2

Basic Probability Inequalities for Sums of Independent Random Variables

In machine learning, the observations contain uncertainty, and to incorporate uncertainty, these observations are modeled as random variables. When we observe many data, a basic quantity of interest is the empirical mean of the observed random variables, which converges to the expectation according to the law of large numbers. We want to upper bound the probability of the event when the empirical mean deviates significantly from the expectation, which is referred to as the tail probability. This chapter studies the basic mathematical tools to estimate tail probabilities by using exponential moment estimates.

Let $X_1, \ldots, X_n$ be n real-valued independent and identically distributed (iid) random variables, with expectation $\mu = \mathbb{E}X_i$. Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.1}$$

Given $\epsilon > 0$, we are interested in estimating the following tail probabilities:

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

In machine learning, we can regard $\bar{X}_n$ as the training error observed on the training data. The unknown mean μ is the test error that we want to infer from the training error. Therefore in machine learning, these tail inequalities can be interpreted as follows: with high probability, the test error is close to the training error. Such results will be used to derive rigorous statements of generalization error bounds in subsequent chapters.

2.1 Normal Random Variable

The general form of tail inequality for the sum of random variables (with relatively light tails) is exponential in ϵ^2. To motivate this general form, we will consider the case of normal random variables. The bounds can be obtained using simple calculus.

Theorem 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 $\epsilon > 0$,*

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

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Proof We first consider a standard normal random variable $X \sim N(0,1)$, which has probability density function

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

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

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

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

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

Since $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0,1)$, by using

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

we obtain the desired result. □

We note that the tail probability of a normal random variable decays exponentially fast, and such an inequality is referred to as an *exponential inequality*. This exponential bound is asymptotically tight as $n \to \infty$ in the following sense. For any $\epsilon > 0$, we have

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

Such a result is also called a large deviation result, which is the regime when the deviation ϵ of the empirical mean from the true mean μ is much larger than the standard deviation $\sigma/\sqrt{n}$ of $\bar{X}_n$ (Deuschel and Stroock, 2001). The analysis of normal random variables can rely on standard calculus. For general random variables with exponentially decaying tail probabilities, we can use the technique of exponential moment to derive similar results. This leads to a general technique to estimate the probability of large deviation of the empirical mean from the true mean.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

2.2 Markov's Inequality

A standard technique to estimate the tail inequality of a random variable is the Markov inequality. Let $X_1, \ldots, X_n$ be n real-valued iid random variables (which are not necessarily normal random variables) with mean μ. Let $\bar{X}_n$ be the empirical mean defined in (2.1). We are interested in estimating the tail bound $\Pr(\bar{X}_n \geq \mu + \epsilon)$, and Markov's inequality states as follows.

Theorem 2.2 (Markov's Inequality) *Given any nonnegative 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 Since $h(x)$ is nonnegative, 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. □

In particular, we may consider the choice of $h(z) = z^2$, which leads to Chebyshev's inequality.

Corollary 2.3 (Chebyshev's Inequality) *We have*

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(X_1)}{n\epsilon^2}. \tag{2.2}$$

Proof 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 $S = \{|\bar{X}_n - \mu| \geq \epsilon\}$. □

Note that Chebyshev's inequality employs $h(z) = z^2$, which leads to a tail inequality that is polynomial in n^{-1} and ϵ. It only requires that the variance of a random variable is bounded. In comparison, the Gaussian tail inequality has a much faster exponential decay. Exponential tail inequality is important for analyzing learning algorithms. In the following, we show that such an inequality can be established for sums of random variables with exponentially decaying tail probabilities.

2.3 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 $\lambda \in \mathbb{R}$. Similar to Chebyshev's inequality, which requires that the variance of a random variable is bounded, we assume exponential moment $\mathbb{E}e^{\lambda X_1} < \infty$ for some $\lambda \neq 0$. This requires that the random variable X_i has tail probability that decays exponentially fast. The following definition is helpful in the analysis.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Definition 2.4 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

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

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

This definition can be used to obtain exponential tail bounds for sums of independent variables as follows.

Theorem 2.5 *For any n and $\epsilon > 0$,*

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

Proof We choose $h(z) = e^{\lambda n z}$ in Theorem 2.2 with $S = \{\bar{X}_n - \mu \geq \epsilon\}$. For $\lambda > 0$, we have

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

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

The first inequality of Theorem 2.5 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. More concrete exponential tail inequalities can be obtained by applying Theorem 2.5 to specific random variables. For example, for Gaussian random variables, we can derive a tail inequality using Theorem 2.5, and compare it to that of Theorem 2.1.

Example 2.6 (Gaussian Random Variable) Assume that $X_i \sim N(\mu, \sigma^2)$, then the exponential moment is

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

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

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

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

This leads to the same probability bound as that of Theorem 2.1 up to a constant factor.

This Gaussian example, together with Theorem 2.1, implies that the exponential inequality derived from Theorem 2.5 is asymptotically tight. This result can be generalized to the large deviation inequality for general random variables. In particular, we have the following theorem.

Theorem 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,

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

Proof We only need to prove the first inequality. Consider $\Pr(X_i \leq x)$ as a function of x, and define a random variable X_i' with density at x as

$$d\Pr(X_i' \leq x) = e^{\lambda x - \Lambda_{X_1}(\lambda)} d\Pr(X_i \leq x).$$

This choice implies that

$$\frac{d}{d\lambda}\Lambda_{X_1}(\lambda) = \frac{\int x e^{\lambda x} d\Pr(X_1 \leq x)}{\int e^{\lambda x} d\Pr(X_1 \leq x)} = \mathbb{E}_{X_1'} X_1'.$$

We now take λ such that

$$\lambda = \arg\max_{\lambda'>0} \left[\lambda'(\mu+\epsilon') - \Lambda_{X_1}(\lambda')\right].$$

By setting the derivative to zero, we obtain

$$\mathbb{E}_{X_1'} X_1' = \frac{d}{d\lambda}\Lambda_{X_1}(\lambda) = \mu+\epsilon', \tag{2.3}$$

$$-\lambda(\mu+\epsilon') + \Lambda(\lambda) = -I(\mu+\epsilon'). \tag{2.4}$$

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Let $\bar{X}'_n = n^{-1}\sum_{i=1}^n X'_i$. Then, by the law of large numbers, we know that for $\epsilon'' > \epsilon'$, we obtain from (2.3)

$$\lim_{n\to\infty} \Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon'']) = 1. \tag{2.5}$$

Since the joint density of $(X'_1, \ldots, X'_n)$ satisfies

$$e^{-\lambda\sum_{i=1}^n x_i + n\Lambda_{X_1}(\lambda)} \prod_i d\Pr(X'_i \le x_i) = \prod_i d\Pr(X_i \le x_i), \tag{2.6}$$

by using $\mathbb{1}(\cdot)$ to denote the set indicator function, we obtain

$$\begin{aligned}
\Pr(\bar{X}_n \ge \mu + \epsilon) &\ge \Pr(\bar{X}_n - \mu \in [\epsilon, \epsilon'']) \\
&= \mathbb{E}_{X_1,\ldots,X_n} \mathbb{1}(\bar{X}_n - \mu \in [\epsilon, \epsilon'']) \\
&= \mathbb{E}_{X'_1,\ldots,X'_n} e^{-\lambda n\bar{X}'_n + n\Lambda(\lambda)} \mathbb{1}(\bar{X}'_n - \mu \in [\epsilon, \epsilon'']) \\
&\ge e^{-\lambda n(\mu+\epsilon'') + n\Lambda(\lambda)} \Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon'']).
\end{aligned}$$

The first equality used the definition of $\Pr(\cdot)$. The second equality used (2.6). The last inequality used Markov's inequality. Now by taking the logarithm, and dividing by n, we obtain

$$\begin{aligned}
&\frac{1}{n}\ln\Pr(\bar{X}_n \ge \mu + \epsilon) \qquad (2.7)\\
&\quad \ge -\lambda(\mu+\epsilon'') + \Lambda(\lambda) + \frac{1}{n}\ln\Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon'']) \\
&\quad = -I(\mu+\epsilon') - \lambda(\epsilon''-\epsilon') + \frac{1}{n}\ln\Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon'']). \qquad (2.8)
\end{aligned}$$

The equality used (2.4). Now we obtain the desired bound by letting $n \to \infty$, applying (2.5), and letting $\epsilon'' \to \epsilon'$ so that $\lambda(\epsilon''-\epsilon') \to 0$ (this is true because λ depends only on ϵ'). □

The combination of Theorem 2.5 and Theorem 2.7 shows that the large deviation tail probability is determined by the rate function. This result is referred to as Cramér's theorem (Cramér, 1938; Deuschel and Stroock, 2001).

For specific cases, one can obtain an estimate of $\Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon''])$ in (2.8) with finite n at $\epsilon' = \epsilon + 2\sqrt{\mathrm{Var}(X_1)/n}$ and $\epsilon'' = \epsilon + 4\sqrt{\mathrm{Var}(X_1)/n}$. Using Chebyshev's inequality, we expect that $\Pr(\bar{X}'_n - \mu \in [\epsilon, \epsilon''])$ is lower bounded by a constant. This means that as $n \to \infty$, the exponential tail inequality of Theorem 2.5 is generally loose by no more than $O(\sqrt{\mathrm{Var}(X_1)/n})$ in terms of deviation ϵ. A concrete calculation will be presented for bounded random variables in Section 2.5.

Before we investigate concrete examples of random variables, we state the following property of the logarithmic generating function of a random variable, which provides intuitions on its behavior. The proof is left as an exercise.

Proposition 2.8 *Given a random variable with finite variance, we have*

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

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

In the application of large deviation bounds, we are mostly interested in the case that deviation ϵ is close to zero. As shown in Example 2.6, the optimal λ we shall choose is $\lambda = O(\epsilon) \approx 0$. It is thus natural to consider the Taylor expansion of the logarithmic moment generating function around $\lambda = 0$. Proposition 2.8 implies that the leading terms of the Taylor expansion are

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

where $\mu = \mathbb{E}[X]$. The first two terms match that of the normal random variable in Example 2.6. When $\epsilon > 0$ is small, to obtain the rate function

$$I_X(\mu+\epsilon) = \sup_{\lambda>0}\left[\lambda(\mu+\epsilon) - \lambda\mu - \frac{\lambda^2}{2}\mathrm{Var}[X] - o(\lambda^2)\right],$$

we should set the optimal λ approximately as $\lambda \approx \epsilon/\mathrm{Var}[X]$, and the corresponding rate function becomes

$$I_X(\mu+\epsilon) \approx \frac{\epsilon^2}{2\mathrm{Var}[X]} + o(\epsilon^2).$$

For specific forms of logarithmic moment generation functions, one may obtain more precise bounds of the rate function. In particular, the following general estimate is useful in many applications. This estimate is what we will use throughout the chapter.

Lemma 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 $\lambda \in [0, \beta^{-1})$,*

$$\Lambda_X(\lambda) \leq \lambda\mu + \frac{\alpha\lambda^2}{2(1-\beta\lambda)}, \tag{2.9}$$

then for $\epsilon > 0$,

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

Proof Note that

$$-I_X(\mu+\epsilon) \leq \inf_{\lambda>0}\left[-\lambda(\mu+\epsilon) + \lambda\mu + \frac{\alpha\lambda^2}{2(1-\beta\lambda)}\right].$$

We can take λ at $\bar{\lambda} = \epsilon/(\alpha+\beta\epsilon)$. This implies that $\alpha\bar{\lambda}/(1-\beta\bar{\lambda}) = \epsilon$. Therefore

$$-I_X(\mu+\epsilon) \leq -\bar{\lambda}\epsilon + \frac{\alpha\bar{\lambda}^2}{2(1-\beta\bar{\lambda})} = -\frac{\bar{\lambda}\epsilon}{2} = -\frac{\epsilon^2}{2(\alpha+\beta\epsilon)}.$$

Moreover, with the same choice of $\bar{\lambda}$, we have

$$-I_X\left(\mu+\epsilon+\frac{\beta}{2\alpha}\epsilon^2\right) \leq -\bar{\lambda}\epsilon\left(1+\frac{\beta}{2\alpha}\epsilon\right) + \frac{\alpha\bar{\lambda}^2}{2(1-\beta\bar{\lambda})} = -\frac{\epsilon^2}{2\alpha}.$$

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

This proves the second desired bound. □

Lemma 2.9 implies the following generic theorem.

Theorem 2.10 *If X_1 has a logarithmic moment generating function that satisfies (2.9) for $\lambda > 0$, then for 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

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

Proof The first inequality of the theorem follows from the first inequality of Lemma 2.9 and Theorem 2.5. The second inequality of the theorem follows from the second inequality of Lemma 2.9 and Theorem 2.5, with $\epsilon = \sqrt{2\alpha t/n}$. □

2.4 Sub-Gaussian Random Variable

The logarithmic moment generating function of a normal random variable is quadratic in λ. More generally, we may define a sub-Gaussian random variable as a random variable with logarithmic moment generating function dominated by a quadratic function in λ. Such random variables have light tails, which implies that they have a tail probability inequality similar to that of a Gaussian random variable.

Definition 2.11 A sub-Gaussian random variable X has quadratic logarithmic moment generating function for all $\lambda \in \mathbb{R}$:

$$\ln \mathbb{E}e^{\lambda X} \leq \lambda\mu + \frac{\lambda^2}{2}b. \tag{2.10}$$

Using (2.10), we can obtain an upper bound of the rate function for sub-Gaussian random variables, which imply the following tail inequality.

Theorem 2.12 *If X_1 is sub-Gaussian as in (2.10), then for all $t > 0$,*

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

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

Proof The result follows from Theorem 2.10 with $\alpha = b$ and $\beta = 0$. □

Common examples of sub-Gaussian random variables include Gaussian random variables and bounded random variables.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Example 2.13 A Gaussian random variable $X_1 \sim N(\mu, \sigma^2)$ is sub-Gaussian with $b = \sigma^2$.

Example 2.14 Consider a bounded random variable: $X_1 \in [\alpha, \beta]$. Then X_1 is sub-Gaussian with $b = (\beta - \alpha)^2/4$.

The tail probability inequality of Theorem 2.12 can also be expressed in a different form. Consider $\delta \in (0, 1)$ such that $\delta = \exp(-t)$, so we have $t = \ln(1/\delta)$. This means that we can alternatively express the first bound of Theorem 2.12 as follows. With probability at least $1 - \delta$, we have

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

This form is often preferred in the theoretical analysis of machine learning algorithms.

2.5 Hoeffding's Inequality

Hoeffding's inequality (Hoeffding, 1963) is an exponential tail inequality for bounded random variables. In the machine learning and computer science literature, it is often referred to as the Chernoff bound.

Lemma 2.15 *Consider a random variable $X \in [0, 1]$ and $\mathbb{E}X = \mu$. We have the following inequality:*

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

Proof 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 proves the second inequality. □

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

The lemma implies that the maximum logarithmic moment generating function of a random variable X taking values in $[0,1]$ is achieved by a $\{0,1\}$-valued Bernoulli random variable with the same mean. Moreover, the random variable X is sub-Gaussian. We can then apply the sub-Gaussian tail inequality in Theorem 2.12 to obtain the following additive form of Chernoff bound.

Theorem 2.16 (Additive Chernoff Bounds) *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 We simply take $b = 1/4$ and $t = 2n\epsilon^2$ in Theorem 2.12 to obtain the first inequality. The second inequality follows from the equivalence of $\bar{X}_n \leq \mu - \epsilon$ and $-\bar{X}_n \leq -\mu + \epsilon$. □

In some applications, one may need to employ a more refined form of Chernoff bound, which can be stated as follows.

Theorem 2.17 *Assume that $X_1 \in [0,1]$. Then for all $\epsilon > 0$, we have*

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq e^{-n\mathrm{KL}(\mu+\epsilon||\mu)},$$
$$\Pr(\bar{X}_n \leq \mu - \epsilon) \leq e^{-n\mathrm{KL}(\mu-\epsilon||\mu)},$$

where $\mathrm{KL}(z||\mu)$ *is the* Kullback–Leibler divergence *(KL-divergence) defined as*

$$\mathrm{KL}(z||\mu) = z \ln \frac{z}{\mu} + (1-z) \ln \frac{1-z}{1-\mu}.$$

Proof Consider the case $z = \mu + \epsilon$. We have

$$-I_{X_1}(z) \leq \inf_{\lambda > 0} [-\lambda z + \ln((1-\mu)e^0 + \mu e^\lambda)].$$

Assume that the optimal value of λ on the right-hand side is achieved at λ_*. By setting the derivative to zero, we obtain the expression

$$z = \frac{\mu e^{\lambda_*}}{(1-\mu)e^0 + \mu e^{\lambda_*}},$$

which implies that

$$e^{\lambda_*} = \frac{z(1-\mu)}{\mu(1-z)}.$$

This implies that $-I_{X_1}(z) \leq -\mathrm{KL}(z||\mu)$. The case of $z = \mu - \epsilon$ is similar. We can thus obtain the desired bound from Theorem 2.5. □

In many applications, we will be interested in the situation $\mu \approx 0$. For example, this happens when the classification error is close to zero. In this case, Theorem 2.17 is superior to Theorem 2.16, and the result implies a simplified form stated in the following corollary.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Corollary 2.18 (Multiplicative Chernoff Bounds) *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*

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

Proof The first and the second results can be obtained from Theorem 2.17 and the inequality $\mathrm{KL}(z||\mu) \geq (z-\mu)^2/\max(2\mu, \mu+z)$ (which is left as an exercise). We then take $z = (1+\epsilon)\mu$ and $z = (1-\epsilon)\mu$, respectively, for the first and the second inequalities.

For the third inequality (which is sharper than the first inequality), we may apply Theorem 2.10. Just observe from Lemma 2.15 that when $\lambda > 0$,

$$\begin{aligned}\Lambda_{X_1}(\lambda) &\leq \ln[(1-\mu)e^0 + \mu e^\lambda] \\ &\leq \mu(e^\lambda - 1) = \mu\lambda + \mu\sum_{k\geq 2}\frac{\lambda^k}{k!} \\ &\leq \mu\lambda + \frac{\mu\lambda^2}{2(1-\lambda/3)}.\end{aligned}$$

In this derivation, the equality used the Taylor expansion of exponential function. The last inequality used $k! \geq 2\cdot 3^{k-2}$ and the sum of infinite geometric series. We may take $\alpha = \mu$ and $\beta = 1/3$ in Theorem 2.10 to obtain the desired bound. □

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 $\gamma \in (0,1)$,

$$\bar{X}_n > (1-\gamma)\mu - \frac{\ln(1/\delta)}{2\gamma n}. \tag{2.11}$$

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 $\gamma > 0$,

$$\bar{X}_n < (1+\gamma)\mu + \frac{(3+2\gamma)\ln(1/\delta)}{6\gamma n}. \tag{2.12}$$

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

For Bernoulli random variables with $X_1 \in \{0,1\}$, the moment generating function achieves equality in Lemma 2.15, and thus the proof of Theorem 2.17 implies that the rate function is given by

$$I_{X_1}(z) = \mathrm{KL}(z||\mu).$$

We can obtain the following lower bound from (2.8), which suggests that the KL formulation of Hoeffding's inequality is quite tight for Bernoulli random variables when n is large.

Corollary 2.19 *Assume that $X_1 \in \{0,1\}$. Then for all $\epsilon > 0$ that satisfy*

$$\epsilon' = \epsilon + 2\sqrt{(\mu+\epsilon)(1-(\mu+\epsilon))/n} < 1-\mu$$

and $n \geq (1-\mu-\epsilon)/(\mu+\epsilon)$, we have

$$\Pr(\bar{X}_n \geq \mu+\epsilon) \geq 0.25\ \exp\left[-n\mathrm{KL}(\mu+\epsilon'||\mu) - \sqrt{n}\Delta I\right],$$

where

$$\Delta I = 2\sqrt{(\mu+\epsilon)(1-\mu-\epsilon)}\ln\frac{(\mu+\epsilon')(1-\mu)}{(1-(\mu+\epsilon'))\mu}.$$

Proof In (2.8), we let $\epsilon'' = 2\epsilon' - \epsilon$. Since $X_i' \in \{0,1\}$ and $\mathbb{E}X_i' = \mu+\epsilon'$, we have $\mathrm{Var}(X_1') = (\mu+\epsilon')(1-\mu-\epsilon')$. Using Chebyshev's inequality, we obtain

$$\begin{aligned}\Pr\left(|\bar{X}_n' - (\mu+\epsilon')| \geq \epsilon'-\epsilon\right) &\leq \frac{(\mu+\epsilon')(1-\mu-\epsilon')}{n(\epsilon'-\epsilon)^2} \\ &= \frac{(\mu+\epsilon')(1-\mu-\epsilon')}{4(\mu+\epsilon)(1-(\mu+\epsilon))} \leq \frac{(\mu+\epsilon')}{4(\mu+\epsilon)} = 0.25 + \frac{\epsilon'-\epsilon}{4(\mu+\epsilon)}.\end{aligned}$$

Therefore

$$\begin{aligned}\Pr\left(\bar{X}_n' \in (\mu+\epsilon, \mu+\epsilon'')\right) &= 1 - \Pr\left(|\bar{X}_n' - (\mu+\epsilon')| \geq \epsilon'-\epsilon\right) \\ &\geq 0.75 - \frac{\epsilon'-\epsilon}{4(\mu+\epsilon)} = 0.75 - 0.5\sqrt{\frac{1-\mu-\epsilon}{n(\mu+\epsilon)}} \geq 0.25.\end{aligned}$$

The choice of λ in (2.4) is given by

$$\lambda = \ln\frac{(\mu+\epsilon')(1-\mu)}{(1-(\mu+\epsilon'))\mu}.$$

By using these estimates, we can obtain the desired bound from (2.8). □

2.6 Bennett's Inequality

In Bennett's inequality, we assume that the random variable is upper bounded and has a small variance. In this case, one can obtain a more refined estimate of the moment generating function by using the variance of the random variable (Bennett, 1962).

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Lemma 2.20 *If $X - \mathbb{E}X \le b$, then $\forall \lambda \ge 0$:*

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

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

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

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

where the first inequality used $\ln z \le z - 1$; the second inequality follows from the fact that the function $\phi(z)$ is non-decreasing (left as an exercise) and $\lambda X' \le \lambda b$. □

Lemma 2.20 gives an estimate of the logarithmic moment generating function, which implies the following result from Theorem 2.5.

Theorem 2.21 (Bennett's Inequality) *If $X_1 \le \mu + b$, for some $b > 0$. Let $\psi(z) = (1+z)\ln(1+z) - z$, then $\forall \epsilon > 0$:*

$$\begin{aligned}
\Pr[\bar{X}_n \ge \mu + \epsilon] &\le \exp\left[\frac{-n\mathrm{Var}(X)}{b^2}\psi\left(\frac{\epsilon b}{\mathrm{Var}(X_1)}\right)\right], \\
\Pr[\bar{X}_n \ge \mu + \epsilon] &\le \exp\left[\frac{-n\epsilon^2}{2\mathrm{Var}(X_1) + 2\epsilon b/3}\right].
\end{aligned}$$

Moreover, for $t > 0$,

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

Proof Lemma 2.20 implies that

$$-I_{X_1}(\mu + \epsilon) \le \inf_{\lambda > 0}\left[-\lambda\epsilon + b^{-2}(e^{\lambda b} - \lambda b - 1)\mathrm{Var}(X_1)\right].$$

We can set the derivative of the objective function on the right-hand side with respect to λ to zero at the minimum solution, and obtain the condition for the optimal λ as follows:

$$-\epsilon + b^{-1}(e^{\lambda b} - 1)\mathrm{Var}(X_1) = 0.$$

This gives the solution $\lambda = b^{-1}\ln(1 + \epsilon b/\mathrm{Var}(X_1))$. Plugging this solution into the objective function, we obtain

$$-I_{X_1}(\mu + \epsilon) \le -\frac{\mathrm{Var}(X_1)}{b^2}\psi\left(\frac{\epsilon b}{\mathrm{Var}(X_1)}\right).$$

The first inequality of the theorem follows from an application of Theorem 2.5.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

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

$$\begin{aligned}\Lambda_{X_1}(\lambda) &\le \mu\lambda + b^{-2}\left[e^{\lambda b} - \lambda b - 1\right]\mathrm{Var}(X_1) \\ &\le \mu\lambda + \frac{\mathrm{Var}(X_1)\lambda^2}{2}\sum_{m=0}^{\infty}(\lambda b/3)^m = \mu\lambda + \frac{\mathrm{Var}(X_1)\lambda^2}{2(1-\lambda b/3)}. \end{aligned} \tag{2.13}$$

The second and the third desired bounds follow from direct applications of Theorem 2.10 with $\alpha = \mathrm{Var}(X_1)$ and $\beta = b/3$. □

Bennett's inequality can be expressed alternatively as follows. Given any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

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

If we apply this to the case that $X_i \in [0, 1]$, then using the variance estimation $\mathrm{Var}(X_1) \le \mu(1-\mu)$, and $b \le 1 - \mu$, the bound implies

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

This is slightly tighter than the corresponding multiplicative Chernoff bound in Corollary 2.18.

Compared to the tail 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 $O(1/n)$. Compared to the additive Chernoff bound, Bennett's inequality is superior when $\mathrm{Var}(X_1)$ is small.

2.7 Bernstein's Inequality

In Bernstein's inequality, we obtain results similar to Bennett's inequality, but using a moment condition (Bernstein, 1924) instead of the boundedness condition. There are several different forms of such inequalities, and we only consider one form, which relies on the following moment assumption.

Lemma 2.22 *If X satisfies the following moment condition with $b, V > 0$ for integers $m \ge 2$:*

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

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

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

Proof We have the following estimation of logarithmic moment generating function:

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

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

This implies the desired bound. □

In general, we may take $c = \mathbb{E}[X]$ and $V = \mathrm{Var}[X]$. The following bound is a direct consequence of Theorem 2.10.

Theorem 2.23 (Bernstein's Inequality) *Assume that X_1 satisfies the moment condition in Lemma 2.22. Then for all $\epsilon > 0$,*

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

and for all $t > 0$,

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

Proof We simply set $\alpha = V$ and $\beta = b/3$ in Theorem 2.10. □

Similar to Bennett's inequality, Bernstein's inequality can be alternatively expressed as follows. With probability at least $1-\delta$,

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

which implies with probability at least $1-\delta$, the following inequality holds for all $\gamma > 0$:

$$\mu < \bar{X}_n + (\gamma/b)V + \frac{b(3+2\gamma)\ln(1/\delta)}{6\gamma n}. \tag{2.14}$$

Example 2.24 If the random variable X is bounded with $|X-\mu| \le b$, then the moment condition of Lemma 2.22 holds with $c = \mu$ and $V = \mathrm{Var}(X)$.

2.8 Nonidentically Distributed Random Variables

If $X_1, \ldots, X_n$ are independent but not identically distributed random variables, then a tail inequality similar to that of Theorem 2.5 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 2.25 *We have for all $\epsilon > 0$,*

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

For sub-Gaussian random variables, we have the following bound.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Corollary 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 $\epsilon > 0$,*

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

The following inequality is a useful application of the aforementioned sub-Gaussian bound for Rademacher average. This bound, also referred to as the Chernoff bound in the literature, is essential for the symmetrization argument of Chapter 4.

Corollary 2.27 *Let $\sigma_i = \{\pm 1\}$ be independent Bernoulli random variables (each takes value ± 1 with equal probability). Let a_i be fixed numbers ($i = 1, \ldots, n$). Then for all $\epsilon > 0$,*

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

Proof Consider $X_i = \sigma_i a_i$ in Corollary 2.26. We can take $\mu = 0$ and $b_i = a_i^2$ to obtain the desired bound. □

One can also derive a Bennett-style tail probability bound.

Corollary 2.28 *If $X_i - \mathbb{E}X_i \leq b$ for all i, then for all $\epsilon > 0$,*

$$\Pr(\bar{X}_n \geq \mu + \epsilon) \leq \exp\left[-\frac{n^2\epsilon^2}{2\sum_{i=1}^n \mathrm{Var}(X_i) + 2nb\epsilon/3}\right].$$

2.9 Tail Inequality for χ^2

Let $X_i \sim N(0,1)$ be iid normal random variables ($i = 1, \ldots, n$), then the random variable

$$Z = \sum_{i=1}^n X_i^2$$

is distributed according to the chi-square distribution with n degrees of freedom, which is often denoted by χ_n^2.

This random variable plays an important role in the analysis of least squares regression. More generally, we may consider the sum of independent sub-Gaussian random variables, and obtain the following tail inequality from Theorem 2.5.

Theorem 2.29 *Let $\{X_i\}_{i=1}^n$ be independent zero-mean sub-Gaussian random variables that satisfy*

$$\ln \mathbb{E}_{X_i} \exp(\lambda X_i) \leq \frac{\lambda^2 b_i}{2},$$

then for $\lambda < 0.5b_i$, we have

$$\ln \mathbb{E}_{X_i} \exp(\lambda X_i^2) \leq -\frac{1}{2}\ln(1 - 2\lambda b_i).$$

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

Let $Z = \sum_{i=1}^n X_i^2$, *then*

$$\Pr\left[Z \geq \sum_{i=1}^n b_i + 2\sqrt{t\sum_{i=1}^n b_i^2} + 2t(\max_i b_i)\right] \leq e^{-t}$$

and

$$\Pr\left[Z \leq \sum_{i=1}^n b_i - 2\sqrt{t\sum_{i=1}^n b_i^2}\right] \leq e^{-t}.$$

Proof Let $\xi \sim N(0,1)$, which is independent of X_i. Then for all $\lambda b_i < 0.5$, we have

$$\begin{aligned}
\Lambda_{X_i^2}(\lambda) &= \ln \mathbb{E}_{X_i} \exp(\lambda X_i^2) \\
&= \ln \mathbb{E}_{X_i} \mathbb{E}_\xi \exp(\sqrt{2\lambda}\xi X_i) \\
&= \ln \mathbb{E}_\xi \mathbb{E}_{X_i} \exp(\sqrt{2\lambda}\xi X_i) \\
&\leq \ln \mathbb{E}_\xi \exp(\lambda \xi^2 b_i) \\
&= -\frac{1}{2}\ln(1-2\lambda b_i),
\end{aligned}$$

where the inequality used the sub-Gaussian assumption. The second and the last equalities can be obtained using Gaussian integration. This proves the first bound of the theorem.

For $\lambda \geq 0$, we obtain

$$\begin{aligned}
\Lambda_{X_i^2}(\lambda) &\leq -0.5\ln(1-2\lambda b_i) \\
&= 0.5\sum_{k=1}^\infty \frac{(2\lambda b_i)^k}{k} \\
&\leq \lambda b_i + (\lambda b_i)^2 \sum_{k\geq 0}(2\lambda b_i)^k \\
&= \lambda b_i + \frac{(\lambda b_i)^2}{1-2\lambda b_i}.
\end{aligned}$$

The first probability inequality of the theorem follows from Theorem 2.10 with $\mu = n^{-1}\sum_{i=1}^n b_i$, $\alpha = (2/n)\sum_{i=1}^n b_i^2$, and $\beta = 2\max_i b_i$.

If $\lambda \leq 0$, then

$$\Lambda_{X_i^2}(\lambda) \leq -0.5\ln(1-2\lambda b_i) \leq \lambda b_i + \lambda^2 b_i^2.$$

The second probability inequality of the theorem follows from the sub-Gaussian tail inequality of Theorem 2.12 with $\mu = n^{-1}\sum_{i=1}^n b_i$ and $b = (2/n)\sum_{i=1}^n b_i^2$. □

From Theorem 2.29, we can obtain the following expressions for χ_n^2 tail bound by taking $b_i = 1$. With probability at least $1-\delta$,

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

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press

and with probability at least $1-\delta$,

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

One may also obtain a tail bound estimate for χ^2_n distributions using direct integration. We leave it as an exercise.

2.10 Historical and Bibliographical Remarks

Chebyshev's inequality is named after the Russian mathematician Pafnuty Chebyshev, and was known in the nineteenth century. The investigation of exponential tail inequalities for sums of independent random variables occurred in the early twentieth century. Bernstein's inequality was one of the first such results. The large deviation principle was established by Cramér, and was later rediscovered by Chernoff (1952). In the following decade, several important inequalities were obtained, such as Hoeffding's inequality and Bennett's inequality. The tail bounds in Theorem 2.29 for χ^2 random variables was first documented in Laurent and Massart (2000), where they were used to analyze least squares regression problems with Gaussian noise. It was later extended to arbitrary quadratic forms of independent sub-Gaussian random variables by Hsu et al. (2012b).

Exercises

2.1 Assume that $X_1, X_2, \ldots, X_n$ are real-valued iid random variables with density function

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

Let $\mu = \mathbb{E}X_1$, and $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

- Estimate $\ln \mathbb{E} \exp(\lambda X_1)$
- Estimate $\Pr(\bar{X}_n \geq \mu + \epsilon)$
- Estimate $\Pr(\bar{X}_n \leq \mu - \epsilon)$

2.2 Prove Proposition 2.8.

2.3 Prove the inequality

$$\mathrm{KL}(z||\mu) \geq \frac{(z-\mu)^2}{\max(z+\mu, 2\mu)},$$

which is needed in the proof of Corollary 2.18.

2.4 Prove that the function $\phi(z) = (e^z - z - 1)/z^2$ is non-decreasing in z.

2.5 Assume that the density function of a distribution $\mathcal{D}$ on $\mathbb{R}$ is $(1-p)U(-1,1)+pU(-1/p,1/p)$ for $p \in (0, 0.5)$, where $U(\cdot)$ denotes the density of the uniform distribution. Let $X_1, \ldots, X_n$ be iid samples from $\mathcal{D}$. For $\epsilon > 0$, estimate the probability

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right)$$

using Bernstein's inequality.

2.6 Write down the density of χ^2 distribution, and use integration to estimate the tail inequalities. Compare the results to those of Theorem 2.29.

2.7 Prove Corollary 2.26 and Corollary 2.28.

https://doi.org/10.1017/9781009093057.003 Published online by Cambridge University Press