CS362 Algorithmic Frontiers April 16, 2014

Lecture 6

Lecturer: Gregory Valiant *Scribe: Hamsa Sridhar & Osbert Bastani*

1 Introduction

In this lecture, we will cover a brief introduction of Stein's method which is used as a general tool for proving central limit theorems.

We have so far derived lower bounds for property testing and estimation by characterizing the generalized multinomial distribution and applying a theorem of Roos to compare it to a Poisson distribution. If we consider $G_M \sim \text{Poisson}(\vec{\lambda})$ as a first-order approximation, we can try to obtain a so-called second-order approximation by comparing $G_M \sim \mathcal{N}(\vec{\lambda}, \Sigma)$ (we use the term second-order in the sense that we are now estimating a second moment). However, unlike the Poisson distribution which is supported on the integers, the Gaussian distribution is supported on the reals and so we must do some discretization.

One standard way to prove the central limit theorem is the "hybridization" approach. Consider n i.i.d. samples $X_1, ..., X_n$ and let $X = \sum_{i=1}^n X_i$. Also, let $G_1, ..., G_n$ be n i.i.d. samples drawn from a Gaussian distribution. Define

$$X^i = G_1 + ... + G_i + X_{i+1} + ... + X_n$$

such that $X^0 = X$ and X^n is distributed according to a Gaussian. Then we can bound the difference between X and the Gaussian random variable X^n by bounding the difference between each X_i and X_{i-1}, i.e. $|X - X^n| = \sum_{i=1}^n |X^i - X^{i-1}|$. Note that this argument only works when the samples are i.i.d. and is not robust to more general settings. Stein's method (1970) allows us to prove central limit theorems more generally.

2 Stein's Method

The goal is to compare X to a random variable $G \sim D$, where D is some "nice" distribution. The intuition behind Stein's method is as follows:

1. Find a transformation T with a unique fixed point G.
2. Look at $|TX - X|$. The hope is that if X is close to G, then TX will be close to G as well.

We will consider $G \sim D = \mathcal{N}(0, 1)$. A natural transformation T that holds G as a fixed point is the **Ornstein-Uhlenbeck process**, which can be thought of as the transformation that adds $\mathcal{N}(0, \epsilon^2)$ noise to the distribution and then scales it back to the origin by multiplying by $\frac{1}{\sqrt{1+\epsilon^2}}$.

Note that for any X with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$, this transformation preserves the first two moments, i.e. we still have $\mathbb{E}[TX] = 0$ and $\mathbb{E}[(TX)^2] = 1$. Thus, we might expect that the bound we derive between X and G would depend on third- and higher-order moments of X.

Our goal is to bound

$$\sup_{f \in H} |\mathbb{E}[f(X)] - \mathbb{E}[f(G)]| .$$

Note that the choice of distance metric we use to compare X and G determines the class of functions H. For instance, if we use earth mover's distance as our metric, H is the class of Lipschitz functions.

To get some intuition for how we shall compare the transformed distribution to the original one, consider an example function f given in Figure 1, where X_1 and X_2 are Gaussians with means at the indicated point. Note that $\mathbb{E}[f(X_1 + noise)] - f(X_1) \approx 0$, as X_1 occurs in a portion of f that is linear, whereas $\mathbb{E}[f(X_2 + noise)] - f(X_2) \propto f''(X_2)$, due to the significant second-derivative at X_2. Similarly,

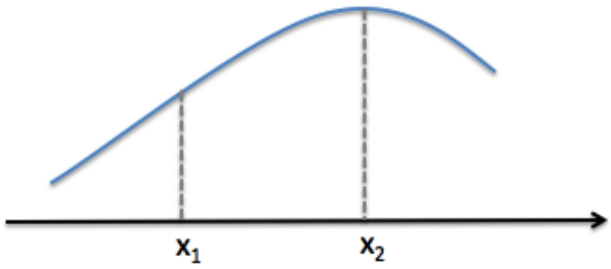

Figure 1: An example function f

the scaling towards the origin will introduce a term linear in the first derivative of f at X. This motivates the following expression:

$$|TG - G| \propto \mathbb{E}[f''(G) - Gf'(G)].$$

On the other hand, because the Gaussian is fixed by this transformation, we know that $|TG-G| = 0$. The following lemma verifies the useful fact that for a standard Gaussian random variable, and any twice-differentiable function f, $\mathbb{E}_{X \leftarrow G}[f''(X) - Xf'(X)] = 0$. Note that in the following formulation, we replace f by its antiderivative. Additionally, we will use the shorthand $\mathbb{E}[f(G)]$ to denote $\mathbb{E}_{X \leftarrow G}[f(X)]$.

Lemma 1 *For all differentiable functions $f : \mathbb{R} \to \mathbb{R}$, letting G denote a standard Gaussian,*

$$\mathbb{E}[f'(G) - G \cdot f(G)] = 0.$$

Proof We evaluate the first term using integration by parts:

$$\begin{aligned}
\mathbb{E}[f'(G)] &= \int_{-\infty}^{\infty} f'(x) e^{-\frac{x^2}{2}} dx \\
&= f(x) e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} x f(x) e^{-\frac{x^2}{2}} dx \\
&= \int_{-\infty}^{\infty} x f(x) e^{-\frac{x^2}{2}} dx \\
&= \mathbb{E}[Gf(G)],
\end{aligned}$$

as claimed. ■

The above lemma guarantees that $\text{Exp}[f(X) - Xf'(X)] = 0$ if X is distributed according to a Gaussian. The next lemma shows that the magnitude of $\text{Exp}[f(X) - Xf'(X)]$ tells us how *far* X is from a Gaussian:

Lemma 2 *Suppose that $h : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with bounded derivative. We have*

$$\mathbb{E}[h(X)] - \mathbb{E}[h(G)] = \mathbb{E}[f_h'(X) - Xf_h(X)],$$

where we define

$$f_h(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} g(t) e^{-\frac{t^2}{2}} dt,$$

where $g(x) = h(x) - \mathbb{E}[h(G)]$.

Proof First, we claim that $f_h'(x) - xf_h(x) = g(x)$. To see this, note that

$$f_h'(x) = e^{\frac{x^2}{2}}\left(g(x)e^{-\frac{x^2}{2}}\right) + \left(xe^{\frac{x^2}{2}}\right)\int_{-\infty}^{x} g(t)e^{-\frac{t^2}{2}}dt = g(x) + xf_h(x),$$

so the claim follows. Then we have

$$\begin{aligned}\mathbb{E}[f_h'(X) - Xf_h(X)] &= \mathbb{E}[g(X)] \\ &= \mathbb{E}\left[h(X) - \mathbb{E}[h(G)]\right] \\ &= \mathbb{E}[h(X)] - \mathbb{E}[h(G)],\end{aligned}$$

as claimed. ■

We can apply Lemma 2 when computing the distance between X and G with respect to some class of functions H:

$$d_H(X,G) = \sup_{h\in H}|\mathbb{E}[h(X)] - \mathbb{E}[h(G)]| = \sup_{h\in H}|\mathbb{E}[f_h'(X) - Xf_h(X)]|.$$

In the following section, we will give an example of how to bound the right hand side in order to prove a central limit theorem.

3 Application to Central Limit Theorems

Now we will show how Stein's method can be applied to proving a typical central limit theorem. Suppose $X_1, ..., X_n$ are i.i.d. with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = 1$, and $|\mathbb{E}[X_i^3]| < \infty$, and let $X = \frac{1}{\sqrt{n}}(X_1 + ... + X_n)$. We will begin by bounding the earth mover's distance, and then transform this into a bound on the total variation distance.

3.1 Bounding Earth Mover's Distance

We will use the following lemma without proof:

Lemma 3 *If $h \in Lip$, then f_h satisfies $\|f_h''\|_\infty \leq 2$.*

Now we will show that $d_W(X,G)$ is small for sufficiently large n. More precisely:

Theorem 4 *We have*

$$d_W(X,G) \leq \frac{3\cdot\mathbb{E}[|X_1|^3]}{\sqrt{n}}.$$

Proof By Lemma 2,

$$d_W(X,G) = \sup_{h\in \mathrm{Lip}}|\mathbb{E}[h(X)] - \mathbb{E}[h(G)]| = \sup_{h\in \mathrm{Lip}}\left|\mathbb{E}[f_h'(X) - Xf_h(X)]\right|.$$

Let $X' = X - \frac{1}{\sqrt{n}}X_1 = \frac{1}{\sqrt{n}}(X_2 + ... + X_n)$. Note that for any differentiable $f : \mathbb{R} \to \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}[Xf(X)] &= \mathbb{E}[\sqrt{n}X_1 f(X)] \\ &= \mathbb{E}\left[\sqrt{n}X_1\left(f(X') + \frac{X_1}{\sqrt{n}}f'(X')\right)\right] + E_1 \\ &= \mathbb{E}[f'(X)] + E_1 + E_2.\end{aligned}$$

The first line follows because

$$\begin{aligned}\mathbb{E}[Xf(X)] &= \frac{1}{\sqrt{n}}\left(\mathbb{E}[X_1 f(X)] + ... + \mathbb{E}[X_n f(X)]\right) \\ &= \frac{1}{\sqrt{n}}\left(\mathbb{E}[X_1 f(X)] + ... + \mathbb{E}[X_1 f(X)]\right) \\ &= \mathbb{E}[\sqrt{n}X_1 f(X)].\end{aligned}$$

In the second line, we claim that

$$|E_1| \leq \frac{\mathbb{E}[|X_1|^3]\|f''\|_\infty}{2\sqrt{n}}.$$

To see this, note that

$$f(X) = f\left(X' + \frac{X_1}{\sqrt{n}}\right) = f(X') + \frac{X_1}{\sqrt{n}}f'(X') + \epsilon_1,$$

where by Taylor's theorem,

$$|\epsilon_1| \leq \frac{X_1^2}{2n}\|f''\|_\infty,$$

so

$$|E_1| = \left|\mathbb{E}\left[\frac{X_1}{\sqrt{n}}\epsilon_1\right]\right| \leq \mathbb{E}\left[\frac{|X_1|}{\sqrt{n}}|\epsilon_1|\right] \leq \frac{\mathbb{E}[|X_1|^3]\|f''\|_\infty}{2\sqrt{n}}.$$

In the third line, we claim that

$$|E_2| \leq \frac{\mathbb{E}[|X_1|^3]\|f''\|_\infty}{\sqrt{n}}.$$

To see this, note that

$$f'(X) = f'\left(X' + \frac{X_1}{\sqrt{n}}\right) = f'(X') + \epsilon_2,$$

where by Taylor's theorem,

$$|\epsilon_2| \leq \frac{X_1}{\sqrt{n}}\|f''\|_\infty.$$

Note that $\mathbb{E}[\sqrt{n}X_1 f(X')] = \sqrt{n}\cdot\mathbb{E}[X_1]\mathbb{E}[f(X')] = 0$, since $\mathbb{E}[X_1] = 0$ by assumption. Hence

$$|E_2| = \left|\mathbb{E}[X_1^2\epsilon_2]\right| \leq \mathbb{E}[X_1^2|\epsilon_2|] \leq \frac{\mathbb{E}[|X_1|^3]\|f''\|_\infty}{\sqrt{n}}.$$

Hence the total error is

$$|E| \leq |E_1| + |E_2| \leq \frac{3\cdot\mathbb{E}[|X_1|^3]\|f''\|_\infty}{2\sqrt{n}},$$

so by Lemma 3, we have

$$d_W(X,G) \leq \sup_{h\in\text{Lip}} \frac{3\cdot\mathbb{E}[|X_1|^3]\|f_h''\|_\infty}{2\sqrt{n}} \leq \frac{3\cdot\mathbb{E}[|X_1|^3]}{\sqrt{n}},$$

as claimed. ■

3.2 Bounding Total Variation Distance

Now we discuss how to bound total variation distance based on earth mover's distance. It will not be possible to obtain a general bound, for example the earth mover's distance between a binomial distribution and a Gaussian distribution is small, but the total variation distance is 1. Instead, we will have two guiding principles:

1. For two distributions A and B, if $d_W(A, B)$ is small, then $d_{\mathrm{TV}}(A \circ f, B \circ f)$ is small, where f is some sufficiently narrow Gaussian curve.

2. As long as A and B are "nice" distributions (in particular, A is unimodal, has discrete support, and $\mathrm{support}(A) = \mathrm{support}(B)$, then we can "deconvolve" to obtain a bound on $d_{\mathrm{TV}}(A, B)$.

We have the following lemma:

Lemma 5 *Suppose that A and B are distributions over $\mathbb{Z}$. Let $f : \mathbb{Z} \to \mathbb{R}^+$ be the PDF of a unimodal distribution on $\mathbb{Z}$, with $\max_{i \in \mathbb{Z}} f(i) = m$. Then $d_{TV}(A \circ f, B \circ f) \leq m \cdot d_W(A, B)$.*

Proof First, we claim that

$$S = \sum_{i \in \mathbb{Z}} |f(i) - f(i-c)| \leq 2mc.$$

Consider the quantities $f(i) - f(i-c)$. Note that for some $k \in \mathbb{Z}$, we have $f(i) - f(i-c) \geq 0$ for $i \leq k$ and $f(i) - f(i-c) \leq 0$ for $i > k$. Hence

$$\begin{aligned}
S &= \sum_{i=-\infty}^{k} (f(i) - f(i-c)) + \sum_{i=k+1}^{\infty} (f(i-c) - f(i)) \\
&= \left(\sum_{i=-\infty}^{k} f(i) - \sum_{i=-\infty}^{k-c} f(i) \right) + \left(\sum_{i=k-c+1}^{\infty} f(i) - \sum_{i=k+1}^{\infty} f(i) \right) \\
&= \sum_{i=k-c+1}^{k} f(i) + \sum_{k-c+1}^{k} f(i) \\
&\leq 2mc.
\end{aligned}$$

The claim follows... ■

The next lemma

Lemma 6 *Suppose that A and B are distributions over $\mathbb{Z}$. Let $f : \mathbb{Z} \to \mathbb{R}^+$ be the PDF of a unimodal distribution over $\mathbb{Z}$, with $\max_{i \in \mathbb{Z}} f(i) = m$. Furthermore, assume that A is unimodal with $\max_{i \in \mathbb{Z}} A(i) = m'$. Then*

$$d_{TV}(A, B \circ f) \leq m \cdot d_W(A, B) + m' \cdot d_W(f, \delta).$$

Proof Note that

$$d_{\mathrm{TV}}(A, B \circ f) \leq d_{\mathrm{TV}}(A, A \circ f) + d_{\mathrm{TV}}(A \circ f, B \circ f),$$

where the first inequality follows from the triangle inequality and the second follows from Lemma 5. Now we have

$$d_{\mathrm{TV}}(A, A \circ f) = d_{\mathrm{TV}}(f \circ A, \delta \circ f) \leq m' \cdot d_W(f, \delta).$$

Together with Lemma 5, this shows that

$$d_{\mathrm{TV}}(A, B \circ f) \leq m \cdot d_W(A, B) + m' \cdot d_W(f, \delta).$$

■

For example, we can take $A = \mathrm{Binomial}(n, k)$ and $B = \mathrm{DiscretizedGaussian}(\mu, \sigma^2)$. Taking $f = \mathrm{Binomial}\left(\sqrt{n}, \frac{1}{2}\right)$ gives a nontrivial bound on $d_{\mathrm{TV}}(A, B)$.