Convergence in Distribution and Stein's Method

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Convergence in Distribution

What should it mean for us to say that two distributions are *close*, or that a sequence μ_n of distributions of some random variables X_n taking values in some state space $\mathcal{X}$ *converges* to another distribution μ of some random variable X? Certainly we'll need to know if X_n and X are close, so we'll restrict ourselves to state spaces $\mathcal{X}$ that are sigma-compact metric spaces (and to avoid needless technical difficulties we'll take those metric spaces to be complete and separable, or "Polish"— so-called because they were first studied by Sierpiński, Kuratowski, Tarski, and other Polish mathematicians). One approach would be to require that the sequence $\mu_n(B)$ should converge to $\mu(B)$ for some class of Borel sets $B \subset \mathcal{X}$, or that integrals $\mathsf{E}h(X_n) = \int_{\mathcal{X}} h(x)\,\mu_n(dx)$ should converge to $\mathsf{E}h(X) = \int_{\mathcal{X}} h(x)\,\mu(dx)$ for some class of Borel measurable functions $h(x) : \mathcal{X} \to \mathbb{R}$. It is seldom useful to ask that $\mu_n(B)$ converge for *all* Borel B (or that $\int_{\mathcal{X}} h(x)\,\mu_n(dx)$ should converge for too wide a class of functions h). In this section we discuss the most successful and most common notion of convergence, simply called "convergence in distribution", the one that arises in most presentations of the Central Limit Theorem; below in Section (2) we will consider some alternatives, and in Section (3) will look at a recent approach to proving new CLT-like results.

A sequence μ_n of distributions on the Borel sets of any Polish space $\mathcal{X}$ (usually $\mathbb{R}$ or $\mathbb{R}^n$) is said to "converge in distribution" to a distribution μ (written: $\mu_n \Rightarrow \mu$) if

$$\int_{\mathcal{X}} h(x)\,\mu_n(dx) \to \int_{\mathcal{X}} h(x)\,\mu(dx) \tag{1}$$

as $n \to \infty$ for every bounded continuous function $h(\cdot)$ on $\mathcal{X}$. This turns out to be equivalent to requiring only the convergence of $\int h(x)\mu_n(dx)$ to $\int h(x)\mu(dx)$ for smaller classes $\mathcal{D}$ of functions $h(\cdot)$ on $\mathcal{X}$, such as the space $\mathcal{D} = C_0^\infty$ of infinitely-differentiable functions that converge to zero at infinity or, in $\mathcal{X} = \mathbb{R}^d$, just the complex exponentials $\mathcal{D} = \big\{h_\omega(x) = e^{i\omega\cdot x} : \omega \in \mathbb{R}^d\big\}$, so Eqn (1) reduces to the requirement that Fourier transforms converge pointwise. One way to quantify the discrepancy between two distributions μ and ν on $\mathcal{X}$ is

$$D_{\mathcal{D}}(\mu,\nu) = \sup_{h\in\mathcal{D}} \Big| \int_{\mathcal{X}} h(x)\,\mu(dx) - \int_{\mathcal{X}} h(x)\,\nu(dx) \Big|$$

for various classes $\mathcal{D}$; we'll see several examples (and alternatives) in a few weeks.

In the special case of distributions of real-valued random variables, so $\mathcal{X} = \mathbb{R}$, convergence in distribution (1) is equivalent to requiring that the distribution functions

$$\mu_n(-\infty, x] = F_n(x) \to F(x) = \mu(-\infty, x]$$

converge at each point $x \in \mathbb{R}$ where $F(x)$ is continuous (or, equivalently, for any countable dense set of points $\{x_j\} \subset \mathcal{X} = \mathbb{R}$). They might fail to converge where $F(x)$ has jumps (can you give an example?) and, even if each $F_n(\cdot)$ is absolutely continuous, the density functions cannot be expected to converge pointwise. In fact, discrete distributions can converge to a continuous one (examples?) and continuous ones can converge to discrete ones (examples?).

Convergence in distribution can be metrized by the *Lévy-Prokhorov* metric. On a complete separable metric space $(\mathcal{X}, d)$, let

$$\begin{aligned}\pi(\mu,\nu) := \inf\Big\{\epsilon > 0 :\ & \mu(A) \le \nu(A^\epsilon) + \epsilon \\ \text{and} \quad & \nu(A) \le \mu(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(\mathcal{X})\Big\}\end{aligned}$$

where $A^\epsilon = \{y \in \mathcal{X} :\ (\exists x \in A)\ \mathrm{d}(x,y) < \epsilon\} = \cup_{x\in A} B_\epsilon(x)$, the union of open ϵ-balls around each point in A. In $d = 1$ dimension this reduces to the *Lévy metric* for the DFs F and G of μ and ν,

$$\pi(F,G) := \inf\{\epsilon > 0 \mid F(x-\epsilon) \le G(x) \le F(x+\epsilon) \text{ for all } x \in \mathbb{R}\}$$

We now show that convergence in distribution is weaker than any of the other forms of convergence we've seen for random variables.

Proposition 1. *If $X_n \to X$ pr. for some $\mathcal{X}$-valued RVs X_n, X on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, then the distributions $\mu_n = P \circ {X_n}^{-1}$ of X_n converge to that $\mu = \mathsf{P} \circ X^{-1}$ of X.*

In this case we often write "$X_n \Rightarrow X$" rather than the more pedantic $\mu_n \Rightarrow \mu$.

Proof. Let $h : \mathcal{X} \to \mathbb{R}$ be a bounded (say, by $|h(x)| \le B < \infty$) and continuous real-valued function on the σ-compact complete separable metric space $(\mathcal{X}, \mathrm{d})$. Fix $\epsilon > 0$ and (by σ-compactness) let $K \subset \mathcal{X}$ be compact with $\mu(K) > 1 - \epsilon/2B$. Since h is *uniformly* continuous on K, find $\delta > 0$ such that $x \in K$ and $y \in \mathcal{X}$ with $\mathrm{d}(x,y) < \delta$ implies $|h(x) - h(y)| < \epsilon$, and find $N_\epsilon \in \mathbb{N}$ such that

$$(\forall n \ge N_\epsilon) \quad \mathsf{P}[\mathrm{d}(X_n, X) \ge \delta] < \epsilon/2B.$$

Then for $n \ge N_\epsilon$,

$$\begin{aligned}
\left| \int_{\mathcal{X}} h(x)\,\mu_n(dx) - \int_{\mathcal{X}} h(x)\,\mu(dx) \right| &= |\mathsf{E}h(X_n) - \mathsf{E}h(X)| \\
&\le \mathsf{E}|h(X_n) - h(X)| \\
&= \quad \mathsf{E}|h(X_n) - h(X)| \mathbf{1}_{\{X \in K,\ \mathrm{d}(X_n,X) < \delta\}} \\
&\quad + \mathsf{E}|h(X_n) - h(X)| \mathbf{1}_{\{X \notin K,\ \mathrm{d}(X_n,X) < \delta\}} \\
&\quad + \mathsf{E}|h(X_n) - h(X)| \mathbf{1}_{\{\mathrm{d}(X_n,X) \ge \delta\}} \\
&\le (\epsilon)\mathsf{P}[\Omega] + (2B)\mathsf{P}[X \notin K] + (2B)\mathsf{P}[\mathrm{d}(X_n,X) \ge \delta] \\
&\le 3\epsilon \qquad \square
\end{aligned}$$

For another (perhaps simpler) proof, in order show:

- $X_n \to X(pr)$ if and only if every increasing sequence $\mathbb{N} \ni n_k \nearrow \infty$ has a subsequence $n_{k_j} \nearrow \infty$ for which $X_{n_{k_j}} \to X(a.s)$ (by contradiction);
- If $X_n \to X(pr)$ and $Y_n = \phi(X_n)$, $Y = \phi(X)$ for any continuous $\phi : \mathbb{R} \to \mathbb{R}$, then $Y_n \to Y(pr)$;
- If $X_n \to X(pr)$ and $Y_n = \phi(X_n)$, $Y = \phi(X)$ for a *bounded* continuous $\phi : \mathbb{R} \to \mathbb{R}$, then $Y_n \to Y(L_p)$ for any $p < \infty$ (by DCT) and, in particular, $\mu_n = \mathsf{P} \circ {X_n}^{-1} \Rightarrow \mu = \mathsf{P} \circ X^{-1}$.

$\square$

For yet another, show $h(X_n) \to h(X)$ *pr.* and apply the DCT.

Since every notion of convergence of random variables we have seen so far (*pr.*, *a.s*, L_∞, L_p, L_1) impies convergence in probability, *all* of them also imply convergence in distribution. Note that the convergence of random variables' distributions $\mu_n(A) = \mathsf{P}[X_n \in A]$ depends only on the distributions μ_n themselves on the Borel sets $\mathcal{B}(\mathcal{X})$ of the state space— since the random variables $X_n : \Omega_n \to \mathcal{X}$ don't even have to be defined on the same probability space, clearly convergence in distribution cannot imply any of the other convergence notions listed above for random variables. For example, we could set $X_n(\omega) = \omega/n$ on

$$\Omega_n = \{1, 2, \dots, n\}, \qquad \mathcal{F}_n = 2^{\Omega_n}, \qquad \mathsf{P}_n(A) = \#(A)/n$$

to find

$$\mu_n(dx) = \frac{1}{n}\sum_{j=1}^{n} \delta_{j/n}(dx) \Rightarrow \mu(dx) = \mathbf{1}_{\{(0,1]\}}(x)\,\lambda(dx)$$

Although $\{X_n\}$ converges in distribution (to the standard uniform), there *is no* "set of ω on which X_n converge" because the $\{X_n\}$ are all defined on different probability spaces $(\Omega_n, \mathcal{F}_n, \mathsf{P}_n)$.

For $\mathcal{X} = \mathbb{R}$ there is a partial converse, however: if $\mu_n \Rightarrow \mu$ then there exists a probabilty space $(\Omega, \mathcal{F}, \mathsf{P})$ (the unit interval with Borel measure will do) and random variables X_n, X on $(\Omega, \mathcal{F}, \mathsf{P})$ with these distributions for which $X_n \to X$ *a.s.* The construction is simple, by the inverse CDF method.

Example 1 (Empirical DF). *Let* $\{X_n\} \overset{\text{iid}}{\sim} \mu(dx)$ *be independent with an arbitrary common distribution* $\mu(A) = \mathsf{P}[X_n \in A]$*; for* $n \in \mathbb{N}$ *the* empirical distribution *is*

$$\begin{aligned}\mu_n(dx) &= \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}(dx), \text{ i.e.,}\\ \mu_n(A) &= \#\{j \in \{1, ..., n\} : \; X_j \in A\}/n\\ &= \frac{1}{n}\sum \mathbf{1}_A(X_j).\end{aligned}$$

then

$$\mu_n \Rightarrow \mu,$$

i.e., *the empirical distribution converges to the true distribution.*

2 Metrics for Convergence

One way to quantify the discrepancy between two distributions μ and ν on $\mathcal{X}$ is

$$D_{\mathcal{D}}(\mu,\nu) := \sup_{h\in\mathcal{D}} \Big| \int_{\mathcal{X}} h(x)\,\mu(dx) - \int_{\mathcal{X}} h(x)\,\nu(dx) \Big|$$

for various classes $\mathcal{D}$; let's consider several examples and some alternatives.

2.1 Total Variation

The *total variation* distance between two distributions μ, ν on any Polish (*i.e.*, complete separable metric) space $\mathcal{X}$ is given by

$$\begin{aligned} \mathrm{TV}(\mu,\nu) &:= \sup\{|\mu(A)-\nu(A)|\ A\in\mathcal{B}(\mathcal{X})\} \\ &= \inf \mathsf{P}[X\neq Y:\ X\sim\mu,\ Y\sim\nu], \end{aligned} \tag{2}$$

so TV is $D_{\mathcal{D}}$ for $\mathcal{D} = \{\text{ Indicators } h=\mathbf{1}_A\}$ or for $\mathcal{D} = \big\{h:\ |h|\le\frac12\big\}$ or $\mathcal{D}=\{h:\ 0\le h\le 1\}$. This is an exceptionally strong notion of 'closeness', too strong for most applications; for example, every discrete distribution has (maximal) distance one from every continuous distribution. If λ is any sigma-finite measure that dominates both μ and ν (so they will each have a density function, by the Radon-Nikodym theorem), then also

$$\mathrm{TV}(\mu,\nu) = \tfrac12 \int_{\mathcal{X}} \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right| d\lambda,$$

half the L_1-distance between their density functions.

2.2 Hellinger

$$\begin{aligned} H(\mu,\nu) &:= \left\{ \frac12 \int_{\mathcal{X}} \Big(\sqrt{d\mu/d\lambda} - \sqrt{d\nu/d\lambda} \Big)^2 \, d\lambda \right\}^{1/2} \\ &= \left\{ 1 - \int_{\mathcal{X}} \sqrt{(d\mu/d\lambda)(d\nu/d\lambda)}\ d\lambda \right\}^{1/2} \end{aligned}$$

for any measure λ that dominates both μ and ν (for example, one can always take $\lambda=\mu+\nu$; the usual choice is Lebesgue measure when μ and ν have densities, or counting measure when they're both discrete. Just as with $\mathrm{TV}(\mu,\nu)$, the value of $H(\mu,\nu)$ doesn't depend on what λ is used). Hellinger and Total Variation determine the same topology, *i.e.*, so a sequence converges in one if and only if it does in the other.

2.3 Kolmogorov-Smirnov

For $\mathcal{X} = \mathbb{R}$,

$$\mathrm{KS}(\mu,\nu) := \sup_{x\in\mathbb{R}} \big\{ \big|\mu\big((-\infty,x]\big) - \nu\big((-\infty,x]\big)\big| \big\}$$

Obviously KS is $D_{\mathcal{D}}$ for $\mathcal{D} = \big\{\mathbf{1}_{(-\infty,x]} : x\in\mathbb{R}\big\}$. Kolmogorov (1933) and Smirnov (1939) famously (and independently) showed that $1/\sqrt{n}$ times the KS distance from any continuous distribution to the empirical distribution for n iid replicates has the same distribution, which converges asymptotically to that of the maximum of the standard Brownian bridge stochastic process, leading to an omnibus non-parametric test of the hypothesis $\{X_i\} \overset{\text{iid}}{\sim} \mu(dx)$. If one of μ, ν has a point mass where the other doesn't, then their K-S distance will be at least the size of that atom; this makes it a poor choice in some applications.

2.4 (Lévy-)Prokhorov

On a complete separable metric space $(\mathcal{X}, \mathrm{d})$,

$$\begin{aligned}\pi(\mu,\nu) := \inf\Big\{\epsilon > 0 :\ & \mu(A) \le \nu(A^\epsilon) + \epsilon \\ \text{and}\quad & \nu(A) \le \mu(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(\mathcal{X})\Big\}\end{aligned}$$

where $A^\epsilon = \{y \in \mathcal{X} :\ (\exists x \in A)\ \mathrm{d}(x,y) < \epsilon\} = \cup_{x\in A} B_\epsilon(x)$, the union of open ϵ-balls around each point in A. This exactly metrizes convergence in distribution (*i.e.*, a sequence $\mu_n \Rightarrow \mu$ if and only if $\pi(\mu_n,\mu) \to 0$); every other metric in this section is strictly stronger, in the sense that convergence of μ_n to μ in that metric implies (but is not implied by) convergence in distribution.

2.5 Wasserstein

On a complete separable metric space $(\mathcal{X}, \mathrm{d})$, for $p \ge 1$ the *Wasserstein* distance between two distributions is

$$\begin{aligned}\mathrm{Wass}_p(\mu,\nu) &:= \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \iint_{\mathcal{X}\times\mathcal{X}} \mathrm{d}(x,y)^p\,\gamma(dx\,dy)\right)^{1/p} \\ &= \inf\{\|X - Y\|_p :\ X\sim\mu,\ Y\sim\nu\}\end{aligned}$$

where $\Gamma(\mu,\nu)$ is the space of probability measures on $\mathcal{X}^2$ with marginals μ and ν. The case $p=1$ is most important:

$$\mathrm{Wass}_1(\mu,\nu) = \sup_{f:\mathcal{X}\to\mathbb{R}} \left\{ \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right| :\ \mathrm{Lip}(f) \le 1 \right\},$$

displaying Wass_1 on $\mathbb{R}$ as $D_{\mathcal{D}}$ for $\mathcal{D} = \{\text{Unit Lipschitz continuous } h(\cdot)\}$. Sometimes called the "transportation metric," this can be interpreted as the minimum cost of moving the support of μ to that of ν if moving cost is proportional to the product of mass times distance.

2.6 Kullback-Leibler Divergence

The "Kullback-Leibler divergence" (Kullback and Leibler, 1951), also called Relative Entropy, from distribution μ to ν on a Polish space $\mathcal{X}$ is:

$$\mathrm{KL}(\mu\|\nu) := \int_{\mathcal{X}} -\log\left[\frac{\nu(dx)}{\mu(dx)}\right]\,\mu(dx),$$

when $\nu \ll \mu$ and the integral is finite (otherwise $\mathrm{KL}(\mu\|\nu)=\infty$). It is non-negative, because $\log y \le y-1$ for all $y>0$ (or by Jensen's inequality), but it is not symmetric in μ and ν and doesn't satisfy the triangle inequality so it can't be a metric. It does determine a topology, though, and hence a notion of convergence. Some authors (including Kullback and Leibler themselves, also Bernardo) construct symmetric analogues, like the "symmetric," "Jensen-Shannon," and "Intrinsic" divergences

$$\begin{aligned}
\mathrm{KL}_{\mathrm{sym}}(\mu,\nu) &:= \mathrm{KL}(\mu\|\nu) + \mathrm{KL}(\nu\|\mu),\\
\mathrm{KL}_{\mathrm{JS}}(\mu,\nu) &:= \tfrac12\left\{\mathrm{KL}\left(\frac{\mu+\nu}{2}\Big\|\mu\right) + \mathrm{KL}\left(\frac{\mu+\nu}{2}\Big\|\nu\right)\right\},\\
\mathrm{KL}_{\mathrm{Int}}(\mu,\nu) &:= \min\{\mathrm{KL}(\mu\|\nu),\ \mathrm{KL}(\nu\|\mu)\}.
\end{aligned}$$

Note $\mathrm{KL}_{\mathrm{JS}}$ is always finite, and $\mathrm{KL}_{\mathrm{Int}}$ is finite if either $\mathrm{KL}(\mu\|\nu)$ or $\mathrm{KL}(\nu\|\mu)$ is, but $\mathrm{KL}(\mu\|\nu)$ (and hence $\mathrm{KL}_{\mathrm{sym}}(\mu,\nu)$) will be infinite unless $\nu \ll \mu$.

2.7 (Fisher) Information Distance

For parametric families of measures $\{\mu_\theta :\ \theta\in\Theta\}$ on some space $\mathcal{X}\subset\mathbb{R}^d$ (particularly within the exponential family), with density functions $\mu_\theta(dx) = f(x\mid\theta)\,\nu(dx)$ with respect to some dominating reference measure ν, let

$$\begin{aligned}
I(\theta) :=& \mathsf{E}\left\{-\nabla^2 \log f(X\mid\theta)\right\}\\
=& \mathsf{E}\left\{(\nabla \log f(X\mid\theta))\,(\nabla\log f(X\mid\theta))'\right\}
\end{aligned}$$

be the Fisher information matrix and construct a Riemannian metric on Θ by

$$\mathrm{d}(\theta_0,\theta_1) := \inf_\gamma \left\{ \int_0^1 \sqrt{\dot\gamma_s' I(\gamma_s)\dot\gamma_s}\; ds \right\}$$

where the infimum is over all differentiable paths $\gamma: [0,1] \to \Theta$ connecting $\gamma_0 = \theta_0$ to $\gamma_1 = \theta_1$ and where $\dot\gamma = d\gamma_s/ds$. In one dimension when $\Theta \subseteq \mathbb{R}$ is a (possibly infinite) interval, this is just

$$=\pi_J\big([\theta_0,\theta_1]\big)$$

for the Jeffreys'-rule prior distribution π_J. In any number of dimensions, the Fisher Information distance on Θ induces a notion of distance for distributions, by

$$\mathrm{FI}\big(\mu_{\theta_0},\mu_{\theta_1}\big) := \mathrm{d}(\theta_0,\theta_1).$$

See Amari (2001) or Amari and Nagoaka (2000, §2.2) for more details.

2.8 Relations among the Metrics

- $\mathrm{TV}(\mu,\nu) \le H(\mu,\nu) \le \sqrt{2\mathrm{TV}(\mu,\nu)}$
- $H^2(\mu,\nu) \le \frac{1}{2}\mathrm{KL}_{\mathrm{Int}}(\nu\|\mu) \le \frac{1}{2}\mathrm{KL}(\nu\|\mu) \le \frac{1}{2}\mathrm{KL}_{\mathrm{sym}}(\nu\|\mu)$
- $\mathrm{KS}(\mu,\nu) \le \mathrm{TV}(\mu,\nu)$
- $\mathrm{KS}(\mu,\nu) \le 2\sqrt{c\,\mathrm{Wass}_1(\mu,\nu)}$ if μ or ν has a pdf bounded by c
- $\mathrm{KL}(\mu,\nu) \approx \frac{1}{2}\mathrm{FI}(\mu,\nu)^2$ for $\mu \approx \nu$ (so their topologies coincide)

Thus $\mu_n \to \mu$ (TV) if and only if $\mu_n \to \mu$ (H). Either of these implies that $\mu_n \to \mu$ (KS), and both are implied by $\mu_n \to \mu$ (KL) or (equivalently, when FI exists) $\mu_n \to \mu$ (FI).

3 Stein's Method

Let $h(\cdot)$ be a continuously differentiable function on $\mathbb{R}$ that doesn't grow too fast as $z \to \pm\infty$ and let $\phi(dz)$ be the standard Normal distribution measure, with Lebesgue density function $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$. Let $Z \sim \mathsf{No}(0,1)$ be a random variable with the standard normal distribution $\phi(dz)$. Then integrating by parts shows

$$\begin{aligned}\int_{-\infty}^{x} h'(z)\phi(dz) &= h(z)\phi(z)\Big|_{-\infty}^{x} - \int_{-\infty}^{x} h(z)\phi'(dz)\\ &= h(x)\phi(x) + \int_{-\infty}^{x} z\, h(z)\phi(dz)\end{aligned}$$

and in the limit as $x \to \infty$ we have

$$\mathsf{E}[h'(Z) - Z\, h(Z)] = 0 \tag{3}$$

for every smooth tame function h. The same argument in reverse shows that if (3) holds for every $h \in C_o^\infty$ then Z must have the $\mathsf{No}(0,1)$ distribution. Perhaps any random variable Z for which Eqn (3) holds *approximately* will have an approximately Normal distribution. In 1972 Charles Stein showed that Eqn (3) characterizes the standard normal distribution completely, and found a way to use this to bound $\big|\int_{\mathbb{R}} h(z)\mu(dz) - \int_{\mathbb{R}} h(z)\phi(dz)\big|$ for various distributions $\mu(dz)$, including that for the sample mean of n iid random variables, leading to a new way of viewing and proving the Central Limit Theorem and to new error bounds for it.

Over the next few decades a wide range of new applications and extensions of the method have emerged (Stein, 1986; Reinert, 2003, 2005; Chatterjee, 2007). For example, an integer-valued random variable Y has the $\mathsf{Po}(\lambda)$ Poisson distribution with mean λ if and only if

$$\mathsf{E}[\lambda h(Y+1) - Y h(Y)] = 0 \tag{4}$$

for all real-valued functions $h(\cdot)$ that don't grow too fast for the expectations to be well-defined. More generally (we'll see that Eqns (3, 4) are special cases) if X_t is a continuous-time stationary Markov process with stationary distribution μ and generator $\mathfrak{A}$, then $X \sim \mu$ if and only if

$$\mathsf{E}[\mathfrak{A} h(X)] = 0 \tag{5}$$

for each h in the domain of $\mathfrak{A}$.

3.1 Stein in Brief

Stein's method for bounding some notion of the distance between the distribution ν of some random variable $W \sim \nu$ and a target distribution μ (and let $Z \sim \mu$ have the target distribution) was summarized by Reinert (2003) as comprising three steps:

1. Find an operator $\mathfrak{A}$ such that $X \sim \mu$ if and only if $\mathsf{E}[\mathfrak{A}h(X)] = 0$ for all suitable h;

2. For each suitable h find a solution $f = f_h$ of *Stein's Equation*

$$h(x) - \mathsf{E}h(Z) = \mathfrak{A}f_h(x) \tag{6}$$

3. Note that this implies

$$\mathsf{E}h(W) - \mathsf{E}h(Z) = \mathsf{E}\mathfrak{A}f_h(W)$$

and hence

$$\Big|\int_{\mathcal{X}} h(x)\mu(dx) - \int_{\mathcal{X}} h(x)\nu(dx)\Big| = \big|\mathsf{E}\mathfrak{A}f_h(W)\big|, \tag{7}$$

so if we can find a way to bound $|\mathfrak{A}f_h(x)|$ uniformly for some suitable class $\mathcal{D} = \{h\}$ of functions, then we can verify distance bounds.

3.2 Continuous-time Markov Chains

Let X_t be a stationary Markov process indexed by $t \in \mathbb{R}_+$ with stationary distribution μ and define a family of operators for $t \geq 0$ and suitable h by:

$$T_t h(x) = \mathsf{E}[h(X_t) \mid X_0 = x].$$

Then $T_0 h(x) = h(x)$ and by the Markov property $\mathsf{E}[T_s(X_t) \mid X_0 = x] = T_{t+s}h(x)$, so $\{T_t\}$ is a continuous *semigroup* of operators. If we define

$$\mathfrak{A}h(x) = \lim_{t \searrow 0} \frac{1}{t}\big[T_t h(x) - h(x)\big],$$

the derivative of T_t at $t = 0$ in the direction h, then formally T_t satisfies

$$T_t h(x) = h(x) + \int_0^t \mathfrak{A}T_s h(x)\, ds \tag{8}$$

and so in some sense we can think of T_t as the exponential $T_t = e^{t\mathfrak{A}}$. Since X_t has stationary distribution μ, then (if LDC or UI apply) we should have

$$\lim_{t\to\infty} T_t h(x) = \lim_{t\to\infty} \mathsf{E} h(X_t) = \int_{\mathcal{X}} h(z)\mu(dz)$$

for every x. If the integral exists the limit of Eqn (8) should give (after rearranging):

$$h(x) - \mathsf{E}h(Z) = \mathfrak{A} \int_0^\infty T_s h(x)\, ds,$$

exactly Stein's Eqn (6) with the formal solution

$$\begin{aligned} f_h(x) &= \int_0^\infty T_s h(x)\, ds \\ &= \lim_{\lambda\searrow 0} \int_0^\infty e^{-\lambda s} T_s h(x)\, ds, \end{aligned}$$

formally the resolvent limit (recall $T_s \approx e^{s\mathfrak{A}}$)

$$= \lim_{\lambda\searrow 0} (-\mathfrak{A} + \lambda I)^{-1} h(x),$$

or something like "$f_h = -\mathfrak{A}^{-1} h$." Let's see some examples.

3.2.1 A Stationary Markov Chain with the Poisson Distribution

For $\lambda > 0$ consider a continuous-time integer-valued process Markov X_t that evolves by

$$\mathsf{P}[X_{t+\epsilon} = j \mid X_t = i] = o(\epsilon) + \begin{cases} \lambda\epsilon & i = j+1 \\ 1 - (\lambda + i)\epsilon & i = j \\ i\epsilon & i = j-1 \end{cases}$$

for non-negative $i, j \in \mathbb{Z}_+$. This "linear death with immigration" process has generator

$$\begin{aligned} \mathfrak{A}h(x) &= \lambda[h(x+1) - h(x)] - x[h(x) - h(x-1)] \\ &= \lambda h(x+1) - (\lambda + x)h(x) + x h(x-1) \end{aligned}$$

A random variable X satisfies $\mathsf{E}[\mathfrak{A}h(X)] = 0$ for each h if and only if it satisfies $\mathsf{E}[\mathcal{A}g(X)] = 0$ with $\mathcal{A}g(x) := [\lambda g(x+1) - x g(x)]$ for each g (including the first-difference $g(x) = h(x) - h(x-1)$), which happens in turn if and only if $\mathsf{P}[X=k]=(\lambda/k)\mathsf{P}[X=k-1]$ for each $k\ge 1$, *i.e.*, if $X\sim\mathsf{Po}(\lambda)$. Stein's Equation is solved recursively by $f_h(0)=0$ and, for integers $x\ge 0$,

$$f_h(x+1)=\frac{1}{\lambda}\left[h(x)+xf_h(x)-e^{-\lambda}\sum_{k=0}^{\infty}h(k)\lambda^k/k!\right]$$

3.2.2 A Stationary Markov Process with the Normal Dist'n

The standard Ornstein-Uhlenbeck Gaussian process with mean zero and covariance $\mathsf{E}[X_sX_t]=e^{-|s-t|}$ is a diffusion with Itô SDE representation

$$dX_t=a(X_t)\,dt+b(X_t)\,dW_t=-X_t\,dt+\sqrt{2}\,dW_t$$

and hence with generator

$$\mathfrak{A}=a(x)\frac{\partial}{\partial x}+\tfrac{1}{2}b^2(x)\frac{\partial^2}{\partial x^2}=-x\frac{\partial}{\partial x}+\frac{\partial^2}{\partial x^2}. \tag{9}$$

A random variable X satisfies "$\mathsf{E}\mathfrak{A}h(X)=0$" for each smooth function h whenever it satisfies "$\mathsf{E}\mathcal{A}g(X)=0$" for each smooth function g (including $g=h'$), for the first-order operator $\mathcal{A}=-x+\frac{\partial}{\partial x}$. Stein's Equation is solved in proving:

Lemma 1. (Stein) *Let $g:\mathbb{R}\to\mathbb{R}$ be bounded and measurable and denote by $Ng=\int g(z)\,\phi(dz)$ the expectation $\mathsf{E}g(Z)$ where $Z\sim\mathsf{No}(0,1)=\phi(dz)$. Then there exists an absolutely-continuous function $f:\mathbb{R}\to\mathbb{R}$ satisfying Stein's Eqn (6)*

$$f'(x)-x\,f(x)=g(x)-Ng$$

and moreover the bounds

$$|f|_\infty\le\sqrt{\pi/2}\;|g-Ng|_\infty\qquad |f'|_\infty\le 2\;|g-Ng|_\infty.$$

Proof. Note $[f(x)\phi(x)]'=[f'(x)-xf(x)]\phi$, so integrating from minus infinity shows the solution must satisfy

$$f(x)\phi(x)=\int_{-\infty}^{x}[g(z)-Ng]\phi(dz),\ \text{so it is}$$

$$f(x)=\frac{\int_{-\infty}^{x}g(z)\phi(dz)-\Phi(x)Ng}{\phi(x)}.$$

The indicated bounds follow from

$$\sup_{x\ge 0}\frac{\Phi(-x)}{\phi(x)}=\sqrt{\pi/2}\qquad\sup_{x\ge 0}\frac{x\Phi(-x)}{\phi(x)}=1$$

(the maxima occur at zero and ∞, respectively). □

3.2.3 CLT as a Stationary Markov Chain

Let $\{\xi_j\}$ be independent L_3 random variables that have been centered and scaled (if necessary) so that $\mathsf{E}\xi_j = 0$ and $\mathsf{E}\xi_j^2 = 1$ for each j; they need not be identically-distributed. First construct a discrete-time Markov chain as follows: For $n \in \mathbb{N}$, set

$$Z_n(0) = (\xi_1, \dots, \xi_n)\,.$$

Pick an index $i \in I_n = \{1, \dots, n\}$ uniformly, a random variable ξ_i^* with the same distribution as ξ_i but independent of everything, and set

$$Z_n(1) = (\xi_1, \dots, \xi_{i-1}, \xi_i^*, \xi_{i+1}, \dots, \xi_n)\,,$$

i.e., $Z_n(0)$ with ξ_i replaced by ξ_i^*. Continue in this fashion to construct a stationary $\mathbb{R}^n$-valued Markov chain. Meanwhile let N_t be a Poisson process with rate n and construct a continuous-time $\mathbb{R}^1$-valued process W_t as:

$$W_t = \frac{1}{\sqrt{n}} \sum_{j \in I_n} Z_n\big(N_t\big).$$

Clearly W_t is a stationary Markov process, whose distribution coincides with that of

$$W_0 = \big(\xi_1 + \cdots + \xi_n\big)/\sqrt{n}.$$

How far is that distribution from the standard Normal? By Taylor's theorem, the generator for this process on a smooth function h is:

$$\begin{aligned}
\mathfrak{A}_n h(w) &= \lim_{t \searrow 0} \frac{n}{t} \mathsf{E}\big\{h(W_t) - h(W_0) \mid W_0 = w\big\} \\
&= \sum_{j \in I_n} \mathsf{E}\Big\{h\big(w + \frac{\xi_j^* - \xi_j}{\sqrt{n}}\big) - h(w)\Big\} \\
&= \sum_{j \in I_n} \mathsf{E}\left\{\frac{\xi_j^* - \xi_j}{\sqrt{n}} h'(w) + \left[\frac{\xi_j^* - \xi_j}{\sqrt{n}}\right]^2 \frac{h''(w)}{2} + \left[\frac{\xi_j^* - \xi_j}{\sqrt{n}}\right]^3 \frac{h'''(w^*)}{6}\right\}
\end{aligned}$$

for some w^* near w; since $\mathsf{E}\xi_j^* = 0$ and $\mathsf{E}\xi_j^{*2} = 1$,

$$\begin{aligned}
&= \sum_{j \in I_n} \Big\{\frac{-\xi_j}{\sqrt{n}} h'(w) + \frac{1 + \xi_j^2}{2n} h''(w) + o(1/n)\Big\} \\
&\approx -w h'(w) + h''(w)
\end{aligned}$$

by the definition of $W_0 = w$ and the LLN. This is the same second-order generator $\mathfrak{A}$ found in Eqn (9); again, we may use the first-order $\mathcal{A} = \frac{\partial}{\partial x} - x$ instead.

3.3 The Central Limit Theorem

Theorem 1. *Suppose $\{\xi_j\}$ are independent with mean zero, variance one, and finite third moments. Set*

$$W_n = (\xi_1 + \cdots + \xi_n)/\sqrt{n}.$$

Then the Wasserstein distance from the distribution μ_n of W_n to the standard normal distribution is bounded by

$$\text{Wass}_1(\mu_n, \phi) \le \frac{3}{n^{3/2}} \sum_{j=1}^{n} \mathsf{E}|\xi_j|^3$$

As an obvious corollary, $\mu_n \Rightarrow \mathsf{No}(0,1)$, and for iid $\{\xi_j\}$ the convergence rate is $n^{-1/2}$. This is similar to (but slightly weaker than) the usual Berry-Esséen bound on $\text{KS}(\mu_n, \phi)$; with a little more work (involving careful estimates of f_h for $h = \mathbf{1}_{(-\infty,x]}$) the traditional Berry-Esséen bounds are available from Stein's Method.

References

Amari, S.-I. (2001), "Information Geometry on Hierarchy of Probability Distributions," *IEEE Transactions on Information Theory*, 47, 1701–1711.

Amari, S.-I. and Nagoaka, H. (2000), *Methods of Information Geometry*, *Translations of mathematical monographs*, volume 191, American Mathematical Society and Oxford University Press, translated from 1993 Japanese version by Daishi Harada.

Barbour, A. D. and Chen, L. H. Y., eds. (2005), *An Introduction to Stein's Method*, *Lecture Note Series, Institute for Mathematical Sciences*, volume 4, Singapore: National University of Singapore.

Chatterjee, S. (2007), *Stein's method and applications*, Berkeley, CA, on-line at `http://www.stat.berkeley.edu/~sourav/stat206Afall07.html`.

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, DE: Springer-Verlag, english translation (1950): *Foundations of the theory of probability*. Chelsea, New York.

Kullback, S. and Leibler, R. A. (1951), "On Information and sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Reinert, G. (2003), "Stein's Method for Chisquare Approximations, Weak Law of Large Numbers, and Discrete Distributions from a Gibbs View Point," Technical report, Dept. of Statistics, University of Oxford, lecture notes for 2003 Program on *Stein's Method and Applications: A Program in Honor of Charles Stein* held Jul 28–Aug 31, 2003 at the National University of Singapore (NUS).

Reinert, G. (2005), "Three general approaches to Stein's method," in Barbour and Chen (2005), pp. 183–222.

Smirnov, V. I. (1939), "On the estimation of the discrepancy between empirical curves of distribution for two independent samples (in Russian)," *Byull. Moskov. Gos. Univ. Ser. A*, 2, 3–16.

Stein, C. (1972), "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, eds. L. M. Le Cam, J. Neyman, and E. L. Scott, Berkeley, CA: University of California Press, volume 2, pp. 583–602.

Stein, C. (1986), *Approximate computation of expectations*, *IMS Lecture Notes-Monograph Series*, volume 7, Hayward, CA: Institute of Mathematical Statistics.