A Short Introduction to Stein’s Method

Gesine Reinert
Department of Statistics
University of Oxford

Michaelmas Term 2011
Outline

Lecture 1: Normal approximation
  Motivation
  Distributional distances
  The Stein equation
  Local dependence
  Couplings

Lecture 2: Poisson approximation and other distributions
  Poisson approximation
    Poisson approximation: the local approach
    Poisson approximation: size bias couplings
  A general framework
  Chisquare distributions
  One-dimensional Gibbs distributions
  Final Remarks
Recall two famous distributional approximations:

Example: $X_1, X_2, \ldots, X_n$ i.i.d., $P(X_i = 1) = p = 1 - P(X_i = 0)$

$$n^{-1/2} \sum_{i=1}^{n} (X_i - p) \approx_d \mathcal{N}(0, p(1 - p))$$

but also

$$\sum_{i=1}^{n} X_i \approx_d \text{Poisson}(np)$$

Which approximation should we choose?
We would like to assess the distance between distributions via explicit bounds.
Often distributional distances are phrased in terms of test functions. This approach is based on weak convergence. For c.d.f.s $F_n, n \geq 0$ and $F$ on the line we say that $F_n$ converges weakly (converges in distribution) to $F$,

$$F_n \xrightarrow{\mathcal{D}} F$$

if

$$F_n(x) \to F(x) \quad (n \to \infty)$$

for all continuity points $x$ of $F$. For the associated probability distributions:

$$P_n \xrightarrow{\mathcal{D}} P.$$
Facts of weak convergence

\[ P_n \xrightarrow{D} P \]

is equivalent to:

1. \( P_n(A) \to P(A) \) for each \( P \)-continuity set \( A \) (i.e. \( P(\partial A) = 0 \))
2. \( \int f dP_n \to \int f dP \) for all functions \( f \) that are bounded, continuous, real-valued
3. \( \int f dP_n \to \int f dP \) for all functions \( f \) that are bounded, infinitely often differentiable, continuous, real-valued.

For a random variable \( X \) with distribution \( \mathcal{L}(X) \),

\[ \mathcal{L}(X_n) \xrightarrow{D} \mathcal{L}(X) \iff \mathbf{E}f(X_n) \to \mathbf{E}f(X) \]

for all functions \( f \) that are bounded, infinitely often differentiable, continuous, real-valued.
Total variation distance

Let $\mathcal{L}(X) = P$, $\mathcal{L}(Y) = Q$; define \textit{total variation distance}

$$d_{TV}(P, Q) = \sup_{A \text{ measurable}} |P(A) - Q(A)|.$$  

If the underlying space is discrete then convergence in total variation is equivalent to convergence in distribution. For probability measures on $\mathbb{Z}^+$ we can write

$$d_{TV}(P, Q) = \frac{1}{2} \sum_{j \geq 0} |P\{j\} - Q\{j\}|.$$  

If the underlying space is continuous then the total variation distance is very strong; for example approximating a discrete integer-valued random variable by a normal random variable we could choose as set $A$ the set of integers; then the total variation distance would be 1.
Consider the set of Lipschitz-continuous functions with Lipschitz constant 1,

\[ \mathcal{L} = \{ g : \mathbb{R} \to \mathbb{R}; |g(y) - g(x)| \leq |y - x| \} \]

and Wasserstein distance

\[
d_W(P, Q) = \sup_{g \in \mathcal{L}} |\mathbb{E}g(Y) - \mathbb{E}g(X)|
\]

\[
= \inf \mathbb{E}|Y - X|,
\]

where the infimum is over all couplings \(X, Y\) such that \(\mathcal{L}(X) = P, \mathcal{L}(Y) = Q\).
Using

\[ \mathcal{F} = \{ f \in \mathcal{L} \text{ absolutely continuous, } f(0) = f'(0) = 0 \} \]

we also have

\[ d_W(P, Q) = \sup_{f \in \mathcal{F}} |E f'(Y) - E f'(X)|. \]
Stein’s Method for Normal Approximation

Stein (1972, 1986)

$Z \sim \mathcal{N}(\mu, \sigma^2)$ if and only if for all smooth functions $f$,

$$
E(Z - \mu)f(Z) = \sigma^2 Ef'(Z).
$$

For a random variable $W$ with $E W = \mu$, $\text{Var} W = \sigma^2$, if

$$
\sigma^2 Ef'(W) - E(W - \mu)f(W)
$$

is close to zero for many functions $f$, then $W$ should be close to $Z$ in distribution.
Sketch of proof for $\mu = 0, \sigma^2 = 1$:

Assume $Z \sim \mathcal{N}(0, 1)$. Integration by parts:

$$
\frac{1}{\sqrt{2\pi}} \int f'(x)e^{-x^2/2} \, dx = \left[ \frac{1}{\sqrt{2\pi}} f(x)e^{-x^2/2} \right] + \frac{1}{\sqrt{2\pi}} \int xf(x)e^{-x^2/2} \, dx
$$

$$
= \frac{1}{\sqrt{2\pi}} \int xf(x)e^{-x^2/2} \, dx
$$

and so $\mathbf{E}f'(Z) = \mathbf{E}Zf(Z)$. 
Conversely, assume $\mathbf{E}Zf(Z) = \mathbf{E}f'(Z)$: We solve the differential equation

$$f'(x) - xf(x) = g(x), \quad \lim_{x \to -\infty} f(x)e^{-x^2/2} = 0$$

for any bounded function $g$, giving

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

Take $g(x) = 1(x \leq x_0) - \Phi(x_0)$, then

$$0 = \mathbf{E}(f'(Z) - Zf(Z)) = \mathbf{P}(Z \leq x_0) - \Phi(x_0)$$

so $Z \sim \mathcal{N}(0, 1)$. 
The Stein equation

Let $\mu = 0$. Given a test function $h$, let $Nh = \mathbf{E} h(Z/\sigma)$, and solve for $f$ in the Stein equation

$$\sigma^2 f'(w) - wf(w) = h(w/\sigma) - Nh$$

giving

$$f(y) = e^{y^2/2} \int_{-\infty}^{y} (h(x/\sigma) - Nh) e^{-x^2/2} dx.$$ 

Now evaluate the expectation of the r.h.s. by the expectation of the l.h.s. We can bound the solution $f$ and its derivatives in terms of the test function $h$ and its derivatives, e.g. $\|f''\| \leq 2 \|h'\|$. 

Example: the sum of i.i.d. random variables

\(X, X_1, \ldots, X_n\) i.i.d. with \(\mathbb{E}X = 0, \text{Var}X = \frac{1}{n}\); put \(W = \sum_{i=1}^{n} X_i\) and put \(W_i = W - X_i = \sum_{j \neq i} X_j\). Then

\[
\mathbb{E}Wf(W) = \sum_{i=1}^{n} \mathbb{E}X_i f(W)
\]

\[
= \sum_{i=1}^{n} \mathbb{E}X_i f(W_i) + \sum_{i=1}^{n} \mathbb{E}X_i^2 f'(W_i) + R
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}f'(W_i) + R.
\]
So

\[ Ef'(W) - EWf(W) = \frac{1}{n} \sum_{i=1}^{n} E\{f'(W) - f'(W_i)\} + R \]

and we can bound the remainder term \( R \).
Theorem: explicit bound on distance

For any smooth $h$

$$|Eh(W) - Nh| \leq \|h'\left(\frac{2}{\sqrt{n}} + \sum_{i=1}^{n} E|X_i^3|\right)\|.$$  

Note: this bound is true for any $n$; nothing goes to infinity.

This theorem extends to local dependence.
Local dependence

Let $X_1, \ldots, X_n$ be mean zero, finite variances, put $W = \sum_{i=1}^{n} X_i$; assume $\text{Var} W = 1$. Suppose that for each $i = 1, \ldots, n$ there exist sets $A_i \subset B_i \subset \{1, \ldots, n\}$ such that $X_i$ is independent of $\sum_{j \notin A_i} X_j$ and $\sum_{j \in A_i} X_j$ is independent of $\sum_{j \notin B_i} X_j$. Define

$$\eta_i = \sum_{j \in A_i} X_j \quad \text{and} \quad \tau_i = \sum_{j \in B_i} X_j.$$ 

**Theorem**

*For any smooth $h$ with $\|h'\| \leq 1$,*

$$|E h(W) - N h| \leq 2 \sum_{i=1}^{n} (E |X_i \eta_i \tau_i| + |E (X_i \eta_i) |E |\tau_i|) + \sum_{i=1}^{n} E |X_i \eta_i^2|. $$
Example: word counts in DNA sequences

Let $B = B_1 B_2 \ldots B_n$ be a string of letters which are i.i.d. from the alphabet $A = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ be a word of length $m$ composed of letters from $A$. Let $V = V(w)$ be the number of times that $w$ occurs in the sequence $B$. Words occur in clumps; for example consider

$$B = T\ G\ A\ A\ C\ A\ A\ A\ C\ A\ A\ A\ C\ A\ A\ A\ G\ A\ A\ C\ A\ A\ A\ A$$

$$w = A\ A\ C\ A\ A$$
Example: word counts in DNA sequences

Let $B = B_1 B_2 \ldots B_n$ be a string of letters which are i.i.d. from the alphabet $\mathcal{A} = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ be a word of length $m$ composed of letters from $\mathcal{A}$. Let $V = V(w)$ be the number of times that $w$ occurs in the sequence $B$. Words occur in clumps; for example consider

$$B = T \ G \ A \ A \ C \ A \ A \ A \ C \ A \ A \ C \ A \ A \ A \ G \ A \ A \ C \ A \ A \ A \ A \ A$$

$$w = A \ A \ C \ A \ A$$
Let $B = B_1 B_2 \ldots B_n$ be a string of letters which are i.i.d. from the alphabet $A = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ be a word of length $m$ composed of letters from $A$. Let $V = V(w)$ be the number of times that $w$ occurs in the sequence $B$. Words occur in clumps; for example consider

\begin{align*}
B &= T \ G \ A \ A \ C \ A \ A \ A \ C \ A \ A \ C \ A \ A \ G \ A \ A \ C \ A \ A \ A \ A \\
w &= A \ A \ C \ A \ A
\end{align*}
Example: word counts in DNA sequences

Let $B = B_1 B_2 \ldots B_n$ be a string of letters which are i.i.d. from the alphabet $A = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ be a word of length $m$ composed of letters from $A$. Let $V = V(w)$ be the number of times that $w$ occurs in the sequence $B$. Words occur in clumps; for example consider

$$B = T \ G \ A \ A \ C \ A \ A \ A \ C \ A \ A \ A \ C \ A \ A \ A \ G \ A \ A \ C \ A \ A \ A \ A \ A$$

$$w = A \ A \ C \ A \ A$$

$V(w) = 4$; there are two clumps of occurrences.
For $j = 1, \ldots, n - m + 1$ define the word indicators

$$I_j = \begin{cases} 
1 & \text{if } B_j B_{j+1} \ldots B_{j+m-1} = w \\
0 & \text{otherwise.} 
\end{cases}$$

Let $p(x) = P(X_1 = x)$ for $x = A, C, G, T$, then

$$E I_j = p(w) := \prod_{i=1}^{m} p(w_i)$$

and

$$EV = (n - m + 1)p(w).$$
The variance

Define, for $\ell = 1, \ldots, m$,

$$p_w(m) = p(w_{m-\ell+1})p(w_{m-\ell+2}) \cdots p(w_m),$$

which is the probability that $w$ occurs, given that $w_1w_2 \cdots w_{m-\ell}$ has occurred. Define the self-overlap indicator

$$\xi(j) = 1(w_{j+1} = w_1, \ldots, w_m = w_{m-j}).$$

This overlap indicator equals 1 if the last $m - j$ letters of $w$ overlap exactly the first $m - j$ letters of $w$. Then the variance $\text{Var}(W) = \sigma^2$ is given by

$$2p(w) \sum_{j=1}^{m-1} (n - m - j + 1)\xi(j)p_w(j) - 2(m - 1)(n - m + 1)p(w)^2.$$

The variance is of order $n$ when $m = o(n)$. 
The neighbourhoods of dependence

Put

\[ X_j = \frac{1}{\sigma}(I_j - p(w)) \]

and

\[ A_i = \{ j \in \{1, \ldots, n\} : |j - i| \leq m - 1 \} \]

as well as

\[ B_i = \{ j \in \{1, \ldots, n\} : |j - i| \leq 2(m - 1) \}. \]

Then \( X_i \) is independent of \( \sum_{j \notin A_i} X_j \) and \( \sum_{j \in A_i} X_j \) is independent of \( \sum_{j \notin B_i} X_j \). Moreover

\[ \eta_i = \sum_{j : |j - i| \leq m - 1} X_j \] and \( \tau_i = \sum_{j : |j - i| \leq 2(m - 1)} X_j \).
Consider the statement of the theorem,

$$|E h(W) - Nh| \leq 2 \sum_{i=1}^{n} (E|X_i \eta_i \tau_i| + E(X_i \eta_i|E|\tau_i|) + \sum_{i=1}^{n} E|X_i \eta_i^2|.$$ 

We bound $|X_i| \leq \frac{1}{\sigma}$ and $|\eta_i| \leq \frac{2m-1}{\sigma}$ as well as $|\tau_i| \leq \frac{4m-1}{\sigma}$ and hence the bound can be bounded by

$$\frac{(2m - 1)(4(4m - 1) + 1)}{\sigma^3} n.$$
The local approach does not work well when there is (weak) global dependence, for example as in simple random sampling. Instead we make use of couplings: changing one random variable should not have much effect on the other random variables if the dependence is weak. Here we consider two such couplings: the size-bias coupling and the zero-bias coupling. There are more, such as the exchangeable pair coupling.
**Size-Bias coupling: \( \mu > 0 \)**

If \( W \geq 0, \mathbb{E}W > 0 \) then \( W^s \) has the \( W \)-size biased distribution if

\[
\mathbb{E}Wf(W) = \mathbb{E}WEf(W^s)
\]

for all \( f \) for which both sides exist.

**Example:** If \( X \sim \text{Bernoulli}(p) \), then \( \mathbb{E}Xf(X) = pf(1) \) and so \( X^s = 1 \).

**Example:** If \( X \sim \text{Poisson}(\lambda) \), then

\[
\mathbb{P}(X^s = k) = \frac{ke^{-\lambda}\lambda^k}{k!\lambda} = \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!}
\]

and so \( X^s = X + 1 \), where the equality is in distribution.
Using

\[ E W f(W) = E W E f(W) \]

with \( f(x) = ? \) we get that

\[ E W_s = \frac{E W^2}{\mu}. \]
Using

\[ EWf(W) = EWf(W^s) \]

with \( f(x) = x \) we get that

\[ EW^s = \frac{EW^2}{\mu}. \]
Using

\[ EWf(W) = EW Ef(W^s) \]

with \( f(x) = x^2 \) we get that

\[ EW(W^s)^2 = \frac{EW^3}{\mu}. \]
Construction

(Goldstein + Rinott 1997) Suppose $W = \sum_{i=1}^{n} X_i$ with $X_i \geq 0$, $EX_i > 0$, all $i$.

Choose index $V$ proportional to the mean, $EX_v$. If $V = v$: replace $X_v$ by $X_v^s$ having the $X_v$-size biased distribution, independent, and if $X_v^s = x$: adjust $\hat{X}_u, u \neq v$, such that

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

Then $W^s = \sum_{u \neq V} \hat{X}_u + X_V^s$ has the $W$-size bias distribution.

Example: sum of independent Bernoullis. Example: $X_i \sim Be(p_i)$ for $i = 1, \ldots, n$

Then $W^s = \sum_{u \neq V} X_u + 1$. 
How does the size-bias coupling help?

\[ X, X_1, \ldots, X_n \text{ i.i.d. non-negative, } \mathbb{E}X = \mu, \text{Var}X = \sigma^2 \text{ and } W = \sum_{i=1}^{n} X_i. \text{ Then} \]

\[
\begin{align*}
\mathbb{E}(W - \mu)f(W) &= \mu \mathbb{E}(f(W^s) - f(W)) \\
&\approx \mu \mathbb{E}(W^s - W)f'(W) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i^s - X_i)f'(W) \\
&\approx \mu \mathbb{E}f'(W)\mathbb{E}(X^s - \mu) \\
&= \mu \mathbb{E}f'(W) \left\{ \frac{1}{\mu} \mathbb{E}X^2 - \mu \right\} \\
&= \sigma^2 \mathbb{E}f'(W).
\end{align*}
\]
As a theorem

\textit{(Goldstein + Rinott 1997)} Let $W \geq 0$ have mean $\mu > 0$ and variance 1. Let $W^s$ have the $W$–size bias distribution and assume that $W^s$ is defined on the same probability space as $W$. Let $h$ be a smooth function. Then

$$|E h(W - \mu) - E h(Z)|$$

$$\leq (\sup h - \inf h) \mu \sqrt{\text{Var}E(W^s - W|W)} + \frac{1}{6} \|h'\|_1 \mu E[(W^s - W)^2].$$
**Example: sum of i.i.d.'s**

\( X, X_1, \ldots, X_n \) i.i.d. non-negative, \( EX = \mu/n \), \( \text{Var}X = \frac{1}{n} \), and 
\( W = \sum_{i=1}^{n} X_i \). Then

\[
\mathbb{E}(W^s - W | W) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i^s - X_i | W)
\]

\[
= \frac{1}{n} \left( \sum_{i=1}^{n} \frac{nEX^2}{\mu} - \sum_{i=1}^{n} X_i \right) = \frac{nEX^2}{\mu} - \frac{1}{n} W
\]

and so

\[
\text{Var}\mathbb{E}(W^s - W | W) = \frac{1}{n^2}.
\]
Moreover

\[ \mathbb{E}[(W^s - W)^2] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i - X_i^s)^2 \]

\[ = \mathbb{E}X^2 - 2\mu \mathbb{E}X^s + \mathbb{E}(X^s)^2 \]

\[ = \mathbb{E}X^2 - 2n\mathbb{E}X^2 + n\frac{\mathbb{E}X^3}{\mu} \]

\[ \leq n \frac{\mathbb{E}X^3}{\mu}. \]

Thus the bound gives

\[ |\mathbb{E}h(W - \mu) - \mathbb{E}h(Z)| \leq \frac{1}{2} \|h''\| \mu \frac{1}{n} + \frac{1}{6} \|h^{(3)}\| \mathbb{E}X^3 n. \]

We think of \( X \approx n^{-\frac{1}{2}} \); then the overall bound is of order \( n^{-\frac{1}{2}} \).
Example: Vertex degrees in Bernoulli random graphs

Let $G(n, p)$ be a random graph on $n$ vertices where each pair of vertices has probability $p$ of making up an edge, independently. The degree $D(v)$ of a vertex $v$ is the number of its neighbours. Let $W$ count the number of vertices with degree $d$. Then

$$W = \sum_{v=1}^{n} I(D(v) = d).$$

The indicators are not independent - if we know that $D(v) = 0$ for $v = 1, \ldots, n-1$, then $D(n) = 0$ also.
Coupling

Take a random graph and pick a vertex \( v \) at random. Set its degree equal to \( d \). If its degree was \( d \) already, nothing needs to be adjusted. If its degree was \( D(v) > d \) then choose \( D(v) - d \) of its edges uniformly at random and delete them. If the degree was \( D(v) < d \) then choose \( d - D(v) \) vertices among the remaining \( n - D(v) - 1 \) non-neighbours of \( v \) and create an edge between each of these and \( v \).

*Goldstein and Rinott (1997)* calculate the bound; see also *Lin and R. (2012)* for a random graph model where \( p = p(u, v) \) may depend on the vertices.
For mean-zero random variables the size bias coupling is not easy to implement. In particular it would not be applicable to normally distributed random variables. This drawback lead to the definition of the zero-bias coupling.
Zero bias coupling

Let $X$ be a mean zero random variable with finite, nonzero variance $\sigma^2$. 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. (Goldstein and R. 1997)

It is easy to verify that $W^*$ has density

$$p^*(w) = \sigma^{-2}E\{W\mathbf{1}(W > w)\}.$$
Examples

Example: If $X \sim \mathcal{N}(0, \sigma^2)$, then $X$ has the $X$-zero bias distribution and the normal is the only fixed point of the zero bias transformation.

Example: If $X \sim \text{Bernoulli}(p) - p$, then

$$
E\{X \mathbf{1}(X > x)\} = p(1 - p) \text{ for } -p < x < 1 - p
$$

and is zero elsewhere, so $X^* \sim \text{Uniform}(-p, 1 - p)$. 

Connection with Wasserstein distance

We have for $W$ mean zero, variance 1,

$$|E h(W) - Nh| = |E [f'(W) - Wf(W)]|$$
$$= |E [f'(W) - f'(W^*)]|$$
$$\leq ||f''|| E |W - W^*|,$$

where $|| \cdot ||$ is the supremum norm. By (Stein 1986) for $h$ absolutely continuous we have $||f''|| \leq 2||h'||$ and hence

$$|E h(W) - Nh| \leq 2||h'|| E |W - W^*|;$$

thus

$$d_W(\mathcal{L}(W), \mathcal{N}(0,1)) \leq 2E |W - W^*|.$$
Construction: sum of i.i.d.s

\[ W = \sum_{i=1}^{n} X_i, \]  
where the \( X_i \)'s are independent mean zero finite variance \( \sigma_i^2 \) variables: Choose an index \( I \) proportional to the variance, zero bias in that variable,  
\[ W^* = W - X_I + X_I^*. \]

Then, for any smooth \( f \),

\[
\mathbb{E} W f(W) = \sum_{i=1}^{n} \mathbb{E} X_i f(W) = \sum_{i=1}^{n} \mathbb{E} X_i f(X_i + \sum_{t \neq i} X_t) \\
= \sum_{i=1}^{n} \sigma_i^2 \mathbb{E} f'(X_i^* + \sum_{t \neq i} X_t) = \sigma^2 \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2} \mathbb{E} f'(W - X_I + X_I^*) \\
= \sigma^2 \mathbb{E} f'(W - X_I + X_I^*) = \sigma^2 \mathbb{E} f'(W^*),
\]

where we have used independence of \( X_i \) and \( X_t, t \neq i \).
And here is the theorem

Let $X_1, \ldots, X_n$ be independent mean zero variables with variances $\sigma_1^2, \ldots, \sigma_n^2$ and finite third moments, and let $W = (X_1 + \ldots + X_n)/\sigma$ where $\sigma^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Then for all absolutely continuous test functions $h$,

$$|E h(W) - Nh| \leq \frac{2\|h'\|}{\sigma^3} \sum_{i=1}^{n} E \left( |X_i| + \frac{1}{2}|X_i|^3 \right) \sigma_i^2.$$

When the variables are i.i.d. with variance 1,

$$|E h(W) - Nh| \leq \frac{3\|h'\| E|X_1|^3}{\sqrt{n}}.$$
The construction is (much) more involved under dependence. When third moments vanish, fourth moments exist: Order $n^{-1}$ bound.
Also: Berry-Esseen bound, combinatorial central limit theorem (Goldstein 2004).
Some remarks

- Left out:
  1. Exchangeable pair couplings, also used for variance reduction in simulations
  2. Multivariate, also coupling approaches
  3. Infinite-dimensional variables (Malliavin calculus), random measures

- Key feature: bounds in the presence of dependence
- In the i.i.d. case: Berry-Esseen inequality not quite recovered
Further reading

More further reading


We first look at Poisson approximation, then we will put the approximations in a general framework. We then apply the general framework to chisquare distributions as well as one-dimensional Gibbs measures.
The standard reference for this section is Barbour, Holst and Janson (1992).
Recall the Poisson($\lambda$) distribution

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots.$$  

This is a discrete distribution, and for approximations on $\mathbb{Z}^+$ we would usually take total variation distance;

$$d_{TV}(P, Q) = \sup_{A \subseteq \mathbb{Z}^+} |P(A) - Q(A)|.$$
A Stein characterisation

Fact: $Z \sim \text{Po}(\lambda) \iff$ for all functions $f : \mathbb{Z}^+ \to \mathbb{R}$ with $\mathbb{E}f(Z) = 0$

$$\mathbb{E}\lambda g(Z + 1) - Zg(Z) = 0.$$

This suggests to use as test functions indicators $I(j \in A)$ for sets $A \subset \mathbb{Z}^+$ and as Stein equation

$$\lambda g(j + 1) - jg(j) = I(j \in A) - \text{Po}\{\lambda\}(A).$$

For any random variable $W$ we hence have

$$\lambda g(W + 1) - Wg(W) = I(W \in A) - \text{Po}\{\lambda\}(A)$$

and taking expectations

$$d_{TV}(\mathcal{L}(W), \text{Po}\{\lambda\}) \leq \sup_g \left\{ \mathbb{E}\lambda g(W + 1) - Wg(W) \right\}.$$

Here the sup is over all functions $g$ which solve the Stein equation.
The solution of the Stein equation

\[ \lambda g(j + 1) - jg(j) = I(j \in A) - Po\{\lambda\}(A) \]

can be calculated iteratively. We only need the results that

\[ \sup_j |g(j)| \leq \min(1, \lambda^{-\frac{1}{2}}) \]

and

\[ \Delta g := \sup_j |g(j + 1) - g(j)| \leq \min(1, \lambda^{-1}). \]

Here \( \lambda^{-1} \) is often referred to as the magic factor.
Example: sum of independent Bernoulli variables

Let $X_1, X_2, \ldots$ be independent Bernoulli($p_i$) and $W = \sum_{i=1}^{n} X_i$. With $\lambda = \sum_{i=1}^{n} p_i$

$$E W g(W) = \sum E X_i g(W) = \sum E X_i g(W - X_i + 1)$$

$$= \sum p_i E g(W - X_i + 1)$$

and thus

$$E \lambda g(W + 1) - W g(W)$$

$$= \sum p_i E (g(W + 1) - g(W - X_i + 1))$$

$$= \sum p_i^2 E (g(W + 1) - g(W - X_i + 1)|X_i = 1)$$

and

$$|E \lambda g(W + 1) - W g(W)| \leq \Delta g \sum p_i^2 \leq \min(1, \lambda^{-1}) \sum p_i^2.$$
The local approach

Again the argument generalises to local dependence.
Let $X_1, \ldots, X_n$ be Bernoulli random variables, $X_i \sim Be(p_i)$, and put $W = \sum_{i=1}^n X_i$. Let $\lambda = \sum_{i=1}^n p_i$. Suppose that for each $i = 1, \ldots, n$ there exists a set $A_i \subset \{1, \ldots, n\}$ such that $X_i$ is independent of $\sum_{j \notin A_i} X_j$. Define

$$\eta_i = \sum_{j \in A_i} X_j.$$

**Theorem**

*For any smooth $h$ with $\|h'\| \leq 1$, and $\lambda = EW$,*

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \sum_{i=1}^n \left[ (p_i E\eta_i + E(X_i\eta_i)) \right] \min (1, \lambda^{-1}).$$
Example: word counts in DNA sequences

Let $\mathbf{B} = B_1 B_2 \ldots B_n$ be a string of letters which are i.i.d. from the alphabet $\mathcal{A} = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ be a word of length $m$ composed of letters from $\mathcal{A}$. Let $V = V(w)$ be the number of times that $w$ occurs in the sequence $\mathbf{B}$. For $j = 1, \ldots, n - m + 1$ define the word indicators

$$X_j = \begin{cases} 1 & \text{if } B_j B_{j+1} \ldots B_{j+m-1} = w \\ 0 & \text{otherwise.} \end{cases}$$

Let $p(x) = \Pr(B_1 = x)$ for $x = A, C, G, T$, then

$$\mathbb{E}X_j = p(w) := \prod_{i=1}^{m} p(w_i).$$
The neighbourhood of dependence

Put $A_i = \{j \in \{1, \ldots, n\} : |j - i| \leq m - 1\}$. Then $X_i$ is independent of $\sum_{j \notin A_i} X_j$. Moreover $\eta_i = \sum_{j : |j - i| \leq m - 1} X_j$. 
Consider the statement of the theorem,

\[ d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \sum_{i=1}^{n} [(p_i \mathbf{E} \eta_i + \mathbf{E}(X_i \eta_i))] \min(1, \lambda^{-1}). \]

Here \( p_i = p(w) \) for all \( i \) and we think of \( \lambda = O(1) \) so that roughly \( p(w) = O(n^{-1}) \). Moreover \( \mathbf{E} \eta_i \leq (2m-1)p(w) \) and \( \mathbf{E}(X_i \eta_i) = \sum_{j:|j-i| \leq m-1} \mathbf{E}X_iX_j \) depends on the amount of self-overlap in \( w \). If \( w \) does not have any self-overlap then we have \( \mathbf{E}X_iX_j = 0 \) for \( |i-j| \leq m-1 \) and hence

\[ d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq n(2m-1)p(w)^2 \min(1, \lambda^{-1}). \]

If \( w \) has considerable self-overlap then \( \mathbf{E}(X_i \eta_i) \) will not be small and indeed a compound Poisson approximation would be advised.
Example: sequencing by hybridisation

(Arratia et al. 1996) Sequencing by hybridisation is a tool to determine a DNA sequence from the unordered list of all $l$-tuples in the sequence; typically $l = 8, 10, 12$. Assume that the multiset of all $l$-tuples is known. The sequence is then uniquely recoverable unless one of the following occurs in the sequence:

- a non-trivial rotation
- a non-trivial transposition using three way repeats
- a non-trivial transposition using two interleaved pairs of repeats.
Using a Poisson approximation for pairs and three way repeats we can bound the probability of the sequence being uniquely recoverable. For example if all letters are equally likely:
The probability of unique recoverability is at least .95 if
\( l = 8 \) and sequence length is at most 85
\( l = 10 \) and sequence length is at most 469
\( l = 12 \) and sequence length is at most 2,288.
Recall: If $W \geq 0$, $\mathbb{E} W = \lambda$ then $W^s$ has the $W$-size biased distribution if

$$\mathbb{E} W f(W) = \lambda \mathbb{E} f(W^s)$$

for all $f$ for which both sides exist.

Example: sum of independent Bernoullis

Example: $X_i \sim \text{Be}(p_i)$ for $i = 1, \ldots, n$; pick $V \propto p_i$, then $W^s = \sum_{u \neq V} X_u + 1$.

**Theorem**

For the sum of possibly dependent Bernoullis, with $\mathcal{L}(|\hat{X}_j, j \neq i) = \mathcal{L}(X_j, j \neq i | X_i = 1)$,

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) = \left| \sum_{i=1}^{n} p_i (\mathbb{E} g(W + 1) - g(\sum_{j \neq i} \hat{X}_j + 1)) \right|$$

$$\leq \Delta g \sum_{i=1}^{n} p_i \mathbb{E} |W - \sum_{j \neq i} \hat{X}_j|.$$
If the Bernoulli random variables are independent then

$$E|W - \sum_{j \neq i} \hat{X}_j| = EX_i = p_i$$

and we obtain the same bound as from the local approach.
A general framework

Start with a target distribution $\mu$.

1. Find characterization: operator $\mathcal{A}$ such that $X \sim \mu$ if and only if for all smooth functions $f$, $E\mathcal{A}f(X) = 0$

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

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

3. Then for any variable $W$,

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

We usually need to bound $f, f', \text{ or } \Delta f$.

Here $h$ is a smooth test function; for nonsmooth functions: see techniques used by Shao, Chen, Rinott and Rotar, Götze.
The generator approach

**Barbour 1989, 1990; Götze 1993**

Choose $\mathcal{A}$ as generator of a Markov process with stationary distribution $\mu$, that is:
Let $(X_t)_{t \geq 0}$ be a homogeneous Markov process
Put $T_t f(x) = E(f(X_t) | X(0) = x)$
The generator is defined as

$$\mathcal{A}f(x) = \lim_{t \downarrow 0} \frac{1}{t} (T_t f(x) - f(x)).$$
Generator facts

(Ethier and Kurtz (1986), for example)

1. $\mu$ stationary distribution then $X \sim \mu$ if and only if $E Af(X) = 0$ for $f$ for which $Af$ is defined

2. $T_t h - h = A \left( \int_0^t T_u h du \right)$ and formally

$$\int h d\mu - h = A \left( \int_0^\infty T_u h du \right)$$

if the r.h.s. exists.
Examples

1. $Ah(x) = h''(x) - xh'(x)$ is the generator of the *Ornstein-Uhlenbeck* process, stationary distribution $\mathcal{N}(0, 1)$.

2. $Ah(x) = \lambda(h(x + 1) - h(x)) + x(h(x - 1) - h(x))$ or

   $$Af(x) = \lambda f(x + 1) - xf(x)$$

   is the generator of an immigration-death process with immigration rate $\lambda$ and unit per capita death rate; the stationary distribution is Poisson($\lambda$).

   Advantage: generalisations to multivariate, diffusions, measure space...
Chisquare distributions

A generator for $\chi^2_p$ is

$$Af(x) = xf''(x) + \frac{1}{2}(p - x)f'(x).$$

*Luk 1994*: A Stein operator for $\text{Gamma}(r, \lambda)$ is

$$Af(x) = xf''(x) + (r - \lambda x)f'(x)$$

and $\chi^2_p = \text{Gamma}(d/2, 1/2)$; and, for $\chi^2_p$, $A$ 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.
The Stein equation is then

\[(\chi^2_p) \quad h(x) - \chi^2_p h = xf''(x) + \frac{1}{2}(p - x)f'(x)\]

where \(\chi^2_p h\) is the expectation of \(h\) under the \(\chi^2_p\)-distribution.

**Lemma**

(Pickett 2002). Suppose \(h : \mathbb{R} \rightarrow \mathbb{R}\) is absolutely bounded, \(|h(x)| \leq ce^{ax}\) for some \(c > 0\ \ a \in \mathbb{R}\), and the first \(k\) derivatives of \(h\) are bounded. Then the equation \((\chi^2_p)\) has a solution \(f = f_h\) such that

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

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

(Improvement over Luk 1994 in \(\frac{1}{\sqrt{p}}\))
Example: squared sum \((R. + Pickett)\)

\[ X_i, i = 1, \ldots, n, \text{ i.i.d. mean zero, variance one, existing } 8^{th} \text{ moment} \]

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

and

\[ W = S^2 \]

\textit{Pickett + R. (2004):} the bound on the distance to Chisquare(1) for smooth test functions is of order \(\frac{1}{n}\)

Also: Pearson’s chisquare statistic.

\textit{Robert Gaunt:} Variance-gamma distributions.
We consider discrete univariate Gibbs measures $\mu$ with $\text{supp}(\mu) = \{0, \ldots, N\}$, where $N \in \mathbb{N}_0 \cup \{\infty\}$. Such a Gibbs measure can be written as

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

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

Conversely, for a given probability distribution $(\mu(k))_{k \in \mathbb{N}_0}$ we can find a representation as a Gibbs measure by choosing

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

with $V(0) = \log \mu(0) + \log Z$. We have some freedom in the choice of $\omega$ and thus of $V$. Fix a representation.
A Markov process

To each Gibbs measure we associate a Markovian birth-death process with unit per-capita death rate $d_k = k$ and birth rate

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

for $k, k + 1 \in \text{supp}(\mu)$. It is easy to see that this birth-death process has invariant measure $\mu$. 
Following the generator approach to Stein’s method, we would therefore choose as generator

\[(Ah)(k) = (h(k+1) - h(k)) \exp\{V(k+1) - V(k)\} \omega + k(h(k-1) - h(k))\]

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

\[(Af)(k) = f(k+1) \exp\{V(k+1) - V(k)\} \omega - kf(k).\]
Examples

1. The *Poisson-distribution* with parameter $\lambda > 0$: We use $\omega = \lambda$, $V(k) = -\lambda$, $Z = 1$. The Stein-operator is

$$(Af)(k) = f(k + 1) \lambda - kf(k)$$

as before.

2. The *Binomial-distribution* with parameters $n$ and $0 < p < 1$: We use $\omega = \frac{p}{1-p}$, $V(k) = -\log((n - k)!)$, and $Z = (n!(1-p)^n)^{-1}$. The Stein-operator is

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

3. The *Hypergeometric distribution*: The Stein-operator is

$$(Af)(k) = (a - k)(n - k) f(k + 1) - (b - n - x) kf(k).$$
For a given function $h$, a solution $f$ of the Stein equation is given by $f(0) = 0$, $f(k) = 0$ for $k \not\in \text{supp}(\mu)$, and

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

We can bound the solution and their first difference $\Delta f$. 
Exploiting size-biasing

Recall that, for $W \geq 0$ with $EW > 0$ we say that $W^*$ has the $W$-size biased distribution if

$$EWg(W) = EW Eg(W^*)$$

for all $g$ for which both sides exist. In particular we obtain that

$$E\left\{\exp\{V(X + 1) - V(X)\} \omega g(X + 1) - X g(X)\right\}$$

$$= E\left\{\exp\{V(X + 1) - V(X)\} \omega g(X + 1) - E X Eg(X^*)\right\}$$

and

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

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

$$\omega E e^{V(X+1)-V(X)} g(X + 1) = \omega E e^{V(X+1)-V(X)} E g(X^*).$$
And here is the theorem

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

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

where $f$ is the solution of the Stein equation.
**Example: lattice approximation in statistical physics**

*Chayes and Klein (1994)*: Assume $A \subset \mathbb{R}^d$ is a rectangle; denote its volume by $|A|$. Consider the intersection of $A$ with the $d$-dimensional lattice $n^{-1}\mathbb{Z}^d$. For each site $m$ in this intersection we associate a Bernoulli random variable $X^n_m$ which takes value 1, with probability $p^n_m$, if a particle is present at $m$ and 0 otherwise. If the collection $(X^n_m)_m$ is independent, the joint distribution can be interpreted as the Gibbs distribution for an ideal gas on the lattice, and we have Poisson convergence.

We provide bounds on the approximation for such particles while allowing for interactions.
More details

The model is as follows. Pick \( n \in \mathbb{N} \) and suppose that \( A \) can be partitioned into a regular array of \( d(n) \) sub-rectangles \( \{S^n_1, \ldots, S^n_{d(n)}\} \) with volumes \( \nu(S^n_m) = \frac{z^n_m}{z^n_n} \). For each \( m, n \) choose a point \( q^n_m \in S^n_m \). We consider a sequence of functions \( (f_k)_k \) satisfying \( f_0 \equiv 1 \) and for each \( k \geq 1 \), \( f_k(x_1, \ldots, x_k) \) is a nonnegative function, Riemann integrable on \( A^k \), such that \( f_k \) is a symmetric function for each \( k \).
Let $X^n_m, 1 \leq m \leq d(n)$, be 0-1 random variables with joint density function

$$P(X^n_1 = a_1, \ldots, X^n_{d(n)} = a_{d(n)}) \propto f_k(q^n_{i_1}, \ldots, q^n_{i_k}) \prod_{m=1}^{d(n)} (z^n_m)^{a_m}$$

where each $a_i \in \{0, 1\}$ and the indices on the right-hand side are determined by $k = \sum_{m=1}^{d(n)} a_m$. Define $S_n = \sum_{m=1}^{d(n)} X^n_m$. 
Then, due to the symmetry of \( f_k \),

\[
P(S_n = k) = \frac{(z_n)^k}{k!} \sum_{i_1=1}^{d(n)} \cdots \sum_{i_k=1}^{d(n)} f_k(q^n_{i_1}, \ldots, q^n_{i_k}) \prod_{m=1}^{k} \nu(S^n_{i_m}) 
\]

\[
\sum_{k=0}^{d(n)} \frac{(z_n)^k}{k!} \sum_{i_1=1}^{d(n)} \cdots \sum_{i_k=1}^{d(n)} f_k(q^n_{i_1}, \ldots, q^n_{i_k}) \prod_{m=1}^{k} \nu(S^n_{i_m})
\]

Note that we can write this distribution as a Gibbs measure \( \mu_n \).
Let $S$ be a nonnegative integer valued random variable defined by

$$P(S = k) = \frac{z^k}{k!} \int_{A^k} f_k(x_1, \ldots, x_k) dx_1 \cdots dx_k \frac{1}{\sum_{k=0}^{\infty} \frac{z^k}{k!} \int_{A^k} f_k(x_1, \ldots, x_k) dx_1 \cdots dx_k}.$$ 

We write this distribution as a Gibbs measure $\mu$. Using our approach we can bound the total variation distance between $\mu_n$ and $\mu$. 
Comparing distributions via generators

Let $\mu$ have generator $A$ and corresponding $(\omega, V)$, and let $\mu_2$ have generator $A_2$ and corresponding $(\omega_2, V_2)$. Assume that both birth-death processes have unit per-capita death rates. Then, for $X \sim \mu_2$, if the solution $f$ of the Stein equation for $\mu$ is such that $A_2 f$ exists,

$$
E h(X) - \mu(h) = E Af(X) = E (A - A_2) f(X).
$$

Using that $X^*$ has the $X$-size bias distribution,

$$
|E h(X) - \mu(h)| \leq \|f\| E(X) \left\{ \frac{|\omega - \omega_2|}{\omega_2} + \frac{\omega}{\omega_2} E \left| e^{(V(X^*)-V(X^*-1))-(V_2(X^*)-V_2(X^*-1))} - 1 \right| \right\}.
$$
For example, for two Poisson distributions $\text{Poisson}(\lambda_1)$ and $\text{Poisson}(\lambda_2)$ this approach gives

$$\left| \mathbb{E} h(X) - \int h \, d\mu \right| \leq \| f \| |\lambda_1 - \lambda_2| .$$

Note that the normalising constant $Z$ 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.
In the first example, \( X_1, X_2, \ldots, X_n \) i.i.d., \( P(X_i = 1) = p = 1 - P(X_i = 0) \), using Stein’s method we can show that

\[
\sup_x \left| P\left( (np(1 - p))^{-1/2} \sum_{i=1}^{n} (X_i - p) \leq x \right) - P\left( \mathcal{N}(0, 1) \leq x \right) \right| \leq 6 \sqrt{\frac{p(1 - p)}{n}}
\]

and

\[
\sup_x \left| P\left( \sum_{i=1}^{n} X_i = x \right) - P(\text{Po}(np) = x) \right| \leq \min(np^2, p).
\]

So, if \( p < \frac{36}{n+36} \), the bound on the Poisson approximation is smaller than the bound on the normal approximation.
Application areas

- sequence analysis
- random graph statistics, and network statistics
- number theory
- MCMC convergence
- exceedances in time series
- interacting particle systems
- epidemics
- permutation tests
Many more distributions

Exchangeable pair couplings, also used for variance reduction in simulations

Multivariate, also coupling approaches

Infinite-dimensional: random measures, functionals of Gaussian random fields

General distributional transformations

Bounds in the presence of dependence

In the i.i.d. case the Berry-Esseen inequality is not quite recovered.
Further reading

