Remark 5.11: Yannaros (1991), Theorem 2.1, gives an estimate which is typically sharper:

$$d_{TV}(\mathrm{Po}\,(\nu_1), \mathrm{Po}\,(\nu_2)) \leq (\nu_1 - \nu_2) \min\left\{1, \frac{1}{\sqrt{\nu_1} + \sqrt{\nu_2}}\right\}.$$

5.3. *Poisson process approximation in total variation*

The following estimates were first obtained by Barbour & Brown (1992a).

Proposition 5.12: *If $h = 1_A$ with $A \in \mathcal{B}(\mathcal{H})$, and f_A is the solution to the Stein equation (5.1), then*

$$\begin{aligned} &(i)\ \ \Delta f_A(\xi) \leq 1 \quad \textit{for all } \xi; \\ &(ii)\ \ \Delta^2 f_A(\xi) \leq 1 \quad \textit{for all } \xi. \end{aligned}$$

Proof: We use Proposition 3.5 to prove the claims. Let τ, τ_1 and τ_2 be independent negative exponentially distributed random variables with mean 1. For (i), we have

$$\begin{aligned} |f_A(\xi + \delta_\alpha) - f_A(\xi)| &= \left|\mathbb{E}\int_0^\infty [1_A(\mathbf{Z}_\xi(t) + \delta_\alpha 1_{\tau > t}) - 1_A(\mathbf{Z}_\xi(t))]dt\right| \\ &= \left|\mathbb{E}\int_0^\infty e^{-t}[1_A(\mathbf{Z}_\xi(t) + \delta_\alpha) - 1_A(\mathbf{Z}_\xi(t))]dt\right| \\ &\leq \int_0^\infty e^{-t}dt = 1. \end{aligned}$$

Apropos of (ii), note that

$$\begin{aligned} \Delta^2 f_A(\xi; x, y) = -\mathbb{E}\int_0^\infty &[1_A(\mathbf{Z}_\xi(t) + \delta_x 1_{\tau_1 > t} + \delta_y 1_{\tau_2 > t}) - 1_A(\mathbf{Z}_\xi(t) + \delta_x 1_{\tau_1 > t}) \\ &- 1_A(\mathbf{Z}_\xi(t) + \delta_y 1_{\tau_2 > t}) + 1_A(\mathbf{Z}_\xi(t))]dt \\ = -\mathbb{E}\int_0^\infty &e^{-2t}[1_A(\mathbf{Z}_\xi(t) + \delta_x + \delta_y) - 1_A(\mathbf{Z}_\xi(t) + \delta_x) \\ &- 1_A(\mathbf{Z}_\xi(t) + \delta_y) + 1_A(\mathbf{Z}_\xi(t))]dt, \end{aligned} \tag{5.15}$$

where the last equation follows because if either $\tau_1 \leq t$ or $\tau_2 \leq t$ occurs, then the integrand is 0. Thus

$$\begin{aligned}|\Delta^2 f_A(\xi; x, y)| &\leq \mathbb{E}\int_0^\infty e^{-2t}|1_A(\mathbf{Z}_\xi(t)+\delta_x+\delta_y) - 1_A(\mathbf{Z}_\xi(t)+\delta_x) \\ &\qquad - 1_A(\mathbf{Z}_\xi(t)+\delta_y) + 1_A(\mathbf{Z}_\xi(t))|dt \\ &\leq 2\int_0^\infty e^{-2t}dt = 1. \end{aligned}$$

■

Remark 5.13: If we consider the Prohorov metric for Poisson process approximation, the Stein factors will be the same as those for the total variation metric [see Xia (1994)].

Theorem 5.3, together with Proposition 5.12, gives the following bounds for Poisson process approximation in terms of the total variation distance.

Theorem 5.14: *We have*

$$\begin{aligned} d_{TV}(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) &\leq \mathbb{E}\int_{\alpha\in\Gamma}(\Xi(A_\alpha)-1)\Xi(d\alpha) \\ &\quad + \min\{\epsilon_1, \epsilon_2\} + \int_{\alpha\in\Gamma}\boldsymbol{\lambda}(A_\alpha)\boldsymbol{\lambda}(d\alpha), \end{aligned} \tag{5.16}$$

where

$$\epsilon_1 = \int_{\alpha\in\Gamma}\mathbb{E}|g(\alpha, \Xi|_{A_\alpha^c}) - \phi(\alpha)|\boldsymbol{\mu}(d\alpha)$$

which is valid if Ξ *is a simple point process, and*

$$\epsilon_2 = \mathbb{E}\int_{\alpha\in\Gamma}\|\Xi|_{A_\alpha^c} - \Xi_\alpha|_{A_\alpha^c}\|\boldsymbol{\lambda}(d\alpha),$$

with $\|\xi_1 - \xi_2\| := \int_\Gamma |\xi_1(dx) - \xi_2(dx)|$, *the variation distance between* ξ_1 *and* ξ_2.

5.4. *Poisson process approximation in the d_2-metric*

In the one-dimensional approximation of $\Xi(B)$, a change in the location of a point makes no difference to $\Xi(B)$ as long as the old and new locations of the point are either both in B or both outside it. However, when considering the total variation metric for the distributions of point processes, even a small shift in the locations of points will result in a totally different configuration, at distance 1 from the original. The metric d_2 lies somewhere between the

two extremes. From our experience in the preceding two sections, if we are to apply Stein's method successfully for the d_2–metric, the first thing that we must do is to "squeeze out" the best Stein's factors for $\Delta f(\xi)$ and $\Delta^2 f(\xi)$, when f is the solution to the Stein Equation (5.1) for any test function $h \in \Psi$, where Ψ, defined in (4.2), is the set of test functions relevant to the d_2–metric. The proofs in this section will be lengthy, but nevertheless, the final result in Theorem 5.27 is widely applicable; see, for example, the beautiful conclusion in Theorem 6.8.

Naturally, the first aim is to find uniform bounds for $\Delta f(\xi)$ and $\Delta^2 f(\xi)$, as considered in Barbour & Brown (1992a) and in Barbour, Holst & Janson (1992), pp. 217-225. The first lemma investigates how much change results from changing the location of a single point.

Lemma 5.15: *If $h \in \Psi$, then*

$$\begin{aligned}&|[f(\xi + \delta_x + \delta_\alpha) - f(\xi + \delta_\alpha)] - [f(\xi + \delta_x + \delta_\beta) - f(\xi + \delta_\beta)]| \\ &\quad \le d_0(\alpha, \beta) \min\left\{1, \frac{1}{\lambda}\left(\frac{1}{\lambda} + 2\ln^+ \lambda\right)\right\}.\end{aligned}$$

This lemma is used in establishing the following three propositions, which are central to the proof of the main theorem. Note that two new Stein factors, appropriate to d_2–approximation, emerge:

$$c_1(\lambda) := \left\{1 \wedge \frac{1.647}{\sqrt{\lambda}}\right\} \quad \text{and} \quad c_2(\lambda) := \left\{1 \wedge \left[\frac{11}{6\lambda}\left(1 + 2\ln^+\left(\frac{6\lambda}{11}\right)\right)\right]\right\}. \tag{5.17}$$

Proposition 5.16: *If $h \in \Psi$, then*

$$\Delta f(\xi) \le c_1(\lambda) \text{ for all } \xi \in \mathcal{H}.$$

Proposition 5.17: *If $h \in \Psi$, then*

$$\Delta^2 f(\xi) \le c_2(\lambda) \text{ for all } \xi \in \mathcal{H}.$$

Proposition 5.18: *If $h \in \Psi$, then*

$$|[f(\xi + \delta_\alpha) - f(\xi)] - [f(\eta + \delta_\alpha) - f(\eta)]| \le c_2(\lambda) d_1'(\xi, \eta).$$

The proofs of all four results make use of Proposition 3.5. Let τ, τ_1 and τ_2 be independent negative exponentially distributed random variables with mean 1, and let $U, U_1, U_2, \cdots$ be independent Γ-valued random elements having distribution $\boldsymbol{\lambda}/\lambda$; assume that the U's, τ's, $\mathbf{D}_\xi$ and $\mathbf{Z}_0$ are all independent as well. As before, we write

$$n := |\xi|, \ Z_n(t) := |\mathbf{Z}_\xi(t)|, \ Z_0(t) := |\mathbf{Z}_0(t)| \text{ and } \lambda_t = \lambda(1 - e^{-t}). \tag{5.18}$$

Proof: *[Lemma 5.15]*. We have

$$\begin{aligned}
&|[f(\xi+\delta_x+\delta_\alpha)-f(\xi+\delta_\alpha)]-[f(\xi+\delta_x+\delta_\beta)-f(\xi+\delta_\beta)]| \\
&\leq \int_0^\infty |\mathbb{E}[h\left(\mathbf{Z}_\xi(t)+\delta_x 1_{\tau_1>t}+\delta_\alpha 1_{\tau_2>t}\right)-h\left(\mathbf{Z}_\xi(t)+\delta_\alpha 1_{\tau_2>t}\right) \\
&\quad -h\left(\mathbf{Z}_\xi(t)+\delta_x 1_{\tau_1>t}+\delta_\beta 1_{\tau_2>t}\right)+h\left(\mathbf{Z}_\xi(t)+\delta_\beta 1_{\tau_2>t}\right)]|dt \\
&= \int_0^\infty e^{-2t}|\mathbb{E}[h\left(\mathbf{Z}_\xi(t)+\delta_x+\delta_\alpha\right)-h\left(\mathbf{Z}_\xi(t)+\delta_\alpha\right) \\
&\quad -h\left(\mathbf{Z}_\xi(t)+\delta_x+\delta_\beta\right)+h\left(\mathbf{Z}_\xi(t)+\delta_\beta\right)]|dt \\
&\leq d_0(\alpha,\beta)\mathbb{E}\int_0^\infty \frac{2e^{-2t}}{Z_n(t)+1}dt \;\leq\; d_0(\alpha,\beta).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\int_0^\infty \frac{e^{-2t}}{Z_n(t)+1}dt \;\leq\; \mathbb{E}\int_0^\infty \frac{e^{-2t}}{Z_0(t)+1}dt \\
&= \int_0^\infty e^{-2t}\cdot\frac{1-e^{-\lambda_t}}{\lambda_t}dt \;=\; \int_0^1 \frac{(1-s)(1-e^{-\lambda s})}{\lambda s}ds \\
&< \int_0^{1/\lambda}(1-s)ds+\int_{1/\lambda}^1 \frac{(1-s)}{\lambda s}ds \;=\; \frac{1}{\lambda}\left(\frac{1}{2\lambda}+\ln\lambda\right),
\end{aligned}$$

for all $\lambda \geq 1$, completing the proof. ∎

Proof: *[Proposition 5.16]*. The proof is essentially the same as that in Barbour & Brown (1992a), with minor modification to simplify the reasoning. The bound of 1 is obvious as the test functions in Ψ are also test functions of the total variation metric. For a λ-dependent bound, we have

$$\begin{aligned}
&|f(\xi+\delta_\alpha)-f(\xi)| \\
&= \left|\mathbb{E}\int_0^\infty [h(\mathbf{Z}_\xi(t)+\delta_\alpha 1_{\tau>t})-h(\mathbf{Z}_\xi(t))]dt\right| \\
&= \left|\int_0^\infty e^{-t}\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha)-h(\mathbf{Z}_\xi(t))]dt\right| \\
&\leq \int_0^\infty e^{-t}\{1\wedge|\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha)-h(\mathbf{Z}_\xi(t))]|\}dt \\
&\leq \int_0^\infty e^{-t}\{1\wedge\{|\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha)-h(\mathbf{Z}_\xi(t)+\delta_U)]| \\
&\qquad +|\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_U)-h(\mathbf{Z}_\xi(t))]|\}\}dt. \qquad (5.19)
\end{aligned}$$

Next,

$$
\begin{aligned}
|\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha)-h(\mathbf{Z}_\xi(t)+\delta_U)]| &\le \mathbb{E}\left\{\frac{1}{Z_n(t)+1}\right\} \\
&= \int_0^1 \mathbb{E}z^{Z_n(t)}dz = \int_0^1 (1-e^{-t}(1-z))^n e^{-\lambda_t(1-z)}dz \\
&\le \int_0^1 e^{-\{ne^{-t}+\lambda_t\}(1-z)}dz = \int_0^1 e^{-\{ne^{-t}+\lambda_t\}z}dz \qquad (5.20) \\
&\le \frac{1-e^{-\lambda_t}}{\lambda_t}, \qquad (5.21)
\end{aligned}
$$

where we take $n=0$ for the last inequality, and, since

$$
\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_U)|Z_0(t)=k] = \mathbb{E}[h(\mathbf{Z}_\xi(t))|Z_0(t)=k+1] := a_{k+1}(t), \quad (5.22)
$$

we have

$$
\begin{aligned}
&|\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_U)-h(\mathbf{Z}_\xi(t)]| \\
&= \left|\sum_{k=0}^{\infty}\{\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_U)|Z_0(t)=k]-\mathbb{E}[h(\mathbf{Z}_\xi(t))|Z_0(t)=k]\}\mathbb{P}(Z_0(t)=k)\right| \\
&= \left|\sum_{k=0}^{\infty} a_k(t)[\mathbb{P}(Z_0(t)=k-1)-\mathbb{P}(Z_0(t)=k)]\right| \\
&= \max_k \mathbb{P}(Z_0(t)=k) \le \frac{1}{\sqrt{2e\lambda_t}}, \qquad (5.23)
\end{aligned}
$$

where the last inequality is again from Barbour, Holst & Janson (1992), p. 262. Finally, combining (5.19), (5.21) and (5.23) gives

$$
\begin{aligned}
&|f(\xi+\delta_\alpha)-f(\xi)| \\
&\le \int_0^\infty e^{-t}\left\{1\wedge\left[\frac{1-e^{-\lambda_t}}{\lambda_t}+\frac{1}{\sqrt{2e\lambda_t}}\right]\right\}dt \\
&\le \int_0^1 \left\{1\wedge\left[\frac{1-e^{-\lambda s}}{\lambda s}+\frac{1}{\sqrt{2e\lambda s}}\right]\right\}ds \le \frac{1}{\lambda}+\int_{\frac{1}{\lambda}}^1\left[\frac{1-e^{-\lambda}}{\lambda s}+\frac{1}{\sqrt{2e\lambda s}}\right]ds \\
&= \frac{1}{\lambda}\left[1+(1-e^{-\lambda})\ln\lambda+\sqrt{\frac{2}{e}}(\sqrt{\lambda}-1)\right] \le \frac{1.647}{\sqrt{\lambda}},
\end{aligned}
$$

where the last inequality is from direct calculation. ∎

Proof: *[Proposition 5.17]*. Replacing h by $1-h$ if necessary, it suffices to show that

$$
f(\xi+\delta_\alpha+\delta_\beta)-f(\xi+\delta_\alpha)-f(\xi+\delta_\beta)+f(\xi) \le c_2(\lambda).
$$

Since $h \in \Psi$ is also a test function for the total variation metric, the bound 1 is obvious. We now establish the λ-dependent uniform bound. To this end, by the same reasoning as for (5.15), we have

$$\begin{aligned} & f(\xi+\delta_\alpha+\delta_\beta) - f(\xi+\delta_\alpha) - f(\xi+\delta_\beta) + f(\xi) \\ &= -\int_0^\infty e^{-2t}\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha+\delta_\beta) - h(\mathbf{Z}_\xi(t)+\delta_\alpha) \\ &\qquad\qquad - h(\mathbf{Z}_\xi(t)+\delta_\beta) + h(\mathbf{Z}_\xi(t))]\,dt. \end{aligned}$$

Next, for $k \geq 0$, let

$$b_k(t) = \frac{1}{2}\left\{\mathbb{E}h\left(\mathbf{D}_\xi(t)+\delta_\alpha+\sum_{i=1}^{k}\delta_{U_i}\right) + \mathbb{E}h\left(\mathbf{D}_\xi(t)+\delta_\beta+\sum_{i=1}^{k}\delta_{U_i}\right)\right\},$$

and define $b_{-1}(t) = \mathbb{E}h\left(\mathbf{D}_\xi(t)\right)$; then we have

$$\begin{aligned} & \mathbb{E}h(\mathbf{Z}_\xi(t)+\delta_\alpha+\delta_\beta) \\ &= \sum_{k=0}^{\infty}\mathbb{E}h\left(\mathbf{D}_\xi(t)+\delta_\alpha+\delta_\beta+\sum_{i=1}^{k}\delta_{U_i}\right)\mathbb{P}(Z_0(t)=k) \\ &= \sum_{k=0}^{\infty}\mathbb{P}(Z_0(t)=k)\left\{b_{k+1}(t)\right. \\ &\qquad \left. + \left[\mathbb{E}h\left(\mathbf{D}_\xi(t)+\delta_\alpha+\delta_\beta+\sum_{i=1}^{k}\delta_{U_i}\right) - b_{k+1}(t)\right]\right\} \\ &\leq \sum_{k=0}^{\infty}\left\{\mathbb{E}\left(\frac{1}{|\mathbf{D}_\xi(t)|+k+2}\right) + b_{k+1}(t)\right\}\mathbb{P}(Z_0(t)=k) \\ &= \mathbb{E}\left\{\frac{1}{Z_n(t)+2}\right\} + \sum_{k=0}^{\infty} b_{k+1}(t)\mathbb{P}(Z_0(t)=k), \end{aligned} \tag{5.24}$$

and

$$\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha) + h(\mathbf{Z}_\xi(t)+\delta_\beta)] = 2\sum_{k=0}^{\infty} b_k(t)\mathbb{P}(Z_0(t)=k), \tag{5.25}$$

and finally

$$\begin{aligned}
\mathbb{E}h(\mathbf{Z}_\xi(t)) &= \sum_{k=0}^{\infty} \mathbb{P}(Z_0(t)=k) \\
&\qquad \times \left\{ b_{k-1}(t) + \left[\mathbb{E}h\left(\mathbf{D}_\xi(t) + \sum_{i=1}^{k} \delta_{U_i} \right) - b_{k-1}(t) \right] \right\} \\
&\leq \sum_{k=1}^{\infty} \mathbb{E}\left\{ \frac{1}{|\mathbf{D}_\xi(t)|+k} \right\} \mathbb{P}(Z_0(t)=k) + \sum_{k=0}^{\infty} b_{k-1}(t)\mathbb{P}(Z_0(t)=k) \\
&= \mathbb{E}\left\{ \frac{1_{Z_n(t)\geq 1}}{Z_n(t)} \right\} + \sum_{k=0}^{\infty} b_{k-1}(t)\mathbb{P}(Z_0(t)=k). \qquad (5.26)
\end{aligned}$$

Combining (5.24), (5.25) and (5.26) gives

$$\begin{aligned}
&\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha+\delta_\beta) - h(\mathbf{Z}_\xi(t)+\delta_\alpha) - h(\mathbf{Z}_\xi(t)+\delta_\beta) + h(\mathbf{Z}_\xi(t))] \\
&\leq \mathbb{E}\left\{ \frac{1}{Z_n(t)+2} \right\} + \mathbb{E}\left\{ \frac{1_{Z_n(t)\geq 1}}{Z_n(t)} \right\} \\
&\qquad + \sum_{k=0}^{\infty} [b_{k+1}(t) - 2b_k(t) + b_{k-1}(t)]\mathbb{P}(Z_0(t)=k) \\
&= \mathbb{E}\left\{ \frac{1}{Z_n(t)+2} \right\} + \mathbb{E}\left\{ \frac{1_{Z_n(t)\geq 1}}{Z_n(t)} \right\} \\
&\quad + \sum_{k=-1}^{\infty} b_k(t)[\mathbb{P}(Z_0(t)=k-1) - 2\mathbb{P}(Z_0(t)=k) + \mathbb{P}(Z_0(t)=k+1)].
\end{aligned}$$

Since $0 \leq b_k(t) \leq 1$, we have

$$\begin{aligned}
&\sum_{k=-1}^{\infty} b_k(t)[\mathbb{P}(Z_0(t)=k-1) - 2\mathbb{P}(Z_0(t)=k) + \mathbb{P}(Z_0(t)=k+1)] \\
&= \sum_{k=-1}^{\infty} b_k(t)\mathbb{P}(Z_0(t)=k+1)\left[\left(\frac{k+1}{\lambda_t} - 1 \right)^2 - \frac{k+1}{(\lambda_t)^2} \right] \\
&\leq \sum_{k=-1}^{\infty} \mathbb{P}(Z_0(t)=k+1)\left(\frac{k+1}{\lambda_t} - 1 \right)^2 = \frac{1}{\lambda_t},
\end{aligned}$$

and, by (5.21),

$$\mathbb{E}\left\{ \frac{1}{Z_n(t)+2} \right\} + \mathbb{E}\left\{ \frac{1_{Z_n(t)\geq 1}}{Z_n(t)} \right\} \leq \frac{8}{3}\mathbb{E}\left\{ \frac{1}{Z_n(t)+1} \right\} \leq \frac{8/3}{\lambda_t}.$$

Hence,

$$\mathbb{E}[h(\mathbf{Z}_\xi(t)+\delta_\alpha+\delta_\beta)-h(\mathbf{Z}_\xi(t)+\delta_\alpha)-h(\mathbf{Z}_\xi(t)+\delta_\beta)+h(\mathbf{Z}_\xi(t))] \le \left\{2\wedge\frac{11/3}{\lambda_t}\right\},$$

and

$$\int_0^\infty e^{-2t}\left(2\wedge\frac{11/3}{\lambda_t}\right)dt = 2\int_0^1(1-s)\left(1\wedge\frac{11/6}{\lambda s}\right)ds = \frac{(11/6)^2}{\lambda^2}+\frac{11/3}{\lambda}\ln\left(\frac{\lambda}{11/6}\right) \le \frac{11}{6\lambda}\left(1+2\ln^+\left(\frac{6\lambda}{11}\right)\right),$$

for all $\lambda \ge 11/6$. ■

Proof: *[Proposition 5.18]*. Write $\xi=\sum_{i=1}^n\delta_{x_i}$ and $\eta=\sum_{i=1}^m\delta_{y_i}$. Without loss of generality, we assume that $m\ge n$ and that

$$d_1'(\xi,\eta)=\sum_{i=1}^n d_0(x_i,y_i)+(m-n).$$

Define $\xi'=\sum_{i=1}^n\delta_{y_i}$; then

$$\begin{aligned}&|[f(\xi+\delta_\alpha)-f(\xi)]-[f(\eta+\delta_\alpha)-f(\eta)]|\\ &\le |[f(\xi+\delta_\alpha)-f(\xi)]-[f(\xi'+\delta_\alpha)-f(\xi')]|\\ &\quad+|[f(\xi'+\delta_\alpha)-f(\xi')]-[f(\eta+\delta_\alpha)-f(\eta)]|.\end{aligned}$$

It follows from (5.6) that

$$|[f(\xi'+\delta_\alpha)-f(\xi')]-[f(\eta+\delta_\alpha)-f(\eta)]|\le c_2(\lambda)(m-n).$$

Next, set $\eta_j=\sum_{i=1}^j\delta_{x_i}+\sum_{i=j+1}^n\delta_{y_i}$, $0\le j\le n$. Then

$$\begin{aligned}&|[f(\xi+\delta_\alpha)-f(\xi)]-[f(\xi'+\delta_\alpha)-f(\xi')]|\\ &\le\sum_{j=0}^{n-1}|[f(\eta_j+\delta_\alpha)-f(\eta_j)]-[f(\eta_{j+1}+\delta_\alpha)-f(\eta_{j+1})]|,\end{aligned}$$

so applying Lemma 5.15 with the fact that

$$\min\left\{1,\frac{1}{\lambda}\left(\frac{1}{\lambda}+2\ln^+\lambda\right)\right\}\le c_2(\lambda),$$

we have

$$|[f(\xi+\delta_\alpha)-f(\xi)]-[f(\xi'+\delta_\alpha)-f(\xi')]|\le c_2(\lambda)\sum_{i=1}^n d_0(x_i,y_i).$$ ■

With these estimates of Stein's factors, we are now ready to state our first main result in terms of the d_2 metric.

Theorem 5.19: *If Ξ is a point process on Γ with mean measure $\boldsymbol{\lambda}$, then*

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}\,(\boldsymbol{\lambda}))$$
$$\leq c_2(\lambda)\mathbb{E}\int_{\alpha\in\Gamma}(\Xi(A_\alpha)-1)\Xi(d\alpha) + \min\{\epsilon_1,\epsilon_2\} + c_2(\lambda)\int_{\alpha\in\Gamma}\boldsymbol{\lambda}(d\alpha)\boldsymbol{\lambda}(A_\alpha),$$

where

$$\epsilon_1 = c_1(\lambda)\mathbb{E}\int_{\alpha\in\Gamma}|g(\alpha,\Xi|_{A_\alpha^c}) - \phi(\alpha)|\boldsymbol{\mu}(d\alpha)$$

which is valid if Ξ is a simple point process, and

$$\epsilon_2 = c_2(\lambda)\mathbb{E}\int_{\alpha\in\Gamma} d_1'(\Xi|_{A_\alpha^c}, \Xi_\alpha|_{A_\alpha^c})\boldsymbol{\lambda}(d\alpha).$$

Remark 5.20: The bound $c_2(\lambda)$ is of the right order and the factor $\ln^+\left(\frac{6\lambda}{11}\right)$ in Proposition 5.17 cannot be removed.

To see this, we use an example from Brown & Xia (1995b). Suppose that (S,d) is a metric space, and that $a, b \notin S$ are two isolated points; that is, that $d(a,b)=1$, and that $d(a,x)=1$ and $d(b,x)=1$ for all $x\in S$. Let $\Gamma = S\cup\{a,b\}$. For $n\geq 1$, $x_1,\cdots,x_n\in S$ and two integers $m_1, m_2 \geq 0$, define a test function

$$h\left(\sum_{i=1}^n \delta_{x_i} + m_1\delta_a + m_2\delta_b\right) = \begin{cases} \frac{1}{n+2}, & \text{if } m_1 = m_2 = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then, by direct verification, h is in the set Ψ of d_2-test functions (4.2). Furthermore, if $supp(\boldsymbol{\lambda}) \subset S$, then $\mathbf{Z}_0(t)$ satisfies $\mathbf{Z}_0(t)(\{a,b\}) = 0$, *a.s.* Note that $Z_0(t) \sim \mathrm{Po}\,(\lambda_t)$, and that

$$\sum_{i=0}^\infty \frac{e^{-\nu}\nu^i}{(i+2)i!} = \frac{e^{-\nu}}{\nu^2}\sum_{i=0}^\infty \frac{1}{i!}\int_0^\nu s^{i+1}ds = \frac{e^{-\nu}}{\nu^2}\int_0^\nu s\sum_{i=0}^\infty \frac{1}{i!}s^i ds$$
$$= \frac{e^{-\nu}}{\nu^2}\int_0^\nu se^s ds = \frac{\nu-1+e^{-\nu}}{\nu^2}.$$

Hence we have

$$\begin{aligned}
&|f(\delta_a+\delta_b)-f(\delta_a)-f(\delta_b)+f(0)| \\
&= \int_0^\infty e^{-2t}\mathbb{E}\{h(\mathbf{Z}_0(t)+\delta_a+\delta_b)-h(\mathbf{Z}_0(t)+\delta_a) \\
&\qquad\qquad -h(\mathbf{Z}_0(t)+\delta_b)+h(\mathbf{Z}_0(t))\}dt \\
&= \int_0^\infty e^{-2t}\mathbb{E}\left\{\frac{1}{Z_0(t)+2}\right\}dt = \int_0^\infty e^{-2t}\sum_{n\geq 0}\frac{e^{-\lambda_t}\lambda_t^n}{(n+2)n!}dt \\
&= \int_0^\infty e^{-2t}\left\{\frac{\lambda_t-1+e^{-\lambda_t}}{\lambda_t^2}\right\}dt \\
&= \frac{1}{\lambda}\int_0^\lambda \frac{r-1+e^{-r}}{r^2}dr - \frac{1}{\lambda^2}\int_0^\lambda \frac{r-1+e^{-r}}{r}dr \asymp \frac{\ln\lambda}{\lambda}, \text{ as } \lambda\to\infty.
\end{aligned}$$

If one takes off the assumption of $supp(\boldsymbol{\lambda}) \subset S$, but supposes that $\lim_{\lambda\to\infty}\boldsymbol{\lambda}(\{a,b\})=0$, then $\Delta^2 f$ still has the order $\frac{\ln\lambda}{\lambda}$.

On the other hand, the example also tells us that the problem causing the trouble is that the number of points in the initial state is small and when we consider Poisson process approximation, one can expect that with more than 95% chance the number of points of Ξ, the process being approximated, is within $3\sqrt{\lambda}$ of λ. Hence, Brown, Weinberg & Xia (2000) proposed a non-uniform bound for the second difference of f, allowing the bound depend on the configurations involved. The bound in Brown, Weinberg & Xia (2000) is rather complicated and it contains 4 terms, leading to unnecessary complication in applications. Brown & Xia (2001) simplified the estimate, and we now follow their ideas to establish a non-uniform bound.

Proposition 5.21: *For each $h\in\Psi$, $x,y\in\Gamma$ and $\xi\in\mathcal{H}$ with $|\xi|=n$, the solution f of (5.1) satisfies*

$$|\Delta^2 f(\xi;x,y)| \leq \frac{3.5}{\lambda}+\frac{2.5}{n+1}. \tag{5.27}$$

This estimate is often used together with the following lemma to extract a Stein factor of order $1/\lambda$.

Lemma 5.22: [Brown, Weinberg & Xia (2000), Lemma 3.1] *For a random variable $X\geq 1$,*

$$\mathbb{E}\left(\frac{1}{X}\right) \leq \frac{\sqrt{\kappa(1+\frac{\kappa}{4})}+1+\frac{\kappa}{2}}{\mathbb{E}(X)}, \tag{5.28}$$

where $\kappa=\frac{\mathrm{Var}(X)}{\mathbb{E}(X)}$.

Proof: Observe that by Jensen's inequality, $\mathbb{E}(\frac{1}{X}) \geq \frac{1}{\mathbb{E}(X)}$. Now define $\Delta := \mathbb{E}\left(\frac{1}{X}\right) - \frac{1}{\mathbb{E}(X)}$. Then, by an application of the Cauchy-Schwarz inequality, and because, if $X \geq 1$ a.s., then $\mathbb{E}(1/X^2) \leq \mathbb{E}(1/X)$, it follows that

$$0 \leq \Delta = \mathbb{E}\left(\frac{\mathbb{E}(X) - X}{X\mathbb{E}(X)}\right) \leq \sqrt{\frac{\kappa}{\mathbb{E}(X)}}\sqrt{\Delta + \frac{1}{\mathbb{E}(X)}}. \tag{5.29}$$

The bound (5.28) follows by squaring (5.29) and solving for Δ. ∎

To prove Proposition 5.21, we need a few technical lemmas. We continue to use the notation in Sections 3.1 and 3.2 with relation (5.18). Let us first consider the change of f when we shift the location of one point.

Lemma 5.23: *If $h \in \Psi$ and $|\xi| = n$, then*

$$|f(\xi + \delta_\alpha) - f(\xi + \delta_\beta)| \leq d_0(\alpha, \beta) \min\left\{1, \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{n}\right), \frac{1}{2\lambda} + \frac{1}{n+1}, \frac{1 + \ln^+ \lambda}{\lambda}\right\}.$$

Proof: [cf Lemma 2.2 of Brown, Weinberg & Xia (2000)] Clearly,

$$\begin{aligned} &|f(\xi + \delta_\alpha) - f(\xi + \delta_\beta)| \\ &\quad \leq \int_0^\infty e^{-t} \left|\mathbb{E}[h(\mathbf{Z}_\xi(t) + \delta_\alpha) - h(\mathbf{Z}_\xi(t) + \delta_\beta)]\right| dt \\ &\quad \leq d_0(\alpha, \beta) \int_0^\infty e^{-t} \mathbb{E}\left\{\frac{1}{Z_n(t) + 1}\right\} dt \;\leq\; d_0(\alpha, \beta), \end{aligned}$$

and, by (5.20), for $n \geq 1$,

$$\begin{aligned} &\int_0^\infty e^{-t} \mathbb{E}\left\{\frac{1}{Z_n(t) + 1}\right\} dt \\ &\leq \int_0^\infty e^{-t} \left[\int_0^1 e^{-[ne^{-t} + \lambda_t]z} dz\right] dt = \int_0^1 \int_0^1 e^{-[n(1-s) + \lambda s]z} dz ds \qquad (5.30) \\ &\leq \int_0^1 \frac{1}{n(1-s) + \lambda s} ds \;=\; \frac{1}{2} \int_0^1 \left\{\frac{1}{ns + \lambda(1-s)} + \frac{1}{n(1-s) + \lambda s}\right\} ds \\ &\leq \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{n}\right) \;\leq\; \frac{1}{2\lambda} + \frac{1}{n+1}, \end{aligned}$$

where the penultimate inequality is due to the fact that, when $n \neq \lambda$, the function

$$\frac{1}{ns + \lambda(1-s)} + \frac{1}{n(1-s) + \lambda s}$$

attains its maximum $\frac{1}{n}+\frac{1}{\lambda}$ at $s=0,\ 1$ and its minimum at $s=0.5$. If $n=0$, the claim is obvious.

For the uniform bound, if $\lambda \le 1$, the claim is obvious; for $\lambda > 1$, taking the worst case of $n=0$ in (5.30), we get

$$\int_0^1\int_0^1 e^{-[n(1-s)+\lambda s]z}dzds \le \left(\int_0^{1/\lambda}+\int_{1/\lambda}^1\right)\frac{1-e^{-\lambda s}}{\lambda s}ds \le \frac{\ln\lambda+1}{\lambda}. \quad \blacksquare$$

Remark 5.24: It might be true that

$$|f(\xi+\delta_\alpha)-f(\xi+\delta_\beta)| \le \frac{d_0(\alpha,\beta)}{2}\left(\frac{1}{\lambda}+\frac{1}{n+1}\right).$$

Lemma 5.25: *Let* $e_n^+ = \mathbb{E}\exp(-\tau_n^+)$ *and* $e_n^- = \mathbb{E}\exp(-\tau_n^-)$; *then*

$$e_n^+ = \frac{\lambda F(n)}{(n+1)F(n+1)},\quad e_n^- = 1+\frac{n}{\lambda}-\frac{\bar{F}(n-1)}{\bar{F}(n)}. \tag{5.31}$$

Proof: By conditioning on the time of the first jump after leaving the initial state, we can produce recurrence relations for e_n^+ and e_n^-:

$$e_n^+ = \frac{\lambda}{\lambda+n+1-ne_{n-1}^+},\ n\ge 1; \qquad e_0^+ = \frac{\lambda}{\lambda+1}; \tag{5.32}$$

$$e_n^- = \frac{n+\lambda}{\lambda}-\frac{n-1}{\lambda e_{n-1}^-},\ n\ge 2, \tag{5.33}$$

[see Wang & Yang (1992), pp 154-156 and pp 174-176]. However, it is rather difficult to get e_1^- and we quote it from [Wang & Yang (1992), pp 174-176]:

$$e_1^- = \frac{\lambda+1}{\lambda}-\frac{1}{\lambda\int_0^\infty p_{00}(t)e^{-t}dt},$$

where $p_{00}(t)=\mathbb{P}(|\mathbf{Z}_0(t)|=0)$. Now, the claim for e_n^+ follows from (5.32) and mathematical induction. For the other claim, since $|\mathbf{Z}_0(t)| \sim \text{Po}\,(\lambda_t)$, we get

$$e_1^- = \frac{\lambda+1}{\lambda}-\frac{1}{1-e^{-\lambda}}.$$

Using mathematical induction in (5.33) gives the claim for e_n^-. $\blacksquare$

Proof: *[Proposition 5.21]*. Write $\xi=\sum_{i=1}^n \delta_{z_i}$. By conditioning on the first jump time after leaving state $\xi+\delta_x$, it follows from (5.4) that

$$\begin{aligned} f(\xi+\delta_x) &= \frac{-[h(\xi+\delta_x)-\text{Po}\,(\boldsymbol{\lambda})(h)]}{n+1+\lambda}+\frac{\lambda}{n+1+\lambda}\mathbb{E}f(\xi+\delta_x+\delta_U) \\ &\quad + \sum_{\alpha\in\{z_1,\cdots,z_n,x\}}\frac{1}{n+1+\lambda}f(\xi+\delta_x-\delta_\alpha), \end{aligned}$$

where $U \sim \boldsymbol{\lambda}/\lambda$. Rearranging the equation gives

$$\mathbb{E}f(\xi+\delta_x+\delta_U) = \frac{h(\xi+\delta_x)-\mathrm{Po}\,(\boldsymbol{\lambda})(h)}{\lambda} + \frac{n+1+\lambda}{\lambda} f(\xi+\delta_x) - \frac{1}{\lambda} \sum_{\alpha\in\{z_1,\cdots,z_n,x\}} f(\xi+\delta_x-\delta_\alpha).$$

This in turn yields

$$\begin{aligned}
&\Delta^2 f(\xi;x,y)\\
&= [f(\xi+\delta_x+\delta_y) - \mathbb{E}f(\xi+\delta_x+\delta_U)] + [f(\xi+\delta_x) - f(\xi+\delta_y)]\\
&\quad + \frac{1}{n+1} \sum_{\alpha\in\{z_1,\cdots,z_n\}} [f(\xi) - f(\xi+\delta_x-\delta_\alpha)] + \frac{h(\xi+\delta_x)-\mathrm{Po}\,(\boldsymbol{\lambda})(h)}{\lambda}\\
&\quad + \frac{n+1-\lambda}{\lambda} \left[f(\xi+\delta_x) - \frac{1}{n+1} \sum_{\alpha\in\{z_1,\cdots,z_n,x\}} f(\xi+\delta_x-\delta_\alpha) \right]. \qquad (5.34)
\end{aligned}$$

Swap x and y to get

$$\begin{aligned}
&\Delta^2 f(\xi;x,y)\\
&= [f(\xi+\delta_x+\delta_y) - \mathbb{E}f(\xi+\delta_y+\delta_U)] + [f(\xi+\delta_y) - f(\xi+\delta_x)]\\
&\quad + \frac{1}{n+1} \sum_{\alpha\in\{z_1,\cdots,z_n\}} [f(\xi) - f(\xi+\delta_y-\delta_\alpha)] + \frac{h(\xi+\delta_y)-\mathrm{Po}\,(\boldsymbol{\lambda})(h)}{\lambda}\\
&\quad + \frac{n+1-\lambda}{\lambda} \left[f(\xi+\delta_y) - \frac{1}{n+1} \sum_{\alpha\in\{z_1,\cdots,z_n,y\}} f(\xi+\delta_y-\delta_\alpha) \right]. \qquad (5.35)
\end{aligned}$$

Adding up (5.34) and (5.35) and then dividing by 2, we obtain

$$\begin{aligned}
&|\Delta^2 f(\xi;x,y)|\\
&\le \max_{z\in\{x,y\}} |f(\xi+\delta_x+\delta_y) - \mathbb{E}f(\xi+\delta_z+\delta_U)|\\
&\quad + \frac{1}{2(n+1)} \sum_{z\in\{x,y\}} \sum_{\alpha\in\{z_1,\cdots,z_n\}} |f(\xi) - f(\xi+\delta_z-\delta_\alpha)|\\
&\quad + \max_{z\in\{x,y\}} \left| \frac{h(\xi+\delta_z)-\mathrm{Po}\,(\boldsymbol{\lambda})(h)}{\lambda} \right| \qquad (5.36)\\
&\quad + \max_{z\in\{x,y\}} \left| \frac{n+1-\lambda}{\lambda} \left[f(\xi+\delta_z) - \frac{1}{n+1} \sum_{\alpha\in\{z_1,\cdots,z_n,z\}} f(\xi+\delta_z-\delta_\alpha) \right] \right|.
\end{aligned}$$

Now, we estimate the four terms in (5.36) individually. First, using the fact that $0 \le h \le 1$, the third term is clearly bounded by $1/\lambda$. Next, we apply

Lemma 5.23 to conclude that the first term is dominated by $\frac{1}{2}\left(\frac{1}{n+1}+\frac{1}{\lambda}\right)$ and the second term is controlled by

$$\frac{1}{2(n+1)}\sum_{z\in\{x,y\}}\sum_{\alpha\in\{z_1,\cdots,z_n\}}\left(\frac{1}{n}+\frac{1}{2\lambda}\right)<\frac{1}{n+1}+\frac{1}{2\lambda}.$$

Finally, we show that, for $z=x$ or y,

$$\frac{n+1-\lambda}{\lambda}\left[f(\xi+\delta_z)-\frac{1}{n+1}\sum_{\alpha\in\{z_1,\cdots,z_n,z\}}f(\xi+\delta_z-\delta_\alpha)\right]\le\frac{1.5}{\lambda}+\frac{1}{n+1} \tag{5.37}$$

so that (5.27) follows from collecting the bounds for the terms in (5.36).

To prove (5.37), it is enough to consider the following two cases with $\lambda\ge 3.5$ since, if $\lambda<3.5$, the bound is at least 1, which is obviously correct.

Case (I): $n+1>\lambda$.

Since $\tau_{n+1}^-=\inf\{t:\ |\mathbf{Z}_{\xi+\delta_z}(t)|=n\}$, the strong Markov property gives

$$f(\xi+\delta_z)=-\mathbb{E}\int_0^{\tau_{n+1}^-}[h(\mathbf{Z}_{\xi+\delta_z}(t))-\mathrm{Po}\,(\boldsymbol{\lambda})(h)]dt+\mathbb{E}f(\eta),$$

where $\eta=\mathbf{Z}_{\xi+\delta_z}(\tau_{n+1}^-)$. Consequently,

$$\begin{aligned}&\frac{n+1-\lambda}{\lambda}\left\{f(\xi+\delta_z)-\frac{1}{n+1}\sum_{\alpha\in\{z_1,\cdots,z_n,z\}}f(\xi+\delta_z-\delta_\alpha)\right\} \qquad (5.38)\\ &\le\frac{n+1-\lambda}{\lambda}\left\{\mathbb{E}\tau_{n+1}^-+\left[\mathbb{E}f(\eta)-\frac{1}{n+1}\sum_{\alpha\in\{z_1,\cdots,z_n,z\}}f(\xi+\delta_z-\delta_\alpha)\right]\right\}.\end{aligned}$$

However, by (3.2) and Proposition A.2.3 of Barbour, Holst & Janson (1992),

$$\frac{n+1-\lambda}{\lambda}\mathbb{E}\tau_{n+1}^-=\frac{(n+1-\lambda)\bar{F}(n+1)}{\lambda(n+1)\mathrm{Po}\,(\lambda)\{n+1\}}\le\frac{(n+1-\lambda)(n+2)}{\lambda(n+1)(n+2-\lambda)}\le\frac{1}{\lambda}. \tag{5.39}$$

To deal with the second term of (5.38), recall the decomposition (3.11), set $\mathbf{D}'(t):=\xi+\delta_z-\mathbf{D}_{\xi+\delta_z}(t)$, which records the individuals who died in $[0,t]$, we can write

$$\mathbf{Z}_{\xi+\delta_z}(t)=\xi+\delta_z+\mathbf{Z}_0(t)-\mathbf{D}'(t).$$

By the time τ_{n+1}^-, at least one of $z_1,\cdots,z_n,z$ has died, so $|\mathbf{D}'(\tau_{n+1}^-)|\ge 1$. Noting that the deaths happen to $z_1,\cdots,z_n,z$ with equal chance, we obtain

$$\mathbb{E}f(\eta)=\frac{1}{n+1}{\sum_\alpha}'\mathbb{E}f\left\{\xi+\delta_z-\delta_\alpha+\mathbf{Z}_0(\tau_{n+1}^-)-[\mathbf{D}'(\tau_{n+1}^-)-\delta_\alpha]\right\},$$

where $\sum'_\alpha$ denotes the sum over all $\alpha \in \{z_1, \cdots, z_n, z\}$ which have died; this, together with Lemma 5.23, yields

$$\begin{aligned}
&\mathbb{E}f(\eta) - \frac{1}{n+1} \sum_{\alpha \in \{z_1, \cdots, z_n, z\}} f(\xi + \delta_z - \delta_\alpha) \\
&\quad \le \mathbb{E}\left[\left|\mathbf{D}'(\tau_{n+1}^-)\right| - 1\right] \left(\frac{1}{2\lambda} + \frac{1}{2(n-1)}\right) \\
&\quad \le \mathbb{E}\left[\left|\mathbf{D}'(\tau_{n+1}^-)\right| - 1\right] \left(\frac{1}{2\lambda} + \frac{1}{n+1}\right), \qquad (5.40)
\end{aligned}$$

where the last inequality is due to the assumption that $n \ge \lambda - 1 \ge 2.5$, so that $n \ge 3$. On the other hand,

$$\mathbb{E}\left[\left|\mathbf{D}'(\tau_{n+1}^-)\right| - 1\right] = n - \mathbb{E}\left|\mathbf{D}_{\xi+\delta_z}(\tau_{n+1}^-)\right| = n - (n+1)\mathbb{P}(\zeta_1 > \tau_{n+1}^-),$$

[see Proposition 3.5]. Noting that $\zeta_1 > \tau_{n+1}^-$ is equivalent to

$$\zeta_1 > \inf\left\{t : \ |\mathbf{Z}_0(t)| + \sum_{i=2}^{n+1} 1_{\zeta_i > t} = n - 1\right\} := \tilde{\tau}_n^-,$$

with ζ_1 independent of $\tilde{\tau}_n^-$ and $\mathcal{L}(\tilde{\tau}_n^-) = \mathcal{L}(\tau_n^-)$, we have from (5.31) that

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbf{D}'(\tau_{n+1}^-)\right| - 1\right] &= n - (n+1)\mathbb{E}\exp(-\tau_n^-) \\
&= -1 + (n+1)\left(\frac{\bar{F}(n-1)}{\bar{F}(n)} - \frac{n}{\lambda}\right). \qquad (5.41)
\end{aligned}$$

Now, we claim that

$$\frac{n+1-\lambda}{\lambda}\mathbb{E}\left[\left|\mathbf{D}'(\tau_{n+1}^-)\right| - 1\right] \le 1. \qquad (5.42)$$

In fact, by (5.41), (5.42) is equivalent to

$$(n+1-\lambda)[\lambda\bar{F}(n-1) - n\bar{F}(n)] - \lambda\bar{F}(n) \le 0. \qquad (5.43)$$

Expanding the formula into a power series of λ, the left hand side of (5.43) becomes

$$\begin{aligned}
&2n\lambda\bar{F}(n) - \lambda^2\bar{F}(n-1) - (n+1)n\bar{F}(n+1) \\
&\quad = e^{-\lambda} \sum_{i=n+1}^{\infty} \frac{\lambda^i}{(i-1)!}[2n+1-i-(n+1)n/i],
\end{aligned}$$

which is obviously negative.

Thus, it follows from (5.40) and (5.42) that the second term of (5.38) is estimated by

$$\frac{n+1-\lambda}{\lambda}\left[\mathbb{E}f(\eta)-\frac{1}{n+1}\sum_{\alpha\in\{z_1,\cdots,z_n,z\}}f(\xi+\delta_z-\delta_\alpha)\right]\le\frac{1}{2\lambda}+\frac{1}{n+1}. \tag{5.44}$$

Combining (5.39) and (5.44) yields (5.37).

Case (II): $n+1\le\lambda$.

Likewise, since $\tau_n^+=\inf\{t:\ |\mathbf{Z}_{\xi+\delta_z-\delta_\alpha}(t)|=n+1\}$, the strong Markov property ensures that

$$f(\xi+\delta_z-\delta_\alpha)=-\mathbb{E}\int_0^{\tau_n^+}[h(\mathbf{Z}_{\xi+\delta_z-\delta_\alpha}(t)-\mathrm{Po}\,(\boldsymbol{\lambda})(h)]dt+\mathbb{E}f(\varsigma),$$

where $\varsigma=\mathbf{Z}_{\xi+\delta_z-\delta_\alpha}(\tau_n^+)$. Hence

$$\begin{aligned}&\frac{n+1-\lambda}{\lambda}[f(\xi+\delta_z)-f(\xi+\delta_z-\delta_\alpha)]\\&\quad\le\frac{\lambda-(n+1)}{\lambda}\mathbb{E}\tau_n^+ +\frac{\lambda-(n+1)}{\lambda}[\mathbb{E}f(\varsigma)-f(\xi+\delta_z)].\end{aligned} \tag{5.45}$$

Using (3.2) and Proposition A.2.3 of Barbour, Holst & Janson (1992) again, we have

$$\frac{\lambda-(n+1)}{\lambda}\mathbb{E}\tau_n^+=\frac{(\lambda-(n+1))F(n)}{\lambda^2\mathrm{Po}\,(\lambda)\{n\}}\le\frac{(\lambda-(n+1))\lambda}{\lambda^2(\lambda-n)}\le\frac{1}{\lambda}. \tag{5.46}$$

To estimate the second term of (5.45), without loss of generality, we may assume $\alpha=z$ and realize

$$\mathbf{Z}_\xi(t)=\mathbf{Z}_0(t)+\mathbf{D}_\xi(t),$$

so that

$$\begin{aligned}\mathbb{E}f(\varsigma)-f(\xi+\delta_z)&=\mathbb{E}[f(\mathbf{Z}_0(\tau_n^+)+\mathbf{D}_\xi(\tau_n^+))-f(\xi+\delta_z)]\\&\le\mathbb{E}(n+1-|\mathbf{D}_\xi(\tau_n^+)|)\left(\frac{1}{2\lambda}+\frac{1}{n+1}\right).\end{aligned} \tag{5.47}$$

But

$$\mathbb{E}(n+1-|\mathbf{D}_\xi(\tau_n^+)|)=n+1-n\mathbb{P}(\zeta_1>\tau_n^+).$$

Since $\zeta_1>\tau_n^+$ is equivalent to

$$\zeta_1>\inf\left\{t:\ |\mathbf{Z}_0(t)|+\sum_{i=2}^n 1_{\zeta_i>t}=n\right\}:=\tilde{\tau}_{n-1}^+,$$

with ζ_1 independent of $\tilde{\tau}^+_{n-1}$ and $\mathcal{L}(\tilde{\tau}^+_{n-1}) = \mathcal{L}(\tau^+_{n-1})$, we get from (5.31) that

$$\mathbb{E}(n+1-|\mathbf{D}_\xi(\tau^+_n)|) = n+1-n\mathbb{E}\exp(-\tau^+_{n-1}) = 1+\frac{nF(n)-\lambda F(n-1)}{F(n)}. \tag{5.48}$$

We now show that

$$\frac{\lambda-(n+1)}{\lambda}\mathbb{E}(n+1-|\mathbf{D}_\xi(\tau^+_n)|) \le 1. \tag{5.49}$$

Using (5.48), (5.49) can be rewritten as

$$(\lambda-(n+1))[nF(n)-\lambda F(n-1)]-(n+1)F(n) \le 0. \tag{5.50}$$

The left hand side of (5.50) can be expanded into the power series of λ as

$$\begin{aligned}
&\lambda nF(n)-(n+1)nF(n)-\lambda^2F(n-1)+(n+1)\lambda F(n-1)-(n+1)F(n)\\
&=e^{-\lambda}\sum_{i=0}^{n-1}\frac{\lambda^{i+1}}{(i+1)!}[(i+1)n-(n+1)n-(i+1)i+(n+1)(i+1)-(n+1)]\\
&\quad-(n+1)^2e^{-\lambda}\\
&=-e^{-\lambda}\sum_{i=0}^{n-1}\frac{\lambda^{i+1}}{(i+1)!}(n-i)^2-(n+1)^2e^{-\lambda}\le 0.
\end{aligned}$$

It now follows from (5.47) and (5.49), that

$$\frac{\lambda-(n+1)}{\lambda}[\mathbb{E}f(\varsigma)-f(\xi+\delta_z)] \le \frac{1}{2\lambda}+\frac{1}{n+1},$$

which, together with (5.45) and (5.46), gives (5.37). ∎

Lemma 5.26: *For $\xi,\eta\in\mathcal{H}$ and $x\in\Gamma$,*

$$\begin{aligned}
&|[f(\xi+\delta_\alpha)-f(\xi)]-[f(\eta+\delta_\alpha)-f(\eta)]|\\
&\quad\le\frac{2}{|\eta|\wedge|\xi|+1}[d'_1(\xi,\eta)-||\eta|-|\xi||]+\left(\frac{3.5}{\lambda}+\frac{2.5}{|\eta|\wedge|\xi|+1}\right)||\eta|-|\xi||\\
&\quad\le\left(\frac{3.5}{\lambda}+\frac{2.5}{|\eta|\wedge|\xi|+1}\right)d'_1(\xi,\eta).
\end{aligned}$$

Proof: Using the setup of the proof of Proposition 5.18, we have from (5.6) and Proposition 5.21 that

$$|[f(\xi'+\delta_\alpha)-f(\xi')]-[f(\eta+\delta_\alpha)-f(\eta)]| \le \left(\frac{3.5}{\lambda}+\frac{2.5}{n+1}\right)(m-n).$$

For the remaining part, let $\eta'_j = \sum_{i=1}^{j} \delta_{x_i} + \sum_{i=j+2}^{n} \delta_{y_i}$, and let τ_1 and τ_2 be independent negative exponential random variables with mean 1 and independent of $\mathbf{Z}_{\eta'_j}$; then

$$
\begin{aligned}
&|[f(\eta_j + \delta_\alpha) - f(\eta_j)] - [f(\eta_{j+1} + \delta_\alpha) - f(\eta_{j+1})]| \\
&\le \int_0^\infty |\mathbb{E}[h(\mathbf{Z}_{\eta'_j}(t) + \delta_\alpha 1_{\tau_1 > t} + \delta_{y_{j+1}} 1_{\tau_2 > t}) - h(\mathbf{Z}_{\eta'_j}(t) + \delta_{y_{j+1}} 1_{\tau_2 > t}) \\
&\quad - h(\mathbf{Z}_{\eta'_j}(t) + \delta_\alpha 1_{\tau_1 > t} + \delta_{x_{j+1}} 1_{\tau_2 > t}) + h(\mathbf{Z}_{\eta'_j}(t) + \delta_{x_{j+1}} 1_{\tau_2 > t})]| dt \\
&= \int_0^\infty e^{-2t} |\mathbb{E}[h(\mathbf{Z}_{\eta'_j}(t) + \delta_\alpha + \delta_{y_{j+1}}) - h(\mathbf{Z}_{\eta'_j}(t) + \delta_{y_{j+1}}) \\
&\quad - h(\mathbf{Z}_{\eta'_j}(t) + \delta_\alpha + \delta_{x_{j+1}}) + h(\mathbf{Z}_{\eta'_j}(t) + \delta_{x_{j+1}}]| dt \\
&\le d_0(x_{j+1}, y_{j+1}) \int_0^\infty e^{-2t} \mathbb{E}\left(\frac{1}{Z_{n-1}(t) + 2} + \frac{1}{Z_{n-1}(t) + 1} \right) dt \\
&= d_0(x_{j+1}, y_{j+1}) \int_0^\infty e^{-2t} \left[\mathbb{E} \int_0^1 (1+z) z^{Z_{n-1}(t)} dz \right] dt \\
&= d_0(x_{j+1}, y_{j+1}) \int_0^\infty e^{-2t} \left[\int_0^1 (1+z) \left(1 - e^{-t} + e^{-t} z\right)^{n-1} e^{-\lambda_t (1-z)} dz \right] dt \\
&< 2 d_0(x_{j+1}, y_{j+1}) \int_0^\infty e^{-2t} \left[\int_0^1 \left(1 - e^{-t} + e^{-t} z\right)^{n-1} dz \right] dt \\
&= \frac{2 d_0(x_{j+1}, y_{j+1})}{n+1},
\end{aligned}
$$

and so

$$
|[f(\xi + \delta_\alpha) - f(\xi)] - [f(\xi' + \delta_\alpha) - f(\xi')]| \le \frac{2}{n+1} \sum_{i=1}^{n} d_0(x_i, y_i);
$$

the proof is completed by applying the triangle inequality. ∎

The estimates in Propositions 5.16, 5.21 and Lemma 5.26 allow us to modify Theorem 5.3, to give another theorem for measuring the errors of Poisson process approximation in terms of d_2 distance. This result is the same as Theorem 3.4 in Chen & Xia (2004), except for smaller constants.

Theorem 5.27: *We have*

$$
\begin{aligned}
&d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \\
&\quad \le \mathbb{E} \int_{\alpha \in \Gamma} \left(\frac{3.5}{\lambda} + \frac{2.5}{\Xi(A_\alpha^c) + 1} \right) (\Xi(A_\alpha) - 1) \Xi(d\alpha) + \min\{\epsilon_1, \epsilon_2\} \\
&\qquad + \mathbb{E} \int_{\alpha \in \Gamma} \left(\frac{3.5}{\lambda} + \frac{2.5}{\Xi(A_\alpha^c) + 1} \right) \boldsymbol{\lambda}(d\alpha) \Xi(A_\alpha), \qquad (5.51)
\end{aligned}
$$

where

$$\epsilon_1 = c_1(\lambda)\mathbb{E}\int_{\alpha\in\Gamma} |g(\alpha, \Xi|_{A_\alpha^c}) - \phi(\alpha)|\boldsymbol{\mu}(d\alpha)$$

which is valid if Ξ *is a simple point process, and*

$$\epsilon_2 = \mathbb{E}\int_{\alpha\in\Gamma}\left(\frac{3.5}{\lambda} + \frac{2.5}{\{\Xi(A_\alpha^c)\wedge\Xi_\alpha(A_\alpha^c)\}+1}\right) d_1'(\Xi|_{A_\alpha^c}, \Xi_\alpha|_{A_\alpha^c})\boldsymbol{\lambda}(d\alpha).$$

6. Applications

6.1. *Bernoulli process*

The simplest example of Poisson random variable approximation is the sum of independent indicator random variables [Barbour, Holst & Janson (1992), Chapter 1]. Correspondingly, for Poisson process approximation, a prototypical example is the Bernoulli process defined in Example 4.5. In terms of the d_2 metric, the example was first discussed in Barbour & Brown (1992a) with a logarithmic factor, and, in Xia (1997a), a bound without the logarithmic factor was obtained for the case with all p_i's are equal; the general case was considered in Xia (1997b). Using Theorems 5.8, 5.14 and 5.27, we can obtain the following estimates.

Theorem 6.1: *For the Bernoulli process* Ξ *on* $\Gamma = [0,1]$ *with mean measure* $\boldsymbol{\lambda}$,

$$d_{TV}(\mathcal{L}(|\Xi|), \mathrm{Po}(\lambda)) \le \frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^n p_i^2, \tag{6.1}$$

$$d_{TV}(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \sum_{i=1}^n p_i^2, \tag{6.2}$$

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le \frac{6}{\lambda - \max_{1\le i\le n} p_i}\sum_{i=1}^n p_i^2. \tag{6.3}$$

Remark 6.2: It is worth pointing out that (6.3) slightly improves the bound in Brown, Weinberg & Xia (2000). However, when λ is large, the estimate in Xia (1997b) is asymptotically better than (6.3) and if all p_i's are equal, the bound in Xia (1997a) is the best, as there is some symmetric structure to be exploited in this case.

Proof: Taking $A_\alpha = \{\alpha\}$ and ϵ_2 in Theorems 5.8, 5.14 and 5.27, we find that the first terms of (5.14), (5.16) and (5.51) are equal to 0. We now set

$\Xi_\alpha = \sum_{1\le i\le n, i/n\neq\alpha} I_i\delta_{i/n} + \delta_\alpha$, so that the second terms of (5.14), (5.16) and (5.51) are also 0. The last term of (5.14) then gives (6.1), the last term of (5.16) yields (6.2), while (6.3) is ensured by the last term of (5.51):

$$\sum_{i=1}^n \left(\frac{3.5}{\lambda} + \mathbb{E}\frac{2.5}{\sum_{1\le j\le n, j\neq i} I_j + 1}\right) p_i^2 \le \frac{6}{\lambda - \max_{1\le i\le n} p_i} \sum_{i=1}^n p_i^2$$

since

$$\begin{aligned}\mathbb{E}\left\{\frac{1}{\sum_{1\le j\le n, j\neq i} I_j + 1}\right\} &= \mathbb{E}\int_0^1 z^{\sum_{1\le j\le n, j\neq i} I_j} dz \\ &= \int_0^1 \prod_{1\le j\le n, j\neq i} [zp_j + (1-p_j)]dz \le \int_0^1 \prod_{1\le j\le n, j\neq i} e^{-p_j(1-z)}dz \\ &= \int_0^1 e^{-(\lambda-p_j)(1-z)}dz \le \frac{1}{\lambda - p_j}.\end{aligned}$$

■

6.2. *2-runs*

The independent increment structure in the Bernoulli process makes life easier when we apply Theorem 5.27. Now, we consider a process with dependence.

Suppose $I_1, I_2, \cdots, I_n$ are independent and identically distributed indicators with

$$\mathbb{P}(I_i = 1) = 1 - \mathbb{P}(I_i = 0) = p.$$

To avoid edge effects, we write $I_{n+1} = I_1$. Let $J_i = I_iI_{i+1}$ for $1 \le i \le n$ and define

$$\Xi = \sum_{i=1}^n J_i\delta_{i/n}.$$

Then Ξ is a point process on Γ with mean measure $\boldsymbol{\lambda} = \sum_{i=1}^n p^2\delta_{i/n}$.

Theorem 6.3: *With the above assumptions, we have*

$$d_{TV}(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le np^3(2-p), \tag{6.4}$$

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \le p(12-p). \tag{6.5}$$

Remark 6.4: The bounds are of the right order. In fact, the bound (6.5) is of the same order as that for the one dimensional Poisson approximation

to the number of 2-runs [see Barbour, Holst & Janson (1992), p. 163]. For (6.4), we take

$$B = \{\xi \in \mathcal{H}: \ \exists i \in \{1, \cdots, n-1\} \text{ such that } \xi\{i/n\} = \xi\{(i+1)/n\} = 1\},$$

then $\mathbb{P}(\Xi \in B) = O(np^3)$ while $\mathrm{Po}\,(\boldsymbol{\lambda})(B) = O(np^4)$.

Remark 6.5: It might be possible to reduce the constant in (6.5) by as much as 10 if one exploits the symmetric structure in Ξ, as done in Xia (1997a).

Proof: In view of Theorems 5.14 and 5.27, we let $A_{j/n} = \{j/n\}$ and realize the Palm processes as

$$\Xi_{j/n} = \sum_{i \neq j-1,j,j+1} J_i \delta_{i/n} + \delta_{j/n} + I_{j-1}\delta_{(j-1)/n} + I_{j+2}\delta_{(j+1)/n}.$$

Then the first terms of (5.16) and (5.51) vanish. Noting that

$$\|\Xi|_{A^c_{j/n}} - \Xi_{j/n}|_{A^c_{j/n}}\| = I_{j-1}(1-I_j) + I_{j+2}(1-I_{j+1}),$$

we see that the second term of (5.16) with ϵ_2 becomes $2np^3(1-p)$ while the last term of (5.16) equals np^4, hence (6.4).

Apropos of (6.5), since

$$d_1'(\Xi|_{A^c_{j/n}}, \Xi_{j/n}|_{A^c_{j/n}}) \leq \|\Xi|_{A^c_{j/n}} - \Xi_{j/n}|_{A^c_{j/n}}\| = I_{j-1}(1-I_j) + I_{j+2}(1-I_{j+1}),$$

and $\Xi(A^c_{j/n}) \leq \Xi_{j/n}(A^c_{j/n})$, in the second term of (5.51) with ϵ_2, the integrand is bounded by

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{3.5}{\lambda} + \frac{2.5}{\Xi(A^c_{j/n})+1}\right)[I_{j-1}(1-I_j) + I_{j+2}(1-I_{j+1})]\right\} \\
&= 2\mathbb{E}\left(\frac{3.5}{\lambda} + \frac{2.5}{\Xi(A^c_{j/n})+1}\right) I_{j-1}(1-I_j) \\
&= \frac{7p(1-p)}{\lambda} + 5p(1-p)\mathbb{E}\left\{\frac{1}{\sum_{i\neq j-2,j-1,j} J_i + I_{j-2} + 1}\right\}. \qquad (6.6)
\end{aligned}$$

Now, instead of using Lemma 5.22, we explain another technique for ex-

tracting Stein's factors:

$$\begin{aligned} 1 &\geq \mathbb{E}\sum_{k=1}^{n}\left(\frac{J_k}{\sum_{i=1}^{n} J_i}1_{|\sum_{i=1}^{n} J_i|\geq 1}\right) \\ &= p^2\sum_{k=1}^{n}\mathbb{E}\left\{\frac{1}{\sum_{i\neq k-1,k,k+1} J_i + 1 + I_{k-1} + I_{k+2}}\right\} \qquad (6.7) \\ &= np^2\mathbb{E}\left\{\frac{1}{\sum_{i\neq -1,0,1} J_i + 1 + I_{-1} + I_2}\right\} \\ &\geq np^2\mathbb{E}\left[\frac{1}{\sum_{i\neq -1,0,1} J_i + 1 + I_{-1} + I_2}\middle| I_2 = 0\right]\mathbb{P}(I_2 = 0) \\ &\geq np^2\mathbb{E}\left\{\frac{1}{\sum_{i\neq -1,0,1} J_i + 1 + I_{-1}}\right\}\mathbb{P}(I_2 = 0), \end{aligned}$$

which implies that

$$\mathbb{E}\left\{\frac{1}{\sum_{i\neq -1,0,1} J_i + 1 + I_{-1}}\right\} \leq \frac{1}{np^2(1-p)}. \qquad (6.8)$$

Therefore, it follows from (6.6) and (6.8) that the second term of (5.51) is bounded by

$$7p(1-p) + 5p = p(12 - 7p).$$

The last term of (5.51), using (6.7), is controlled by

$$\begin{aligned} &\mathbb{E}\int_{\Gamma}\left(\frac{3.5}{\lambda} + \frac{2.5}{\Xi(A_\alpha^c) + 1}\right)\boldsymbol{\lambda}(d\alpha)\Xi(A_\alpha) \\ &= \sum_{k=1}^{n} p^4\left(\frac{3.5}{\lambda} + \frac{2.5}{\sum_{i\neq k-1,k,k+1} J_i + I_{k-1} + I_{k+2} + 1}\right) \leq 6p^2, \end{aligned}$$

completing the proof. ∎

6.3. *Matérn hard core process*

Hard core processes were first introduced in statistical mechanics to model the distribution of particles with repulsive interactions [see Ruelle (1969), p. 6]. Hard core processes can be obtained by deleting certain points in a Poisson point process; hence, when the number of points deleted is relatively small, the hard core process should be close to a Poisson process. Barbour & Brown (1992a) investigated how well a particular type of hard

core process is approximated by a Poisson process with the same mean measure, a corrected estimate being given in Brown & Greig (1994).

Here, we consider the Matérn hard core process that was studied in Chen & Xia (2004). This Matérn hard core process is a particular example of a distance model [see Matérn (1986), p. 37] and is also a model for underdispersion [see Daley & Vere-Jones (1988), p. 366].

To start with, let Z be a homogeneous Poisson process on Γ, where Γ is a compact subset of R^d with volume $V(\Gamma) \neq 0$. Such a process can be represented as

$$Z = \sum_{i=1}^{N} \delta_{X_i},$$

where $X_1, X_2, \ldots$ are independent uniform random variables on Γ, and $N \sim \mathrm{Po}(\mu)$ is independent of $\{X_i; i \geq 1\}$. Let $B(x,r)$ be the r-neighborhood of x, $B(x,r) = \{y \in \Gamma : 0 < d_0(y,x) < r\}$, where $d_0(x,y) = |x-y| \wedge 1$. The Matérn hard core process Ξ is produced by deleting any point within distance r of another point, irrespective of whether the latter point has itself already been deleted [see Cox & Isham (1980), page 170]:

$$\Xi = \sum_{i=1}^{N} \delta_{X_i} 1_{\{Z(B(X_i,r))=0\}}.$$

In other words, if $\{\alpha'_n\}$ is a realization of points of the Poisson process, then the points deleted are

$$\{\alpha''_n\} = \{x \in \{\alpha'_n\} : |x-y| < r \text{ for some } y \neq x, y \in \{\alpha'_n\}\}$$

and $\{\alpha_n\} := \{\alpha'_n\} \backslash \{\alpha''_n\}$ constitutes a realization of the Matérn hard core process Ξ [see Daley & Vere-Jones (1988)]. Also,

$$\Xi(d\alpha) = \sum_{i=1}^{N} \delta_{X_i}(d\alpha) 1_{\{Z(B(X_i,r))=0\}} = 1_{\{Z(B(\alpha,r))=0\}} Z(d\alpha).$$

Let κ_d be the volume of the unit ball in R^d.

Theorem 6.6: *The mean measure of Ξ is given by*

$$\boldsymbol{\lambda}(d\alpha) = e^{-\mu V(\alpha,r)/V(\Gamma)} \mu V(\Gamma)^{-1} d\alpha,$$

and

$$d_{TV}(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \leq 2\vartheta\lambda, \tag{6.9}$$

$$d_2(\mathcal{L}(\Xi), \mathrm{Po}(\boldsymbol{\lambda})) \leq 7\vartheta + 5\vartheta[3 + (1 - e^{-2^{-d}\vartheta})\vartheta]/(1 + (1-2\vartheta)/\lambda), \tag{6.10}$$

where $V(\alpha,r)$ is the volume of $B(\alpha,r)$ and $\vartheta = \mu\kappa_d(2r)^d/V(\Gamma)$.

Proof: [See Chen & Xia (2004)] We prove (6.10) first. The Poisson property of Z implies that the counts of points in disjoint sets are independent. So

$$\boldsymbol{\lambda}(d\alpha) = \mathbb{E}(\Xi(d\alpha)) = \mathbb{E}1_{\{Z(B(\alpha,r))=0\}}\mathbb{E}Z(d\alpha) = e^{-\mu V(\alpha,r)/V(\Gamma)}\mu V(\Gamma)^{-1}d\alpha.$$

Also, whether a point outside $B(\alpha,2r)\cup\{\alpha\}$ is deleted or not is independent of the behavior of Z in $B(\alpha,r)\cup\{\alpha\}$. Hence we choose $A_\alpha = B(\alpha,2r)\cup\{\alpha\}$ so that, for any α and β, we have

$$\mathbb{E}\left\{\frac{1}{\Xi(\Gamma_{\alpha\beta})+1}\right\}\Xi(d\alpha)\Xi(d\beta) = \mathbb{E}\left\{\frac{1}{\Xi(\Gamma_{\alpha\beta})+1}\right\}\mathbb{E}\Xi(d\alpha)\Xi(d\beta),$$

where $\Gamma_{\alpha\beta} = \Gamma\setminus(A_\alpha\cup A_\beta)$. Applying Theorem 5.27 gives

$$\begin{aligned} &d_2(\mathcal{L}(\Xi),\mathrm{Po}(\boldsymbol{\lambda})) \\ &\le \int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha\setminus\{\alpha\}}\left(\frac{3.5}{\lambda}+\mathbb{E}\left\{\frac{2.5}{\Xi(\Gamma_{\alpha\beta})+1}\right\}\right)\mathbb{E}\Xi(d\alpha)\Xi(d\beta) \\ &\quad+\int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha}\left(\frac{3.5}{\lambda}+\mathbb{E}\left\{\frac{2.5}{\Xi(\Gamma_{\alpha\beta})+1}\right\}\right)\boldsymbol{\lambda}(d\alpha)\boldsymbol{\lambda}(d\beta). \qquad (6.11)\end{aligned}$$

Now,

$$\mathbb{E}\Xi(\Gamma_{\alpha\beta}) = \int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx,$$

where $\mu_\Gamma = \mu/V(\Gamma)$. On the other hand,

$$\mathbb{E}\Xi(d\alpha)\Xi(d\beta) = \begin{cases} e^{-\mu_\Gamma(V(\alpha,r)+V(\beta,r))}\mu_\Gamma^2 d\alpha d\beta, & \text{if } |\alpha-\beta|\ge 2r, \\ e^{-\mu_\Gamma(V(\alpha,r)+V(\beta,r)-V(\alpha,\beta,r))}\mu_\Gamma^2 d\alpha d\beta, & \text{if } r\le|\alpha-\beta|<2r, \\ 0, & \text{if } 0<|\alpha-\beta|<r, \\ e^{-\mu_\Gamma V(\alpha,r)}\mu_\Gamma d\alpha, & \text{if } \alpha=\beta; \end{cases}$$

where $V(\alpha,\beta,r)$ is the volume of $B(\alpha,r)\cap B(\beta,r)$. Hence,

$$\begin{aligned} \mathbb{E}[\Xi(\Gamma_{\alpha\beta})^2] &= \mathbb{E}\int\int_{x,y\in\Gamma_{\alpha\beta}}\Xi(dx)\Xi(dy) \\ &= \int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx \\ &\quad+\int\int_{x,y\in\Gamma_{\alpha\beta},|x-y|\ge 2r} e^{-\mu_\Gamma(V(x,r)+V(y,r))}\mu_\Gamma^2 dxdy \\ &\quad+\int\int_{x,y\in\Gamma_{\alpha\beta},r\le|x-y|<2r} e^{-\mu_\Gamma(V(x,r)+V(y,r)-V(x,y,r))}\mu_\Gamma^2 dxdy. \end{aligned}$$

Writing

$$[\mathbb{E}\Xi(\Gamma_{\alpha\beta})]^2 = \int\int_{x,y\in\Gamma_{\alpha\beta}} e^{-\mu_\Gamma(V(x,r)+V(y,r))}\mu_\Gamma^2 dxdy,$$

we have

$$\begin{aligned}
&\mathrm{Var}\,(\Xi(\Gamma_{\alpha\beta}))\\
&= \int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx\\
&\quad + \int\int_{x,y\in\Gamma_{\alpha\beta}, r\le|x-y|<2r} e^{-\mu_\Gamma(V(x,r)+V(y,r)-V(x,y,r))}\mu_\Gamma^2 dxdy\\
&\quad - \int\int_{x,y\in\Gamma_{\alpha\beta},|x-y|<2r} e^{-\mu_\Gamma(V(x,r)+V(y,r))}\mu_\Gamma^2 dxdy\\
&\le \int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx\\
&\quad + \int\int_{x,y\in\Gamma_{\alpha\beta},|x-y|<2r} e^{-\mu_\Gamma V(x,r)}[1-e^{-\mu_\Gamma V(y,r)}]\mu_\Gamma^2 dxdy\\
&\le \{1+(1-e^{-\mu_\Gamma\kappa_d r^d})\mu_\Gamma\kappa_d(2r)^d\}\int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx.
\end{aligned}$$

Thus,

$$\kappa = \frac{\mathrm{Var}\,(\Xi(\Gamma_{\alpha\beta})+1)}{\mathbb{E}(\Xi(\Gamma_{\alpha\beta})+1)} \le \frac{\mathrm{Var}\,(\Xi(\Gamma_{\alpha\beta}))}{\mathbb{E}(\Xi(\Gamma_{\alpha\beta}))} \le 1+(1-e^{-\mu_\Gamma\kappa_d r^d})\mu_\Gamma\kappa_d(2r)^d,$$

which, together with Lemma 5.22, yields

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{\Xi(\Gamma_{\alpha\beta})+1}\right\} &\le \frac{2+\kappa}{\int_{\Gamma_{\alpha\beta}} e^{-\mu_\Gamma V(x,r)}\mu_\Gamma dx+1}\\
&\le \frac{3+(1-e^{-\mu_\Gamma\kappa_d r^d})\mu_\Gamma\kappa_d(2r)^d}{\lambda+1-2\mu_\Gamma\kappa_d(2r)^d}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha\setminus\{\alpha\}} \mathbb{E}\Xi(d\alpha)\Xi(d\beta) &\le \int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha} e^{-\mu_\Gamma V(\alpha,r)}\mu_\Gamma^2 d\alpha d\beta\\
&\le \mu_\Gamma\kappa_d(2r)^d\lambda, \qquad (6.12)
\end{aligned}$$

and

$$\int_{\alpha\in\Gamma}\int_{\beta\in A_\alpha} \boldsymbol{\lambda}(d\alpha)\boldsymbol{\lambda}(d\beta) \le \mu_\Gamma\kappa_d(2r)^d\lambda. \qquad (6.13)$$

Applying these inequalities to the relevant terms in (6.11) gives (6.10).

To prove (6.9), we use (5.16) with ϵ_2, then the second term is 0 and the remaining two terms are estimated by (6.12) and (6.13). ∎

6.4. *Networks of queues*

Melamed (1979) proved that for an open migration process, a necessary and sufficient condition for the equilibrium flow along a link to be Poisson is the absence of loops: no customer can travel along the link more than once. Barbour & Brown (1996) quantified the statement by allowing the customers a small probability of travelling along the link more than once and proved Poisson process approximation theorems analogous to Melamed's theorem. The d_2 bound was later improved by Brown, Weinberg & Xia (2000) and the following context is taken from Brown, Fackrell & Xia (2005).

We generally use the same notation as in Barbour & Brown (1996), with adjustment to fit our presentation, and the argument is adapted from Brown, Weinberg & Xia (2000). We consider an open migration process consisting of a system of J queues, in which individuals are allowed to move from one queue to another, and to enter or exit the system from any queue, with the following specifications. Arrivals at each of the queues are assumed to be independent Poisson streams with rates ν_j, for $1 \leq j \leq J$. Service requirements are assumed to be independent negative exponential random variables with mean 1. The total service effort at queue j is specified by a service function $\phi_j(m)$, where m is the number of customers in the queue. We assume that $\phi_j(0) = 0$, that $\phi_j(1) > 0$ and that the service functions are non-decreasing. We also assume that the individuals in the same queue all receive the same service effort. Let λ_{jk} be the probability that an individual moves to queue k on leaving queue j, and let μ_j be the exit probability from queue j; then the sum of all transition probabilities from any state in the system equals one: $\mu_j + \sum_{k=1}^{J} \lambda_{jk} = 1$ for all $1 \leq j \leq J$. The process $\{N(t) : t \in \mathbb{R}\}$ of the numbers of customers in line at various queues at time t is a pure jump Markov process on $\{0, 1, \cdots\}^J$ with transition rate ν_j from state $n = (n_1, \cdots, n_J)$ to $n + e_j$, and, if $n_j \geq 1$, the process has transition rate $\lambda_{jk}\phi_j(n_j)$ from n to $n - e_j + e_k$ and rate $\mu_j\phi_j(n_j)$ from n to $n - e_j$, where e_j is the jth coordinate vector in $\{0, 1, \cdots\}^J$.

For each $1 \leq j, k \leq J$ we define a point process Ξ^{jk} counting the number of transitions from queue j to queue k, and set $\Xi = \{\Xi^{jk}, 0 \leq j, k \leq J\}$, where departures are interpreted as transitions to 0, and arrivals as transitions from 0. Let $\mathcal{S} = \{(j,k) | 0 \leq j, k \leq J\}$ be the set of all possible direct links, and let C be a subset of $\mathcal{S}$. Fix any $t > 0$, and take as the carrier space $\Gamma = [0, t] \times C$. Thus an element of Γ is of the form $(s, (j, k))$, representing a transition from queue j to queue k at time s, with $(s, (0, k))$ representing an arrival to queue k at time s, and $(s, (j, 0))$ representing an

exit from queue j at time s. Let d_0 be the metric on Γ defined by

$$d_0((s_1,(j,k)),(s_2,(l,m))) = I_{[(j,k)\neq(l,m)]} + (|s_1 - s_2| \wedge 1) I_{[(j,k)=(l,m)]},$$

for $s_1, s_2 \in [0,t]$ and links $(j,k),(l,m) \in C$. Let Ξ_C be the restricted process $\{\Xi^{jk} | (j,k) \in C\}$ and define $\Xi^{C,t}$ to be the restriction of Ξ_C to $[0,t]$. The mean measure of the process $\Xi^{C,t}$ is $\boldsymbol{\lambda}^{C,t}$, where $\boldsymbol{\lambda}^{C,t}$ denotes the restriction to $[0,t]$ of $\boldsymbol{\lambda}^C$, defined by

$$\boldsymbol{\lambda}^C(ds,(j,k)) := \begin{cases} \rho_{jk} ds, & s \geq 0,\ (j,k) \in C, \\ 0, & \text{otherwise}, \end{cases} \tag{6.14}$$

where ρ_{jk} is the steady state flow along the link (j,k). A simple calculation gives

$$|\boldsymbol{\lambda}^{C,t}| = t \sum_{(j,k)\in C} \rho_{jk} =: t\mathcal{R}^C. \tag{6.15}$$

The path that a customer takes through the network is tracked by a forward customer chain X, which is a Markov chain with state space $\{0,1,\ldots,J\}$, where state 0 represents the point of arrival and departure of an individual into and from the system, and the other states represent the queues. This chain has nonzero transition probabilities p_{jk} given by

$$p_{0k} = \frac{\nu_k}{\sum_{j=1}^J \nu_j},\ p_{j0} = \mu_j,\ \text{and}\ p_{jk} = \lambda_{jk},$$

where $1 \leq j,k \leq J$. We assume that the parameters λ_{jk}, μ_j and ν_j allow an individual to have positive probability of reaching any queue from outside the system and of leaving the system from any queue. Since the state space is finite, the forward chain X is irreducible and every state is persistent.

We now summarise a few facts about the forward customer chain. Interested readers should refer to Barbour & Brown (1996) for more details. First, the expected number of visits to queue j between returns to 0 in the forward chain is given by $\alpha_j / \sum_k \nu_k$, with $\{\alpha_j,\ 1 \leq j \leq J\}$ the unique solution to the equations

$$\alpha_j = \nu_j + \sum_{i=1}^J \alpha_i \lambda_{ij}.$$

In terms of the quantity α_j, we have $\rho_{jk} = \lambda_{jk}\alpha_j$.

Provided that

$$\sum_{n=0}^{\infty} \frac{\alpha_j^n}{\prod_{r=1}^n \phi_j(r)} < \infty \text{ for all } j,$$

then N has a unique stationary distribution, under which, for each $t > 0$, $N_j(t)$, $j = 1, \cdots, J$ are independent with

$$\mathbb{P}(N_j(t) = k) = \frac{\alpha_j^k / \prod_{r=1}^k \phi_j(r)}{\sum_{n=0}^{\infty} \left\{ \frac{\alpha_j^n}{\prod_{r=1}^n \phi_j(r)} \right\}}, \qquad k \in \mathbf{Z}_+.$$

Also of use is the backward customer chain X^*, which is the forward customer chain for the time-reversal of the above queueing network. The key point is that these chains allow us to define the following random variables, which essentially control the approximations.

Let $X^{(k)}$ be a realization of X started in state k, and let $X^{*(j)}$ be a realization of X^* started in state j. We define

$$\alpha_C^k := \sum_{i=0}^{\infty} I[(X_i^{(k)}, X_{i+1}^{(k)}) \in C] \quad \text{and} \quad \beta_C^j := \sum_{i=0}^{\infty} I[(X_{i+1}^{*(j)}, X_i^{*(j)}) \in C].$$

Observe that α_C^k is the number of transitions along links in C yet to be made by a customer currently in state k, while β_C^j is the number of transitions along links in C already made by a customer currently in state j. Let $\theta_C^{jk} := \mathbb{E}(\alpha_C^k + \beta_C^j)$, which can be interpreted as the expected number of past and future transitions along links in C, for an individual currently moving from queue j to queue k. Then define

$$\overline{\theta}_C := \sum_{(j,k) \in C} \theta_C^{jk} \frac{\rho_{jk}}{\sum_{(j',k') \in C} \rho_{j'k'}},$$

the average extra number of visits to links in C of a customer who is on a link in C, the average being weighted by the steady state traffic flow along the link.

Theorem 6.7: [Barbour & Brown (1996)] *With the above setup and (6.14), we have*

$$d_{TV}(\mathcal{L}(\Xi^{C,t}), \mathrm{Po}\,(\boldsymbol{\lambda}^{C,t})) \le t \sum_{(j,k) \in C} \rho_{jk} \theta_C^{jk}.$$

They also show that, in the stationary state,

$$d_2(\mathcal{L}(\Xi^{C,t}), \mathrm{Po}\,(\boldsymbol{\lambda}^{C,t})) \le 2.5 \left(1 + 2 \log^+ \left(\frac{2t\mathcal{R}^C}{5}\right)\right) \overline{\theta}_C, \qquad (6.16)$$

where $\mathcal{R}^C$ is as defined in (6.15).

This bound is small if $\overline{\theta}_C$ is small, but the logarithmic factor causes the bound to increase with time. Brown, Weinberg & Xia (2000) removed the

logarithmic factor and, by directly applying Theorem 5.27, we can obtain a sharper estimate as follows.

Theorem 6.8: [Brown, Fackrell & Xia (2005)] *With the above setup and (6.14), we have*

$$d_2(\mathcal{L}(\Xi^{C,t}), \mathrm{Po}\,(\boldsymbol{\lambda}^{C,t})) \leq (11 + 5\overline{\theta}_C)\overline{\theta}_C,$$

for $t > 1/\mathcal{R}^C$ *and* $\overline{\theta}_C < 1/11$.

The bound of Theorem 6.8 is small if $\overline{\theta}_C$ is small, and it does not increase with time t. It is smaller than the bound given in Theorem 6.7 for large values of t. It is also more natural than (6.16) because it is small when $\overline{\theta}_C$ is small, however large the value of t.

To prove the claim, we need the following lemma.

Lemma 6.9: [Barbour & Brown (1996), Lemma 1] *For the open queueing network, the reduced Palm distribution for the network given a transition from queue* j *to queue* k *at time 0* $(1 \leq j, k \leq J)$ *is the same as that for the original network, save that the network on* $(0, \infty)$ *behaves as if there were an extra individual at queue* k *at time 0 and the network on* $(-\infty, 0)$ *behaves as if there were an extra individual in queue* j *at time 0.*

Proof: *[Theorem 6.8].* Lemma 6.9 implies that

$$|\Xi^{C,t}| \wedge |(\Xi^{C,t})_\alpha - \delta_\alpha| = |\Xi^{C,t}|,$$

because the reduced Palm process is statistically equivalent to the process $\Xi^{C,t}$ with the addition of extra individuals. Furthermore, these individuals are independent of the original process, and so the difference $[(\Xi^{C,t})_\alpha - \delta_\alpha] - \Xi^{C,t}$ is independent of $|\Xi^{C,t}|$.

By applying the bound of Theorem 5.27 with $A_\alpha = \{\alpha\}$ and ϵ_2, the first and last terms of (5.51) vanish, so we can reduce the bound to the following integral:

$$\mathbb{E}\int_\Gamma \left(\frac{3.5}{|\boldsymbol{\lambda}^{C,t}|} + \frac{2.5}{|\Xi^{C,t}| + 1}\right) \|[(\Xi^{C,t})_\alpha - \delta_\alpha] - \Xi^{C,t}\| \boldsymbol{\lambda}^{C,t}(d\alpha).$$

By an application of the properties of the process $\Xi^{C,t}$ and its reduced Palm process,

$$\mathbb{E}\int_\Gamma \frac{3.5}{|\boldsymbol{\lambda}^{C,t}|} \|[(\Xi^{C,t})_\alpha - \delta_\alpha] - \Xi^{C,t}\| \boldsymbol{\lambda}^{C,t}(d\alpha) = 3.5\overline{\theta}_C. \tag{6.17}$$

Next, by applying (5.28),

$$\mathbb{E}\int_{\Gamma}\left[\frac{2.5}{|\Xi^{C,t}|+1}\right]\|[(\Xi^{C,t})_{\alpha}-\delta_{\alpha}]-\Xi^{C,t}\|\boldsymbol{\lambda}^{C,t}(d\alpha)$$
$$\leq 2.5\left(1+\frac{\gamma}{2}+\sqrt{\gamma\left(\frac{\gamma}{4}+1\right)}\right)\overline{\theta}_C\leq(5+2.5\gamma)\overline{\theta}_C, \qquad (6.18)$$

where $\gamma=1+2\overline{\theta}_C$ [see Brown, Weinberg & Xia (2000), p. 163].

The proof is completed by adding the estimates of (6.17) and (6.18). ■

7. Further developments

Research on Poisson process approximation is far from over. First of all, in some applications, the bounds obtained are still too large to justify the use of the Poisson process approximation [see Leung et al. (2004)]. It may well be that the factor in Proposition 5.21 could be reduced to $\frac{1}{\lambda}+\frac{1}{n+1}$, which would be of some help. The class of d_1-Lipschitz functions of configurations also needs to be more fully investigated, and related to the kinds of information required in applications.

A particular case of Poisson process approximation is multivariate Poisson approximation, as studied in Barbour (1988). Here, it is usually more appropriate to use the total variation metric instead of the d_2 metric. Roos (1999) proved the optimal bound for the total variation distance between the distribution of the sum of independent non-identically distributed Bernoulli random vectors and a multivariate Poisson distribution. The main idea is an adaptation of a method originally used by Kerstan (1964) in the univariate case, which is based on an appropriate expansion of the difference of the probability generating functions. This means that it may prove difficult to generalize the method beyond the sum of independent random vectors. On the other hand, Stein's method has already been shown to be very versatile in handling dependence, and it is a tantalizing problem to use Stein's method to prove the optimal bounds in this setting.

The extension to compound Poisson process approximation was initiated in Barbour & Månsson (2002). However, even for compound Poisson random variable approximation, there are severe technical difficulties involved in the direct approach using Stein's method, and the optimal general results have not yet been found. New ideas are needed to make further progress even in compound Poisson random variable approximation, let alone process approximation. Thus, the paper of Barbour & Månsson (2002) is just a beginning in this area.

Acknowledgement: This work was partly supported by ARC Discovery project number DP0209179.

References

1. D. J. ALDOUS (1989) *Probability approximations via the Poisson clumping heuristic.* Springer, New York.
2. P. K. ANDERSEN, Ø. BORGAN, R. D. GILL & N. KEIDING (1993) *Statistical Models Based on Counting Processes.* Springer-Verlag , New York.
3. BARBOUR, A. D. (1988) Stein's method and Poisson process convergence. *J. Appl. Probab.* **25** (A), 175-184.
4. A. D. BARBOUR & T. C. BROWN (1992a) Stein's method and point process approximation. *Stoch. Procs. Applics* **43**, 9–31.
5. A. D. BARBOUR & T. C. BROWN (1992b) Stein-Chen method, point processes and compensators. *Ann. Probab.* **20**, 1504–1527.
6. A. D. BARBOUR & T. C. BROWN (1996) Approximate versions of Melamed's theorem. *J. Appl. Probab.* **33**, 472–489.
7. A. D. BARBOUR, T.C. BROWN & A. XIA (1998) Point processes in time and Stein's method. *Stochastics and Stochastics Reports* **65**, 127–151.
8. A. D. BARBOUR & G. K. EAGLESON (1983) Poisson approximation for some statistics based on exchangeable trials. *Adv. Appl. Prob.* **15**, 585-600.
9. A. D. BARBOUR, L. HOLST & S. JANSON (1992) *Poisson Approximation.* Oxford Univ. Press.
10. A. D. BARBOUR & M. MÅNSSON (2002) Compound Poisson process approximation. *Ann. Probab.* **30**, 1492–1537.
11. P. BILLINGSLEY (1968) *Convergence of probability measures.* John Wiley & Sons.
12. T. C. BROWN (1983) Some Poisson approximations using compensators. *Ann. Probab.* **11**, 726–744.
13. T. C. BROWN, M. FACKRELL & A. XIA (2005) Poisson process approximation in Jackson networks. *COSMOS* **1** (to appear).
14. T. C. BROWN & D. GREIG (1994) Correction to: "Stein's method and point process approximation" [*Stoch. Procs Applics* **43**, 9–31 (1992)] by A. D. Barbour & T. C. Brown, *Stoch. Procs. Applics* **54**, 291–296.
15. T. C. BROWN & M. J. PHILLIPS (1999) Negative Binomial Approximation with Stein's Method. *Methodology and Computing in Applied Probability* **1**, 407-421.
16. T. C. BROWN, G. V. WEINBERG & A. XIA (2000) Removing logarithms from Poisson process error bounds. *Stoch. Procs. Applics* **87**, 149–165.
17. T. C. BROWN & A. XIA (1995a) On metrics in point process approximation. *Stochastics and Stoch. Reports* **52**, 247–263.
18. T. C. BROWN & A. XIA (1995b) On Stein-Chen factors for Poisson approximation. *Statis. Prob. Lett.* **23**, 327–332.
19. T. C. BROWN & A. XIA (2001) Stein's method and birth-death processes. *Ann. Probab.* **29**, 1373–1403.

20. T. C. BROWN & A. XIA (2002) How many processes have Poisson counts? *Stoch. Procs. Applics* **98**, 331–339.
21. L. H. Y. CHEN (1975) Poisson approximation for dependent trials. *Ann. Probab.* **3**, 534–545.
22. L. H. Y. CHEN & A. XIA (2004) Stein's method, Palm theory and Poisson process approximation. *Ann. Probab.* **32**, 2545–2569.
23. D. R. COX & V. ISHAM (1980) *Point Processes.* Chapman & Hall.
24. D. J. DALEY & D. VERE-JONES (1988) *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.
25. C. DELLACHERIE & P. A. MEYER (1982) *Probabilities and Potential B.* North-Holland.
26. W. EHM (1991) Binomial approximation to the Poisson binomial distribution. *Statistics and Probability Letters* **11**, 7–16.
27. S. N. ETHIER & T. G. KURTZ (1986) *Markov processes: characterization and convergence.* John Wiley & Sons.
28. G. S. FISHMAN (1996) *Monte Carlo, concepts, algorithms and applications.* Springer, New York.
29. P. FRANKEN, D. KÖNIG, U. ARNDT & V. SCHMIDT (1982) *Queues and Point Processes.* John Wiley & Sons.
30. D. FREEDMAN (1974) The Poisson approximation for dependent events. *Ann. Probab.* **2**, 256–269.
31. J. JACOD & A. N. SHIRYAEV (1987) *Limit Theorems for Stochastic Processes.* Springer-Verlag, Berlin.
32. L. JANOSSY (1950) On the absorption of a nucleon cascade. *Proc. R. Irish Acad. Sci. Sec. A* **53**, 181–188.
33. YU. M. KABANOV, R. S. LIPTSER & A. N. SHIRYAEV (1983) Weak and strong convergence of distributions of counting processes. *Theor. Probab. Appl.* **28**, 303–335.
34. O. KALLENBERG (1976) *Random Measures. Academic Press* .
35. A. F. KARR (1986) *Point Processes and Their Statistical inference.* Marcel Dekker.
36. J. KEILSON (1979) *Markov Chain Models - Rarity and Exponentiality.* Springer-Verlag.
37. J. KERSTAN (1964) Verallgemeinerung eines Satzes von Prochorow und Le Cam. *Z. Wahrsch. verw. Geb.* **2**, 173–179.
38. J. F. C. KINGMAN (1993) *Poisson processes.* Oxford Univ. Press.
39. M. Y. LEUNG, K. P. CHOI, A. XIA & L. H. Y. CHEN (2004) Nonrandom Clusters of Palindromes in Herpesvirus Genomes. *Journal of Computational Biology* (to appear).
40. M. N. M. VAN LIESHOUT (2000) *Markov point processes and their applications.* Imperial College Press.
41. R. S. LIPTSER & A. N. SHIRYAEV (1978) *Statistics of Random Processes* (II). Springer-Verlag .
42. B. MATÉRN (1986) *Spatial variation.* 2nd edn, Lecture Notes in Statistics **36**, Springer-Verlag.

43. B. MELAMED (1979) Characterizations of Poisson traffic streams in Jackson queueing networks. *Adv. Appl. Prob.* **11**, 422–438.
44. C. PALM (1943) Intensitätsschwankungen im Fernsprechverkehr. *Ericsson Techniks* **44**, 1–189.
45. D. PFEIFER (1987) On the distance between mixed Poisson and Poisson distributions. *Statistics and Decisions* **5**, 367–379.
46. YU. V. PROHOROV (1956) Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1**, 157–214.
47. C. J. PRESTON (1975) Spatial birth-and-death processes. *Bull. ISI* **46**, 371–391.
48. S. T. RACHEV (1991) *Probability metrics and the Stability of Stochastic Models.* John Wiley & Sons.
49. R. D. REISS (1993) *A course on point processes.* Springer, New York.
50. A. RENYI (1967) Remarks on the Poisson process. *Stud. Sci. Math. Hungar.* **5**, 119–123.
51. B. ROOS (1999) On the rate of multivariate Poisson convergence. *J. Multivariate Anal.* **69**, 120–134.
52. D. RUELLE (1969) *Statistical mechanics: Rigorous results.* W. A. Benjamin.
53. R. J. SERFLING (1975) A general Poisson approximation theorem. *Ann. Probab.* **3**, 726–731.
54. A. N. SHIRYAEV (1984) *Probability* (translated by R. P. Boas). Springer-Verlag .
55. A. SOKAL (1989) *Monte Carlo methods in statistical mechanics: foundations and new algorithms.* Cours de troisième cycle de la physique en Suisse romande.
56. C. STEIN (1971) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **3**, 583–602.
57. L. TIERNEY (1994) Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1728.
58. T. VISWANATHAN (1992) *Telecommunication switching systems and networks.* Prentice-Hall of India.
59. Z. WANG & X. YANG (1992) *Birth and death processes and Markov chains.* Springer-Verlag.
60. A. XIA (1994) *Poisson Approximations and Some Limit Theorems for Markov Processes.* PhD thesis, University of Melbourne, Australia.
61. A. XIA (1997a) On the rate of Poisson process approximation to a Bernoulli process. *J. Appl. Probab.* **34**, 898–907.
62. A. XIA (1997b) On using the first difference in Stein-Chen method. *Ann. Appl. Probab.* **7**, 899–916.
63. A. XIA (1999) A probabilistic proof of Stein's factors. *J. Appl. Probab.* **36**, 287–290.
64. A. XIA (2000) Poisson approximation, compensators and coupling. *Stoch. Analysis and Applics* **18**, 159–177.
65. N. YANNAROS (1991) Poisson approximation for random sums of Bernoulli random variables. *Statist. Probab. Lett.* **11**, 161–165.

Three general approaches to Stein's method

Gesine Reinert

Department of Statistics, University of Oxford
1 South Parks Road, Oxford OX1 3TG, UK
E-mail: reinert@stats.ox.ac.uk

Stein's method is in no way restricted to the particular settings discussed in the previous chapters: normal, Poisson and compound Poisson approximation for random variables, and Poisson point process approximation. This chapter is intended to illustrate that the method can be very generally exploited, and that one can set up and apply Stein's method in a wide variety of contexts. Three main approaches for achieving this are described: the generator method, density equations and coupling equations. These are applied to chi-squared approximation, to weak laws of large numbers in general spaces, including spaces of measures, and to discrete Gibbs distributions. The S-I-R epidemic model, which is discussed in some detail, serves to illustrate the extent of the possibilities.

Contents

1. Introduction

The goal of this chapter is to illustrate how Stein's method can be applied to a variety of distributions. We shall employ three different approaches, namely the generator method (see also Chapter 2, Section 2.2), density equations, and coupling equations. Two main examples to bear in mind are

(1) The standard normal distribution $\mathcal{N}(0,1)$: Let N denote the expectation under $\mathcal{N}(0,1)$. The Stein characterization for $\mathcal{N}(0,1)$ is that $X \sim \mathcal{N}(0,1)$ if and only if for all continuous and piecewise continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ with $N(|f'|) < \infty$, we have

$$\mathbb{E}f'(X) - \mathbb{E}Xf(X) = 0.$$

See Stein (1986) and Chapter 1 for a thorough treatment.

(2) The Poisson distribution with mean λ, $\mathrm{Po}(\lambda)$, with the corresponding Stein characterization that $X \sim \mathrm{Po}(\lambda)$ if and only if for all real-valued functions f for which both sides of the equation exist we have that

$$\lambda\mathbb{E}f(X+1) - \mathbb{E}Xf(X) = 0.$$

For a detailed treatment, see for example Chen (1975), Arratia, Goldstein & Gordon (1989), Barbour, Holst & Janson (1992), and Chapter 2, Section 2.

Stein's method for a general target distribution μ can be sketched as follows.

(1) Find a suitable characterization, namely an operator $\mathcal{A}$ such that $X \sim \mu$ if and only if for all smooth functions f, $\mathbb{E}\mathcal{A}f(X) = 0$ holds.

(2) For each smooth function h find a solution $f = f_h$ of the *Stein equation*

$$h(x) - \int h d\mu = \mathcal{A}f(x). \tag{1.1}$$

(3) Then for any variable W, it follows that

$$\mathbb{E}h(W) - \int h d\mu = \mathbb{E}\mathcal{A}f(W),$$

where f is the solution of the Stein equation for h.

Usually, in order to yield useful results, it is necessary to bound f, f' or, for discrete distributions, $\Delta f = \sup_x |f(x+1) - f(x)|$. In what follows, we shall always assume that the test function h is smooth; for nonsmooth functions, the reader is referred to the techniques used in Chapter 1, by Rinott & Rotar (1996), and by Götze (1991).

In Section 2, the generator approach is briefly described. The flexibility of this approach is illustrated by the applications to chi-squared approximations in Section 3, to laws of large numbers for empirical measures in Section 4, and to Gibbs measures in Section 5. In Section 6, we give a more involved example, the mean-field behaviour of the general stochastic epidemic. Section 7 explains the density approach, and in Section 8 the approach via coupling equations, viewed as distributional transformations, is given. The powerful approach of exchangeable pairs is described in detail in Stein (1986); for reasons of space, it is omitted here. This chapter is meant as an accessible overview; essentially none of the results given here are new, but rather it is hoped that this compilation will draw the reader to see the variety of possible approaches currently used in Stein's method.

2. The generator approach

The generator approach was introduced in Barbour (1988,1990), and was also developed in Götze (1991). The basic idea is to choose the operator $\mathcal{A}$ to be the generator of a Markov process with stationary distribution μ. That is, for a homogeneous Markov process $(X_t)_{t\geq 0}$, put $T_t f(x) = \mathbb{E}(f(X_t)|X(0) = x)$. The *generator* of the Markov process is defined as $\mathcal{A}f(x) = \lim_{t\downarrow 0} \frac{1}{t}(T_t f(x) - f(x))$. Standard results for generators (Ethier & Kurtz (1986) pp. 9–10, Proposition 1.5) yield

(1) If μ is the stationary distribution of the Markov process then $X \sim \mu$ if and only if $\mathbb{E}\mathcal{A}f(X) = 0$ for all real-valued functions f for which $\mathcal{A}f$ is defined.

(2) $T_t h - h = \mathcal{A}\left(\int_0^t T_u h du\right)$, and, formally taking limits,

$$\int h d\mu - h = \mathcal{A}\left(\int_0^\infty T_u h du\right),$$

if the right-hand side exists.

Thus the generator approach gives both a Stein equation and a candidate for its solution. One could hence view this approach as a Markov process perturbation technique.

Example 2.1: The operator

$$\mathcal{A}h(x) = h''(x) - xh'(x) \tag{2.1}$$

is the generator of the *Ornstein–Uhlenbeck* process with stationary distribution $\mathcal{N}(0,1)$. Putting $f = h'$ gives the classical Stein characterization for $\mathcal{N}(0,1)$.

Example 2.2: The operator $\mathcal{A}h(x) = \lambda(h(x+1)-h(x))+x(h(x-1)-h(x))$ or, for $f(x) = h(x+1) - h(x)$,

$$\mathcal{A}f(x) = \lambda f(x+1) - xf(x), \tag{2.2}$$

is the generator of an immigration–death process with immigration rate λ and unit per capita death rate. Its stationary distribution is $\mathrm{Po}(\lambda)$ (see Chapter 2, Section 2.2 and Chapter 3, Section 3). Again it yields the classical Stein characterization of the Poisson distribution.

A main advantage of the generator approach is that it easily generalizes to multivariate distributions and to distributions on more complex spaces, such as the distributions of path-valued random elements, or of measure-valued elements. However, the generator approach is not always easily set up; see the problems associated with the compound Poisson distribution as described in Chapter 2, Section 3. Also note that there is no unique choice of generator for any given target distribution — just as there is no unique choice of Stein equation for a given target distribution.

In some instances there is a useful heuristic for finding a suitable generator. Let us assume that the target distribution is based on the distributional limit of some function $\Phi_n(X_1, \ldots, X_n)$ as $n \to \infty$, where $X_1, \ldots, X_n$ are independent and identically distributed; furthermore, assume that $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$ for all i. Using ideas from the method of exchangeable pairs, we can construct a reversible Markov chain as follows.

(1) Start with $Z_n(0) = (X_1, \ldots, X_n)$.
(2) Pick an index $I \in \{1, \ldots, n\}$ independently uniformly at random; if $I = i$, replace X_i by an independent copy X_i^*.
(3) Put $Z_n(1) = (X_1, \ldots, X_{I-1}, X_I^*, X_{I+1}, \ldots\ X_n)$.
(4) Draw another index uniformly at random, throw out the corresponding random variable and replace it by an independent copy.
(5) Repeat the procedure.

This Markov chain is then converted into a continuous-time Markov process by letting $N(t)$, $t \geq 0$, be a Poisson process with rate 1, and setting

$$W_n(t) = Z_n(N(t)).$$

The generator $\mathcal{A}_n$ for Markov process is given by the expression

$$\mathcal{A}_n f(\Phi_n(\mathbf{x})) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E} f(\Phi_n(x_1, \ldots, x_{i-1}, X_i^*, x_{i+1}, \ldots\ x_n)) - f(\Phi_n(\mathbf{x})),$$

where $\mathbf{x} = (x_1, \ldots\ x_n)$, and f is a smooth function. Taylor's expansion then yields

$$\begin{aligned}
\mathcal{A}_n f(\Phi_n(\mathbf{x})) &\approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i^* - x_i) f'(\Phi_n(\mathbf{x})) \frac{\partial}{\partial x_i}\Phi_n(\mathbf{x}) \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n} \mathbb{E}(X_i^* - x_i)^2 \left\{ f''(\Phi_n(\mathbf{x})) \left(\frac{\partial}{\partial x_i}\Phi_n(\mathbf{x})\right)^2 + f'(\Phi_n(\mathbf{x})) \frac{\partial^2}{\partial x_i^2}\Phi_n(\mathbf{x}) \right\} \\
&= -\frac{1}{n}\sum_{i=1}^{n} x_i f'(\Phi_n(\mathbf{x})) \frac{\partial}{\partial x_i}\Phi_n(\mathbf{x}) \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n} (1 + x_i^2) \left\{ f''(\Phi_n(\mathbf{x})) \left(\frac{\partial}{\partial x_i}\Phi_n(\mathbf{x})\right)^2 + f'(\Phi_n(\mathbf{x})) \frac{\partial^2}{\partial x_i^2}\Phi_n(\mathbf{x}) \right\}.
\end{aligned}$$

Letting $n \to \infty$, with a suitable scaling, would then give a generator for the target distribution.

Example 2.3: Suppose that we are interested in a central limit theorem, with target distribution $\mathcal{N}(0,1)$, the standard normal distribution. In the above setting, put $\Phi_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i$. Then $\frac{\partial}{\partial x_i}\Phi_n(\mathbf{x}) = \frac{1}{\sqrt{n}}$ and $\frac{\partial^2}{\partial x_i^2}\Phi_n(\mathbf{x}) = 0$; hence

$$\begin{aligned}
\mathcal{A}_n f(\Phi_n(\mathbf{x})) &\approx -\frac{1}{n^{3/2}}\sum_{i=1}^{n} x_i f'(\Phi_n(\mathbf{x})) + \frac{1}{2n^2}\sum_{i=1}^{n} (1 + x_i^2) f''(\Phi_n(\mathbf{x})) \\
&= -\frac{1}{n}\Phi_n(\mathbf{x}) f'(\Phi_n(\mathbf{x})) + \frac{1}{2n} f''(\Phi_n(\mathbf{x})) \left(1 + \frac{1}{n}\sum_{i=1}^{n} x_i^2\right) \\
&\approx \frac{1}{n}\left\{ f''(\Phi_n(\mathbf{x})) - \Phi_n(\mathbf{x}) f'(\Phi_n(\mathbf{x})) \right\},
\end{aligned}$$

where we have applied the law of large numbers for the last approximation. If we choose a Poisson process with rate n instead of rate 1, then the factor $\frac{1}{n}$ vanishes, and we obtain as limiting generator $\mathcal{A}f(x) = f''(x) - xf'(x)$, the operator from (2.1).

3. Chi-squared distributions

Following the heuristic, we first find a generator for the chi-squared distribution with p degrees of freedom. Let $\mathbf{X}_1, \ldots, \mathbf{X}_p$ be independent and identically distributed random vectors, where $\mathbf{X}_i = (X_{i,1}, \ldots X_{i,n})$; the components $X_{i,j}$ are again assumed to be independent and identically distributed, having zero means, unit variances, and finite fourth moments. We then set

$$\Phi_n(\mathbf{x}) = \sum_{i=1}^{p} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} x_{i,j} \right)^2 .$$

Choose an index uniformly from $\{1, \ldots, p\} \times \{1, \ldots, n\}$. We have $\frac{\partial}{\partial x_{i,j}} \Phi_n(\mathbf{x}) = \frac{2}{n} \sum_{k=1}^{n} x_{i,k}$ and $\frac{\partial^2}{\partial x_{i,j}^2} \Phi_n(\mathbf{x}) = \frac{2}{n}$, giving

$$\begin{aligned}
\mathcal{A}_n f(\Phi_n(\mathbf{x})) &\approx -\frac{2}{pn} f'(\Phi_n(\mathbf{x})) \sum_{i=1}^{p} \sum_{j=1}^{n} x_{i,j} \frac{1}{n} \sum_{k=1}^{n} x_{i,k} \\
&\quad + \frac{1}{2dn} f''(\Phi_n(\mathbf{x})) \sum_{i=1}^{p} \sum_{j=1}^{n} (1 + x_{i,j}^2) \frac{4}{n^2} \left(\sum_{k=1}^{n} x_{i,k} \right)^2 \\
&\quad + \frac{1}{2dn} f'(\Phi_n(\mathbf{x})) \sum_{i=1}^{p} \sum_{j=1}^{n} (1 + x_{i,j}^2) \frac{2}{n} \\
&\approx -\frac{2}{pn} f'(\Phi_n(\mathbf{x})) \Phi_n(\mathbf{x}) + \frac{4}{pn} f''(\Phi_n(\mathbf{x})) \Phi_n(\mathbf{x}) + \frac{2}{n} f'(\Phi_n(\mathbf{x}));
\end{aligned}$$

for the last approximation, we have again applied the law of large numbers. This suggests as generator

$$\mathcal{A}f(x) = \frac{4}{p} x f''(x) + 2 \left(1 - \frac{x}{p} \right) f'(x).$$

It is more convenient to choose

$$\mathcal{A}f(x) = x f''(x) + \frac{1}{2}(p - x) f'(x) \tag{3.1}$$

as generator for χ_p^2; this is just a rescaling.

Luk (1994) chooses as Stein operator for the Gamma distribution $\mathrm{Ga}(r, \lambda)$ the operator

$$\mathcal{A}f(x) = x f''(x) + (r - \lambda x) f'(x). \tag{3.2}$$

As $\chi_p^2 = \mathrm{Ga}(d/2, 1/2)$, this agrees with our generator. Luk showed that, for χ_p^2, $\mathcal{A}$ is the generator of a Markov process given by the solution of the

stochastic differential equation

$$X_t = x + \frac{1}{2}\int_0^t (p - X_s)ds + \int_0^t \sqrt{2X_s}\, dB_s,$$

where B_s is standard Brownian motion. Luk also found the transition semigroup, which can be used to solve the Stein equation

$$h(x) - \chi_p^2(h) = xf''(x) + \frac{1}{2}(p - x)f'(x), \tag{3.3}$$

where $\chi_p^2(h)$ is the expectation of h under the χ_p^2-distribution. His bound has recently been improved as follows.

Lemma 3.1: [Pickett 2002]. *Suppose that $h : \mathbb{R} \to \mathbb{R}$ is absolutely bounded, with $|h(x)| \leq ce^{ax}$ for some $c > 0$ and $a \in \mathbb{R}$, and that the first k derivatives of h are bounded. Then the equation (3.3) has a solution $f = f_h$ such that*

$$\|f^{(j)}\| \leq \sqrt{\frac{2\pi}{p}}\, \|h^{(j-1)}\|,$$

with $h^{(0)} = h$.

(Note the improvement over Luk (1994), in gaining the factor $\frac{1}{\sqrt{p}}$). To put this approach into use, we consider a basic example.

Example 3.2: (The squared sum of i.i.d. random variables).
Let X_i, $i = 1, \ldots, n$, be independent and identically distributed, with mean zero, variance one, and with finite 8^{th} moment. Define

$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i; \qquad W = S^2.$$

To show that $\mathcal{L}(W)$ is close to χ_1^2, we bound the quantity

$$2\mathbb{E}\mathcal{A}f(W) = 2\mathbb{E}Wf''(W) + \mathbb{E}(1 - W)f'(W),$$

where $\mathcal{A}$ is the generator given in (3.1) with $p = 1$. The full argument is to be found in Pickett & Reinert (2004); we give an indication of the salient features.

Defining $g(s) = sf'(s^2)$, it follows that

$$g'(s) = f'(s^2) + 2s^2 f''(s^2)$$

and that

$$\begin{aligned} 2\mathbb{E}Wf''(W) + \mathbb{E}(1 - W)f'(W) &= \mathbb{E}g'(S) - \mathbb{E}f'(W) + \mathbb{E}(1 - W)f'(W) \\ &= \mathbb{E}g'(S) - \mathbb{E}Sg(S). \end{aligned}$$

Now proceed as for standard normal approximation. Set

$$S_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j,$$

and observe that

$$\begin{aligned}\mathbb{E}Sg(S) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}X_i g(S) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}X_i g(S_i) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i^2 g'(S_i) + R_1,\end{aligned}$$

where

$$R_1 = \frac{1}{n^{3/2}} \sum_i \mathbb{E}X_i^3 g''(S_i) + \frac{1}{2n^2} \sum_i \mathbb{E}X_i^4 g^{(3)}\left(S_i + \theta_1 \frac{X_i}{\sqrt{n}}\right),$$

by Taylor's expansion, for some $0 < \theta_1 = \theta_1(S_i, X_i) < 1$. From independence, it follows that

$$\mathbb{E}Sg(S) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}g'(S_i) + R_1 = \mathbb{E}g'(S) + R_1 + R_2,$$

where

$$\begin{aligned}R_2 &= \frac{1}{n^{3/2}} \sum_i \mathbb{E}X_i g''(S_i) + \frac{1}{2n^2} \sum_i \mathbb{E}X_i^2 g^{(3)}\left(S_i + \theta_2 \frac{X_i}{\sqrt{n}}\right) \\ &= \frac{1}{2n^2} \sum_i \mathbb{E}X_i^2 g^{(3)}\left(S_i + \theta_2 \frac{X_i}{\sqrt{n}}\right),\end{aligned}$$

again by Taylor's expansion, for some $0 < \theta_2 = \theta_2(S_i, X_i) < 1$.

To bound R_1 and R_2, we calculate

$$g''(s) = 6sf''(s^2) + 4s^3 f^{(3)}(s^2)$$

and

$$g^{(3)}(s) = 24s^2 f^{(3)}(s^2) + 6f''(s^2) + 8s^4 f^{(4)}(s^2).$$

With $\beta_i = \mathbb{E}X_1^i$, we hence obtain

$$\begin{aligned}&\frac{1}{2n^2} \sum_i \mathbb{E}X_i^2 \left| g^{(3)}\left(S_i + \theta_2 \frac{X_i}{\sqrt{n}}\right)\right| \\ &\leq \frac{24}{n} \|f^{(3)}\| \left(1 + \frac{\beta_4}{n}\right) + \frac{6}{n} \|f''\| + \frac{8}{n} \|f^{(4)}\| \left(6 + \frac{\beta_4}{n} + 4\frac{\beta_3^2}{\sqrt{n}} + 6\frac{\beta_4}{n} + \frac{\beta_6}{n^2}\right) \\ &= c(f)\frac{1}{n},\end{aligned}$$

where c is a constant depending on f and on the distribution of the X_i, but not upon n; it can be calculated explicitly. Similarly, we can obtain a bound of order $1/n$ for $\frac{1}{2n^2}\sum_i \mathbb{E}X_i^4 \left|g^{(3)}(S_i + \theta_1 \frac{X_i}{\sqrt{n}})\right|$, where we now use β_8.

For the expression $\frac{1}{n^{3/2}}\sum_i \mathbb{E}X_i^3 g''(S_i)$, we use an antisymmetry argument. We have, for some constant $c(f)$,

$$\frac{1}{n^{3/2}}\sum_i \mathbb{E}X_i^3 g''(S_i) = \frac{1}{\sqrt{n}}\beta_3 \mathbb{E}g''(S) + c(f)\frac{1}{n}$$

and

$$\mathbb{E}g''(S) = 6\mathbb{E}Sf''(S^2) + 4\mathbb{E}S^3 f^{(3)}(S^2).$$

Here we have again used Taylor's expansion. Note that g'' is antisymmetric, $g''(-s) = -g''(s)$, so that for $Z \sim \mathcal{N}(0,1)$ we have $\mathbb{E}g''(Z) = 0$. Thus $|\mathbb{E}g''(S) - \mathbb{E}g''(Z)| = |\mathbb{E}g''(S)|$, and it is (almost) routine now to apply Stein's method for normal approximation to show that $|\mathbb{E}g''(S)| \leq c(f)/\sqrt{n}$ for some constant $c(f)$ that depends on f and on the distribution of the X_i, but not on n.

Combining these bounds shows that the bound on the distance to χ_1^2 for smooth test functions is of order $\frac{1}{n}$. This result can also be extended to χ_p^2-approximations; see Pickett & Reinert (2004).

4. The weak law of large numbers

Using the generator method and the above heuristic, it is straightforward to find a generator for the target distribution δ_0, point mass at 0, namely,

$$\mathcal{A}f(x) = -xf'(x),$$

and the corresponding transition semigroup is given by

$$T_t h(x) = h\left(xe^{-t}\right);$$

see Reinert (1994). A Stein equation for point mass at 0 is hence given by

$$h(x) - h(0) = -xf'(x). \tag{4.1}$$

Details of the treatment of the weak law of large numbers as presented here can be found in Reinert (1994).

Lemma 4.1: [Reinert (1994)]. *If $h \in C_b^2(\mathbb{R})$, then the Stein equation (4.1) has a solution $f = f_h \in C_b^2$ satisfying*

$$\|f'\| \leq \|h'\|, \text{ and } \|f''\| \leq \|h''\|.$$

Proof: The proof illustrates briefly how to bound solutions of the Stein equation using the generator method. First note that we may assume that $h(0) = 0$. The generator method gives as candidate solution

$$f(x) = -\int_0^\infty h\left(xe^{-t}\right) dt = -\int_0^x \frac{h(t)}{t} dt,$$

so, for $x \neq 0$, we have

$$|f'(x)| = \left|\frac{h(x)}{x}\right| \leq \|h'\|,$$

and for $x = 0$ we have $f'(0) = -h'(0)$, giving the first assertion. For the second assertion, for $x \neq 0$,

$$|f''(x)| = \left|\frac{h(x)}{x^2} - \frac{h'(x)}{x}\right| \leq \|h''\|,$$

and for $x = 0$ we have $f''(0) = -h''(0)$. This completes the proof. ∎

Example 4.2: (Weak law of large numbers).
Suppose that $X_1, \ldots, X_n$ are (dependent) random variables having zero mean and finite variance, and put

$$W = W_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then, by Taylor's expansion, for some $0 < \theta = \theta(W) < 1$, we have

$$\mathbb{E}\mathcal{A}f(W) = -\mathbb{E}Wf'(W) = -\mathbb{E}Wf'(0) + \mathbb{E}W^2 f''(\theta W) = \mathbb{E}W^2 f''(\theta W),$$

and

$$|\mathbb{E}\mathcal{A}f(W)| \leq \|f''\| \mathrm{Var}\,(W).$$

In particular, if $\mathrm{Var}\,(W_n) \to 0$ as $n \to \infty$, then the weak law of large numbers holds. This argument can easily be generalized to point mass at μ, with generator $\mathcal{A}f(x) = (\mu - x)f'(x)$. Note that the above bound is explicit; there is no need for $n \to \infty$.

The above argument for the weak law of large numbers could, of course, be replaced by a simple use of Chebyshev's inequality. However, its generalization to measure-valued random elements yields effective bounds on approximations for empirical measures.

4.1. *Empirical measures*

First we need the set-up for empirical measures; this is slightly technical. Denote by E a locally compact Hausdorff space with countable basis, for example, $E = \mathbb{R}$, $\mathbb{R}^d$ or $\mathbb{R}_+$. Thus we can define a metric on E, and it makes sense to talk about the Borel sets $\mathcal{B}$ of E. For a signed measure μ on E, that is, a measure that would be allowed to take on negative values, define the norm

$$\|\mu\| = \sup_{A \in \mathcal{B}} |\mu(A)|.$$

Then the space of all bounded signed measures

$$M_b(E) = \{\mu : \|\mu\| < \infty\}$$

is a linear space. We equip this space with the vague topology, which is defined as follows. Put

$$C_c(E) = \{f : E \to \mathbb{R} \text{ continuous with compact support}\}.$$

For $(\nu_n)_{n \geq 1}$ and ν in $M_b(E)$, we say that ν_n converges to ν vaguely,

$$\nu_n \stackrel{v}{\Rightarrow} \nu \iff \text{for all } f \in C_c(E) : \int f d\nu_n \to \int f d\nu.$$

Note the difference to weak convergence, the latter being defined via continuous bounded test functions. For example, the set of point masses $\delta_n \stackrel{v}{\Rightarrow} 0$ for $n \to \infty$, but it does not converge weakly.

For Stein's method, we would prefer a class of test functions from $C_c(E)$ that allows for Taylor expansion. It suffices to find a suitable convergence-determining subclass of $C_c(E)$. We consider functions of the type

$$F(\nu) = f\left(\int \phi_i \, d\nu, i = 1, \ldots, m\right) \tag{4.2}$$

for some m, $f \in C_b^\infty(\mathbb{R}^m)$ and $\phi_i, \ldots, \phi_m \in C_c(E)$, and we define

$$\mathcal{F} = \{F \in C_c(E) : F \text{ satisfies (4.2)}\}.$$

Using the Stone–Weierstrass Theorem, the following lemma is straightforward.

Lemma 4.3: *$\mathcal{F}$ is convergence-determining for vague convergence. So is the restricted class $\mathcal{F}_0$ of functions such that $\|f'\| \leq 1, \|f''\| \leq 1, \|\phi_i\| \leq 1$ for $i = 1, \ldots, m$. Also, if $E = \mathbb{R}^d$ or some connected open or closed subset of $\mathbb{R}^d$, instead of $C_c(E)$ we could use $C_b^\infty(E)$, the space of all bounded continuous functions on E that are infinitely often boundedly differentiable.*

Here, we use the notation $\|f'\| = \sum_{j=1}^m \|f_{(j)}\|$, where $f_j = \frac{\partial}{\partial x_j} f$.

4.2. *Weak law of large numbers for empirical measures*

The above framework now allows us to formulate and prove a weak law of large numbers for empirical measures. Let $X_1, \ldots, X_n$ be random elements taking values in E, and let μ_i be the law of X_i. Put

$$\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i$$

and define the *empirical measure*

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_a denotes the point mass at a. Then, for all smooth functions f, we have

$$\mathbb{E} \int f \, d\xi_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} f(X_i) = \int f \, d\bar{\mu}.$$

Suppose that we would like to bound the distance between $\mathcal{L}(\xi_n)$ and some δ_μ, say, where typically μ should be close to $\bar{\mu}$. As in the real-valued case, for F of the form (4.2), the generator for point mass at μ is

$$\mathcal{A}F(\nu) = \sum_{j=1}^{m} f_{(j)} \left(\int \phi_i \, d\nu, i = 1, \ldots, m \right) \left(\int \phi_j \, d\mu - \int \phi_j \, d\nu \right).$$

We could also describe this generator in terms of a Gateaux derivative, $\mathcal{A}F(\nu) = F'(\nu)[\mu - \nu]$. With a proof very similar to the real-valued case, it is easy to show that the following bounds on the solution of the Stein equations hold.

Lemma 4.4: *For every H of the form*

$$H(\nu) = h \left(\int \phi_i \, d\nu, i = 1, \ldots, m \right) \tag{4.3}$$

for some $m \geq 1$, $h \in C_b^\infty(\mathbb{R}^m)$ and $\phi_i, \ldots, \phi_m \in C_c(E)$, the solution $F = F_H$ of the Stein equation is of the form (4.2), with the same ϕ_i's. Furthermore, $\|f'\| \leq \|h'\|$ and $\|f''\| \leq \|h''\|$.

This immediately leads to the following theorem.

Theorem 4.5: (Weak law of large numbers for empirical measures). *For all $H \in \mathcal{F}$, we have*

$$|\mathbb{E}H(\xi_n) - H(\mu)| \leq \sum_{j=1}^{m} \|h_{(j)}\| \left| \int \phi_j \, d\bar{\mu} - \int \phi_j \, d\mu \right|$$
$$+ \sum_{j,k=1}^{m} \|h_{(j,k)}\| \left\{ \max_{1 \leq j \leq m} \left[\int \phi_j \, d\bar{\mu} - \int \phi_j \, d\mu \right]^2 + \mathrm{Var} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) \right) \right\}.$$

If $\bar{\mu} = \mu$, then we recover the usual variance bound. This result is very general; we shall now consider some typical situations where it can be applied.

4.3. *Mixing random elements*

Let $X_1, \ldots, X_n$ be random elements taking values in E. Set

$$\mathcal{B}_{i,j} = \{A, B \in \mathcal{B} : \mathbb{P}(X_i \in A) \neq 0, \mathbb{P}(X_j \in B) \neq 0\}$$

and

$$\rho_n = \frac{1}{n^2} \sum_{i,j=1}^{m} \sup_{A,B \in \mathcal{B}_{i,j}} |\mathrm{Corr}\,(\mathbf{I}(X_i \in A), \mathbf{I}(X_j \in B))|$$

Then, if $\|\phi_i\| \leq 1$ for $i = 1, \ldots, m$, it is easy to see that

$$\mathrm{Var} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_i(X_i) \right) \leq 4\rho_n.$$

Thus the bound in Theorem 4.5 can be expressed in terms of this mixing coefficient. A similar approach is possible for other definitions of mixing.

4.4. *Locally dependent random elements*

Assume that for all $i \in I = \{1, \ldots, n\}$ there is a set $\Gamma_i \subset I$ not containing i such that X_i is independent of $(X_j, j \notin \Gamma_i)$. Then, if $\|\phi_j\| \leq 1$ for each $i = 1, \ldots, m$, we have as bound on the variance in Theorem 4.5

$$\mathrm{Var} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) \right) \leq \frac{1}{n} + \frac{2}{n^2} \sum_{i=1}^{n} |\Gamma_i|. \tag{4.4}$$

The approach could also be extended to the more general case in which there is a relatively small neighbourhood of strong dependence, whereas the dependence outside this small neighbourhood is weak. To illustrate this approach, we consider the following example.

Example 4.6: (Dissociated families). Let $(Y_i)_{i\in\mathbb{N}}$ be a family of independent and identically distributed random elements on a space $\mathcal{X}$, let $k \in \mathbb{N}$ be fixed, and define the set of multi-indices

$$\Gamma^{(n)} = \{(j_1, \ldots, j_k) \in \Gamma : \ j_1, \ldots, j_k \in \{1, \ldots, n\}, \ j_r \neq j_s \ \text{ for } \ r \neq s\}.$$

Let ψ be a measurable function $\mathcal{X}^k \to E$, and, for $(j_1, \ldots, j_k) \in \Gamma^{(n)}$, put

$$X_{j_1,\ldots,j_k} = \psi(Y_{j_1}, \ldots, Y_{j_k}).$$

Then $(X_{j_1,\ldots,j_k})_{(j_1,\ldots,j_k)\in\Gamma^{(n)}}$ is a dissociated family of identically distributed elements. Define $\mu := \mu_{j_1,\ldots,j_k} := \mathcal{L}(X_{j_1,\ldots,j_k})$; in view of the construction, this measure does not depend on the chosen multi-index. Note that, if $J \in \Gamma^{(n)}$ and $K \in \Gamma^{(n)}$ are disjoint multi-indices, then X_J and X_K are independent. Fix any enumeration of the $r(n) := n(n-1)\cdots(n-k+1)$ elements of the set $\Gamma^{(n)}$. Then our empirical measure of interest can be written as

$$\xi_n = \frac{1}{r(n)} \sum_{i=1}^{r(n)} \delta_{X_{i,n}}.$$

We now derive the following result.

Theorem 4.7: *For the above dissociated family, we have*

$$|\mathbb{E}H(\xi_n) - H(\mu)| \leq \frac{2k+1}{n(n-k+1)}$$

for each $H \in \mathcal{F}_0$.

Proof: For a multi-index $J \in \Gamma^{(n)}$, we set

$$\Gamma(J) := \{L \in \Gamma^{(n)} : \ J \neq L, \, L \cap J \neq \emptyset\};$$

then $\Gamma(J)$ is the dependence neighbourhood for X_J, with

$$|\Gamma(J)| = k\left(\frac{(n-1)!}{(n-k+1)!} - 1\right)$$

for $k \geq 2$, and $|\Gamma(J)| = 0$ for $k = 1$, and thus

$$\frac{1}{r(n)^2} \sum_{J\in\Gamma^{(n)}} |\Gamma(J)| < k\,\frac{(n-k)!(n-1)!}{n!(n-k+1)!} = \frac{k}{n(n-k+1)}.$$

With (4.4), this yields as bound for the variance in Theorem 4.5

$$\operatorname{Var}\left(\frac{1}{r(n)}\sum_{i=1}^{r(n)}\phi_j(X_i)\right) \le \frac{1}{n(n-1)\cdots(n-k+1)} + \frac{2k}{n(n-k+1)}$$
$$\le \frac{2k+1}{n(n-k+1)}.$$

As the $X_{j_1,\ldots,j_k}$'s are identically distributed, we have $\bar{\mu} = \mu$, so the variance term is the only contribution to the bound in Theorem 4.5. This finishes the proof. ∎

The above result can be extended to cover any family of functions $(\psi_{j_1,\ldots,j_k})_{(j_1,\ldots,j_k)\in\Gamma^{(n)}}$, in which the functions at each multi-index may be different.

4.5. *The size-bias coupling*

As often in the context of Stein's method, couplings can be very useful for weak laws of large numbers for empirical measures. For a better understanding of the Palm–measure related coupling that we introduce, we first recall the size-bias coupling for real-valued random variables.

Let $W \ge 0$ be a nonnegative real-valued random variable and assume that $\mathbb{E}W > 0$. Then a random variable W^* is said to have the W-size-biased distribution if the equation

$$\mathbb{E}Wg(W) = \mathbb{E}W\mathbb{E}g(W^*) \tag{4.5}$$

is satisfied for all g for which both sides of the equation exist. This implicit characterization is equivalent to requiring that the ratio of the densities of $\mathcal{L}(W^*)$ and $\mathcal{L}(W)$ at x is proportional to x. In many examples, the distribution of W^* is particularly easy to determine.

Example 4.8: If $W \sim \mathrm{Be}(p)$ is a Bernoulli random variable with parameter $0 < p \le 1$, then we have $\mathbb{E}Wg(W) = pg(1)$ for all functions g. Since $\mathbb{E}W = p$, it follows that $W^* = 1$ a.s.; that is, W^* is deterministic and takes the value 1. As this conclusion does not depend on the value of p, it also follows that the size-bias transformation is not one-to-one.

Example 4.9: If $W \sim \mathrm{Po}(\lambda)$ has the Poisson distribution with mean $\lambda > 0$, then $\mathbb{E}W = \lambda$, and it follows from the Stein–Chen equation that $\mathbb{E}Wg(W) = \lambda\mathbb{E}g(W+1)$ for all functions g for which both sides of the equation exist. From this we see that $W^* = W + 1$ in distribution; in other

words, W^* has the distribution of a Poisson random variable with mean λ, which is shifted by 1.

In the weak law of large numbers setting, the size-bias coupling leads to the equation

$$\mathbb{E}\mathcal{A}f(W) = \mathbb{E}(\mathbb{E}W - W)f'(W) = \mathbb{E}W(\mathbb{E}f'(W) - \mathbb{E}f'(W^*)),$$

where W^* has the W-size-biased distribution. Here we assume that $W \geq 0$ with $\mathbb{E}W > 0$. Thus the distance between the distribution of W and of W^* gives a measure for the distance between the distribution of W and point mass at $\mathbb{E}W$.

For W a sum of random variables, Goldstein & Rinott (1996) give the following construction of W^*. Suppose that $W = \sum_{i=1}^n X_i$, with $X_i \geq 0$ real-valued random variables such that $\mathbb{E}X_i > 0$ for all $i = 1, \ldots, n$. First choose a random index V according to

$$\mathbb{P}(V = v) = \frac{\mathbb{E}X_v}{\mathbb{E}W},$$

that is, proportional to the mean of X_v, independently of the other random variables. If $V = v$, then we replace X_v by an independent random variable X_v^* having the X_v-size-biased distribution. Given X_v^*, the other random variables in the sum W are adjusted as follows. If $X_v^* = x$, then choose the random variables $(\hat{X}_u,\, u \neq v)$ in such a way that

$$\mathcal{L}(\hat{X}_u,\, u \neq v) = \mathcal{L}(X_u,\, u \neq v \,|\, X_v = x)$$

is satisfied. Then

$$W^* = \sum_{u \neq V} \hat{X}_u + X_V^*$$

has the W-size-biased distribution.

Example 4.10: A classical example for this construction is the case when $W = \sum_{i=1}^n X_i$, where $X_i \sim \mathrm{Be}(p_i)$ and $0 < p_i \leq 1$ for $i = 1, \ldots, n$; the Bernoulli variables may be dependent. To construct W^*, we choose an index V as above, proportional to the means; then, if $V = v$, we choose $(\hat{X}_u,\, u \neq v)$ in such a way that

$$\mathcal{L}(\hat{X}_u, u \neq v) = \mathcal{L}(X_u, u \neq v \,|\, X_v = 1).$$

Since $X_v^* = 1$ a.s.,

$$W^* = \sum_{u \neq V} \hat{X}_u + 1$$

has the W-size-biased distribution. This coupling is used extensively for Poisson approximation; see Barbour, Holst & Janson (1992, Chapter 2).

In light of size-biasing for real-valued random variables, we can the define a size-biased distribution for random measures. This distribution is again defined in terms of test functions.

Definition 4.11: Let ξ be a random measure on E, with mean measure $\mathbb{E}[\xi] = \mu$, and let $\phi \in C_c$ be a nonnegative real-valued continuous function having compact support, with $\int \phi \, d\mu > 0$. We say that ξ^ϕ has the ξ size-biased distribution in direction ϕ if

$$\mathbb{E}G(\xi) \int \phi \, d\xi = \int \phi \, d\mu \, \mathbb{E}G(\xi^\phi)$$

for all G for which the expectations on both sides exist.

This implicit definition is related to size-biasing the real-valued random variables $\int \phi \, d\xi$, but the formulation in terms of admissible test functions G is more general than a reduction to the real-valued case.

Example 4.12: Suppose that $\xi = \delta_X$ with $\mathcal{L}(X) = \mu$, and that $\phi \geq 0$ is a real-valued function. Then, for any test function G for which the left-hand side exists,

$$\mathbb{E}G(\xi) \int \phi \, d\xi = \mathbb{E}\phi(X) G(\delta_X) = \int \phi \, d\mu \, \mathbb{E}G(\delta_X^\phi).$$

If $X \geq 0$ and $\phi(x) = x$, then for $G(\nu) = \int g \, d\nu$ we obtain

$$\mathbb{E}G(\xi) \int \phi \, d\xi = \mathbb{E}Xg(X) = (\mathbb{E}X)\mathbb{E} \int g(\nu) \, d(\delta_X^{\phi(x)=x})(\nu).$$

Using the real-valued size-bias coupling with X^* having the X-size-bias distribution, we obtain that

$$\mathbb{E} \int g(\nu) \, d(\delta_X)^{\phi(x)=x}(\nu) = \mathbb{E}g(X^*),$$

and hence

$$(\delta_X)^{\phi(x)=x} = \delta_{X^*}.$$

Thus real-valued size-biasing can be viewed as a special case of size-biasing for random measures.

As in the real-valued case, we can construct a size-biased empirical measure as follows. Let $\xi_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical measure of interest, and denote its mean measure by $\mathbb{E}[\xi_n] = \bar{\mu}_n$. Fix a non-negative real-valued function ϕ. Pick an index $V \in \{1, \ldots, n\}$ proportionally to the mean of $\phi(X_v)$,

$$\mathbb{P}(V = v) = \frac{\mathbb{E}\phi(X_v)}{n \int \phi \, d\bar{\mu}_n},$$

independently of all other random elements. If $V = v$, take $(\delta_{X_v})^\phi$ to have the δ_{X_v}-size-biased distribution in direction ϕ. If $\delta^\phi_{X_v} = \eta$, then choose the remaining variables $(\hat{\delta}_{X_u;\phi}, u \neq v)$ according to

$$\mathcal{L}(\hat{\delta}_{X_u;\phi}, u \neq v) = \mathcal{L}(\delta_{X_u}, u \neq v \,|\, \delta^\phi_{X_v} = \eta).$$

This construction depends on the choice of ϕ, but when the random variables $X_1, \ldots, X_n$ are independent, it shows that we need to adjust only one of the Dirac measures involved in the empirical measure. When the function ϕ is an indicator function, this construction reduces to constructing a Palm measure; see for example Kallenberg (2002), page 207, Theorem 11.5 and the preceding paragraph.

The size-bias coupling in connection with the Stein equation leads to considering, for F of the form (4.2), the generator expression

$$\begin{aligned}\mathbb{E}\mathcal{A}F[\xi] &= \sum_{j=1}^m \int \phi_j \, d\mu \\ &\times \mathbb{E}\left\{ f_{(j)}\left(\int \phi_i \, d\xi_n^{\phi_j}, i = 1, \ldots, n\right) - f_{(j)}\left(\int \phi_i, \; d\xi_n, i = 1, \ldots, n\right)\right\}.\end{aligned}$$

Since, in the independent case, the above construction results in a change to only one of the Dirac measures, the right-hand side should also be small in the case of weakly dependent random elements.

Some final remarks on this approach for weak laws of large numbers.

(1) The above approach uses test functions, and hence only gives results in terms of vague convergence (or weak convergence when all the measures involved have total mass 1). Further argument is needed to obtain almost sure convergence; see, for example, Van der Vaart & Wellner (1996, pp. 122–126) for possible techniques.
(2) In view of the test functions used, the above could also be viewed as a shorthand for multivariate laws of large numbers.

(3) We could also have formulated our results in terms of a Zolotarev-type distance,

$$\zeta(\mu,\nu) := \sup_{g\in\mathcal{F}_0} \left| \int g d\mu - \int g d\nu \right|.$$

We shall see a more involved example on how to apply this technique in Section 6.

5. Discrete distributions from a Gibbs view point

This section presents the recent work of Eichelsbacher & Reinert (2004a,b). Gibbs distributions provide a general framework for discrete univariate distributions. Thus a Stein approach to Gibbs measures can be applied to any univariate discrete distribution. To start with, we review some more examples for Stein operators for univariate discrete distributions.

Example 5.1: The Binomial(n,p)-distribution has Stein operator $\mathcal{A}$ with

$$\mathcal{A}f(k) = (n-k)pf(k+1) - k(1-p)f(k) \tag{5.1}$$

for $0 \le k \le n$; see Ehm (1991).

Example 5.2: The hypergeometric distribution with parameters (n,a,b), having probabilities

$$p_k = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}, \qquad k = 0,\dots,a,$$

has Stein operator $\mathcal{A}$ with

$$\mathcal{A}f(k) = (n-k)(a-k)f(k+1) - k(b-n+k)f(k), \tag{5.2}$$

see Künsch (1998), Reinert & Schoutens (1998) and Schoutens (2000).

Example 5.3: The geometric distribution with parameter p, starting at 0, having probabilities

$$p_k = p(1-p)^k, \qquad k \ge 0.$$

For functions f with $f(0) = 0$, a Stein operator is

$$\mathcal{A}f(k) = (1-p)f(k+1) - f(k), \tag{5.3}$$

for $k \ge 0$; see Peköz (1996).

In what follows we shall exploit the connection between discrete univariate distributions and birth-death processes, as studied also by Brown & Xia (2001), Holmes (2004) and Weinberg (2000).

We consider discrete univariate Gibbs measures μ having support $\text{supp}(\mu) = \{0, \ldots, N\}$, where $N \in \mathbb{Z}_+ \cup \{\infty\}$. By definition, such a Gibbs measure can be written as

$$\mu(k) = \frac{1}{\mathbb{Z}} \exp(V(k)) \frac{\omega^k}{k!}, \quad k = 0, 1, \ldots, N, \tag{5.4}$$

with $\mathbb{Z} = \sum_{k=0}^{N} \exp(V(k)) \frac{\omega^k}{k!}$, where $\omega > 0$ is fixed. We shall assume that the normalizing constant $\mathbb{Z}$ exists.

Note that the assignment of V and w in the representation (5.4) of a Gibbs measure is not unique. For example, if μ denotes the Poisson distribution $\text{Po}(\lambda)$, we could choose $\omega = \lambda, V(k) = -\lambda$ for all $k \geq 0$, and $\mathbb{Z} = 1$, or $V(k) = 0$ for all $k, \omega = \lambda$, and $\mathbb{Z} = e^{\lambda}$.

Conversely, for a given probability distribution $(\mu(k))_{k \in \{0,\ldots,N\}}$ we can find a representation (5.4) as a Gibbs measure by choosing

$$V(k) = \log \mu(k) + \log k! + \log \mathbb{Z} - k \log \omega, \quad k = 0, 1, \ldots, N,$$

with $V(0) = \log \mu(0) + \log \mathbb{Z}$. Again, we have some freedom in the choice of ω, and thus of V. Fix a representation (5.4). To each such Gibbs measure we associate a Markovian birth-death process with unit per-capita death rate $d_k = k$ and birth rate

$$b_k = \omega \exp\{V(k+1) - V(k)\} = (k+1) \frac{\mu(k+1)}{\mu(k)}, \tag{5.5}$$

for $k, k+1 \in \text{supp}(\mu)$. It is easy to see that this birth-death process has invariant measure μ. Following the generator approach to Stein's method, we would therefore choose as generator

$$(\mathcal{A}h)(k) = (h(k+1) - h(k)) \exp\{V(k+1) - V(k)\}\omega + k(h(k-1) - h(k))$$

or, with the simplification $f(k) = h(k) - h(k-1)$,

$$(\mathcal{A}f)(k) = f(k+1) \exp\{V(k+1) - V(k)\}\omega - kf(k). \tag{5.6}$$

In our approach, we typically choose unit per-capita death rates, as used for Poisson processes, see Barbour, Holst & Janson (1992, Chapter 10). Other choices of birth and death rates may be advantageous in some situations, see Brown & Xia (2001) and Holmes (2004). To illustrate the approach, we consider some standard examples.

Example 5.4: The Poisson distribution with mean $\lambda > 0$. We use $\omega = \lambda$, $V(k) = -\lambda$, $\mathcal{Z} = 1$. The Stein operator resulting from (5.6) is the same as the operator (2.2).

Example 5.5: The Binomial distribution with parameters $0 < p < 1$ and n. We use $\omega = \frac{p}{1-p}$, $V(k) = -\log((n-k)!)$ and $\mathcal{Z} = (n!(1-p)^n)^{-1}$. The Stein operator resulting from (5.6) is

$$(\mathcal{A}f)(k) = f(k+1)\frac{p(n-k)}{(1-p)} - kf(k).$$

This differs from the operator (5.1) only by a factor $1-p$, hence bounds on these two operators are equivalent.

Example 5.6: The hypergeometric distribution. The Stein operator resulting from (5.6) is the same as (5.2).

Example 5.7: The negative binomial or Pascal distribution with parameters $\gamma > 0$ and $0 < p < 1$, for which $\mu(k) = \binom{k+\gamma-1}{k} p^\gamma (1-p)^k$ for $k = 0, 1, \ldots$. We obtain the Stein operator

$$(\mathcal{A}f)(k) = f(k+1)(1-p)(k+\gamma) - kf(k).$$

A special case of this is the geometric distribution with parameter p, starting at 0, corresponding to taking $\gamma = 1$ in the Pascal distribution; this gives $\mu(k) = p(1-p)^k$ for $k = 0, 1, \ldots$. The Stein operator resulting from (5.6) is now

$$(\mathcal{A}f)(k) = f(k+1)(1-p)(k+1) - kf(k),$$

which differs from (5.3).

5.1. *Bounds on the solution of the Stein equation*

In order to implement Stein's method in this context, we need to derive bounds on the solutions of the Stein equation (1.1) for the birth-death process generator (5.6). It is straightforward to verify that, for a given function h, a solution f of the Stein equation (1.1) is given by $f(0) = 0$, $f(k) = 0$ for $k \notin \mathrm{supp}(\mu)$, and

$$\begin{aligned} f(j+1) &= \frac{j!}{\omega^{j+1}} e^{-V(j+1)} \sum_{k=0}^{j} e^{V(k)} \frac{\omega^k}{k!} (h(k) - \mu(h)) \\ &= -\frac{j!}{\omega^{j+1}} e^{-V(j+1)} \sum_{k=j+1}^{N} e^{V(k)} \frac{\omega^k}{k!} (h(k) - \mu(h)). \end{aligned}$$

The following non-uniform bounds on f and Δf are derived in Eichelsbacher & Reinert (2004a) and also in Brown & Xia (2001), Weinberg (2000) and Holmes (2004), using different methods.

Lemma 5.8:

(1) Put $M := \sup_{0 \le k \le N-1} \max\left(e^{V(k)-V(k+1)}, e^{V(k+1)-V(k)}\right)$, *and assume that* $M < \infty$. *Then for every* $j \in \mathbb{Z}_+$ *we have*

$$|f(j)| \le 2 \min\left\{1, \sqrt{\frac{M}{\omega}}\right\}.$$

(2) Assume that the birth rates (5.5) are non-increasing, that is,

$$\exp\big(V(k+1) - V(k)\big) \le \exp\big(V(k) - V(k-1)\big),$$

and that death rates are unit per capita. For every $j \in \mathbb{Z}_+$ *we then have*

$$|\Delta f(j)| \le \frac{1}{j} \wedge \frac{e^{V(j)}}{\omega e^{V(j+1)}}.$$

Indeed, Brown & Xia (2001) give bounds for Δf for a wide class of birth-death processes satisfying some monotonicity condition on the rates.

Example 5.4: (continued). For the Poisson distribution with mean $\lambda > 0$, the non-uniform bound gives

$$|\Delta f(k)| \le \frac{1}{k} \wedge \frac{1}{\lambda};$$

see also Barbour, Holst & Janson (1992, p. 8, (1.22)). It is shown in Eichelsbacher & Reinert (2004a) that the non-uniform bound may yield some slight improvement on the bounds for sums of independent but not identically distributed indicator variables. The bound $\|f\| \le 2 \min\left(1, \frac{1}{\sqrt{\lambda}}\right)$ recovers the bound from Barbour, Holst & Janson (1992, Lemma 1.1.1).

Example 5.7: (continued). For the Pascal distribution with parameter $\gamma \in \{1, 2, \ldots\}$ and $0 < p < 1$, the non-uniform bounds give

$$|\Delta f(k)| \le \frac{1}{k} \wedge \frac{1}{(1-p)(k+\gamma)},$$

leading to the uniform bound $1 \wedge \frac{1}{(1-p)\gamma}$; it is not difficult to see that $M = \infty$, so that we do not obtain a bound for $|f(j)|$. In the case of a geometric distribution starting at 0, it is however possible to obtain a bound for $|f(j)|$ using sightly different calculations.

5.2. *The size-bias coupling*

Again we can employ the size-bias coupling to derive bounds on the distance to Gibbs measure. Recall from (4.5) in Section 4.5 that, for $W \geq 0$ with $\mathbb{E}W > 0$, we say that W^* has the W-size-biased distribution if $\mathbb{E}Wg(W) = \mathbb{E}W\mathbb{E}g(W^*)$ for all g for which both sides exist. In particular, we deduce that

$$\mathbb{E}\Big\{\exp\{V(X+1)-V(X)\}\,\omega\, g(X+1) - X\, g(X)\Big\}$$
$$= \mathbb{E}\Big\{\exp\{V(X+1)-V(X)\}\,\omega\, g(X+1) - \mathbb{E}X\mathbb{E}g(X^*)\Big\}.$$

Note that

$$\mathbb{E}X = \omega\mathbb{E}e^{V(X+1)-V(X)}.$$

This immediately leads to the following lemma, which provides a characterization of discrete univariate Gibbs measures on the non-negative integers in terms of their size-biased distributions.

Lemma 5.9: *Let $X \geq 0$ be such that $0 < \mathbb{E}(X) < \infty$, and let μ be a discrete univariate Gibbs measure on the non-negative integers as in (5.4). Then $X \sim \mu$ if and only if for all bounded g we have that*

$$\omega\,\mathbb{E}e^{V(X+1)-V(X)}g(X+1) = \omega\,\mathbb{E}e^{V(X+1)-V(X)}\mathbb{E}g(X^*).$$

Thus for any $W \geq 0$ with $0 < \mathbb{E}W < \infty$ we have

$$\mathbb{E}h(W) - \mu(h) = \omega\{\mathbb{E}e^{V(W+1)-V(W)}f(W+1) - \mathbb{E}e^{V(W+1)-V(W)}\mathbb{E}f(W^*)\},$$

where f is the solution of the Stein equation (1.1) with generator (5.6).

We can also compare two discrete Gibbs distributions by comparing their birth rates and their death rates; a similar idea has been employed in Holmes (2004). Let μ as in (5.4) have generator $\mathcal{A}$ and corresponding (ω, V), and let μ_2 have generator $\mathcal{A}_2$ and corresponding (ω_2, V_2). Assume that both birth-death processes have unit per capita death rates. Then, for $X \sim \mu_2$ and $f \in \mathcal{B}$, if the solution f of the Stein equation (1.1) for μ is

such that $\mathcal{A}_2 f$ exists, we calculate

$$\begin{aligned}
&\mathbb{E}h(X) - \mu(h) \\
&= \mathbb{E}\mathcal{A}f(X) \\
&= \mathbb{E}\{(\mathcal{A} - \mathcal{A}_2)f(X)\} \\
&= \mathbb{E}\left\{f(X+1)\left(\omega e^{V(X+1)-V(X)} - \omega_2 e^{V_2(X+1)-V_2(X)}\right)\right\} \\
&= \omega\mathbb{E}\left\{f(X+1)e^{V_2(X+1)-V_2(X)}e^{V(X+1)-V(X)-(V_2(X+1)-V_2(X))}\right\} \\
&\qquad - \mathbb{E}(X)\mathbb{E}f(X^*) \\
&= \frac{\omega}{\omega_2}\mathbb{E}(X)\mathbb{E}\left\{f(X^*)e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))}\right\} \\
&\qquad - \mathbb{E}(X)\mathbb{E}f(X^*) \\
&= \frac{\omega-\omega_2}{\omega_2}\mathbb{E}(X)\mathbb{E}f(X^*) \\
&\qquad + \frac{\omega}{\omega_2}\mathbb{E}(X)\mathbb{E}f(X^*)\left\{e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - 1\right\},
\end{aligned}$$

where X^* has the X-size-biased distribution. Thus we obtain the bound

$$\begin{aligned}
&\left|\mathbb{E}h(X) - \int h d\mu\right| \\
&\leq \|f\|\mathbb{E}(X)\left\{\frac{|\omega-\omega_2|}{\omega_2} + \frac{\omega}{\omega_2}\mathbb{E}\left|e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - 1\right|\right\}.
\end{aligned}$$

For example, for two Poisson distributions $\mathrm{Po}(\lambda_1)$ and $\mathrm{Po}(\lambda_2)$, this approach gives

$$\left|\mathbb{E}h(X) - \int h d\mu\right| \leq \|f\|\,|\lambda_1 - \lambda_2|.$$

In Eichelsbacher & Reinert (2004a), the approach is employed to bound the distance of summary statistics in an example from statistical physics. In Eichelsbacher & Reinert (2004b), it is generalized to point processes, with the aim of applying it to interacting particle systems.

Note that the normalising constant $\mathbf{Z}$ in the Gibbs distribution, which is often difficult to calculate, is nowhere explicitly needed above. This is one of the main advantages of this approach using Stein's method.

6. Example: an S-I-R epidemic

The general stochastic epidemic (GSE) was introduced in Bartlett (1949) in its most basic form; see also Bailey (1975) for a thorough description.

We shall employ a construction suggested in Sellke (1983). The results presented here are from Reinert (1995,2001).

This model presupposes a population of total size K at time $t = 0$. At any time $t > 0$, an individual in the population may be susceptible (S) to a certain disease, infected (I) by the disease, or removed (R). We assume that an individual is infectious when infected, that any individual can be infected only once, and we ignore births and deaths due to other causes. We suppose that at time $t = 0$ the population consists of aK infected and bK susceptible individuals, with $a + b = 1$.

To construct the GSE, let $(l_i, r_i)_{i \in \mathbb{N}}$ be positive i. i. d. random vectors, and let $(\hat{r}_i)_{i \in \mathbb{N}}$ be positive i. i. d. random variables, independent of $(l_i, r_i)_{i \in \mathbb{N}}$. The r_i's and the $\hat{r}_i$'s give the length of the infectious period, as follows. An individual i, if already infected at time 0, stays infected for a period of length $\hat{r}_i$, and is then removed. If the individual i is susceptible at time 0, then it becomes infected at time $A_i^K = F_K^{-1}(l_i)$, stays infected for a period of length r_i, and is then removed. Here, l_i can be viewed as individual i's resistance to infection; more precisely, if $Z_K(t)$ denotes the proportion of infectives present in the population at time t, and if $\lambda(t, (x(s))_{s \leq t})$ denotes the current force of infection, taken with considerable generality to be a function acting on a one-dimensional parameter (time) and on right-continuous functions (the proportions of infected individuals at all times in the past), then the accumulated pressure of infection is given by

$$F_K(t) = \int_{(0,t]} \lambda(s, Z_K)\, ds.$$

The time at which individual i becomes infected is then assumed to be given by

$$A_i^K = \inf \left\{ t \in \mathbb{R}_+ : \int_{(0,t]} \lambda(s, Z_K)\, ds = l_i \right\};$$

that is, the first time that the accumulated pressure of infection in the population exceeds the individual's resistance.

This formulation includes the classical case of Bartlett's (1949) GSE, where $\lambda(t, x) = x(t)$, the resistances l_i are assumed to be i. i. d. negative exponentially NE (1) distributed with mean 1, and r_i and $\hat{r}_i$ are assumed to be i. i. d. negative exponential with the same mean $1/\rho$, independent of the l_i. This results in a Markovian model, where standard Markov techniques can be applied. A generalization was studied by Wang (1975,1977), with force of infection $\lambda(t, x) = \lambda(x(t))$, the resistances l_i again being i. i. d.

NE (1)-distributed, and for each i, the random variables l_i and r_i are independent. This still results in some Markovian structure. Also, classically, only the vector of the proportions of susceptibles, infected and removed individuals over time was considered. In contrast, we study the empirical measure

$$\xi_K = \frac{1}{K}\sum_{i=1}^{aK} \delta_{[0,\hat{r}_i)} + \frac{1}{K}\sum_{i=1}^{bK} \delta_{[A_i^K, A_i^K + r_i)}.$$

Note that

$$\xi_K([0,t] \times (t,\infty)) = \frac{1}{K}\sum_{i=1}^{aK} 1_{[0,\hat{r}_i)}(t) + \frac{1}{K}\sum_{i=1}^{bK} 1_{[A_i^K, A_i^K + r_i)}(t)$$

gives the proportion of infected individuals at time t, and we recover the proportions of susceptibles and removed individuals at time t in similar fashion. Moreover,

$$\xi_K([0,s] \times (t,\infty]), \quad t > s,$$

gives the proportion of individuals that were infected before time s, but not removed before time t. This quantity, not covered by the classical approach, is of particular interest if public health policy has changed during the course of the epidemic, for example in reaction to the discovery of the epidemic.

We are interested in the limiting behaviour of the empirical measure as $K \to \infty$; in the context of statistical physics this is often called a mean-field approximation. First we have to make some assumptions. Let D_+ denote the space of all functions $x : \mathbb{R}_+ \to [-1,1]$ that are right-continuous with left-hand limits.

(1) Assume that $\lambda : \mathbb{R}_+ \times D_+ \to \mathbb{R}_+$ is uniformly bounded by a constant γ, and is Lipschitz in $x \in D_+$ with Lipschitz constant α. Assume also that λ non-anticipating, in the sense that $\lambda(t,x)$ depends on the function x only through $(x_s)_{0 \le s \le t}$, and assume also that, for all $t \in \mathbb{R}_+$,

$$\lambda(t,x) = 0 \iff x(t) = 0.$$

(2) Assume that there is a constant $\beta > 0$ such that, for each $x \in \mathbb{R}_+$, the conditional cumulative distribution function $\Psi_x(t) := \mathbb{P}[l_1 \le t \,|\, r_1 = x]$ has a density $\psi_x(t)$ that is uniformly bounded from above by β;

$$\psi_x(t) \le \beta \text{ for all } x \in \mathbb{R}_+ \text{ and all } t \in \mathbb{R}_+.$$

(3) Assume that the l_i have a distribution function Ψ which possesses a density ψ.

(4) Assume that r_i and $\hat{r}_i$ have distribution function Φ such that $\Phi(0) = 0$, so that infected individuals are not immediately removed.

To determine the limiting mean measure for the empirical measure, consider the following heuristic. As $F_K(t) = \int_0^t \lambda(s, Z_K)\, ds$, the expression

$$Z_K(t) = \frac{1}{K} \sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i > t) + \frac{1}{K} \sum_{j=1}^{bK} \mathbf{1}(F_K^{-1}(l_j) \leq t < F_K^{-1}(l_j) + r_j)$$

gives the proportion of infected at time t. For $f \in C(\mathbb{R}_+, \mathbb{R})$, $t \in \mathbb{R}_+$, define the operators

$$\mathcal{Z}f(t) = a(1 - \Phi(t)) + b\mathbb{P}(f(t - r_1) \leq l_1 < f(t))$$

and

$$Lf(t) = \int_{(0,t]} \lambda(s, \mathcal{Z}f)\, ds.$$

Due to the law of large numbers, $Z_K \approx \mathcal{Z}F_K$, and thus

$$F_K \approx LF_K.$$

Thus F_K is close to being a fixed point of L. It turns out that this fixed point exists and is unique on every finite time interval, and hence can be used to describe the limiting mean measure.

We restrict all quantities to a finite time interval $[0, T]$, where $T > 0$ is arbitrary. These restrictions are denoted by a superscript T or a subscript T. The following theorem confirms the heuristics.

Theorem 6.1: *For $T \in \mathbb{R}_+$, the operator L is a contraction on $[0, T]$, and the equation*

$$f(t) = \int_{(0,t]} \lambda(s, \mathcal{Z}f)\, ds, \quad 0 \leq t \leq T, \tag{6.1}$$

has a unique solution G_T.

For $T \in \mathbb{R}_+$, let G_T be the solution of (6.1), and let $\tilde{\mu}^T$ be given by

$$\tilde{\mu}^T([0, r] \times [0, s]) = \mathbb{P}^T[l_1 \leq G_T(r), l_1 \leq G_T(s - r_1)], \qquad r, s \in (0, T].$$

Put

$$\mu^T = a(\delta_0 \times d\Phi)^T + b\tilde{\mu}^T.$$

This gives our limiting mean measure. Indeed the following theorem gives a bound on the mean-field approximation for the empirical measure, using Stein's method.

Theorem 6.2: *For all $H \in \mathcal{F}_0$ of the form (4.3) and for all $T \in \mathbb{R}_+$, we have*

$$|\mathbb{E}H(\mathcal{L}(\xi_n^T) - \mathbb{E}H(\delta_{\mu^T})| \leq \frac{\sqrt{a}+\sqrt{b}}{\sqrt{K}} + \alpha b \beta T(T+2)\exp(b\lceil 2\alpha\beta T\rceil)\left\{\frac{1+b}{\sqrt{K}} + \frac{2}{K}\right\}.$$

Here, $\lceil x \rceil$ is the smallest integer larger than x.

The proof of this theorem is rather involved, and can be found in Reinert (2001), so it is only sketched here. The main arguments used are the Glivenko–Cantelli theorem, to justify the law of large numbers argument, the Banach contraction theorem for Theorem 6.1, and a coupling argument to disentangle the dependence between the infection times. Note that F_K and l_1 are not independent, but if $F_{K,1}$ denotes the same infection time as F_K, except that individual 1 from the susceptible population is left out, then $F_{K,1}$ and l_1 are independent.

Proof: (Sketch of the proof of Theorem 6.2).

We assume that Theorem 6.1 is proved already. We abbreviate

$$\zeta_K = \frac{1}{bK}\sum_{i=1}^{bK} \delta_{(A_i^K, A_i^K + r_i)};$$

this is the part of ξ_K that contains much dependence. We use the notation $\langle \phi, \nu \rangle = \int \phi \, d\nu$. For F of the form (4.2), we bound $\mathbb{E}\mathcal{A}F$, where $\mathcal{A}$ is the operator associated with the Dirac measure at μ^T. Then we have

$$\sum_{j=1}^m \mathbb{E}\left\{ f_{(j)}(\langle {\xi_K}^T, \phi_k\rangle,\, k=1,\ldots,m)\langle \mu^T - {\xi_K}^T, \phi_j\rangle \right\}$$

$$= a\sum_{j=1}^m \mathbb{E}\left\{ f_{(j)}(\langle {\xi_K}^T, \phi_k\rangle,\, k=1,\ldots,m) \left\langle (\delta_0 \times \hat{\mu})^T - \frac{1}{aK}\sum_{i=1}^{aK} \delta^T_{(0,\hat{r}_i)}, \phi_j \right\rangle \right\} \quad (6.2)$$

$$+ b\sum_{j=1}^m \mathbb{E} f_{(j)}(\langle {\xi_K}^T, \phi_k\rangle,\, k=1,\ldots,m)\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle. \quad (6.3)$$

For the first summand (6.2), we employ the Cauchy-Schwarz inequality and

the bounds on the functions involved in the form (4.2), to obtain

$$\begin{aligned}
&\left| a\sum_{j=1}^{m} \mathbb{E} f_{(j)}(\langle {\xi_K}^T, \phi_k\rangle,\ k=1,\dots,m) \left\langle (\delta_0 \times \hat{\mu})^T - \frac{1}{aK}\sum_{i=1}^{aK} \delta^T_{(0,\hat{r}_i)}, \phi_j \right\rangle \right| \\
&\quad \le\ a\sum_{j=1}^{m} \|f_{(j)}\| \mathbb{E}\left| \frac{1}{aK}\sum_{i=1}^{aK} (\phi_j(0,\hat{r}_i) - \mathbb{E}\phi_j(0,\hat{r}_i)) \right| \\
&\quad \le\ a\sum_{j=1}^{m} \|f_{(j)}\| \sqrt{\operatorname{Var}\left(\frac{1}{aK}\sum_{i=1}^{aK} (\phi_j(0,\hat{r}_i) - \mathbb{E}\phi_j(0,\hat{r}_i)) \right)} \\
&\quad \le\ \frac{\sqrt{a}}{\sqrt{K}}.
\end{aligned}$$

Similarly, for the second summand (6.3), we obtain

$$\begin{aligned}
&b\sum_{j=1}^{m} \mathbb{E} f_{(j)}(\langle {\xi_K}^T, \phi_k\rangle,\ k=1,\dots,m)\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle \\
&= b\sum_{j=1}^{m} \mathbb{E} f_{(j)}(\langle a(\delta_0 \times \hat{\mu})^T + b\zeta_K^T, \phi_k\rangle,\ k=1,\dots,m)\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle + R_1,
\end{aligned}$$

where, by Taylor's expansion, $|R_1| \le 2b\sqrt{a/K}$. It thus remains to bound

$$\begin{aligned}
&b\sum_{j=1}^{m} \mathbb{E} f_{(j)}(\langle a(\delta_0 \times \hat{\mu})^T + b\zeta_K^T, \phi_k\rangle,\ k=1,\dots,m)\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle \\
&\le b\sum_{i=1}^{m} \|f_{(j)}\| \mathbb{E}\left| \langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle \right| \\
&= b\sum_{j=1}^{m} \|f_{(j)}\| \mathbb{E}\left| \frac{1}{bK}\sum_{i=1}^{bK} \phi_i((F_K^T)^{-1}(l_i), (F_K^T)^{-1}(l_i) + r_i) \right. \\
&\qquad\qquad \left. - \phi_j(G_T^{-1}(l_i), G_T^{-1}(l_i) + r_i) \right| \\
&\le b\sum_{j=1}^{m} \|f_{(j)}\| \|\phi_j'\| \mathbb{E}\left| (F_K^T)^{-1}(l_1) - (G_T)^{-1}(l_1) \right|.
\end{aligned}$$

To tackle the problem that F_K and l_1 are dependent, we couple the process to the same process with susceptible individual 1 omitted; denote by $F_{K,1}$ the accumulated pressure of infection in this new process. Then $F_K^{-1}(l_1) = F_{K,1}^{-1}(l_1)$ and

$$\mathbb{E}\left| (F_K^T)^{-1}(l_1) - G_T^{-1}(l_1) \right| = \mathbb{E}\left| (F_{K,1}^T)^{-1}(l_1) - G_T^{-1}(l_1) \right|.$$

In order to describe the new process, for $h \in D([0,T])$, the space of right-continuous functions $[0,T] \to [-1,1]$ with left-hand limits, we define the operators

$$\mathcal{Z}_{K,1}h(t) = \frac{1}{K}\sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i > t) + \frac{1}{K}\sum_{j=2}^{bK} \mathbf{1}(h(t-r_j) < l_j \leq h(t))$$

and

$$L_{K,1}h(t) = \int_{(0,t]} \lambda(s, \mathcal{Z}_{K,1}h)\, ds.$$

Note that $F_K^{-1}(l_1) = F_{K,1}^{-1}(l_1)$ by construction, and, for all $t \leq T$, that

$$\begin{aligned} \|F_{K,1} - G_T\|_t &= \|L_{k,1}F_{K,1} - LG_T\|_t \\ &\leq \sup_h \|L_{k,1}h - Lh\|_t + \|LF_{K,1} - LG_T\|_t. \end{aligned}$$

For each $h \in D([0,T])$, we have

$$\begin{aligned} \|L_{K,1}h - Lh\|_T &\leq \alpha \int_0^T \sup_{s \leq x} |\mathcal{Z}_{K,1}h(s) - \mathcal{Z}h(s)|\, ds \\ &\leq \alpha T \left(aR_1 + 2bR_2 + \frac{2}{K} \right), \end{aligned}$$

where

$$R_1 := \sup_s \left| \frac{1}{aK} \sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i \leq s) - \Phi(s) \right|$$

and

$$R_2 := \sup_s \left| \frac{1}{bK-1} \sum_{i=2}^{bK} \mathbf{1}(l_i \leq s) - \Psi(s) \right|.$$

Results from Massart (1990) enable us to derive the bounds $\mathbb{E}R_1 \leq \frac{1}{\sqrt{aK}}$ and $\mathbb{E}R_2 \leq \frac{1}{\sqrt{bK}}$ for these remainder terms. Thus

$$\mathbb{E} \sup_h \|L_{K,1}h - Lh\|_T \leq \alpha T \left\{ (1+b)\frac{1}{\sqrt{K}} + \frac{2}{K} \right\} =: S(K).$$

To bound $\mathbb{E}\|LF_{K,1} - LG_T\|_t$ we use again a contraction argument. We have

$$|LF_{K,1}(t) - LG_T(t)| \leq \alpha b \beta \int_0^t \|F_{K,1} - G_T\|_x (1 + \Phi(x))\, dx$$

and hence

$$\mathbb{E}\|LF_{K,1} - LG_T\|_t \leq S(K) + \alpha b\beta \int_0^t \|F_{K,1} - G_T\|_x (1 + \Phi(x))\, dx.$$

Fix some $c \geq b$ and put $\eta = 1/(2c\alpha\beta)$; then

$$\mathbb{E}\|LF_{K,1} - LG_T\|_\eta \leq \frac{c}{c-b} S(K).$$

By induction we can show that

$$\mathbb{E}\|LF_{K,1} - LG_T\|_{k\eta} \leq \left(\frac{c}{c-b}\right)^k S(K).$$

As we consider the process only restricted to $[0, T]$, the largest k we have to take into account is $k = \lceil \frac{T}{\eta} \rceil$, yielding

$$\mathbb{E}\|LF_{K,1} - LG_T\|_{k\eta} \leq \exp(\lceil 2c\alpha\beta T \rceil)(\log c - \log(c-b)) S(K).$$

Letting $c \to \infty$ gives the assertion. ∎

Some concluding remarks:

(1) The bound derived is the first known bound on the distance, and furthermore it is explicit. Unfortunately, we do not obtain an almost sure result, and we assume smoothness for our test functions. Also, for parameter estimation, it turns out that the bounds are not very useful; instead, a Gaussian approximation would be needed.

(2) The model used here is more realistic than the Markovian model, and the results are more informative, from using the empirical measure. To make the model even more realistic, we could assume for the initially infected that the $\hat{r}_i$ are not identically distributed; this does not entail much further complication.

(3) The factor $\frac{1}{\sqrt{K}}$ in the bounds seems to be optimal. This suggests the validity of a Gaussian approximation. For the vector of susceptibles, infected and removed in the Markovian setting, such an approximation was derived by Barbour (1974).

(4) In Barbour (1975), it is shown that the waiting time until the epidemic dies out is, very roughly, of order $\log K$, in the Markovian model. This indicates that a deterministic approximation may not be good for the whole time course of the epidemic. Much of the fluctuation is due to the initial variation in the epidemic process, which is quite similar to that of a branching process when only few infected and many susceptible individuals are present. When restricting to a time interval in which the proportion of infectives present is always substantial, the bound on

the approximation is much improved, growing only linearly in time; see Reinert (2001).

(5) It would be interesting to investigate how the model would behave in a spatial setting, where we do not assume homogeneous mixing.

7. The density approach

This approach was first suggested by Stein (2004); it provides an alternative to the generator method, in particular when a generator is not easily available. Suppose that we are in the following situation. Let p be a strictly positive density on the whole real line, having a derivative p' in the sense that, for all x,

$$p(x) = \int_{-\infty}^{x} p'(y)dy = -\int_{x}^{\infty} p'(y)dy,$$

and assume that

$$\int_{-\infty}^{\infty} |p'(y)|dy < \infty.$$

Let

$$\psi(x) = \frac{p'(x)}{p(x)}.$$

Proposition 7.1: *In order that a random variable Z be distributed according to the density p, it is necessary and sufficient that, for all functions f that have a derivative f' and such that*

$$\int_{-\infty}^{\infty} |f'(z)|\, p(z)dz < \infty,$$

we have

$$\mathbb{E}(f'(Z) + \psi(Z)f(Z)) = 0.$$

Example 7.2: The standard normal distribution $\mathcal{N}(0,1)$. Here, we have $\psi(x) = -x$, and the above conditions are satisfied. The characterization results in the classical Stein equation.

Example 7.3: The Gamma distribution $\mathrm{Ga}(a,\lambda)$, with density given on $x > 0$ by $p_{\lambda,a}(x) = \lambda^a e^{-\lambda x} x^{a-1}/\Gamma(a)$. Although the density is not positive on the whole real line, Proposition 7.1 gives an indication of how to obtain

a Stein characterization. Here, $\psi(x) = x^{-1}\{a - 1 - \lambda x\}$, for $x > 0$. This would yield a characterization of type

$$\mathbb{E}\left\{f'(X) + X^{-1}\{a - 1 - \lambda X\}f(X)\right\} = 0.$$

Comparing this with the characterization (3.2) we see that the two characterizations are equivalent, by putting $g(x) = xf(x)$.

For convenience, write $\phi(x) = -\psi(x)$. Then the following Stein theorem is derived in Stein (2004).

Theorem 7.4: *Suppose that Z has probability density function p satisfying the assumptions of the above proposition. Let (W, W') be an exchangeable pair such that $\mathbb{E}(\phi(W))^2 = \sigma^2 < \infty$, and let*

$$\lambda = \frac{\mathbb{E}(\phi(W') - \phi(W))^2}{2\sigma^2}.$$

Then, for all piecewise continuous functions $h : \mathbb{R} \to \mathbb{R}$ that satisfy $\mathbb{E}|h(Z)| < \infty$, we have

$$\begin{aligned}
&\mathbb{E}h(W) - \mathbb{E}h(Z) \\
&= \mathbb{E}(Vh)(W) - \frac{1}{\lambda\sigma^2}\mathbb{E}\left\{(\phi(W') - \phi(W))((Uh)(W') - (Uh)(W))\right\} \\
&\quad - \mathbb{E}\left\{\mathbb{E}^W\left(\frac{\phi(W') - (1-\lambda)\phi(W)}{\lambda}\right)(Uh)(W)\right\},
\end{aligned}$$

where Uh and Vh are defined by

$$(Uh)(w) = \frac{\int_{-\infty}^{w}\left(h(x) - \int_{-\infty}^{\infty} h(y)p(y)dy\right)p(x)\,dx}{p(w)}$$

and

$$(Vh)(w) = (Uh)'(w);$$

we use $\mathbb{E}^W$ to denote expectation conditional on W.

Theorem 7.4 can be employed to assess the error in a normal approximation via simulations, replacing the expectations by sample means.

8. Distributional transformations

This is joint work with Larry Goldstein and can be found in Goldstein & Reinert (2005). We have already seen, in connection with Gibbs measures, how the size-biased distribution can be used to characterize a target distribution. The work presented here takes this idea further.

First, the size-biased distribution is defined only for non-negative random variables. For random variables with zero mean, we instead define the zero bias distributional transformation as follows (see Goldstein & Reinert 1997).

Definition 8.1: Let X be a mean zero random variable with finite, nonzero variance σ^2. We say that X^* has the X-zero biased distribution if for all differentiable f for which $\mathbb{E}Xf(X)$ exists,

$$\mathbb{E}Xf(X) = \sigma^2 \mathbb{E}f'(X^*).$$

The zero bias distribution X^* exists for all X that have mean zero and finite variance. In Goldstein & Reinert (1997), this coupling is used to obtain bounds of order $1/n$ for smooth test functions, under symmetry assumptions on the underlying distribution. The following theorem shows that it is indeed possible to define much more general functional biasing.

Theorem 8.2: *Let X be a random variable, $m \in \{0, 1, 2, \ldots\}$ and let P be a function on the support of X such that P has exactly m sign changes, is positive on its rightmost interval and is such that*

$$\frac{1}{m!}\mathbb{E}X^j P(X) = \alpha \delta_{j,m} \quad j = 0, \ldots, m,$$

for some $\alpha > 0$. Then there exists a unique distribution for a random variable $X^{(m)}$ such that

$$\mathbb{E}P(X)G(X) = \alpha \mathbb{E}G^{(m)}(X^{(m)})$$

for all m times differentiable functions G for which $\mathbb{E}P(X)G(X)$ exists.

The $X^{(m)}$ distribution is named the X–P biased distribution.

For example, with $P(x) = x$, we obtain that for any random variable X such that $\mathbb{E}X = 0$ and $\sigma^2 = \mathbb{E}X^2 < \infty$, there exists a random variable $X^{(1)}$ such that, for all smooth G, we have $\mathbb{E}XG(X) = \sigma^2 \mathbb{E}G'(X^{(1)})$. Thus we recover the zero bias distribution.

Theorem 8.2 allows us to define much more general distributional transformations. To illustrate this, consider infinitely divisible random variables $\{Z_\lambda\}_{\lambda>0}$ with moments of all orders. Assume that there is a collection $\{P_m^\lambda\}_{m\geq 1}$ of monic polynomials, where P_m^λ has m distinct roots, is positive on its rightmost interval, and the collection is orthogonal with respect to the law of Z_λ. Define

$$\alpha_m^\lambda = \frac{1}{m!}\mathbb{E}Z_\lambda^m P_m^\lambda(Z_\lambda)$$

and the set

$$\mathcal{M}_\lambda^m = \{X : \mathbb{E}X^{2m} < \infty, \quad \mathbb{E}X^j = \mathbb{E}Z_\lambda^j, \quad 0 \le j \le 2m\}.$$

For every $X \in \mathcal{M}_\lambda^m$, for $j = 0, \ldots, m$, we then have

$$\frac{1}{m!}\mathbb{E}X^j P_m^\lambda(X) = \frac{1}{m!}\mathbb{E}Z_\lambda^j P_m^\lambda(Z_\lambda) = \alpha_m^\lambda \delta_{j,k}.$$

Theorem 8.2 yields that for all $X \in \mathcal{M}_\lambda^m$ there exists a random variable X_m^λ such that

$$\mathbb{E}P_m^\lambda(X)G(X) = \alpha_m^\lambda \mathbb{E}G^{(m)}(X_m^\lambda).$$

As with the size-biased distributon, there is a construction for sums of independent random variables available. Let $m \in \{0, 1, \ldots\}$. Let $X_1, \cdots, X_n$ be independent random variables with

$$X_i \in \mathcal{M}_{\lambda_i}^m$$

for some $\lambda_1, \ldots, \lambda_n$, and

$$W = \sum_{i=1}^n X_i.$$

For the transformation we shall bias different summands to different degrees; hence we introduce the multi-index $\mathbf{m} = (m_1, \ldots, m_n)$, and let

$$m = |\mathbf{m}| = \sum_{i=1}^n m_i.$$

Furthermore, for $\Lambda = (\lambda_1, \ldots, \lambda_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$, let

$$\alpha_\Lambda^{(\mathbf{m})} = \prod_{i=1}^n \alpha_{\lambda_i}^{(m_i)} \quad \text{and} \quad P_\Lambda^{\mathbf{m}}(\mathbf{x}) = \prod_{i=1}^n P_{\lambda_i}^{m_i}(x_i).$$

Assume that $\{P_\lambda^m(x)\}_{m \ge 0}$ satisfies the conditions of Theorem 8.2, and let $\lambda = \lambda_1 + \cdots + \lambda_n$. In this setting, Goldstein & Reinert (2005) derive the following result.

Theorem 8.3: *Suppose that for some weights $c_{\mathbf{m}}$, the family of orthogonal polynomials $\{P_\lambda^m(x)\}_{m \ge 0}$ satisfies the identity*

$$P_\lambda^m(w) = \sum_{\mathbf{m}} c_{\mathbf{m}} P_\Lambda^{\mathbf{m}}(\mathbf{x}),$$

with $w = x_1 + \cdots + x_n$, where $\sum_{\mathbf{m}}$ denotes the sum over all $\mathbf{m}$ such that $|\mathbf{m}| = m$. Then

$$\alpha_\lambda^{(m)} = \sum_{\mathbf{m}} c_{\mathbf{m}} \alpha_\Lambda^{(\mathbf{m})},$$

and we may consider the variable $\mathbf{I}$*, independent of all other variables, with distribution*

$$\mathbb{P}(\mathbf{I} = \mathbf{m}) = \frac{c_{\mathbf{m}} \alpha_{\Lambda}^{(\mathbf{m})}}{\alpha_{\lambda}^{(m)}}. \tag{8.1}$$

Furthermore, the variable

$$W_{\lambda}^{(m)} = \sum_{i=1}^{n} (X_i)_{\lambda_i}^{(I_i)}$$

has the W*–*P_{λ}^{m} *distribution.*

Example 8.4: (Hermite biasing). For $\sigma^2 = \lambda > 0$, define the collection of Hermite polynomials $\{H_n^{\lambda}\}_{n \geq 0}$ through the generating function

$$e^{xt - \frac{1}{2}\lambda t^2} = \sum_{n=0}^{\infty} H_n^{\lambda}(x) \frac{t^n}{n!},$$

In this case $\alpha_m^{\lambda} = \lambda^m$, and the index I in (8.1) is chosen according to the multinomial distribution Mult (m, Λ). Denoting the Hermite polynomials for $\lambda = 1$ by $H_m^1 = H_m$, we obtain as Stein-type equations for the standard normal distribution

$$h(x) - \mathcal{N}h = \phi'(x) H_{m-1}(x) - H_m(x)\phi(x)$$

and

$$h(x) - \mathcal{N}h = \phi^{(m)}(x) - H_m(x)\phi(x).$$

The standard normal distribution is the unique fixed point of the Hermite-bias transformation of any order, hence this gives an infinite number of Stein characterisations for the standard normal distribution.

Example 8.5: (Charlier biasing). The Charlier polynomials correspond to the Poisson distribution with mean λ; here again we obtain $\alpha_m^{\lambda} = \lambda^m$, and $I \sim$ Mult (m, Λ). As the Poisson distribution can be shown to be the fixed point of the Charlier-bias transformation for any order, again we obtain an infinite number of characterizations for the Poisson distribution.

Example 8.6: (Laguerre biasing). The monic Laguerre polynomials are orthogonal for the Gamma distribution. However, the Gamma distribution with fixed parameter is not a fixed point of the Laguerre-bias transformation.

Connections between distributions and orthogonal polynomials in the context of Stein's method were also studied by Diaconis & Zabell (1991); there one can also find more examples of orthogonal polynomials.

Note that there are many other applications of Stein's method to other distributions; see for example the work on Markov chain Monte Carlo methods in Diaconis (2004), and on the uniform distribution in Diaconis (1989). The above tutorial lectures are merely an introduction to a very rich field with many open problems.

Acknowledgement: The author would like to thank the organizers of this excellent workshop.

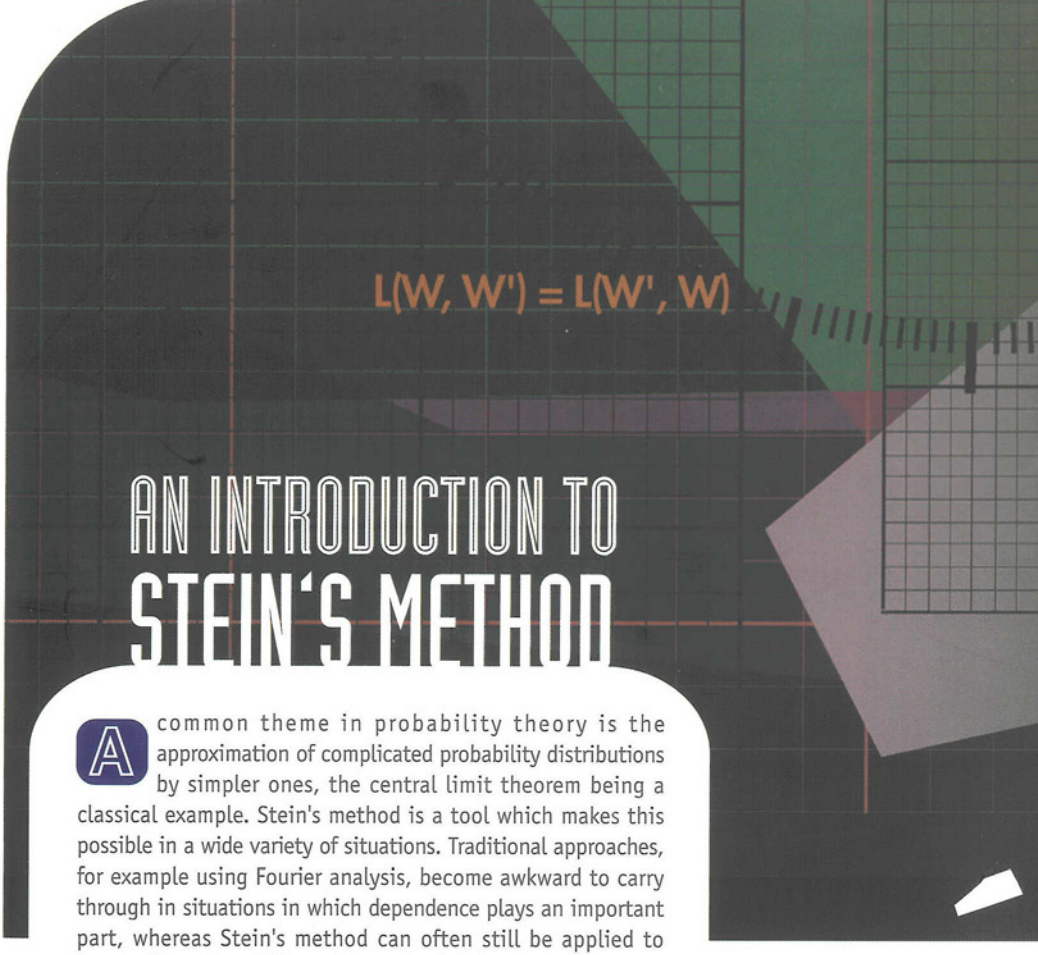
References

1. R. ARRATIA, L. GOLDSTEIN & L. GORDON (1989) Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* 17, 9–25.
2. N. T. J. BAILEY (1975) *The mathematical theory of infectious diseases and its applications.* (2nd ed.) Griffin, London.
3. A. D. BARBOUR (1974) On a functional central limit theorem for Markov population processes. *Adv. Appl. Probab.* 6, 21–39.
4. A. D. BARBOUR (1975) The duration of the closed stochastic epidemic. *Biometrika* 62, 477–482.
5. A. D. BARBOUR (1988) Stein's method and Poisson process convergence. *J. Appl. Probab.* 25A, 175–184.
6. A. D. BARBOUR (1990) Stein's method for diffusion approximations. *Probability Theory and Related Fields* 84, 297–322.
7. A. D. BARBOUR, L. HOLST & S. JANSON (1992) *Poisson Approximation.* Oxford Science Publications.
8. M. S. BARTLETT (1949) Some evolutionary stochastic processes. *J. Roy. Statist. Soc., Ser. B* 11, 211–229.
9. T. C. BROWN & A. XIA (2001) Stein's method and birth-death processes. *Ann. Probab.* 29, 1373–1403.
10. L. H. Y. CHEN (1975) Poisson approximation for dependent trials. *Ann. Probab.* 3, 534–545.
11. P. DIACONIS (1989) An example for Stein's method. Stanford Stat. Dept. Technical Report.
12. P. DIACONIS (2004) Stein's method for Markov chains: first examples. In: *Stein's Method: Expository Lectures and Applications*, Eds: P. Diaconis & S. Holmes, pp. 27–43. IMS Lecture Notes **46**, Beachwood, Ohio.
13. P. DIACONIS & S. ZABELL (1991) Closed form summation for classical distributions: variations on a theme of de Moivre. *Statistical Science* 6, 284-302.
14. P. EICHELSBACHER & G. REINERT (2004a) Stein's method for discrete Gibbs measures. (In revision).

15. P. EICHELSBACHER & G. REINERT (2004b) Stein's method for spatial Gibbs measures. Preprint.
16. W. EHM (1991) Binomial approximation to the Poisson binomial distribution. *Statist. Probab. Lett.* 11, 7–16.
17. S. N. ETHIER & T. G. KURTZ (1986) *Markov Processes, Characterization and Convergence.* Wiley, New York.
18. L. GOLDSTEIN & G. REINERT (1997) Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* 7, 935–952.
19. L. GOLDSTEIN & G. REINERT (2005) Distributional transformations, orthogonal polynomials, and Stein characterizations. *J. Theor. Probab.* (to appear).
20. L. GOLDSTEIN & Y. RINOTT (1996) On multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Prob.* 33, 1–17.
21. F. GÖTZE (1991) On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19, 724–739.
22. S. HOLMES (2004) Stein's method for birth and death chains. In: *Stein's Method: Expository Lectures and Applications*, Eds: P. Diaconis & S. Holmes, pp. 45–67. IMS Lecture Notes **46**, Beachwood, Ohio.
23. O. KALLENBERG (2002) *Foundations of Modern Probability,* 2nd edn. Springer, New York.
24. H.-R. KÜNSCH (1998) Personal communication.
25. H. M. LUK (1994) Stein's method for the gamma distribution and related statistical applications. Ph.D. thesis, University of Southern California, Los Angeles, USA.
26. P. MASSART (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* 18, 1269–1283.
27. E. PEKÖZ (1996) Stein's method for geometric approximation. *J. Appl. Probab.* 33, 707–713.
28. A. PICKETT (2002) *Stein's method for chisquare approximations.* Transfer thesis, University of Oxford.
29. A. PICKETT & G. REINERT (2004) Stein's method for chi-squared approximations. Preprint.
30. G. REINERT (1994) A weak law of large numbers for empirical measures via Stein's method. *Ann. Probab.* 23, 334–354.
31. G. REINERT (1995) The asymptotic evolution of the General Stochastic Epidemic. *Ann. Appl. Probab.* 5, 1061–1086.
32. G. REINERT (2001) Stein's method in application for epidemic processes. In *Complex Stochastic Systems,* Eds: O. E. Barndorff-Nielsen, D. R. Cox & C. Klüppelberg, pp. 235–275. Chapman and Hall, Boca Raton.
33. G. REINERT & W. SCHOUTENS (1998) Stein's method for the hypergeometric distribution. Preprint.
34. Y. RINOTT & V. ROTAR (1996) A multivariate CLT for local dependence with $n^{-1/2} \log n$ rate and applications to multivariate graph related statistics. *J. Multivariate Analysis* 56, 333–350.
35. W. SCHOUTENS (2000) *Stochastic Processes and Orthogonal Polynomials.* Springer, New York.

36. T. SELLKE (1983) On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* 20, 390–394.
37. C. STEIN (1986) *Approximate Computation of Expectations.* IMS, Hayward, California.
38. C. STEIN, WITH P. DIACONIS, S. HOLMES & G. REINERT (2004) Use of exchangeable pairs in the analysis of simulations. In: *Stein's Method: Expository Lectures and Applications*, Eds: P. Diaconis & S. Holmes, pp. 1–26. IMS Lecture Notes **46**, Beachwood, Ohio.
39. A. W. VAN DER VAART & J. A. WELLNER (1996) *Weak Convergence and Empirical Processes.* Springer, New York.
40. F. S. J. WANG (1975) Limit theorems for age and density dependent stochastic population models. *J. Math. Bio.* 2, 373–400.
41. F. S. J. WANG (1977) A central limit theorem for age-and-density-dependent population processes. *Stoch. Proc. Appl.* 5, 173–193.
42. G. V. WEINBERG (2000) Stein factor bounds for random variables. *J. Appl. Probab.* 37, 1181-1187.

INDEX



AN INTRODUCTION TO STEIN'S METHOD

A common theme in probability theory is the approximation of complicated probability distributions by simpler ones, the central limit theorem being a classical example. Stein's method is a tool which makes this possible in a wide variety of situations. Traditional approaches, for example using Fourier analysis, become awkward to carry through in situations in which dependence plays an important part, whereas Stein's method can often still be applied to great effect. In addition, the method delivers estimates for the error in the approximation, and not just a proof of convergence. Nor is there in principle any restriction on the distribution to be approximated; it can equally well be normal, or Poisson, or that of the whole path of a random process, though the techniques have so far been worked out in much more detail for the classical approximation theorems.

This volume of lecture notes provides a detailed introduction to the theory and application of Stein's method, in a form suitable for graduate students who want to acquaint themselves with the method. It includes chapters treating normal, Poisson and compound Poisson approximation, approximation by Poisson processes, and approximation by an arbitrary distribution, written by experts in the different fields. The lectures take the reader from the very basics of Stein's method to the limits of current knowledge.

5792 hc
ISBN 981-256-280-X

SINGAPORE UNIVERSITY PRESS
World Scientific