Stat 928: Statistical Learning Theory **Lecture: 5**

The Moment Method; Convex Duality; and Large/Medium/Small Deviations

Instructor: Sham Kakade

1 The Exponential Inequality and Convex Duality

The 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.

Theorem 1.1. *For any n and $\epsilon > 0$:*

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

Similarly

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

Another way to write tail bound is

Corollary 1.2. *We have that*

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

where $I(z)$ defined as

$$-I(z) = \inf_{\lambda > 0} [-\lambda z + \ln E e^{\lambda X}]$$

is the rate function.

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

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

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

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

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

1.1 The Fenchel Conjugate

Assume we have a vector space $\mathcal{X}$, equipped with an inner product $< \cdot, \cdot >$ (and a dual space $\mathcal{X}^*$ — for all practical purposes, think of $\mathcal{X}^* = \mathcal{X}$).

Let f be a function:

$$f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$$

i.e. f takes values on the extended real number line. The convex conjugate is f^*, where

$$f^* : \mathcal{X}^* \to \mathbb{R} \cup \{+\infty\},$$

is defined as:

$$f^*(\lambda) = \sup_{x \in \mathcal{X}} (< \lambda, x > - f(x))$$

Equivalently,

$$f^*(\lambda) = - \inf_{x \in \mathcal{X}} (- < \lambda, x > + f(x))$$

1.2 A Variational interpretation of the Rate Function

The function

$$\Gamma(\lambda) = \ln E e^{\lambda X}$$

is called the *cumulant generating function* (i.e. it is the logarithmic moment generating function) of a random variable X.

A slightly modified function is to constrain λ to bigger than 0. So a modified function is:

$$\Gamma_+(\lambda) = \begin{cases} \Gamma(\lambda) & \text{if } \lambda > 0 \\ \infty & \text{else} \end{cases}$$

This modification is so that when we take an $\inf$ over λ, we effectively are only considering the positive λ.

Corollary 1.3. *We have that*

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

and that

$$\Gamma_+^*(z) = I(z)$$

Let us now understand Γ_+^*

Let P be the original measure on X. Let us now define a different measure, P_λ on X as follows:

$$dP_\lambda(X \geq x) = e^{\lambda x - \Gamma(\lambda)} dP(X \geq x)$$

Note that it is normalized.

Given some ϵ, suppose we find the λ such that:

$$\mathbb{E}_{X \sim P_\lambda}[X] = \mu + \epsilon$$

Slightly abusing notation let $P_{\mu+\epsilon}$ be this distribution (e.g. $P_{\mu+\epsilon}$ is P_λ for the λ which solves the above). Intuitively, $P_{\mu+\epsilon}$ is the "perturbed" distribution which shifts the mean by ϵ.

Lemma 1.4. *Assume $P_{\mu+\epsilon}$ exists. Then:*

$$I(\mu + \epsilon) = -KL(P_{\mu+\epsilon} || P)$$

Proof. Left as a homework problem. □

2 "Small", "Medium", and "Large" Deviations

The tail probability we are interested in is:

$$P(\bar{X}_n \geq \mu + \epsilon)$$

For large n, we seek to understand the behavior of:

$$P(\bar{X}_n \geq \mu + \epsilon_n)$$

where ϵ_n is a function of n.

We can think of "three" natural asymptotics. If ϵ_n behaves as $1/\sqrt{n}$, e.g.

$$\epsilon_n = \Theta\left(z\frac{\sigma^2}{\sqrt{n}}\right)$$

then the CLT capture the tail probability (under the tail probability with respect z, under the standard normal).

If ϵ is a constant, then this "large deviation" tail probability tends to 0. Here, we can characterize

$$\underline{\lim}_{n\to\infty} n^{-1} P(\bar{X}_n \geq \mu + \epsilon)$$

As we will see, the rate function governs this asymptote.

In practice, we are often interested in "medium" deviations. By this, we often use

$$\epsilon_n \to 0 \text{ and } \sqrt{n}\epsilon_n \to \infty$$

e.g. we typically are interested in ϵ_n going to 0 at a rate slower than $\frac{1}{\sqrt{n}}$ (this is because as n increases, we also increase the model complexity).

As we now see, the the rate function accurately captures both the large and small regime, so we expect it to naturally capture the medium deviation regime.

2.1 The Small Deviation Limit

Note that

$$P(\bar{X}_n \geq \mu + z\frac{\sigma^2}{\sqrt{n}}) \leq -nI(\mu + z\frac{\sigma^2}{\sqrt{n}})$$

Theorem 2.1. *We have that:*

$$\lim_{n\to\infty} -nI(\mu + z\frac{\sigma^2}{\sqrt{n}}) = -\frac{z^2}{2}$$

Proof. HW exercise □

Note this is upper bound is sharp approximation to the true Gaussian tail probability (as seen in the last lecture). Hence, in the small deviation regime, we have not lost much.

2.2 The Large Deviation Limit

For large deviation, the exponential Markov inequality is asymptotically tight in the following sense:

Theorem 2.2. *For all $\epsilon' > \epsilon > 0$:*

$$\lim_{n\to\infty} n^{-1} \ln P(\bar{X}_n \geq \mu + \epsilon) \geq -I(\mu + \epsilon').$$

Proof. We only prove the first inequality. Let X_i have CDF at x as $P(X_i \geq x)$, and define distribution of X_i' with density at x as $dP_\lambda(X_i' \geq x) = e^{\lambda x - \Gamma(\lambda)} dP(X_i \geq x)$ (e.g. under P_λ). Let $\bar{X}_n' = n^{-1} \sum_{i=1}^n X_i'$. Then $e^{-n\lambda \sum_i x_i + n\Gamma(\lambda)} \prod_i dP(X_i' \geq x_i) = \prod_i d(P(X_i \geq x_i)$. We use I to denote indicator function:

$$\begin{aligned}
P(\bar{X}_n \geq \mu + \epsilon) \geq & P(\bar{X}_n - \mu \in [\epsilon, \epsilon']) \\
= & E_{X,\ldots,X_n} I(\bar{X}_n - \mu \in [\epsilon, \epsilon']) \\
= & E_{X',\ldots,X_n'} e^{-\lambda n \bar{X}_n' + n\Gamma(\lambda)} I(\bar{X}_n' - \mu \in [\epsilon, \epsilon']) \\
\geq & e^{-\lambda n(\mu+\epsilon') + n\Gamma(\lambda)} E_{\bar{X}_n'} I(\bar{X}_n' - \mu \in [\epsilon, \epsilon']) \\
\geq & e^{-\lambda n(\mu+\epsilon') + n\Gamma(\lambda)} (1 - o(1)) \geq e^{\inf_{\lambda>0}(-\lambda n(\mu+\epsilon') + n\Gamma(\lambda)} (1 - o(1)).
\end{aligned}$$

Note that $\Gamma'(\lambda) = E_{\bar{X}_n'} \bar{X}_n'$. Therefore with $\lambda = \arg\min \inf_\lambda [-\lambda(\mu + (\epsilon' + \epsilon)/2) + \Gamma(\lambda)]$, we have $E_{\bar{X}_n'} \bar{X}_n' = \Gamma'(\lambda) = \mu + (\epsilon' + \epsilon)/2$, and thus by the law of large numbers $E_{\bar{X}_n'} I(\bar{X}_n' - \mu \in [\epsilon, \epsilon']) = 1 - o(1)$ as $n \to \infty$. This proves the lower bound. □

Compare with Gaussian for similarity. Generally, with a more careful estimate, one can obtain a tight lower bound for finite n at $\epsilon' = \epsilon + O(\sqrt{Var(X)/n})$. That is, for exponential inequality, the looseness in terms of deviation ϵ is only $O(\sqrt{Var(X)/n})$.

3 Sub-Gaussian Random Variables

Recall that for a Gaussian random variable:

$$\ln E e^{\lambda(X-\mu)} = \frac{\lambda^2}{2}\sigma^2.$$

We say X is a sub-Gaussian random variable if it has quadratically bounded logarithmic moment generating function,e.g.

$$\ln E e^{\lambda(X-\mu)} \leq \frac{\lambda^2}{2} b.$$

For this case, we have for $z > \mu$:

$$-I(z) = \inf_{\lambda>0} (-\lambda z + \lambda\mu + \frac{\lambda^2}{2} b).$$

Taking derivative at optimal λ_*:

$$-z + \mu = \lambda_* b,$$

which implies that $\lambda_* = (\mu - z)/b$ and

$$I(z) = \frac{(z-\mu)^2}{2b}.$$

That is, we have

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-n\epsilon^2/2b}.$$

Similarly,

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-n\epsilon^2/2b}.$$

Bounded Random Variables

Clearly, a Gaussian variable $N(\mu, \sigma^2)$ is sub-Gaussian with $b = \sigma^2$.

Now let us consider the case of a bounded random variable.

Lemma 3.1. *(Hoeffding's Lemma) Suppose that $X \in [b_-, b_+]$ with probability 1, then X is sub-Gaussian with $b = (b_+ - b_-)^2/4$.*

To prove this, first let us observe the following:

Lemma 3.2. *We have the following equality for the second derivative of Γ*

$$\Gamma''(\lambda) = \mathrm{Var}_{P_\lambda}(X)$$

which is the variance of X under P_λ.

Proof. Left as a HW exercise. □

Now we can prove Hoeffding's lemma:

Proof. First observe that:

$$|X - \frac{b_+ - b_-}{2}| \leq \frac{b_+ - b_-}{2}$$

Hence:

$$\mathrm{Var}_{P_\lambda}(X) = \mathrm{Var}_{P_\lambda}(X - \frac{b_+ - b_-}{2}) \leq \frac{(b_+ - b_-)^2}{4}$$

Since $\Gamma'(0) = \mu$, we have that:

$$\Gamma'(\lambda) \leq \lambda\mu + \lambda^2 \frac{(b_+ - b_-)^2}{8}$$

(by integration and the fact the upper bound has the same first derivative as Γ at $\lambda = 0$). □

Corollary 3.3. *Suppose that $X \in [b_-, b_+]$ with probability 1, then*

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-2n\epsilon^2/(b_+ - b_-)^2}.$$

Similarly,

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-n2\epsilon^2/(b_+ - b_-)^2}.$$