Stein's Method for Chisquare Approximations, Weak Law of Large Numbers, and Discrete Distributions from a Gibbs View Point

Gesine Reinert
Department of Statistics
University of Oxford

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 [18]), 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 $f : \mathbf{R} \to \mathbf{R}$ with $N|f'| < \infty$, we have

$$Ef'(X) - EXf(X) = 0.$$

   See [39] and [11] for a thorough treatment.

2. The Poisson distribution with parameter λ, Poisson(λ) with the corresponding Stein characterization that $X \sim Poisson(\lambda)$ if and only if for all real-valued functions f for which both sides of the equation exist we have that

$$\lambda Ef(X+1) - EXf(X) = 0.$$

   For a detailed treatment, see for example [10], [1], [7], and [18].

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, $E\mathcal{A}f(X) = 0$ holds.

2. For each smooth function h find a solution $f = f_h$ of the *Stein equation*

$$h(x) - \int h d\mu = \mathcal{A}f(x). \tag{1}$$

3. Then for any variable W, it holds that

$$Eh(W) - \int h d\mu = E\mathcal{A}f(W),$$

   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, in case of discrete distributions, $\Delta f = \sup_x |f(x+1) - f(x)|$. In the following we shall always assume that the test function h is smooth; for nonsmooth functions, the reader is referred to the techniques used by [11], by [36], and by [23].

In Section 1, the generator approach is briefly described. The flexibility of this approach is illustrated by the applications to Chisquare approximations in Section 2, to laws of large numbers for empirical measures in

Section 3, and to Gibbs measures in Section 4. In Section 5 we give a more involved example, the mean-field behaviour of the general stochastic epidemic. Section 6 explains the density approach, and in Section 7 the approach via coupling equations, viewed as distributional transformations, is given. The powerful approach of exchangeable pairs is described in detail in [39]; for space reasons 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.

1 The generator approach

The generator approach is developed by Barbour in [5], [6], and also by Götze [23]. The idea is to choose the operator $\mathcal{A}$ as a generator of a Markov process with stationary distribution μ. That is, for a homogeneous Markov process $(X_t)_{t\geq 0}$, put $T_t f(x) = E(f(X_t)|X(0) = x)$. 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 yield (see [19])

1. If μ is the stationary distribution of the Markov process then $X \sim \mu$ if and only if $E\mathcal{A}f(X) = 0$ 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.

Examples

1. The operator

$$\mathcal{A}h(x) = h''(x) - xh'(x) \tag{2}$$

   is the generator of *Ornstein-Uhlenbeck* process, with stationary distribution $\mathcal{N}(0,1)$. Putting $f = h'$ gives the classical Stein characterization for $\mathcal{N}(0,1)$.

2. The operator $\mathcal{A}h(x) = \lambda(h(x+1) - h(x)) + x(h(x-1) - h(x))$ or, for $f(x) = h(x+1) - h(x)$,

$$\mathcal{A}f(x) = \lambda f(x+1) - xf(x), \tag{3}$$

   is the generator of an immigration-death process with immigration rate λ and unit per capita death rate. Its stationary distribution is Poisson(λ) (see [18], [7]). Again it yields the classical Stein characterization of the Poisson distribution.

A main advantage of the generator approach is that it easily generalizes to to multivariate distributions and distributions on more complex spaces, such as distributions of path-valued random elements, or distributions 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 [18]. Also note that there is not a unique choice of generator for any given target distribution - just as there is not a unique choice of Stein equation for a given target distribution.

In some instances there is a useful heuristic to find 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 $X_1, \ldots, X_n$ i.i.d.; furthermore assume that $EX_i = 0$ and $EX_i^2 = 1$ for all i. Using ideas from 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 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 following expression, where $\mathbf{x} = (x_1, \ldots\ x_n)$, and f is a smooth function,

$$\mathcal{A}_n f(\Phi_n(\mathbf{x})) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} E f(\Phi_n(x_1, \ldots, x_{i-1}, X_i^*, x_{i+1}, \ldots\ x_n)) - f(\Phi_n(\mathbf{x})).$$

Taylor expansion yields

$$\begin{aligned}
\mathcal{A}_n f(\Phi_n(\mathbf{x})) \quad &\approx \quad \frac{1}{n} \sum_{i=1}^{n} E(X_i^* - x_i) f'(\Phi_n(\mathbf{x})) \frac{\partial}{\partial x_i} \Phi_n(\mathbf{x}) \\
&\quad + \frac{1}{2n} \sum_{i=1}^{n} E(X_i^* - x_i)^2 \left\{ f''(\Phi_n(\mathbf{x})) \left(\frac{\partial}{\partial x_i} \Phi_n(\mathbf{x}) \right)^2 + f'(\Phi_n(\mathbf{x})) \frac{\partial^2}{\partial x_i^2} \Phi_n(\mathbf{x}) \right\} \\
&= \quad -\frac{1}{n} \sum_{i=1}^{n} x_i f'(\Phi_n(\mathbf{x})) \frac{\partial}{\partial x_i} \Phi_n(\mathbf{x}) + \frac{1}{2n} \sum_{i=1}^{n} (1 + x_i^2) \left\{ f''(\Phi_n(\mathbf{x})) \left(\frac{\partial}{\partial x_i} \Phi_n(\mathbf{x}) \right)^2 \right. \\
&\quad \left. + f'(\Phi_n(\mathbf{x})) \frac{\partial^2}{\partial x_i^2} \Phi_n(\mathbf{x}) \right\}.
\end{aligned}$$

Letting $n \to \infty$, with a suitable scaling, would then give a generator for the target distribution.

Example: Suppose 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

$$\begin{aligned}
\mathcal{A}_n f(\Phi_n(\mathbf{x})) \quad &\approx \quad -\frac{1}{n^{3/2}} \sum_{i=1}^{n} x_i f'(\Phi_n(\mathbf{x})) + \frac{1}{2n^2} \sum_{i=1}^{n} (1 + x_i^2) f''(\Phi_n(\mathbf{x})) \\
&= \quad -\frac{1}{n} \Phi_n(\mathbf{x}) f'(\Phi_n(\mathbf{x})) + \frac{1}{2n} f''(\Phi_n(\mathbf{x})) \left(1 + \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) \\
&\approx \quad \frac{1}{n} \left(f''(\Phi_n(\mathbf{x})) - \Phi_n(\mathbf{x}) f'(\Phi_n(\mathbf{x})) \right),
\end{aligned}$$

where we 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).

2 Chisquare distributions

Following the heuristic, we first find a generator for the Chisquare distribution with p degrees of freedom. Let $\mathbf{X}_1, \ldots, \mathbf{X}_p$ be i.i.d. random vectors, where $\mathbf{X}_i = (X_{i,1}, \ldots X_{i,n})$; the components $X_{i,j}$ are again assumed to be i.i.d., mean zero, with $EX_{i,j}^2 = 1$, and finite fourth moment. We put

$$\Phi_n(\mathbf{x}) = \sum_{i=1}^{p} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} x_{i,j} \right)^2 .$$

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 $\frac{\partial^2}{\partial x^2_{i,j}} \Phi_n(\mathbf{x}) = \frac{2}{n}$, 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 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 as generator for χ_p^2

$$\mathcal{A}f(x) = x f''(x) + \frac{1}{2}(p - x) f'(x);$$

this is just a rescaling.

Luk [27] chooses as Stein operator for $Gamma(r, \lambda)$ the operator

$$\mathcal{A}f(x) = x f''(x) + (r - \lambda x) f'(x). \tag{4}$$

As $\chi_p^2 = Gamma(d/2, 1/2)$, this agrees with our generator. [27] 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. [27] also found the transition semigroup, which can be used to solve the Stein equation

$$h(x) - \chi_p^2 h = x f''(x) + \frac{1}{2}(p - x) f'(x), \tag{5}$$

where $\chi^2_p h$ is the expectation of h under the χ^2_p-distribution. This bound has recently been improved as follows.

Lemma 1 *[30] Suppose $h : \mathbf{R} \to \mathbf{R}$ is absolutely bounded, $|h(x)| \leq ce^{ax}$ for some $c > 0$ $a \in \mathbf{R}$, and the first k derivatives of h are bounded. Then the equation (5) has a solution $f = f_h$ such that*

$$\| f^{(j)} \| \leq \frac{\sqrt{2\pi}}{\sqrt{p}} \| h^{(j-1)} \|$$

with $h^{(0)} = h$.

(Note the improvement over [27] in gaining the factor $\frac{1}{\sqrt{p}}$). To put this approach into use, we consider a basic example.

Example: squared sum of i.i.d. random variables ([31])
Let $X_i, i = 1, \ldots, n$, be i.i.d. mean zero, variance one, with exisiting 8^{th} moment. Put

$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

and

$$W = S^2.$$

We would like to bound

$$2EWf''(W) + E(1 - W)f'(W).$$

Put

$$g(s) = sf'(s^2),$$

then

$$g'(s) = f'(s^2) + 2s^2 f''(s^2)$$

and

$$\begin{aligned} 2EWf''(W) + E(1 - W)f'(W) &= Eg'(S) - Ef'(W) + E(1 - W)f'(W) \\ &= Eg'(S) - ESg(S) \end{aligned}$$

Now proceed as for standard normal approximation: Put

$$S_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j.$$

Then

$$\begin{aligned} ESg(S) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} EX_i g(S) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} EX_i g(S_i) + \frac{1}{n} \sum_{i=1}^{n} EX_i^2 g'(S_i) + R_1, \end{aligned}$$

where

$$R_1 = \frac{1}{n^{3/2}} \sum_i EX_i^3 g''(S_i) + \frac{1}{2n^2} \sum_i EX_i^4 g^{(3)} \left(S_i + \theta \frac{X_i}{\sqrt{n}} \right),$$

by Taylor expansion, for some $0 < \theta < 1$. From independence it follows that

$$ESg(S) = \frac{1}{n}\sum_{i=1}^{n} Eg'(S_i) + R_1 = Eg'(S) + R_1 + R_2,$$

where

$$\begin{aligned} R_2 &= \frac{1}{n^{3/2}}\sum_i EX_i g''(S_i) + \frac{1}{2n^2}\sum_i EX_i^2 g^{(3)}\left(S_i + \theta\frac{X_i}{\sqrt{n}}\right) \\ &= \frac{1}{2n^2}\sum_i EX_i^2 g^{(3)}\left(S_i + \theta\frac{X_i}{\sqrt{n}}\right), \end{aligned}$$

again by Taylor expansion, for some $0 < \theta < 1$ (possibly different to the one in R_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 $\beta_i = EX_1^i$ we hence obtain

$$\begin{aligned} &\frac{1}{2n^2}\sum_i EX_i^2 \left|g^{(3)}(S_i + \theta\frac{X_i}{\sqrt{n}})\right| \\ &\leq \frac{24}{n} \parallel f^{(3)} \parallel \left(1 + \frac{\beta_4}{n}\right) + \frac{6}{n} \parallel f'' \parallel + \frac{8}{n} \parallel f^{(4)} \parallel \left(6 + \frac{\beta_4}{n} + 4\frac{\beta_3^2}{\sqrt{n}} + 6\frac{\beta_4}{n} + \frac{\beta_6}{n^2}\right) \\ &= c(f)\frac{1}{n}, \end{aligned}$$

where c is a constant depending on f, and on the distribution of the X_i, but it does not depend on n; it can be calculated explicitly. Similarly we can obtain a bound of order $1/n$ for $\frac{1}{2n^2}\sum_i EX_i^4 \left|g^{(3)}(S_i + \theta\frac{X_i}{\sqrt{n}})\right|$, where we now employ β_8. The details can be found in [31].

For the expression $\frac{1}{n^{3/2}}\sum_i EX_i^3 g''(S_i)$ we use an antisymmetric argument. We have, for some constant $c(f)$,

$$\frac{1}{n^{3/2}}\sum_i EX_i^3 g''(S_i) = \frac{1}{\sqrt{n}}\beta_3 Eg''(S) + c(f)\frac{1}{n}$$

and

$$Eg''(S) = 6ESf''(S^2) + 4ES^3 f^{(3)}(S^2).$$

Here we again used Taylor expansion. Note that g'' is antisymmetric, $g''(-s) = -g''(s)$, so for $Z \sim \mathcal{N}(0,1)$ we have $Eg''(Z) = 0$. Thus $|Eg''(S) - Eg''(Z)| = |Eg''(S)|$, and it is (almost) routine now to apply Stein's method for normal approximation to show that $|Eg''(S)| \leq c(f)/\sqrt{n}$ 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 Chisquare(1) for smooth test functions is of order $\frac{1}{n}$. This result can be extended to χ_p^2-approximations; see [31].

3 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 [32]. A Stein equation for point mass at 0 is hence given by

$$h(x) - h(0) = -xf'(x). \tag{6}$$

Details of the treatment of the weak law of large numbers as presented here can be found in [32].

Lemma 2 *[32] If $h \in C_b^2(\mathbf{R})$, then the Stein equation (6) has solution $f = f_h \in C_b^2$ such that*

$$\| f' \| \leq \| h' \|, \text{ and } \| f'' \| \leq \| h'' \|.$$

Proof. The proof illustrates briefly how to bound solutions of the Stein equation using the generator method. First note that we may assume $h(0) = 0$. 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: Weak law of large numbers

Suppose that $X_1, \ldots, X_n$ are mean zero, finite variance, and put

$$W = W_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then, by Taylor expansion, for some $0 < \theta < 1$, we have that

$$E\mathcal{A}f(W) = -EWf'(W) = -EWf'(0) + EW^2 f''(\theta W) = EW^2 f''(\theta W),$$

and

$$|E\mathcal{A}f(W)| \leq \| f'' \| Var(W).$$

In particular, if $Var(W_n) \to 0$ as $n \to \infty$ then the weak law of large numbers holds. This 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 main use for the weak law of large numbers in connection with Stein's method is its generalization to measure-valued random elements, thus yielding bounds on approximations for empirical measures.

3.1 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, $E = \mathbf{R}, \mathbf{R}^d$, or $\mathbf{R}_+$. 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

$$C_c(E) = \{f : E \to \mathbf{R} \text{ continuous with compact support}\}.$$

For $(\nu_n)_n, \nu \in M_b(E)$, 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

$$F(\nu) = f\left(\int \phi_i d\nu, i = 1, \ldots, m\right) \tag{7}$$

for some $m, f \in C_b^\infty(\mathbf{R}^m)$ and $\phi_i, \ldots, \phi_m \in C_c(E)$ and we define

$$\mathcal{F} = \{F \in C_c(E) : F \text{ satisfies (7)}\}.$$

Using the Stone-Weierstrass Theorem the following lemma is straightforward.

Lemma 3 *$\mathcal{F}$ is convergence-determining for vague convergence. So is the restricted class $\mathcal{F}_0$ that assumes that $\| f' \| \leq 1, \| f'' \| \leq 1, \| \phi_i \| \leq 1$ for $i = 1, \ldots, m$. Also, if $E = \mathbf{R}^d$ or some connected open or closed subset of $\mathbf{R}^d$, 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 bounded differentiable.*

Here, we use the notation $\| f' \| = \sum_{j=1}^m \| f_{(j)} \|$, where $f_j = \frac{\partial}{\partial x_j} f$.

3.2 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*

$$\xi_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Then, for all smooth functions f, we have

$$E\int f d\xi_n = \frac{1}{n}\sum_{i=1}^{n} Ef(X_i) = \int f d\bar{\mu}.$$

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 $\bar{\mu}$. Similarly to the real-valued case, for F of the form (7) the generator for point mass at μ is

$$\mathcal{A}F(\nu) \quad = \quad \sum_{j=1}^{m} f_{(j)}\left(\int \phi_i d\nu, i=1,\ldots,m\right)\left(\int \phi_j d\mu - \int \phi_j d\nu\right).$$

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 *For every H of the form*

$$H(\nu) = h\left(\int \phi_i d\nu, i=1,\ldots,m\right) \tag{8}$$

for some $m, h \in C_b^\infty(\mathbf{R}^m)$ and $\phi_i,\ldots,\phi_m \in C_c(E)$ the solution $F = F_H$ of the Stein equation is of the form (7) with the same ϕ_i's. Furthermore, $\|f'\| \leq \|h'\|$, and $\|f''\| \leq \|h''\|$.

This immediately leads to the following theorem.

Theorem 1 *(Weak law of large numbers for empirical measures) For all $H \in \mathcal{F}$ we have*

$$\begin{aligned} |EH(\xi_n) - H(\mu)| \quad \leq \quad & \sum_{j=1}^{m} \| h_{(j)} \| \left| \int \phi_j d\bar{\mu} - \int \phi_j d\mu \right| \\ & + \sum_{j,k=1}^{m} \| h_{(j,k)} \| \left\{ \max_{1\leq j\leq m} \left[\int \phi_j d\bar{\mu} - \int \phi_j d\mu \right]^2 \right. \\ & \left. + Var\left(\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right)\right\}. \end{aligned}$$

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.

3.3 Mixing random elements

Let $X_1,\ldots,X_n$ be random elements taking values in E. Put $\mathcal{B}_{i,j} = \{A, B \in \mathcal{B} : \mathbf{P}(X_i \in A) \neq 0, \mathbf{P}(X_j \in B) \neq 0\}$ and

$$\rho_n = \frac{1}{n^2}\sum_{i,j=1}^{m} \sup_{A,B\in\mathcal{B}_{i,j}} |Corr(\mathbf{I}(X_i \in A), \mathbf{I}(X_j \in B))|$$

Then, if $\| \phi_j \| \leq 1$ for $i = 1,\ldots,m$, it is easy to see that

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\right) \leq 4\rho_n.$$

Thus the bound in Theorem 1 can be expressed in terms of this mixing coefficient. A similar approach is possible for other definitions of mixing.

3.4 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 $i = 1, \ldots, m$, we have as bound on the variance in Theorem 1

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)\right) \leq \frac{1}{n} + \frac{2}{n^2}\sum_{i=1}^{n}|\Gamma_i|. \tag{9}$$

This approach could be extended to the more general case that 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: A dissociated family. Let $(Y_i)_{i\in\mathbf{N}}$ be a family of i.i.d. random elements on a space $\mathcal{X}$, let $k \in \mathbf{N}$ be fixed, and define the set of multi-indices

$$\Gamma^{(n)} = \{(j_1, \ldots, j_k) \in \Gamma : j_1, \ldots, j_k \in \{1, \ldots, n\}, j_r \neq j_s \text{ for } r \neq s\}.$$

Suppose ψ is a measurable functions $\mathcal{X}^k \to E$, 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. Let $\mu = \mu_{j_1,\ldots,j_k} = \mathcal{L}(X_{j_1,\ldots,j_k})$; due to 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. For $n \in \mathbf{N}$ fixed, the set $\Gamma^{(n)}$ has $n(n-1)\cdots(n-k+1)$ elements; fix an enumeration of these. Let $r(n) = n(n-1)\cdots(n-k+1)$, then our empirical measure of interest is thus

$$\xi_n = \frac{1}{r(n)}\sum_{i=1}^{r(n)}\delta_{X_{i,n}}.$$

We shall derive the following result.

Theorem 2 *For the above dissociated family, we have for $H \in \mathcal{F}_0$*

$$|EH(\xi_n) - H(\mu)| \leq \frac{2k+1}{n(n-k+1)}.$$

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

$$|\Gamma(J)| \quad = \quad k\left(\frac{(n-1)!}{(n-k+1)!} - 1\right)$$

for $k \geq 2$, and $|\Gamma(J)| = 0$ for $k = 1$, and thus

$$\frac{1}{r(n)^2}\sum_{J\in\Gamma^{(n)}}|\Gamma(J)| \quad < \quad k\frac{(n-k)!(n-1)!}{n!(n-k+1)!} = \frac{k}{n(n-k+1)}.$$

With (9) this yields as bound for the variance in Theorem 1

$$Var\left(\frac{1}{r(n)}\sum_{i=1}^{r(n)}\phi_j(X_i)\right) \leq \frac{1}{n(n-1)\cdots(n-k+1)} + \frac{2k}{n(n-k+1)} \leq \frac{2k+1}{n(n-k+1)}.$$

As the $X_{j_1,\dots,j_k}$'s are identically distributed, we have $\bar{\mu} = \mu$, so the variance term is the only contribution to the bound in Theorem 1. This finishes the proof.

The above result can be extended to family of functions $(\psi_{j_1,\dots,j_k})_{(j_1,\dots,j_k)\in\Gamma^{(n)}}$, where for each multi-index we may have a different function.

3.5 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 suggested Palm-measure type coupling, we first recall the size-bias coupling for real-valued random variables.

Let $W \geq 0$ be a nonnegative real-valued random variable and assume $EW > 0$. Then a random variable W^* is said to have the W-size biased distribution if the equation

$$EWg(W) = EWEg(W^*) \tag{10}$$

is satisfied for all g for which both sides of the equation exist. This implicit characterization can be translated into distributions. In some examples these are particularly easy to determine.

Examples:

1. If $W \sim Be(p)$ is a Bernoulli random variable with parameter $0 < p \leq 1$, then for all functions g for which both sides of the equation exist we have $EWg(W) = pg(1)$, so we must have $W^* = 1$, that is, W^* is deterministic and takes on the value 1. As this argument does not depend on the value of p, this illustrates also that the size-bias transformation is not one-to-one.

2. If $W \sim \mathrm{Po}(\lambda)$ has the Poisson distribution with parameter $\lambda > 0$, then from the Stein-Chen equation we obtain for all functions g for which both sides of the equation exist that $EWg(W) = \lambda Eg(W+1)$. From this we see that $W^* = W + 1$ in distribution, that is, W^* has the distribution of a Poisson(λ) random variable which is shifted by 1.

In the weak law of large numbers setting, the size bias coupling leads to the equation

$$E\mathcal{A}f(W) = E(EW - W)f'(W) = EW(Ef'(W) - Ef'(W^*)),$$

where W^* has the W-size bias distribution. Here we assume that $W \geq 0$ with $EW > 0$. 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 EW.

For W being the sum of random variables, [22] give the following construction of W^*. Suppose $W = \sum_{i=1}^n X_i$ with $X_i \geq 0$ being real-valued random variables such that $EX_i > 0$ for all $i = 1, \dots, n$. First we choose a random index V according to

$$P(V = v) = \frac{EX_v}{EW},$$

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 $\hat{X}_u, u \neq v$ such 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 bias distribution.

Example: A classical example for this construction is the case that $W = \sum_{i=1}^n X_i$, where $X_i \sim Be(p_i)$ 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. If $V = v$ we choose $\hat{X}_u, u \neq v$, such that

$$\mathcal{L}(\hat{X}_u, u \neq v) = \mathcal{L}(X_u, u \neq v | X_v = 1).$$

As $X_v^* = 1$,

$$W^* = \sum_{u \neq V} \hat{X}_u + 1$$

has the W-size bias distribution. This coupling is used extensively for Poisson approximation, see [7].

In light of size biasing for real-valued random variables, we can the define a size bias distribution for random measures. This distribution is again defined in terms of test functions.

Definition 1 *Let ξ be a random measure on E, with mean measure $E[\xi] = \mu$, let $\phi \in C_c$ be a nonnegative real-valued continuous funtion having compact support, with $\int \phi d\mu > 0$. We say that ξ^ϕ has the ξ size biased distribution in direction ϕ if*

$$EG(\xi) \int \phi d\xi = \int \phi d\mu EG(\xi^\phi)$$

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: Suppose $\xi = \delta_X$ with $\mathcal{L}(X) = \mu$, and $\phi \geq 0$ is a real-valued function. Then for any test function G for which the left-hand side exists,

$$EG(\xi) \int \phi d\xi \quad = \quad E\phi(X) G(\delta_X) = \int \phi d\mu \, EG(\delta_X^\phi).$$

If $X \geq 0$ and $\phi(x) = x$, then for $G(\nu) = \int g d\nu$ we obtain

$$EG(\xi) \int \phi d\xi \quad = \quad EXg(X) = (EX) E \int g(\nu) d(\delta_X^{\phi(x)=x})(\nu).$$

Using the real-valued size bias coupling with X^* having the X-size bias distribution, we obtain that

$$E \int g(\nu) d(\delta_X)^{\phi(x)=x}(\nu) = Eg(X^*),$$

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.

Similarly to 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 $E[\xi_n] = \bar{\mu}_n$. Fix a non-negative real-valued function ϕ. Pick an index $V \in \{1, \ldots, n\}$ proportionally to the mean of $\phi(X_v)$,

$$P(V = v) = \frac{E\phi(X_v)}{n \int \phi d\bar{\mu}_n},$$

independently of all other random elements. If $V = v$, take $(\delta_{X_v})^\phi$ to have the δ_{X_v}-size bias 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 [25].

The size-bias coupling in connection with the Stein equation leads to considering, for F of the form (7), the generator expression

$$E\mathcal{A}F[\xi] \quad = \quad \sum_{j=1}^m \int \phi_j d\mu E\left\{ f_{(j)}\left(\int \phi_i d\xi_n^{\phi_j}, i = 1, \ldots, n \right) - f_{(j)}\left(\int \phi_i d\xi_n, i = 1, \ldots, n \right) \right\}.$$

As in the independent case the above construction will result in changing only one of the Dirac measures, the right-hand side should 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 involved measures have total mass 1). More arguments are needed to obtain almost sure convergence; see for example [41] for possible techniques.

2. In view of the test functions used, the above could also be viewed as a shorthand for multivariate law 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 5.

4 Discrete distributions from a Gibbs view point

This section presents recent joint work with Peter Eichelsbacher, [15] and [16]. 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, we review some more examples for Stein operators for univariate discrete distributions.

1. The Binomial(n, p)-distribution has Stein operator $\mathcal{A}$ with

$$\mathcal{A}f(k) = (n - k)pf(k + 1) - k(1 - p)f(k) \tag{11}$$

for $0 \leq k \leq n$, see [17].

2. The hypergeometric distribution with parameters (n, a, b), that is,

$$p_k = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}$$

for $k = 0, \ldots a$, has Stein operator $\mathcal{A}$ with

$$\mathcal{A}f(k) = (n-k)(a-k)f(k+1) - k(b-n+k)f(k), \tag{12}$$

see [26], [35] or [37].

3. The geometric distribution with parameter (p) with start at 0: for functions f with $f(0) = 0$ a Stein operator is

$$\mathcal{A}f(k) = (1-p)f(k+1) - f(k), \tag{13}$$

for $k \geq 0$, see [29].

In what follows we shall exploit the connection between discrete univariate distribution and birth-death processes, as studied also by [9], [24] and [44].

We consider discrete univariate Gibbs measures μ with $\mathrm{supp}(\mu) = \{0, \ldots, N\}$, where $N \in \mathbf{N}_0 \cup \{\infty\}$. By definition, a such a Gibbs measure can be written as

$$\mu(k) = \frac{1}{\mathbf{Z}} \exp(V(k)) \frac{\omega^k}{k!}, \quad k = 0, 1, \ldots, N, \tag{14}$$

with $\mathbf{Z} = \sum_{k=0}^{N} \exp(V(k)) \frac{\omega^k}{k!}$, where $\omega > 0$ is fixed. We shall assume that the normalizing constant $\mathbf{Z}$ exists.

Note that the assignment of V and of w in the representation (14) of a Gibbs measure is not unique. For example, if μ denotes the Poisson distribution $Po(\lambda)$, we could choose $\omega = \lambda, V(k) = -\lambda$ for all $k \geq 0$, and $\mathbf{Z} = 1$, or $V(k) = 0$ for all $k, \omega = \lambda$, and $\mathbf{Z} = e^{\lambda}$.

Conversely, for a given probability distribution $(\mu(k))_{k \in \{0,\ldots,N\}}$ we can find a representation (14) as a Gibbs measure by choosing

$$V(k) = \log \mu(k) + \log k! + \log \mathbf{Z} - k \log \omega, \quad k = 0, 1, \ldots, N,$$

with $V(0) = \log \mu(0) + \log \mathbf{Z}$. Again, we have some freedom in the choice of ω and thus of V. Fix a representation (14). To each such Gibbs measure we associate a Markovian birth-death process with unit per-capita death rate $d_k = k$ and birth rate

$$b_k = \omega \exp\{V(k+1) - V(k)\} = (k+1) \frac{\mu(k+1)}{\mu(k)}, \tag{15}$$

for $k, k+1 \in \mathrm{supp}(\mu)$. 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

$$(\mathcal{A}h)(k) \quad = \quad (h(k+1) - h(k)) \exp\{V(k+1) - V(k)\}\omega + k(h(k-1) - h(k))$$

or, with the simplification $f(k) = h(k) - h(k-1)$,

$$(\mathcal{A}f)(k) \quad = \quad f(k+1) \exp\{V(k+1) - V(k)\}\omega - kf(k). \tag{16}$$

In our approach we would typically choose unit per-capita death rates, as used for Poisson processes, see [7]. Other choices of birth and death rates may be advantageous in some situations, see [9] and [24]. To illustrate the approach, we consider some standard examples.

1. The Poisson distribution with parameter $\lambda > 0$: We use $\omega = \lambda, V(k) = -\lambda, \mathcal{Z} = 1$. The Stein operator resulting from (16) is the same as the operator (3).

2. The Binomial distribution with parameters n and $0 < p < 1$: 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 (16) is

$$(\mathcal{A}f)(k) = f(k+1)\,\frac{p(n-k)}{(1-p)} - kf(k).$$

   This differs from the operator (11) only by a factor $1-p$, hence bounds on these two operators are equivalent.

3. The hypergeometric distribution: The Stein operator resulting from (16) is the same as (12).

4. The Pascal distribution with parameter $\gamma \in \{1, 2, \ldots\}$ and $0 < p < 1$, that is, $\mu(k) = \binom{k+\gamma-1}{k} p^\gamma (1-p)^k$ 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, shifted by one, namely $\gamma = n = 1$ in Pascal; $\mu(k) = p(1-p)^k$ for $k = 0, 1, \ldots$. The Stein operator resulting from (16) is now

$$(\mathcal{A}f)(k) = f(k+1)\,(1-p)(k+1) - kf(k),$$

   which deviates from (13).

4.1 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) for the birth-death process generator (16). It is straightforward to verify that for a given function h, a solution f of the Stein equation (1) is given by $f(0) = 0$, $f(k) = 0$ for $k \notin \mathrm{supp}(\mu)$, and

$$\begin{aligned} f(j+1) &= \frac{j!}{\omega^{j+1}} e^{-V(j+1)} \sum_{k=0}^{j} e^{V(k)} \frac{\omega^k}{k!} \left(h(k) - \mu(h)\right) \\ &= -\frac{j!}{\omega^{j+1}} e^{-V(j+1)} \sum_{k=j+1}^{N} e^{V(k)} \frac{\omega^k}{k!} \left(h(k) - \mu(h)\right). \end{aligned}$$

The following non-uniform bounds on f and Δf are derived in [15] and also in [9], [44] as well as in [24], using different methods.

Lemma 5 *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* $j \in \mathbf{N}_0$ *we have that*

$$|f(j)| \le 2 \min\left\{1, \frac{\sqrt{M}}{\sqrt{\omega}}\right\}.$$

2. Assume that the birth rates (15) are non-increasing, that is,

$$\exp\bigl(V(k+1) - V(k)\bigr) \le \exp\bigl(V(k) - V(k-1)\bigr),$$

and that death rates are unit per capita. For every $j \in \mathbf{N}_0$ *we then have*

$$|\Delta f(j)| \le \frac{1}{j} \wedge \frac{e^{V(j)}}{\omega e^{V(j+1)}}.$$

Indeed [9] give bounds for Δf for a wide class of birth-death processes satisfying some monotonicity condition on the rates.

Examples

1. For the Poisson distribution with parameter $\lambda > 0$, the non-uniform bound gives

$$|\Delta f(k)| \leq \frac{1}{k} \wedge \frac{1}{\lambda},$$

recovering the uniform bound $1 \wedge 1/\lambda$, see [7]. In general it does not compare favourably to the uniform $1/\lambda(1-e^{-\lambda})$ in [7], but in [15] it is shown 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 $\| f \| \leq 2 \min\left(1, \frac{1}{\sqrt{\lambda}}\right)$ recovers the bound from [7].

2. For the Pascal distribution with parameter $\gamma \in \{1, 2, \ldots\}$ and $0 < p < 1$, the non-uniform bounds give

$$|\Delta f(k)| \leq \frac{1}{k} \wedge \frac{1}{(1-p)(k+\gamma)},$$

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 shifted by 1 it is however possible to obtain a bound for $|f(j)|$ using sightly different calculations.

4.2 The size-bias coupling

Again we can employ the size-bias coupling to derive bounds on the distance to Gibbs measure. Recall from (10) in Subsection 3.5 that, for $W \geq 0$ with $EW > 0$ we say that W^* has the W-size biased distribution if $EWg(W) = EWEg(W^*)$ for all g for which both sides exist. In particular we obtain that

$$\begin{aligned} & E\Big\{\exp\{V(X+1) - V(X)\}\, \omega\, g(X+1) - X\, g(X)\Big\} \\ &= E\Big\{\exp\{V(X+1) - V(X)\}\, \omega\, g(X+1) - EX Eg(X^*)\Big\}. \end{aligned}$$

Note that

$$EX = \omega E e^{V(X+1)-V(X)}.$$

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-bias distribution.

Lemma 6 *Let $X \geq 0$ be such that $0 < E(X) < \infty$, and let μ be a discrete univariate Gibbs measure on the non-negative integers as in (14). Then $X \sim \mu$ if and only if for all bounded g we have that*

$$\omega\, E e^{V(X+1)-V(X)} g(X+1) \quad = \quad \omega\, E e^{V(X+1)-V(X)} Eg(X^*).$$

Thus for any $W \geq 0$ with $0 < EW < \infty$ we have that

$$Eh(W) - \mu(h) \quad = \quad \omega\{E e^{V(W+1)-V(W)} f(W+1) - E e^{V(W+1)-V(W)} Ef(W^*)\},$$

where f is the solution of the Stein equation (1) with generator (16).

We can also compare two discrete Gibbs distributions by comparing their birth rates and their death rates; a similar idea has been employed in [24]. Let μ as in (14) 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) for μ is such that $\mathcal{A}_2 f$ exists, we calculate

$$\begin{aligned}
&Eh(X) - \mu(h) \\
&= E\mathcal{A}f(X) \\
&= E(\mathcal{A} - \mathcal{A}_2)f(X) \\
&= Ef(X+1)\left(\omega e^{V(X+1)-V(X)} - \omega_2 e^{V_2(X+1)-V_2(X)}\right) \\
&= \omega Ef(X+1)e^{V_2(X+1)-V_2(X)}e^{V(X+1)-V(X)-(V_2(X+1)-V_2(X))} - E(X)Ef(X^*) \\
&= \frac{\omega}{\omega_2}E(X)Ef(X^*)e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - E(X)Ef(X^*) \\
&= \frac{\omega-\omega_2}{\omega_2}E(X)Ef(X^*) + \frac{\omega}{\omega_2}E(X)Ef(X^*)\left\{e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - 1\right\},
\end{aligned}$$

where X^* has the X-size bias distribution. Thus we obtain the bound

$$\begin{aligned}
&\left|Eh(X) - \int h d\mu\right| \\
&\leq \| f \| E(X)\left\{\frac{|\omega-\omega_2|}{\omega_2} + \frac{\omega}{\omega_2}E\left|e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - 1\right|\right\}.
\end{aligned}$$

For example, for two Poisson distributions Poisson(λ_1) and Poisson(λ_2) this approach gives

$$\left|Eh(X) - \int h d\mu\right| \leq \| f \| |\lambda_1 - \lambda_2|.$$

In [15] this approach is employed to bound the distance of summary statistics in an example from statistical physics. In [16] the above approach 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 not needed explicitly in the Stein approach. This is one of the main advantages of the above approach.

5 Example: an S-I-R epidemic

The general stochastic epidemic (GSE) was introduced in [8] in its most basic form; see also [2] for a thorough description. We shall employ a construction suggested in [38]. The results presented here are from [33] and [34].

This model considers a population of total size K at time $t = 0$. At any time $t > 0$ an individual in the population can be susceptible (S) to a certain disease, infected (I) by this 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. Let us say that at time $t = 0$ the population consists of aK infected and bK susceptible individuals, such that $a + b = 1$.

To construct the GSE, let $(l_i, r_i)_{i\in\mathbf{N}}$ be positive i.i.d. random vectors, and let $(\hat{r}_i)_{i\in\mathbf{N}}$ positive i.i.d. random variables, independent of $(l_i, r_i)_{i\in\mathbf{N}}$. The r_i's and the $\hat{r}_i$'s give the length of the infectious period, as follows. An individual i, if was already infected at time 0, will stay infected for a period of length $\hat{r}_i$, then gets 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, then gets removed. Here l_i can be viewed as individual i's resistance to infection, namely, if $Z_K(t)$ denotes the proportion of infectives present in the population at time t, and if $\lambda(t,(x(s))_{s\leq t})$ is a function acting on a one-dimensional parameter (time) and right-continuous functions (the proportion of infected individuals in the past) that will be viewed as an accumulation function, then the infectious pressure is given by

$$F_K(t) = \int_{(0,t]} \lambda(s, Z_K) ds.$$

The time of infection of individual i is then assumed to be given by

$$A_i^K = \inf \left\{ t \in \mathbf{R}_+ : \int_{(0,t]} \lambda(s, Z_K) ds = l_i \right\},$$

that is, the first time that the infectious pressure in the population exceeds the individual's resistance.

This includes the classical case of Bartlett's GSE [8], where $\lambda(t,x) = x(t)$, the resistances $(l_i)_i$ are assumed to be i. i. d. exponentially distributed with parameter 1, and $(r_i)_i, (\hat{r}_i)_i$ are assumed to be i. i. d. exponential with same parameter ρ, independent of $(l_i)_i$. This results in a Markovian model, where standard Markov techniques can be applied. A generalization was studied by [42], [43], with the accumulation function $\lambda(t,x) = \lambda(x(t))$, the resistances $(l_i)_i$ being i. i. d. $exp(1)$-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 at time t, and similarly we recover the proportion of susceptibles at time t and the proportion of removed at time t. Moreover,

$$\xi_K([0,s] \times (t,\infty]), \ t > s,$$

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 particularly interesting if public health policy changed during the course of the epidemic, for example as 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. Firstly we shall have to make some assumptions. Let D_+ denote the space of all functions $x : \mathbf{R}_+ \to [-1,1]$ that are right-continuous with left-hand limits.

1. Assume that $\lambda : \mathbf{R}_+ \times D_+ \to \mathbf{R}_+$ is uniformly bounded by a constant γ, Lipschitz in $x \in D_+$ with Lipschitz constant α. Assume that λ non-anticipating in the sense that $\lambda(t,x)$ depends on the function x only through $(x_s)_{0\leq s\leq t}$, and assume also that for all $t \in \mathbf{R}_+$

$$\lambda(t,x) = 0 \iff x(t) = 0.$$

2. Assume that there is a constant $\beta > 0$ such that, for each $x \in \mathbf{R}_+$, the conditional cumulative distribution function $\Psi_x(t) := \mathbf{P}[l_1 \leq t | r_1 = x]$ has a density $\psi_x(t)$ that is uniformly bounded from above by β;

$$\psi_x(t) \leq \beta \text{ for all } x \in \mathbf{R}_+ \text{ and all } t \in \mathbf{R}_+.$$

3. Assume that the $(l_i)_i$ 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

$$Z_K(t) \quad = \quad \frac{1}{K}\sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i > t) + \frac{1}{K}\sum_{j=1}^{bK} \mathbf{1}(F_K^{-1}(l_j) \leq t < F_K^{-1}(l_j) + r_j)$$

gives the proportion of infected at time t. For $f \in C(\mathbf{R}_+, \mathbf{R}), t \in \mathbf{R}_+$, define the operators

$$\mathcal{Z}f(t) \quad = \quad a(1 - \Phi(t)) + b\mathbf{P}(f(t - r_1) \leq l_1 < f(t))$$

and

$$Lf(t) \quad = \quad \int_{(0,t]} \lambda(s, \mathcal{Z}f)ds.$$

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 will turn 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 3 *For $T \in \mathbf{R}_+$, the operator L is a contraction on $[0, T]$, and the equation*

$$f(t) = \int_{(0,t]} \lambda(s, \mathcal{Z}f)ds, \quad 0 \leq t \leq T, \tag{17}$$

has a unique solution G_T.

For $T \in \mathbf{R}_+$, let G_T be the solution of (17) and let $\tilde{\mu}^T$ be given for $r, s \in (0, T]$ by

$$\tilde{\mu}^T([0, r] \times [0, s]) = \mathbf{P}^T[l_1 \leq G_T(r), l_1 \leq G_T(s - r_1)].$$

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 4 *Let For all $H \in \mathcal{F}_0$ of the form (8) and for all $T \in \mathbf{R}_+$, we have that*

$$|EH(\mathcal{L}(\xi_n^T) - EH(\delta_{\mu^T})| \quad \leq \quad \frac{\sqrt{a} + \sqrt{b}}{\sqrt{K}} + \alpha b\beta T(T + 2)\exp(b\lceil 2\alpha\beta T\rceil)\left\{(1 + b)\sqrt{\frac{1}{K}} + \frac{2}{K}\right\}.$$

Here, $\lceil x \rceil$ is the smallest integer larger than x.

The proof of this theorem is rather involved and can be found in [34], so it only be sketched here. The main arguments used are the Glivenko-Cantelli theorem to justify the law of large numbers argument, the Contraction Theorem for Theorem 3, 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 similar infection time with individual 1 from the susceptible population left out, then $F_{K,1}$ and l_1 are independent.

Sketch of the proof of Theorem 4.

We assume that Theorem 3 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 (7) we bound $E\mathcal{A}F$, where $\mathcal{A}$ is the operator associated with the Dirac measure at μ^T. Then we have

$$\begin{aligned}
&\sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle {\xi_K}^T, \phi_k \rangle, k = 1, \ldots, m) \langle \mu^T - {\xi_K}^T, \phi_j \rangle \\
&= a \sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle {\xi_K}^T, \phi_k \rangle, k = 1, \ldots, m) \left\langle (\delta_0 \times \hat{\mu})^T - \frac{1}{aK} \sum_{i=1}^{aK} \delta^T_{(0,\hat{r}_i)}, \phi_j \right\rangle \qquad (18) \\
&\quad + b \sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle {\xi_K}^T, \phi_k \rangle, k = 1, \ldots, m) \langle \tilde{\mu}^T - \zeta_K^T, \phi_j \rangle. \qquad (19)
\end{aligned}$$

For the first summand (18) we employ the Cauchy-Schwarz inequality and the bounds on the functions involved in the form (7) to obtain

$$\begin{aligned}
&\left| a \sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle {\xi_K}^T, \phi_k \rangle, k = 1, \ldots, m) \langle (\delta_0 \times \hat{\mu})^T - \frac{1}{aK} \sum_{i=1}^{aK} \delta^T_{(0,\hat{r}_i)}, \phi_j \rangle \right| \\
&\leq a \sum_{j=1}^{m} \| f_{(j)} \| \, \mathbf{E} \left| \frac{1}{aK} \sum_{i=1}^{aK} (\phi_j(0, \hat{r}_i) - \mathbf{E}\phi_j(0, \hat{r}_i)) \right| \\
&\leq a \sum_{j=1}^{m} \| f_{(j)} \| \left(\mathrm{Var} \left(\frac{1}{aK} \sum_{i=1}^{aK} (\phi_j(0, \hat{r}_i) - \mathbf{E}\phi_j(0, \hat{r}_i)) \right)^{\frac{1}{2}} \right) \\
&\leq \frac{\sqrt{a}}{\sqrt{K}}.
\end{aligned}$$

Similarly, for the second summand (19) we obtain

$$\begin{aligned}
&b \sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle {\xi_K}^T, \phi_k \rangle, k = 1, \ldots, m) \langle \tilde{\mu}^T - \zeta_K^T, \phi_j \rangle \\
&= b \sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle a(\delta_0 \times \hat{\mu})^T + b\zeta_K^T, \phi_k \rangle, k = 1, \ldots, m) \langle \tilde{\mu}^T - \zeta_K^T, \phi_j \rangle + R_1,
\end{aligned}$$

where by Taylor expansion $|R_1| \leq 2b\frac{\sqrt{a}}{\sqrt{K}}$. It remains to bound

$$\begin{aligned} & b\sum_{j=1}^{m} \mathbf{E} f_{(j)}(\langle a(\delta_0 \times \hat{\mu})^T + b\zeta_K^T, \phi_k\rangle, k = 1, \ldots, m)\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle \\ \leq \quad & b\sum_{i=1}^{m} \| f_{(j)} \| \mathbf{E} \left|\langle \tilde{\mu}^T - \zeta_K^T, \phi_j\rangle\right| \\ = \quad & b\sum_{j=1}^{m} \| f_{(j)} \| \mathbf{E} \left| \frac{1}{bK}\sum_{i=1}^{bK} \phi_i((F_K^T)^{-1}(l_i), (F_K^T)^{-1}(l_i) + r_i) - \phi_j(G_T^{-1}(l_i), G_T^{-1}(l_i) + r_i)\right| \\ \leq \quad & b\sum_{j=1}^{m} \| f_{(j)} \| \| \phi_j' \| \mathbf{E} \left|(F_K^T)^{-1}(l_1) - (G_T)^{-1}(l_1)\right|. \end{aligned}$$

To tackle the problem that F_K and l_1, we couple the process to the same process with susceptible individual 1 omitted; denote by $F_{K,1}$ the infectious pressure in this new process. Then $F_K^{-1}(l_1) = F_{K,1}^{-1}(l_1)$ and

$$\mathbf{E}\left|(F_K^T)^{-1}(l_1) - G_T^{-1}(l_1)\right| \quad = \quad \mathbf{E}\left|(F_{K,1}^T)^{-1}(l_1) - G_T^{-1}(l_1)\right|.$$

In order to desribe this new process, for $h \in D([0,T])$, the space of right-continuous function $[0,T] \to [-1,1]$ with left-hand limits, we define the operators

$$\mathcal{Z}_{K,1}h(t) \quad = \quad \frac{1}{K}\sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i > t) + \frac{1}{K}\sum_{j=2}^{bK} \mathbf{1}(h(t - r_j) < l_j \leq h(t))$$

and

$$L_{K,1}h(t) \quad = \quad \int_{(0,t]} \lambda(s, \mathcal{Z}_{K,1}h)\, ds.$$

Note that $F_K^{-1}(l_1) = F_{K,1}^{-1}(l_1)$ by construction, and, for all $t \leq T$,

$$\| F_{K,1} - G_T \|_t = \| L_{k,1}F_{K,1} - LG_T \|_t \leq \sup_h \| L_{k,1}h - Lh \|_t + \| LF_{K,1} - LG_T \|_t .$$

For each $h \in D([0,T])$ we have that

$$\begin{aligned} \| L_{K,1}h - Lh \|_T \quad & \leq \quad \alpha \int_0^T \sup_{s \leq x} |\mathcal{Z}_{K,1}h(s) - \mathcal{Z}h(s)| ds \\ & \leq \quad \alpha T \left(aR_1 + 2bR_2 + \frac{2}{K}\right), \end{aligned}$$

where

$$R_1 = \sup_s \left| \frac{1}{aK}\sum_{i=1}^{aK} \mathbf{1}(\hat{r}_i \leq s) - \Phi(s)\right|$$

and

$$R_2 = \sup_s \left| \frac{1}{bK - 1}\sum_{i=2}^{bK} \mathbf{1}(l_i \leq s) - \Psi(s)\right|.$$

[28] enables us to derive as bounds for these remainder terms that $\mathbf{E}R_1 \leq \frac{1}{\sqrt{aK}}$ and $\mathbf{E}R_2 \leq \frac{1}{\sqrt{bK}}$. Thus

$$\mathbf{E}\sup_h \| L_{K,1}h - Lh \|_T \quad \leq \quad \alpha T \left\{ (1 + b)\frac{1}{\sqrt{K}} + \frac{2}{K}\right\} =: S(K).$$

To bound $\mathbf{E} \parallel LF_{K,1} - LG_T \parallel_t$ we use again a contraction argument. We have that

$$|LF_{K,1}(t) - LG_T(t)| \quad \leq \quad \alpha b \beta \int_0^t \parallel F_{K,1} - G_T \parallel_x (1 + \Phi(x)) dx$$

and hence

$$\mathbf{E} \parallel LF_{K,1} - LG_T \parallel_t \quad \leq \quad S(K) + \alpha b \beta \int_0^t \parallel F_{K,1} - G_T \parallel_x (1 + \Phi(x)) dx.$$

Fix some $c \geq b$ and put

$$\eta = \frac{1}{2c\alpha\beta}$$

then

$$\mathbf{E} \parallel LF_{K,1} - LG_T \parallel_\eta \quad \leq \quad \frac{c}{c-b} S(K).$$

By induction we can show that

$$\mathbf{E} \parallel LF_{K,1} - LG_T \parallel_{k\eta} \quad \leq \quad \left(\frac{c}{c-b}\right)^k S(K).$$

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

$$\mathbf{E} \parallel LF_{K,1} - LG_T \parallel_{k\eta} \quad \leq \quad \exp(\lceil 2c\alpha\beta T \rceil)(\log c - \log(c-b)) S(K).$$

Letting $c \to \infty$ gives the assertion.

Some concluding remarks:

1. The derived bound 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, due to using the empirical measure. To make the model even more realistic, we could assume for the initially infected that $(\hat{r}_i)_i$ 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 [3].

4. In [4] it is shown that the waiting time until 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 instead considering only a time interval when there is already a substantial proportion of infectives present, then the bound on the approximation much improves, growing only linearly in time, see [34].

5. It would be interesting to investigate how the model would behave in a spatial setting, where we do not assume homogeneous mixing.

6 The density approach

This approach was first suggested by Stein in [40]; 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 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 for which*

$$\int_{-\infty}^{\infty} |f'(z)|\, p(z)dz < \infty,$$

we have

$$\mathbf{E}(f'(Z) + \psi(Z)f(Z)) = 0.$$

Some examples are

1. The standard normal distribution $\mathbf{N}(0,1)$. Here, $\psi(x) = -x$, and the above conditions are satisfied. The characterization results in the classical Stein equation.

2. The Gamma distribution with density $p_{\lambda,a}(x) = \frac{\lambda^a e^{-\lambda x} x^{a-1}}{\Gamma(a)}$. Although the density is not positive on the whole real line, Proposition 1 gives an indication on how to obtain a Stein characterization. Here $\psi(x) = \frac{a-1-\lambda x}{x}$, for $x > 0$. This would yield the characterization of type

$$\mathbf{E}f'(X) + \frac{a-1-\lambda X}{X} f(X) = 0.$$

Comparing this with the characterization (4) we see by putting $g(x) = xf(x)$ that the two characterizations are equivalent.

Let for convenience

$$\phi(x) = -\psi(x).$$

The following Stein theorem is derived in [40].

Theorem 5 *Suppose Z has probability density function p satisfying the assumptions of the above proposition. Let (W, W') be an exchangeable pair such that $\mathbf{E}(\phi(W))^2 = \sigma^2 < \infty$, and let*

$$\lambda = \frac{\mathbf{E}(\phi(W') - \phi(W))^2}{2\sigma^2}.$$

Then, for all piecewise continuous functions h on $\mathbf{R}$ to $\mathbf{R}$ for which $\mathbf{E}|h(Z)| < \infty$, we have that

$$\begin{aligned}
&\mathbf{E}h(W) - \mathbf{E}h(Z) \\
&= \mathbf{E}(Vh)(W) \\
&\quad -\frac{1}{\lambda\sigma^2}\mathbf{E}(\phi(W') - \phi(W))((Uh)(W') - (Uh)(W)) - \mathbf{E}\mathbf{E}^W\left(\frac{\phi(W') - (1-\lambda)\phi(W)}{\lambda}\right)(Uh)(W),
\end{aligned}$$

where Uh and Vh are defined by

$$(Uh)(w) = \frac{\int_{-\infty}^{z}\left(h(x) - \int_{-\infty}^{\infty} h(y)p(y)dy\right)p(x)dx}{p(z)}$$

and

$$(Vh)(w) = (Uh)'(w).$$

Theorem 5 can be employed to assess the error in a normal approximation via simulations, replacing the expectations by sample means.

7 Distributional transformations

This is joint work with Larry Goldstein and can be found in [21]. We already saw in connection with Gibbs measures how the size-bias distribution can be used to characterize a target distribution. The work presented here takes this idea further.

Firstly, the size bias 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 [20]).

Definition 2 *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 $EXf(X)$ exists,*

$$EXf(X) = \sigma^2 Ef'(X^*).$$

The zero bias distribution X^* exists for all X that have mean zero and finite variance. In [20] 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 6 *Let X be a random variable, $m \in \{0, 1, 2, \ldots\}$ and let P a function on the support of X such that P has exactly m sign changes, is positive on its rightmost interval and*

$$\frac{1}{m!}EX^jP(X) = \alpha\delta_{j,m} \quad j = 0, \ldots, m,$$

for some $\alpha > 0$. Then there exists a unique distribution for a random variable $X^{(m)}$ such that such that

$$EP(X)G(X) = \alpha EG^{(m)}(X^{(m)})$$

for all m times differentiable functions G for which $EP(X)G(X)$ exists.

The $X^{(m)}$ distribution is named the $X-P$ biased distribution.

For example, with $P(x)=x$, we obtain that for any variable X such that $\sigma^2=EX^2<\infty$ and so that $EX=0$, there exists a random variable $X^{(1)}$ such that, for all smooth G, we have $EXG(X)=\sigma^2EG'(X^{(1)})$. Thus we recover the zero bias distribution.

Theorem 6 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

$$\alpha_m^\lambda=\frac{1}{m!}EZ_\lambda^m P_m^\lambda(Z_\lambda)$$

and the set

$$\mathcal{M}_\lambda^m=\{X: EX^{2m}<\infty,\quad EX^j=EZ_\lambda^j,\quad 0\leq j\leq 2m\}.$$

For every $X\in\mathcal{M}_\lambda^m$, for $j=0,\ldots,m$, we then have

$$\frac{1}{m!}EX^jP_m^\lambda(X) \quad = \quad \frac{1}{m!}EZ_\lambda^jP_m^\lambda(Z_\lambda)=\alpha_m^\lambda\delta_{j,k}.$$

Theorem 6 yields that for all $X\in\mathcal{M}_\lambda^m$ there exists a random variable X_m^λ such that

$$EP_m^\lambda(X)G(X)=\alpha_m^\lambda EG^{(m)}(X_m^\lambda).$$

Similarly to the size bias distributon, there is a construction for sums of independent variables available. Let $m\in\{0,1,\ldots\}$. Let $X_1,\cdots,X_n$ be independent 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 summand to different degrees; hence let us 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 $\{P_\lambda^m(x)\}_{m\geq 0}$ satisfies the conditions of Theorem 6, and let $\lambda=\lambda_1+\cdots+\lambda_n$. In this setting, [21] derive the following result

Theorem 7 *Suppose that for some weights $c_{\mathbf{m}}$ the family of orthogonal polynomials $\{P_\lambda^m(x)\}_{m\geq 0}$ satisfies the identity*

$$P_\lambda^m(w) \quad = \quad \sum_{\mathbf{m}} c_{\mathbf{m}}P_\Lambda^{\mathbf{m}}(\mathbf{x}),$$

with $w=x_1+\cdots+x_n$. 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*

$$P(\mathbf{I}=\mathbf{m})=\frac{c_{\mathbf{m}}\alpha_{\Lambda}^{(\mathbf{m})}}{\alpha_{\lambda}^{(m)}}. \tag{20}$$

Furthermore, the variable

$$W_{\lambda}^{(m)}=\sum_{\mathbf{m}}(X_i)_{\lambda_i}^{(I_i)}$$

has the $W-P_{\lambda}^{m}$ *distribution.*

Finally, some examples.

1. *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 (20) is chosen according to the multinomial distribution $Mult(m,\Lambda)$. 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.

2. *Charlier biasing:* The Charlier polynomials correspond to the Poisson distribution with parameter λ; here again we obtain $\alpha_m^{\lambda}=\lambda^m$, and $I\sim Mult(m,\Lambda)$. 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.

3. *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 [14]; there one can also find more examples for 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 [13] and on the uniform distribution [12]. 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] Arratia, R., Goldstein, L., and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* **17**, 9–25.

[2] Bailey, N.T.J. (1975). *The mathematical theory of infectious diseases and its applications.*(2nd ed.) Griffin, London.

[3] Barbour, A.D. (1974). On a functional central limit theorem for Markov population processes. *Adv. Appl. Probab.* **6**, 21–39.

[4] Barbour, A.D. (1975). The duration of the closed stochastic epidemic. *Biometrika* **62**, 477–482.

[5] Barbour, A.D. (1988). Stein's method and Poisson process convergence. *J. Appl. Probab.* **25A**, 175–184.

[6] Barbour, A.D. (1990). Stein's method for diffusion approximations. *Probability Theory and Related Fields* **84**, 297–322.

[7] Barbour, A.D., Holst, L., and Janson, S. (1992). *Poisson Approximation.* Oxford Science Publications.

[8] Bartlett, M.S. (1949) Some evolutionary stochastic processes. *J. R. Statist. Soc., Ser. B* **11**, 211–229.

[9] Brown, T., and Xia, A. (2001). Stein's method and birth-death processes. *Ann. Probab.* **29**, 1373-1403.

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

[11] Chen, L.H.Y., and Shao, Q. (2004). *Stein's method and normal approximation.* In this volume.

[12] Diaconis, P. (1989). An example for Stein's method. Stanford Stat. Dept. Technical Report.

[13] Diaconis, P. (2003). Stein's method for Markov chains: first examples. To appear in *Proceedings of the Stein seminar 1998.* Stanford, S. Holmes ed.

[14] Diaconis, P., and Zabell, S. (1991). Closed form summation for classical distributions: variations on a theme of de Moivre. *Statistical Science* **6**, 284-302.

[15] Eichelsbacher, P., and Reinert, G. (2004). Stein's method for discrete Gibbs measures. In revision.

[16] Eichelsbacher, P., and Reinert, G. (2004). Stein's method for spatial Gibbs measures. Preprint.

[17] Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statist. Probab. Lett.* **11**, 7-16.

[18] Erhardsson, T. (2004). *Stein's method for Poisson and compound Poisson approximations.* In this volume.

[19] Ethier, S.N., and Kurtz, T. (1986). *Markov Processes, Characterization and Convergence.* Wiley, New York.

[20] GOLDSTEIN, L., AND REINERT, G. (1997). Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* **7**, 935-952.

[21] GOLDSTEIN, L., AND REINERT, G. (2003). Distributional transformations, orthogonal polynomials, and Stein characterizations. Submitted.

[22] GOLDSTEIN, L., AND RINOTT, Y. (1996). On multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Prob.* **33**, 1–17.

[23] GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19**, 724–739.

[24] HOLMES, S. (2003). Stein's method for birth and death chains. Preprint.

[25] KALLENBERG, O. (2002). *Foundations of Modern Probability.* Springer, New York etc., 2^{nd} edition.

[26] KÜNSCH, H.-R. (1998) Personal communication.

[27] LUK, H.M. (1994). Stein's method for the gamma distribution and related statistical applications. Ph.D. thesis. University of Southern California, Los Angeles, USA.

[28] MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269–1283.

[29] PEKÖZ, E. (1996). Stein's method for geometric approximation. *J. Appl. Probab.* **33**, 707-713.

[30] PICKETT, A. (2002). *Stein's method for chisquare approximations.* Transfer thesis, University of Oxford.

[31] PICKETT, A., AND REINERT, G. (2004). Stein's method for chisquare approximations. Preprint.

[32] REINERT, G. (1994). A weak law of large numbers for empirical measures via Stein's method. *Ann. Probab.* **23**, 334–354.

[33] REINERT, G. (1995) The asymptotic evolution of the General Stochastic Epidemic. *Ann. Appl. Probab.* **5**, 1061–1086.

[34] REINERT, G. (2001). Stein's method in application for epidemic processes. In *Complex Stochastic Systems.* O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg, eds., Chapman and Hall, Boca Raton etc., 235-275.

[35] REINERT, G., AND SCHOUTENS, W. (1998). Stein's method for the hypergeometric distribution. Preprint.

[36] RINOTT, Y., AND ROTAR, V. (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.

[37] SCHOUTENS, W. (2000) *Stochastic Processes and Orthogonal Polynomials.* Springer, New York etc.

[38] SELLKE, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* **20**, 390–394.

[39] STEIN, C. (1986). *Approximate Computation of Expectations.* IMS, Hayward, California.

[40] STEIN, C. (2003). Use of Normal Approximations in the Analysis of Simulations. To appear in *Proceedings of the Stein seminar 1998.* Stanford, S. Holmes ed.

[41] Van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York etc.

[42] Wang, F.S.J. (1975). Limit theorems for age and density dependent stochastic population models. *J. Math. Bio.* **2**, 373–400.

[43] Wang, F.S.J. (1977). A central limit theorem for age-and-density-dependent population processes. *Stoch. Proc. Appl.* **5**, 173–193.

[44] Weinberg, G.V. (2000). Stein factor bounds for random variables. *J. Appl. Probab.* **37**, 1181-1187.