CSE 546: Machine Learning **Lecture 5 extras**

Optional Reading: Large deviations and the χ^2 tail bound

Instructor: Sham Kakade

1 The Central Limit Theorem

While true under more general conditions, the following is a rather simple proof of the central limit theorem. This proof provides some insight into our theory of large deviations (e.g. how far away a random variable is form its mean).

Recall that $M_X(\lambda) = \mathbb{E}e^{\lambda X}$ is the moment generating function of a random variable X.

Theorem 1.1. *Suppose $X_1, X_2, \ldots X_n$ is a sequence of independent, identically distributed (i.i.d.) random variables with mean μ and variance σ_2. Suppose that the $M_X(\lambda)$ exists for all λ in a neighborhood of 0. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then for all x,*

$$\lim_{n\to\infty} \Pr\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

where $\Phi(\cdot)$ is the standard normal CDF.

Roughly, this says that, as $n \to \infty$, $\bar{X}_n$ is distributed according to a Gaussian with mean μ and variance σ^2/n.

Proof. Without loss of generality, assume $\mu = 0$. Define $\bar{Z}_n = \frac{\bar{X}_n}{\sigma/\sqrt{n}} = \frac{\sum_i X_i}{\sigma\sqrt{n}}$. By independence and properties of the MGF, we have:

$$M_{\bar{Z}_n}(\lambda) = \mathbb{E}e^{\lambda\frac{\sum_i X_i}{\sigma\sqrt{n}}} = \mathbb{E}e^{\lambda\frac{X_1}{\sigma\sqrt{n}}}\mathbb{E}e^{\lambda\frac{X_2}{\sigma\sqrt{n}}}\ldots\mathbb{E}e^{\lambda\frac{X_n}{\sigma\sqrt{n}}} = (M_X(\frac{\lambda}{\sigma\sqrt{n}}))^n$$

where we have used independence of X_i in the first step.

As the moment generating function exists around 0 (and the derivatives of the moment generating function are the moments), Taylor's theorem implies:

$$\begin{aligned} M_X(s) &= M_X(0) + M_X'(0)s + \frac{1}{2}M_X''(0)s^2 + \frac{1}{3!}M_X'''(0)s^3\ldots \\ &= 1 + 0 + \frac{1}{2}M_X''(0)s^2 + o(s^2) \end{aligned}$$

where a function $g(s) = o(s^2)$ if $g(s)/s^2 \to 0$ as $s \to 0$. Hence,

$$M_X(\frac{\lambda}{\sigma\sqrt{n}}) = 1 + \frac{1}{2}\frac{\lambda^2}{n} + o(\frac{\lambda^2}{n})$$

where the last term is with respect to $n \to \infty$. Hence,

$$M_{\bar{Z}_n}(\lambda) = \left(1 + \frac{1}{2}\frac{\lambda^2}{n} + o(\frac{\lambda^2}{n})\right)^n \to \exp(\frac{\lambda^2}{2})$$

Thus the limiting moment generating function of $\bar{Z}_n$ is identical to that of a standard normal (in a neighborhood of 0 for λ). This proves they have identical CDFs (using properties of the MGF). □

2 Large Deviations

Note that the CLT says

$$\lim_{n\to\infty} \Pr\left(\bar{X}_n \leq \mu + z\sigma\sqrt{n}\right) = \Phi(z)$$

or, equivalently,

$$\lim_{n\to\infty} \Pr\left(\bar{X}_n \geq \mu + z\sigma\sqrt{n}\right) = 1 - \Phi(z)$$

which is the (asymptotic) probability in the tail.

Instead, suppose we seek the following probability

$$\Pr\left(\bar{X}_n \geq \mu + \epsilon\right) = ??$$

, where ϵ is fixed. Does the central limit theorem say anything useful? It is easy to see that, for any ϵ

$$\lim_{n\to\infty} \Pr\left(\bar{X}_n \geq \mu + \epsilon\right) = 0$$

Instead, we seek a more meaningful limit. In particular, we will examine:

$$\frac{1}{n} \ln \Pr\left(\bar{X}_n \geq \mu + \epsilon\right) = ??$$

Does the CLT provide the limit of the above quantity? Why have we chosen $\frac{1}{n}$?

The answer to the former question is "no" since the key difference is that ϵ is fixed in the above (while the CLT only quantifies a limit for ϵ which scales as $1/\sqrt{n}$.

2.1 Large Deviations for a Gaussian random variable

Let X be a standard Gaussian random variable: $X \sim N(0,1)$, with density function

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

For $\epsilon > 0$, what is the probability $P(X \geq \epsilon)$?

We have the following upper bound

$$\begin{aligned} P(X \geq \epsilon) &= \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \leq \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x^2+\epsilon^2)/2} dx \\ &= 0.5 e^{-\epsilon^2/2} dx \end{aligned}$$

and lower bound

$$\begin{aligned} P(X \geq \epsilon) &= \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \\ &\geq \int_{0}^{1} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \geq 0.34 e^{-(2\epsilon+\epsilon^2)/2} dx \\ &\geq 0.5 e^{-(\epsilon+1)^2/2} dx. \end{aligned}$$

Therefore we have

$$0.5e^{-(\epsilon+1)^2/2} \le P(X \ge \epsilon) \le 0.5e^{-\epsilon^2/2}.$$

Equivalently,

$$P(X \ge \epsilon) \le 0.5e^{-\epsilon^2/2} \le P(X \ge \epsilon - 1)$$

which shows that our upper bound is sandwiched between the tail probabilities within one deviation.

Now let $X_1, \ldots, X_n$ be n iid Gaussians $X_i \sim N(\mu, \sigma^2)$, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then since $P(\bar{X}_n \ge \mu + \epsilon) = P(\sqrt{n}(\bar{X}_n - \mu)/\sigma \ge \sqrt{n}\epsilon/\sigma)$, where $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0,1)$, the above bound becomes

$$0.5e^{-n(\epsilon+\sigma/\sqrt{n})^2/2\sigma^2} \le P(\bar{X}_n \ge \mu + \epsilon) \le 0.5e^{-n\epsilon^2/2\sigma^2}.$$

The tail probability decays exponentially fast. The bound is tight, meaning that any fixed ϵ:

$$\lim_{n\to\infty} n^{-1} \ln P(\bar{X}_n \ge \mu + \epsilon) = -\epsilon^2/2\sigma^2.$$

This is a large deviation result (meaning fixed deviation ϵ from the mean is much larger than standard deviation $\sigma/\sqrt{n}$ of X).

3 Markov Inequality

More generally, let $X_1, \ldots, X_n$ be n iid random variables (not necessarily Gaussian) with mean μ, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, we are interested in estimating the tail bound $P(\bar{X}_n \ge \mu + \epsilon)$ and $P(\bar{X}_n \le \mu - \epsilon)$ for some $\epsilon > 0$.

Generally, this is achieved through Markov inequality:

Lemma 3.1. *Suppose $Z \ge 0$, with probability 1. For all $t \ge 0$,*

$$\Pr(Z \ge t) \le \mathbb{E}[Z]/t$$

Proof. Observe that

$$\mathbb{E}(Z) \ge \mathbb{E}[Z\, \mathrm{I}[Z > t]] \ge t\mathbb{E}[\mathrm{I}[Z > t]] = t \Pr(Z \ge t)$$

which proves the result. □

Corollary 3.2. *Suppose $g(x) \ge 0$ and that $\mu = \mathbb{E}[X]$. Then:*

$$P(\bar{X}_n \ge \mu + \epsilon) \le \frac{Eg(\bar{X}_n - \mu)}{\inf_{z \ge \epsilon} g(z)}.$$

One can use moment inequality with $g(z) = |z|^m$ for some m. However, one needs to estimate $Eg(\bar{X}_n - \mu)$. In particular, Chebyshev inequality picks $g(z) = z^2$, which is easy to estimate:

$$Eg(\bar{X}_n - \mu) = E(\bar{X}_n - \mu)^2 = \frac{1}{n} Var(X_1).$$

Therefore

$$P(\bar{X}_n \ge \mu + \epsilon) \le \frac{Var(X_1)}{n\epsilon^2}.$$

4 Exponential Inequality

The following technique gives us a method to derivate tail bounds for a much larger class of distributions.

In order to get exponential tail bounds, we choose $g(z) = e^{\lambda n z}$ for some tuning parameter $\lambda > 0$. Then Markov inequality becomes

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \frac{Ee^{\lambda n(\bar{X}_n - \mu)}}{e^{\lambda n \epsilon}} = \frac{Ee^{\lambda \sum_{i=1}^n (X_i - \mu)}}{e^{\lambda n \epsilon}} = e^{-\lambda n \epsilon} E^n e^{\lambda(X_1 - \mu)}.$$

Note that in order to use this estimate, we have to assume that $Ee^{\lambda(X_1-\mu)} < \infty$ for some $\lambda > 0$. Taking logarithm, it follows that we have the following theorem

Theorem 4.1. *For any n and $\epsilon > 0$:*

$$n^{-1} \ln P(\bar{X}_n \geq \mu + \epsilon) \leq \inf_{\lambda > 0} [-\lambda \epsilon + \ln Ee^{\lambda(X_1 - \mu)}].$$

Similarly

$$n^{-1} \ln P(\bar{X}_n \leq \mu - \epsilon) \leq \inf_{\lambda < 0} [\lambda \epsilon + \ln Ee^{\lambda(X_1 - \mu)}].$$

The function $\Gamma(\lambda) = \ln Ee^{\lambda X_1}$ is called logarithmic moment generating function of a random variable X_1. Exponential inequality for sum of independent random variables is very easy to apply because independence allows us to change the problem of estimating the exponential moment of the sum of independent random variables into the estimating of the exponential moment of a single random variable. Another way to write tail bound is

Corollary 4.2. *We have that*

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \exp[-nI(\mu + \epsilon)],$$

where $I(z)$ defined as

$$-I(z) = \inf_{\lambda > 0} [-\lambda z + \ln Ee^{\lambda X_1}]$$

is the rate function.

Example: Gaussian random variable $X_i \sim N(\mu, \sigma^2)$, then

$$Ee^{\lambda(X_1 - \mu)} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\lambda x} e^{-x^2/2\sigma^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\lambda^2\sigma^2/2} e^{-(x/\sigma - \lambda\sigma)^2/2} dx/\sigma = e^{\lambda^2\sigma^2/2}.$$

Therefore (with optimal $\lambda = \epsilon/\sigma^2$ below)

$$\inf_{\lambda > 0} [-\lambda \epsilon + \ln Ee^{-\lambda(X_1 - \mu)}] = \inf_{\lambda > 0} [-\lambda \epsilon + \lambda^2 \sigma^2/2] = -\epsilon^2/2\sigma^2.$$

Exactly the same (and tight) estimate of Gaussian tail inequality derived by integration.

5 χ^2 Tail Bound

Let $X_i \sim \mathcal{N}(0, 1)$ be independent Gaussians, then the distribution of $Z = \sum_{i=1}^n X_i^2$ is χ^2 with n degrees of freedom.

This variable is important for analyzing least squares regression.

Theorem 5.1. *Let $X_i \sim \mathcal{N}(0,1)$ be independent Gaussians, then the distribution of $Z = \sum_{i=1}^n X_i^2$ is χ^2. We have that (for the upper tail):*

$$P(Z/n \geq 1+\epsilon) \leq \exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right]$$

One useful upper bound (for obtaining sharp constants) is:

$$\exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right] \leq \exp\left[-\frac{n}{2}(1+\epsilon - \sqrt{1+2\epsilon})\right]$$

A bound that is more comparable to the Bennet-style bound is:

$$\exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right] \leq \exp[-n\epsilon^2/(4+4\epsilon)].$$

(note the difference between the upper and lower tail).

For the lower tail:

$$P(Z/n \leq 1-\epsilon) \leq \exp[-n\epsilon^2/4].$$

Hence, with probability $1-\delta$:

$$Z/n \leq 1 + 2\sqrt{\ln(1/\delta)/n} + 2\frac{\ln(1/\delta)}{n}$$

and with probability $1-\delta$:

$$Z/n \geq 1 - 2\sqrt{\ln(1/\delta)/n}.$$

The logarithmic moment generating function of X_i^2 for $\lambda < 0.5$ is

$$\Gamma(\lambda) = \ln Ee^{\lambda X_i^2} = -0.5\ln(1-2\lambda),$$

and $EX_i^2 = 1$.

Proof. We only prove the upper tail. The lower tail is simpler to prove in that we can use the bound $\log(1+x) > 1 + x - x^2/2$ for $x > 0$.

From the moment method, we must constrain $\lambda < -.5$, or, equivalently, set $\Gamma(\lambda) = \infty$ for $\lambda \geq 0.5$. Hence,

$$I(1+\epsilon) = \inf_{0.5>\lambda>0}[-\lambda(1+\epsilon) - 0.5\ln(1-2\lambda)] = -\frac{1}{2}(\epsilon - \log(1+\epsilon))$$

where the inf is achieved at $1+\epsilon = \frac{1}{1-2\lambda}$ or equivalently $\lambda = \frac{\epsilon}{2(1+\epsilon)}$.

The first claim is completed by noting that $\log(1+\epsilon) \leq \sqrt{1+2\epsilon} - 1$, for $\epsilon > 0$. To see this, first note equality at $\epsilon = 0$. Also, note that derivative on the left hand side is:

$$\frac{1}{1+\epsilon} =\leq \frac{1}{\sqrt{1+2\epsilon}}$$

where the right hand side is the derivative of $\sqrt{1+2\epsilon}$.

For the second claim, the proof is completed by noting that the function $f(x) = (x - \log(1+x)) * (1+x)$. Note that $f'(x) = 2x - \log(1+x)$, $f''(x) = (1+2x)/(1+x)$, and $f'''(x) = 1/(1+x)^2 >= 0$. So $f(x) >= x^2/2$.

The rest of the proof is straight forward. □