Basic Probability Inequalities

Mathematical Analysis of Machine Learning Algorithms
(Chapter 2)

Basic Probability Inequalities

We derive exponential tail probability inequalities for sums of independent random variables. These inequalities are the basic tools to analyze machine learning algorithms.

Let $X_1, \ldots, X_n$ be n iidrandom variables with mean

$$\mu = \mathbb{E} X_i.$$

Let the empirical mean be

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Given $\epsilon > 0$, we are interested in estimating the tail probability

$$\Pr\left(\bar{X}_n \geq \mu + \epsilon\right), \qquad \Pr\left(\bar{X}_n \leq \mu - \epsilon\right).$$

Gaussian Random Variables

Theorem 1 (Thm 2.1)

Let $X_1, \ldots, X_n$ be n iid Gaussian random variables $X_i \sim N(\mu, \sigma^2)$, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then given any $\epsilon > 0$:

$$0.5e^{-n(\epsilon+\sigma/\sqrt{n})^2/2\sigma^2} \leq \Pr(\bar{X}_n \geq \mu + \epsilon) \leq 0.5e^{-n\epsilon^2/2\sigma^2}.$$

- Exponential inequality: the tail probability of a normal random variable decays exponentially fast as ϵ increases.
- The result is asymptotically tight as $n \to \infty$. For any $\epsilon > 0$:

$$\lim_{n\to\infty} \frac{1}{n} \ln \Pr(|\bar{X}_n - \mu| \geq \epsilon) = -\frac{\epsilon^2}{2\sigma^2}.$$

Proof of Theorem 1 (Upper-bound)

Consider a standard normal random variable $X \sim N(0, 1)$:

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

Given $\epsilon > 0$, we can upper bound the tail probability $\Pr(X \geq \epsilon)$.

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

Therefore we have

$$\Pr(X \geq \epsilon) \leq 0.5 e^{-\epsilon^2/2}.$$

Since $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0, 1)$, we obtain the bound from

$$\Pr(\bar{X}_n \geq \mu + \epsilon) = \Pr(\sqrt{n}(\bar{X}_n - \mu)/\sigma \geq \sqrt{n}\epsilon/\sigma).$$

Proof f of Theorem 1 (Lower-bound)

We also have the following lower bound:

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

Therefore we have

$$0.5 e^{-(\epsilon+1)^2/2} \leq \Pr(X \geq \epsilon).$$

Since $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0, 1)$, we obtain the bound from

$$\Pr(\bar{X}_n \geq \mu + \epsilon) = \Pr(\sqrt{n}(\bar{X}_n - \mu)/\sigma \geq \sqrt{n}\epsilon/\sigma).$$

Markov's Inequality

More generally, we can derive tail inequality using Markov' inequality.

Theorem 2 (Markov's Inequality, Thm 2.2)

Given any non-negative function $h(x) \geq 0$*, and a set* $S \subset \mathbb{R}$*, we have*

$$\Pr(\bar{X}_n \in S) \leq \frac{\mathbb{E}\, h(\bar{X}_n)}{\inf_{x \in S} h(x)}.$$

Proof of Theorem 2.

Since $h(x)$ is non-negative, we have

$$\mathbb{E}\, h(\bar{X}_n) \geq \mathbb{E}_{\bar{X}_n \in S}\, h(\bar{X}_n) \geq \mathbb{E}_{\bar{X}_n \in S}\, h_S = \Pr(\bar{X}_n \in S)\, h_S,$$

where $h_S = \inf_{x \in S} h(x)$. This leads to the desired bound. □

Example: Chebyshev's Inequality

Corollary 3 (Chebyshev's Inequality)

We have

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(X_1)}{n\epsilon^2}. \tag{1}$$

Proof of Corollary 3.

Let $h(x) = x^2$, then

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

The desired bound follows from the Markov inequality with $S = \{|\bar{X}_n - \mu| \geq \epsilon\}$. □

Exponential Tail Inequality

In order to obtain exponential tail bounds, we will need to choose

$$h(z) = e^{\lambda n z}$$

in Markov's inequality with some tuning parameter $\lambda \in \mathbb{R}$.

Given $\epsilon > 0$, then the Markov inequality of Theorem 2 with $S = \{\bar{X}_n - \mu \geq \epsilon\}$ becomes

$$\begin{aligned} \Pr(\bar{X}_n \geq \mu + \epsilon) \leq & \frac{\mathbb{E} e^{\lambda n \bar{X}_n}}{e^{\lambda n(\mu+\epsilon)}} = \frac{\mathbb{E} e^{\lambda \sum_{i=1}^n X_i}}{e^{\lambda n(\mu+\epsilon)}} \\ = & \frac{\mathbb{E} \prod_{i=1}^n e^{\lambda X_i}}{e^{\lambda n(\mu+\epsilon)}} = e^{-\lambda n(\mu+\epsilon)} \left[\mathbb{E} e^{\lambda X_1}\right]^n . \end{aligned}$$

Note that in order to use this estimate, we have to assume that $\mathbb{E} e^{\lambda(X_1-\mu)} < \infty$ for some $\lambda > 0$.

Rate Function

Definition 4

Given a random variable X, we may define its logarithmic moment generating function as

$$\Lambda_X(\lambda) = \ln \mathbb{E} e^{\lambda X}.$$

Moreover, given $z \in \mathbb{R}$, the rate function $I_X(z)$ is defined as

$$I_X(z) = \begin{cases} \sup_{\lambda > 0} \left[\lambda z - \Lambda_X(\lambda)\right] & z > \mu \\ 0 & z = \mu \\ \sup_{\lambda < 0} \left[\lambda z - \Lambda_X(\lambda)\right] & z < \mu, \end{cases}$$

where $\mu = \mathbb{E}[X]$.

Upper Bound

Theorem 5 (Thm 2.5)

For any n and $\epsilon > 0$*:*

$$\frac{1}{n}\ln \Pr(\bar{X}_n \geq \mu + \epsilon) \leq -\, I_{X_1}(\mu + \epsilon) = \inf_{\lambda > 0} \left[-\lambda(\mu + \epsilon) + \ln \mathbb{E} e^{\lambda X_1} \right],$$
$$\frac{1}{n}\ln \Pr(\bar{X}_n \leq \mu - \epsilon) \leq -\, I_{X_1}(\mu - \epsilon) = \inf_{\lambda < 0} \left[-\lambda(\mu - \epsilon) + \ln \mathbb{E} e^{\lambda X_1} \right].$$

The first inequality of Theorem 5 can be rewritten as

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \exp[-n I_{X_1}(\mu + \epsilon)].$$

It shows that the tail probability of the empirical mean decays exponentially fast, if the rate function $I_{X_1}(\cdot)$ is finite.

Proof

We choose $h(z) = e^{\lambda n z}$ in Theorem 2 with $S = \{\bar{X}_n - \mu \geq \epsilon\}$. For $\lambda > 0$, we have

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \frac{\mathbb{E} e^{\lambda n \bar{X}_n}}{e^{\lambda n(\mu+\epsilon)}} = \frac{\mathbb{E} e^{\lambda \sum_{i=1}^n X_i}}{e^{\lambda n(\mu+\epsilon)}}$$
$$= \frac{\mathbb{E} \prod_{i=1}^n e^{\lambda X_i}}{e^{\lambda n(\mu+\epsilon)}} = e^{-\lambda n(\mu+\epsilon)} \left[\mathbb{E} e^{\lambda X_1}\right]^n .$$

The last equation used the independence of X_i as well as they are identically distributed. Therefore by taking logarithm, we obtain

$$\ln \Pr(\bar{X}_n \geq \mu + \epsilon) \leq n \left[-\lambda(\mu + \epsilon) + \ln \mathbb{E}\, e^{\lambda X_1}\right] .$$

Taking $\inf$ over $\lambda > 0$ on the right hand side, we obtain the first desired bound. Similarly, we can obtain the second bound.

Example: Gaussian Random Variable

Assume that $X_i \sim N(\mu, \sigma^2)$, then the exponential moment is

$$\begin{aligned}\mathbb{E}e^{\lambda(X_1-\mu)} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\lambda x} e^{-x^2/2\sigma^2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\lambda^2\sigma^2/2} e^{-(x/\sigma-\lambda\sigma)^2/2} dx/\sigma = e^{\lambda^2\sigma^2/2}.\end{aligned}$$

Therefore,

$$I_{X_1}(\mu+\epsilon) = \sup_{\lambda>0}\left[\lambda\epsilon - \ln \mathbb{E}e^{\lambda(X_1-\mu)}\right] = \sup_{\lambda>0}\left[\lambda\epsilon - \frac{\lambda^2\sigma^2}{2}\right] = \frac{\epsilon^2}{2\sigma^2},$$

where the optimal λ is achieved at $\lambda = \epsilon/\sigma^2$. Therefore

$$\Pr(\bar{X}_n \geq \mu+\epsilon) \leq \exp[-nI_{X_1}(\mu+\epsilon)] = \exp\left[\frac{-n\epsilon^2}{2\sigma^2}\right].$$

Lower Bound

In the large deviation situation, the exponential Markov inequality is asymptotically tight in the following sense.

Theorem 6 (Thm 2.7)

For all $\epsilon' > \epsilon > 0$*:*

$$\underline{\lim}_{n\to\infty}\frac{1}{n}\ln\Pr(\bar{X}_n \geq \mu+\epsilon) \geq -I_{X_1}(\mu+\epsilon').$$

Similarly,

$$\underline{\lim}_{n\to\infty}\frac{1}{n}\ln\Pr(\bar{X}_n \leq \mu-\epsilon) \geq -I_{X_1}(\mu-\epsilon').$$

Rate Function: Some Intuitions

Proposition 7

Given a random variable with finite variance. We have:

$$\Lambda_X(\lambda)\Big|_{\lambda=0} = 0, \ \frac{d\Lambda_X(\lambda)}{d\lambda}\Big|_{\lambda=0} = \mathbb{E}[X], \ \frac{d^2\Lambda_X(\lambda)}{d\lambda^2}\Big|_{\lambda=0} = \mathrm{Var}[X].$$

$$\Lambda_X(\lambda) = \lambda\mu + \frac{\lambda^2}{2}\mathrm{Var}[X] + o(\lambda^2),$$

where $\mu = \mathbb{E}[X]$. Therefore

$$\begin{aligned} I_X(\mu+\epsilon) &= \sup_{\lambda>0}\left[\lambda(\mu+\epsilon) - \lambda\mu - \frac{\lambda^2}{2}\mathrm{Var}[X] - o(\lambda^2)\right] \\ &\approx \frac{\epsilon^2}{2\mathrm{Var}[X]} - o(\epsilon^2). \end{aligned}$$

$$\Pr(\bar{X}_n \geq \mu+\epsilon) \lesssim \exp\left[-\frac{n\epsilon^2}{2\mathrm{Var}[X_1]} + o(n\epsilon^2)\right]. \quad (2)$$

Sub-Gaussian Random Variables

Definition 8

A sub-Gaussian random variable X_1 has quadratic logarithmic moment generating function:

$$\ln \mathbb{E} e^{\lambda X_1} \leq \lambda \mu + \frac{\lambda^2}{2} b. \tag{3}$$

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

$$-I_{X_1}(z) = \inf_{\lambda > 0} (-\lambda z + \lambda \mu + \frac{\lambda^2}{2} b).$$

We have the following condition at the optimal λ_*:

$$-z + \mu = \lambda_* b \quad \Longrightarrow \quad \lambda_* = (\mu - z)/b,$$

which implies that

$$I_{X_1}(z) = \frac{(z - \mu)^2}{2b}.$$

Tail Inequality for Sub-Gaussians

Theorem 9 (Thm 2.12)

If X_1 is sub-Gaussian as in (3)*, then for all $\epsilon > 0$:*

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

Example 10 (Gaussian)

Gaussian random variable $X_1 \sim N(\mu, \sigma^2)$ is sub-Gaussian with $b = \sigma^2$.

Example 11 (from Chernoff bound)

Consider a bounded random variable: $X_1 \in [\alpha, \beta]$. Then X_1 is sub-Gaussian with $b = (\beta - \alpha)^2/4$.

Alternative Expression of Tail Bounds

The tail probability inequality of Theorem 9 can also be expressed in a different form. Consider $\delta \in (0, 1)$ such that

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \exp(-n\epsilon^2/2b) = \delta.$$

We can solve for

$$\epsilon = \sqrt{(2b/n)\ln(1/\delta)}.$$

This implies that we can alternatively express the bound of Theorem 9 as follows.

Alternative Expression

With probability at least $1 - \delta$, we have

$$\bar{X}_n < \mu + \sqrt{\frac{2b\ln(1/\delta)}{n}}.$$

A Generic Estimate on Rate Function

Lemma 12 (Lem 2.9)

Consider a random variable X so that $\mathbb{E}[X] = \mu$. Assume that there exists $\alpha > 0$ and $\beta \geq 0$ such that for $\lambda \in [0, \beta^{-1})$:

$$\Lambda_X(\lambda) \leq \lambda\mu + \frac{\alpha\lambda^2}{2(1 - \beta\lambda)}, \tag{4}$$

then for $\epsilon > 0$:

$$-I_X(\mu + \epsilon) \leq -\frac{\epsilon^2}{2(\alpha + \beta\epsilon)},$$
$$-I_X\left(\mu + \epsilon + \frac{\beta\epsilon^2}{2\alpha}\right) \leq -\frac{\epsilon^2}{2\alpha}.$$

Tail Probability Bound

Lemma 12 implies the following generic theorem.

Theorem 13 (Thm 2.10)

If X_1 has a logarithmic moment generating function that satisfies (4) *for $\lambda > 0$, then all $\epsilon > 0$:*

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \exp\left[\frac{-n\epsilon^2}{2(\alpha + \beta\epsilon)}\right].$$

Moreover, for $t > 0$, we have

$$\Pr\left(\bar{X}_n \geq \mu + \sqrt{\frac{2\alpha t}{n}} + \frac{\beta t}{n}\right) \leq e^{-t}.$$

Chernoff Bound

We consider a random variable $X \in [0, 1]$ and $\mathbb{E}X = \mu$. Chernoff bound, or Hoeffding's inequality, is an exponential tail inequality for bounded random variables.

Theorem 14 (Additive Chernoff bounds, Thm 2.16)

Assume that $X_1 \in [0, 1]$. *Then for all* $\epsilon > 0$*:*

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

Proof: Moment Generation Function

Lemma 15 (Lem 2.15)

Consider a random variable $X \in [0, 1]$ and $\mathbb{E}X = \mu$. We have the following inequality:

$$\ln \mathbb{E}e^{\lambda X} \leq \ln[(1-\mu)e^0 + \mu e^\lambda] \leq \lambda\mu + \lambda^2/8.$$

This lemma shows that the random variable X_1 is sub-Gaussian. We can thus apply the sub-Gaussian tail-inequality in Theorem 9 to obtain the Chernoff bound.

Proof of Lemma 15

Let $h_L(\lambda) = \mathbb{E}e^{\lambda X}$ and $h_R(\lambda) = (1-\mu)e^0 + \mu e^\lambda$. We know that $h_L(0) = h_R(0)$. Moreover, when $\lambda \geq 0$:

$$h_L'(\lambda) = \mathbb{E}Xe^{\lambda X} \leq \mathbb{E}Xe^\lambda = \mu e^\lambda = h_R'(\lambda),$$

and similarly $h_L'(\lambda) \geq h_R'(\lambda)$ when $\lambda \leq 0$. This proves the first inequality. Now we let

$$h(\lambda) = \ln[(1-\mu)e^0 + \mu e^\lambda].$$

It implies that

$$h'(\lambda) = \frac{\mu e^\lambda}{(1-\mu)e^0 + \mu e^\lambda},$$

and

$$\begin{aligned} h''(\lambda) =& \frac{\mu e^\lambda}{(1-\mu)e^0 + \mu e^\lambda} - \frac{(\mu e^\lambda)^2}{[(1-\mu)e^0 + \mu e^\lambda]^2} \\ =& |h'(\lambda)|(1 - |h'(\lambda)|) \leq 1/4. \end{aligned}$$

Using Taylor expansion, we obtain the inequality $h(\lambda) \leq h(0) + \lambda h'(0) + \lambda^2/8$, which implies the second inequality.

Multiplicative Chernoff Bounds

Corollary 16 (Multiplicative Chernoff Bounds, Cor 2.18)

Assume that $X_1 \in [0, 1]$*. Then for all* $\epsilon > 0$*:*

$$\Pr\left(\bar{X}_n \geq (1+\epsilon)\mu\right) \leq \exp\left[\frac{-n\mu\epsilon^2}{2+\epsilon}\right],$$

$$\Pr\left(\bar{X}_n \leq (1-\epsilon)\mu\right) \leq \exp\left[\frac{-n\mu\epsilon^2}{2}\right].$$

Moreover, for $t > 0$*, we have:*

$$\Pr\left(\bar{X}_n \geq \mu + \sqrt{\frac{2\mu t}{n}} + \frac{t}{3n}\right) \leq e^{-t}.$$

Alternative Expressions

The multiplicative form of Chernoff bound can be expressed alternatively as follows. With probability at least $1 - \delta$:

$$\mu < \bar{X}_n + \sqrt{\frac{2\mu \ln(1/\delta)}{n}}.$$

It implies that for any $\gamma \in (0, 1)$:

$$\bar{X}_n > (1 - \gamma)\mu - \frac{\ln(1/\delta)}{2\gamma n}. \tag{5}$$

Moreover, with probability at least $1 - \delta$:

$$\bar{X}_n < \mu + \sqrt{\frac{2\mu \ln(1/\delta)}{n}} + \frac{\ln(1/\delta)}{3n}.$$

It implies that for any $\gamma > 0$:

$$\bar{X}_n < (1 + \gamma)\mu + \frac{(3 + 2\gamma)\ln(1/\delta)}{6\gamma n}. \tag{6}$$

Bennett's Inequality

From (2), we know that the leading term of the tail inequality is

$$\frac{-n\epsilon^2}{2\mathrm{Var}(X_1)},$$

which is superior to Chernoff bound when variance is small.

Theorem 17 (Bennett's Inequality, simplification of Thm 2.21)

If $X_1 \leq \mu + b$, for some $b > 0$. Then $\forall \epsilon > 0$:

$$\Pr[\bar{X}_n \geq \mu + \epsilon] \leq \exp\left[\frac{-n\epsilon^2}{2\mathrm{Var}(X_1) + 2\epsilon b/3}\right].$$

Moreover, for $t > 0$:

$$\Pr\left[\bar{X}_n \geq \mu + \sqrt{\frac{2\mathrm{Var}(X_1)t}{n}} + \frac{bt}{3n}\right] \leq e^{-t}.$$

Alternative Form

Bennett's Inequality: Alternative Expression

Given any $\delta \in (0, 1)$, with probability larger than $1 - \delta$, we have

$$\bar{X}_n \leq \mu + \sqrt{\frac{2\mathrm{Var}(X_1)\ln(1/\delta)}{n}} + \frac{b\ln(1/\delta)}{3n}.$$

Compared to the bound for Gaussian random variables, this form of Bennett's inequality has an extra term $b\ln(1/\delta)/(3n)$, which is of higher order in $1/n$. It vanishes asymptotically.

Compared to the Chernoff bound, the Bennett's inequality is superior when $\mathrm{Var}(X_1)$ is small.

Proof of Theorem 17 (I/II)

Lemma 18 (Lem 2.20)

If $X - \mathbb{E}X \leq b$, then $\forall\lambda \geq 0$:

$$\ln \mathbb{E}e^{\lambda X} \leq \lambda \mathbb{E}X + \lambda^2 \phi(\lambda b)\mathrm{Var}(X),$$

where $\phi(z) = (e^z - z - 1)/z^2$.

Proof of Lemma 18.

Let $X' = X - \mathbb{E}X$. We have

$$\begin{aligned}
\ln \mathbb{E}e^{\lambda X} =& \lambda \mathbb{E}X + \ln \mathbb{E}e^{\lambda X'} \leq \lambda \mathbb{E}X + \mathbb{E}e^{\lambda X'} - 1 \\
=& \lambda \mathbb{E}X + \lambda^2 \mathbb{E}\frac{e^{\lambda X'} - \lambda X' - 1}{(\lambda X')^2}(X')^2 \\
\leq& \lambda \mathbb{E}X + \lambda^2 \mathbb{E}\phi(\lambda b)(X')^2.
\end{aligned}$$

The first inequality used $\ln z \leq z - 1$; the second inequality used the fact that the function $\phi(z)$ is non-decreasing and $\lambda X' \leq \lambda b$. □

Proof of Theorem 17 (II/II)

Given $\lambda \in (0, 3/b)$, it is easy to verify the following inequality using the Taylor expansion of the exponential function

$$\begin{aligned}\Lambda_{X_1}(\lambda) \leq& \mu\lambda + b^{-2}\left[e^{\lambda b} - \lambda b - 1\right]\mathrm{Var}(X_1)\\ \leq& \mu\lambda + \frac{\mathrm{Var}(X_1)\lambda^2}{2}\sum_{m=0}^{\infty}(\lambda b/3)^m = \mu\lambda + \frac{\mathrm{Var}(X_1)\lambda^2}{2(1-\lambda b/3)}.\end{aligned} \tag{7}$$

The desired bound follow from a direct application of Theorem 13 with $\alpha = \mathrm{Var}(X_1)$ and $\beta = b/3$.

Bernstein's Inequality: Moment Condition

Lemma 19 (Lem 2.22)

If X satisfies the following moment condition for integers $m \geq 2$:

$$\mathbb{E}[X - c]^m \leq m!(b/3)^{m-2} V/2,$$

where $b, V > 0$ and c is arbitrary. Then when $\lambda \in (0, 3/b)$:

$$\ln \mathbb{E} e^{\lambda X} \leq \lambda \mathbb{E} X + \frac{\lambda^2 V}{2(1 - \lambda b/3)}.$$

Proof of Theorem 19.

It follows from the logarithmic moment generating function estimate below:

$$\begin{aligned} \ln \mathbb{E} e^{\lambda X} \leq \lambda c + \mathbb{E} e^{\lambda(X-c)} - 1 \leq & \lambda \mathbb{E} X + 0.5 V \lambda^2 \sum_{m=2} (b/3)^{m-2} \lambda^{m-2} \\ = & \lambda \mathbb{E} X + 0.5 \lambda^2 V (1 - \lambda b/3)^{-1}. \end{aligned}$$

□

Bernstein's Inequality

Theorem 20 (Thm 2.23)

Assume that X_1 satisfies the moment condition in Lemma 19. Then for all $\epsilon > 0$:

$$\Pr[\bar{X}_n \geq \mu + \epsilon] \leq \exp\left[\frac{-n\epsilon^2}{2V + 2\epsilon b/3}\right],$$

and for all $t > 0$:

$$\Pr\left[\bar{X}_n \geq \mu + \sqrt{\frac{2Vt}{n}} + \frac{bt}{3n}\right] \leq e^{-t}.$$

Proof of Theorem 20.

We simply set $\alpha = V$ and $\beta = b/3$ in Theorem 13. □

Note that if the random variable X is bounded with $|X| \leq M$, then the moment condition holds with $b = M/3$ and $V = \mathrm{Var}(X)$.

Non-IID Case

If $X_1, \ldots, X_n$ are independent but not identically distributed random variables, then a similar tail inequality holds.

Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}\bar{X}_n$, then we have the following bound.

Theorem 21 (Thm 2.25)

We have for all $\epsilon > 0$:

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \inf_{\lambda > 0} \left[-\lambda n(\mu + \epsilon) + \sum_{i=1}^{n} \ln \mathbb{E} e^{\lambda X_i} \right].$$

Sub-Gaussian Bound

Corollary 22 (Cor 2.26)

If $\{X_i\}$ *are independent sub-Gaussian random variables with* $\ln \mathbb{E} e^{\lambda X_i} \leq \lambda \mathbb{E} X_i + 0.5\lambda^2 b_i$, *then for all* $\epsilon > 0$*:*

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \exp\left[-\frac{n^2\epsilon^2}{2\sum_{i=1}^n b_i}\right].$$

Example

The following is a useful example for Rademacher average using sub-Gaussian bound.

Corollary 23 (Cor 2.27)

Let $\sigma_i = \{\pm 1\}$ be independent random Bernoulli variables, and let a_i be fixed numbers ($i = 1, \ldots, n$). Then for all $\epsilon > 0$:

$$\Pr(n^{-1} \sum_{i=1}^{n} \sigma_i a_i \geq \epsilon) \leq \exp\left[-\frac{n\epsilon^2}{2n^{-1} \sum_{i=1}^{n} a_i^2} \right].$$

Summary (Chapter 2)

- Exponential Tail Inequalities can be used to bound the difference of true mean and the observed empirical mean.
- Gaussian case: direct calculation
- Markov Inequality: upper bounds
- Nearly matching lower bounds.
- Chernoff bound: useful for the generation situation, with deviation of order $O(1/\sqrt{n})$.
- Bennett/Bernstein's inequality: refined bound which is useful when variance is small, and when we want to achieve faster than $O(1/\sqrt{n})$ convergence.