Lecture 2

Lecture date: Aug 29, 2007 *Scribe: David Rosenberg*

1 Distances between probability measures

Stein's method often gives bounds on how close distributions are to each other.

A typical distance between probability measures is of the type

$$d(\mu, \nu) = \sup \left\{ \left| \int f d\mu - \int f d\nu \right| \, : \, f \in \mathcal{D} \right\},$$

where $\mathcal{D}$ is some class of functions.

1.1 Total variation distance

Let $\mathcal{B}$ denote the class of Borel sets. The total variation distance between two probability measures μ and ν on $\mathbb{R}$ is defined as

$$\mathrm{TV}(\mu, \nu) := \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)| \,.$$

Here

$$\mathcal{D} = \{1_A \, : \, A \in \mathcal{B}\} \,.$$

Note that this ranges in $[0, 1]$. Clearly, the total variation distance is not restricted to the probability measures on the real line, and can be defined on arbitrary spaces.

1.2 Wasserstein distance

This is also known as the Kantorovich-Monge-Rubinstein metric.

Defined only when probability measures are on a metric space.

$$\mathrm{Wass}(\mu, \nu) := \sup \left\{ \left| \int f \, d\mu - \int f \, d\nu \right| : f \text{ is 1-Lipschitz} \right\},$$

i.e. sup over all f s.t. $|f(x) - f(y)| \leq d(x, y)$, d being the underlying metric on the space. The Wasserstein distance can range in $[0, \infty]$.

1.3 Kolmogorov-Smirnov distance

Only for probability measures on $\mathbb{R}$.

$$\begin{aligned}\mathrm{Kolm}(\mu,\nu) &:= \sup_{x\in\mathbb{R}} |\mu((-\infty,x]) - \nu((-\infty,x])| \\ &\le \mathrm{TV}(\mu,\nu).\end{aligned}$$

1.4 Facts

- All three distances defined above are stronger than weak convergence (i.e. convergence in distribution, which is weak* convergence on the space of probaility measures, seen as a dual space). That is, if any of these metrics go to zero as $n\to\infty$, then we have weak convergence. But converse is not true. However, weak convergence is metrizable (e.g. by the Prokhorov metric).

- Important coupling interpretation of total variation distance:

  $$TV(\mu,\nu) = \inf\{P(X\neq Y) \,:\, (X,Y) \text{ is a r.v. s.t. } X\sim\mu, Y\sim\nu\}$$

  (i.e. infimum over all joint distributions with given marginals.)

- Similarly, for μ,ν on the real line,

  $$\mathrm{Wass}(\mu,\nu) = \inf\{\mathbf{E}\,|X-Y| : (X,Y) \text{ is a r.v. s.t. } X\sim\mu, Y\sim\nu\}$$

  (So it's often called the Wass_1, because of L_1 norm.)

- TV is a very strong notion, often too strong to be useful. Suppose $X_1, X_2, \dots$ iid ±1. $S_n = \sum_1^n X_i$. Then

  $$\frac{S_n}{\sqrt{n}} \Longrightarrow N(0,1)$$

  But $TV(\frac{S_n}{\sqrt{n}}, Z) = 1$ for all n. Both Wasserstein and Kolmogorov distances go to 0 at rate $1/\sqrt{n}$.

Lemma 1 *Suppose W,Z are two r.v.'s and Z has a density w.r.t. Lebesgue measure bounded by a constant C. Then* $\mathrm{Kolm}(W,Z) \le 2\sqrt{C\mathrm{Wass}(W,Z)}$.

Proof: Consider a point t, and fix an ϵ. Define two functions g_1 and g_2 as follows. Let $g_1(x) = 1$ on $(-\infty,t)$, 0 on $[t+\epsilon,\infty)$ and linear interpolation in between. Let $g_2(x) = 1$ on $(-\infty, t-\epsilon]$, 0 on $[t,\infty)$, and linear interpolation in between. Then g_1 and g_2 form upper and lower 'envelopes' for $1_{(-\infty,t]}$. So

$$P(W\le t) - P(Z\le t) \le \mathbf{E}\,g_1(W) - \mathbf{E}\,g_1(Z) + \mathbf{E}\,g_1(Z) - P(Z\le T).$$

Now $\mathbf{E}\, g_1(W) - \mathbf{E}\, g_1(Z) \leq \frac{1}{\epsilon}\text{Wass}(W,Z)$ since g_1 is $(1/\epsilon)$-Lipschitz, and $\mathbf{E}\, g_1(Z) - P(Z \leq t) \leq C\epsilon$ since Z has density bdd by C.

Now using g_2, same bound holds for the other side: $P(Z \leq t) - P(W \leq t)$. Optimize over ϵ to get the required bound. □

1.5 A stronger notion of distance

Exercise 1: S_n a simple random walk (SRW). $S_n = \sum_1^n X_i$, with X_i iid ± 1. Then

$$\frac{S_n}{\sqrt{n}} \Longrightarrow Z \sim N(0,1).$$

The Berry-Esseen bound: Suppose $X_1, X_2, \ldots$ iid $\mathbf{E}(X_1) = 0, \mathbf{E}(X_1^2) = 1, \mathbf{E}\,|X|^3 < \infty$. Then

$$\text{Kolm}\left(\frac{S_n}{\sqrt{n}}, Z\right) \leq \frac{3\,\mathbf{E}\,|X_1|^3}{\sqrt{n}}$$

Can also show that for SRW,

$$\text{Wass}\left(\frac{S_n}{\sqrt{n}}, Z\right) \leq \frac{Const}{\sqrt{n}}$$

This means that it is possible to construct $\frac{S_n}{\sqrt{n}}$ and Z on the same space such that

$$\mathbf{E}\left|\frac{S_n}{\sqrt{n}} - Z\right| \leq \frac{C}{\sqrt{n}}$$

Can we do it in the strong sense? That is:

$$P\left(\left|\frac{S_n}{\sqrt{n}} - Z\right| > \frac{t}{\sqrt{n}}\right) \leq Ce^{-ct}.$$

This is known as Tusnády's Lemma. Will come back to this later.

2 Integration by parts for the gaussian measure

The following result is sometimes called 'Stein's Lemma'.

Lemma 2 *If $Z \sim N(0,1)$, and $f : \mathbb{R} \to \mathbb{R}$ is an absolutely continuous function such that $\mathbf{E}\,|f'(Z)| < \infty$, then $\mathbf{E}\, Zf(Z) = \mathbf{E}\, f'(Z)$.*

Proof: First assume f has compact support contained in (a,b). Then the result follows from integration by parts:

$$\int_a^b xf(x)e^{-x^2/2}dx = \left[f(x)e^{-x^2/2}\right]_a^b + \int_a^b f'(x)e^{-x^2/2}\,dx.$$

Now take any f s.t. $\mathbf{E}\,|Zf(Z)| < \infty, \mathbf{E}\,|f'(Z)| < \infty, \mathbf{E}\,|f(Z)| < \infty$.

Take a piecewise linear function g that takes value 1 in $[-1,1]$, 0 outside $[-2,2]$, and between 0 and 1 elsewhere. Let

$$f_n(x) := f(x)g(x/n).$$

Then clearly,

$$|f_n(x)| \leq |f(x)| \text{ for all } x \text{ and } f_n(x) \to f(x) \text{ pointwise.}$$

Similarly, $f_n' \to f'$ pointwise. Rest follows by DCT. The last step is to show that the finiteness of $\mathbf{E}\,|f'(Z)|$ implies the finiteness of the other two expectations.

Suppose $\mathbf{E}\,|f'(Z)| < \infty$. Then

$$\begin{aligned}\int_0^\infty |xf(x)|\,e^{-x^2/2}\,dx &\leq \int_0^\infty x\int_0^x \left|f'(y)\right|\,dy\,e^{-x^2/2}\,dx\\ &= \int_0^\infty \left|f'(y)\right| \underbrace{\int_y^\infty xe^{-x^2/2}dx}_{e^{-y^2/2}}\,dy.\end{aligned}$$

Finiteness of $\mathbf{E}\,|f(Z)|$ follows from the inequality $|f(x)| \leq \sup_{|t|\leq 1}|f(t)| + |xf(x)|$. $\square$

Exercise 2: Find f s.t. $\mathbf{E}\,|Zf(Z)| < \infty$ but $\mathbf{E}\,|f'(Z)| = \infty$.

Next time, Stein's method. Sketch:

Suppose you have a r.v. W and $Z \sim N(0,1)$ and you want to bound

$$\sup_{g\in\mathcal{D}} |\mathbf{E}\,g(W) - \mathbf{E}\,g(Z)| \leq \sup_{f\in\mathcal{D}'} \left|\mathbf{E}\left(f'(W) - Wf(W)\right)\right|$$

Main difference between stein's method and characteristic functions is that Stein's method is a *local* technique. We transfer a *global* problem to a local problem. It's a theme that is present in many branches of mathematics.