Tail and Concentration Inequalities

From here on, we use $\mathbf{1}_A$ to denote the indicator variable for event A, i.e. $\mathbf{1}_A = 1$ if A holds and $\mathbf{1}_A = 0$ otherwise. Our presentation follows closely the first chapter of [**?**].

1 Markov Inequality

Theorem 1.1. *If X is a r.v. taking only non-negative values, $\mu = \mathrm{E}[X]$, then $\forall a > 0$*

$$\text{Prob}[X \geq a] \leq \frac{\mu}{a}. \tag{1}$$

Proof. From the simple fact that $a\mathbf{1}_{\{X\geq a\}} \leq X$, taking expectation on both sides we get $a\mathrm{E}\left[\mathbf{1}_{\{X\geq a\}}\right] \leq \mu$, which implies (1). □

Problem 1. Use Markov inequality to prove the following. Let $c \geq 1$ be an arbitrary constant. If n people have a total of d dollars, then there are at least $(1 - 1/c)n$ of them each of whom has less than cd/n dollars.

(You can easily prove the above statement from first principle. However, please set up a probability space, a random variable, and use Markov inequality to prove it. It is instructive!)

2 Chebyshev Inequality

Theorem 2.1 (Two-sided Chebyshev's Inequality)**.** *If X is a r.v. with mean μ and variance σ^2, then $\forall a > 0$,*

$$\text{Prob}\big[|X-\mu| \geq a\big] \leq \frac{\sigma^2}{a^2}$$

Proof. Let $Y = (X-\mu)^2$, then $\mathrm{E}[Y] = \sigma^2$ and Y is a non-negative r.v.. From Markov inequality (1) we have

$$\text{Prob}\big[|X-\mu| \geq a\big] = \text{Prob}\big[Y \geq a^2\big] \leq \frac{\sigma^2}{a^2}.$$

□

The one-sided versions of Chebyshev inequality are sometimes called Cantelli inequality.

Theorem 2.2 (One-sided Chebyshev's Inequality)**.** *Let X be a r.v. with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$, then for all $a > 0$,*

$$\begin{aligned} \text{Prob}[X \geq \mu + a] &\leq \frac{\sigma^2}{\sigma^2 + a^2} && (2) \\ \text{Prob}[X \leq \mu - a] &\leq \frac{\sigma^2}{\sigma^2 + a^2}. && (3) \end{aligned}$$

Proof. Let $Y = X - \mu$, then $\mathrm{E}[Y] = 0$ and $\mathrm{Var}[Y] = \mathrm{Var}[X] = \sigma^2$. (Why?) Thus, for any t such that $t + a > 0$ we have

$$\begin{aligned}
\text{Prob}[Y \geq a] &= \text{Prob}[Y + t \geq a + t] \\
&= \text{Prob}\left[\frac{Y+t}{a+t} \geq 1\right] \\
&\leq \text{Prob}\left[\left(\frac{Y+t}{a+t}\right)^2 \geq 1\right] \\
&\leq \mathrm{E}\left[\left(\frac{Y+t}{a+t}\right)^2\right] \\
&= \frac{\sigma^2 + t^2}{(a+t)^2}
\end{aligned}$$

The second inequality follows from Markov inequality. The above analysis holds for any t such that $t + a > 0$. We pick t to minimize the right hand side, which is $t = \sigma^2/a > 0$. That proves (2). □

Problem 2. Prove (3).

3 Bernstein, Chernoff, Hoeffding

3.1 The basic bound using Bernstein's trick

Let us consider the simplest case, and then relax assumptions one by one. For $i \in [n]$, let X_i be i.i.d. random variables which are all Bernoulli with parameter p. Let $X = \sum_{i=1}^n X_i$. Then, $\mathrm{E}[X] = np$. We will prove that, as n gets large X is "far" from $\mathrm{E}[X]$ with exponentially low probability.

Let m be such that $np < m < n$, we want to bound $\text{Prob}[X \geq m]$. For notational convenience, let $q = 1 - p$. Bernstein taught us the following trick. For any $t > 0$ the following holds.

$$\begin{aligned}
\text{Prob}[X \geq m] &= \text{Prob}\left[tX \geq tm\right] \\
&= \text{Prob}\left[e^{tX} \geq e^{tm}\right] \\
&\leq \frac{\mathrm{E}\left[e^{tX}\right]}{e^{tm}} \\
&= \frac{\mathrm{E}\left[\prod_{i=1}^n e^{tX_i}\right]}{e^{tm}} \\
&= \frac{\prod_{i=1}^n \mathrm{E}\left[e^{tX_i}\right]}{e^{tm}} \quad \text{(because the } X_i \text{ are independent)} \\
&= \frac{\prod_{i=1}^n (pe^t + q)}{e^{tm}} \\
&= \frac{(pe^t + q)^n}{e^{tm}}.
\end{aligned}$$

The inequality on the third line follows from Markov inequality (1). Naturally, we set t to minimize the right hand side, which is

$$t_0 = \ln \frac{mq}{(n-m)p} > 0.$$

Plugging t_0 in, we obtain the following after simple algebraic manipulations:

$$\text{Prob}[X \geq m] \leq \left(\frac{pn}{m}\right)^m \left(\frac{qn}{n-m}\right)^{n-m}. \tag{4}$$

This is still quite a mess. But there's a way to make it easier to remember. The *relative entropy* (or Kullberg-Leibler distance) between two Bernoulli distributions with parameters p and p' is defined to be

$$\text{RE}(p\|p') := p \ln \frac{p}{p'} + (1-p) \ln \frac{1-p}{1-p'}.$$

There are several different interpretations of the relative entropy function. You can find them from the Wikipedia entry on relative entropy. It can be shown that $\text{RE}(p\|p') \geq 0$ for all $p, p' \in (0,1)$. Anyhow, we can rewrite (4) simply as

$$\text{Prob}[X \geq m] \leq e^{-n \cdot \text{RE}(m/n \| p)}. \tag{5}$$

Next, suppose the X_i are still Bernoulli variables but with different parameters p_i. Let $q_i = 1 - p_i$, $p = (\sum_i p_i)/n$ and $q = 1 - p$. Note that $\text{E}[X] = np$ as before. A similar analysis leads to

$$\text{Prob}[X \geq m] \leq \frac{\prod_{i=1}^n (p_i e^t + q_i)}{e^{tm}} \leq \frac{(pe^t + q)^n}{e^{tm}}.$$

The second inequality is due to the *geometric-arithmetic means* inequality, which states that, for any non-negative real numbers $a_1, \cdots, a_n$ we have

$$a_1 \cdots a_n \leq \left(\frac{a_1 + \cdots + a_n}{n}\right)^n.$$

Thus, (5) holds when the X_i are Bernoulli and they don't have to be identically distributed.

Finally, consider a fairly general case when the X_i do not even have to be discrete variables. Suppose the X_i are independent random variables where $\text{E}[X_i] = p_i$ and $X_i \in [0,1]$ for all i. Again, let $p = \sum_i p_i/n$ and $q = 1 - p$. Bernstein's trick leads us to

$$\text{Prob}[X \geq m] \leq \frac{\prod_{i=1}^n \text{E}\left[e^{tX_i}\right]}{e^{tm}}.$$

The problem is, we no longer can compute $\text{E}\left[e^{tX_i}\right]$ because we don't know the X_i's distributions. Hoeffding taught us another trick. For $t > 0$, the function $f(x) = e^{tx}$ is convex. Hence, the curve of $f(x)$ inside $[0,1]$ is below the linear segment connecting the points $(0, f(0))$ and $(1, f(1))$. The segment's equation is

$$y = (f(1) - f(0))x + f(0) = (e^t - 1)x + 1 = e^t x + (1 - x).$$

Hence,

$$\text{E}\left[e^{tX_i}\right] \leq \text{E}\left[e^t X_i + (1 - X_i)\right] = p_i e^t + q_i.$$

We thus obtain (4) as before. Overall, we just proved the following theorem.

Theorem 3.1 (Bernstein-Chernoff-Hoeffding)**.** *Let $X_i \in [0,1]$ be independent random variables where $\text{E}[X_i] = p_i, i \in [n]$. Let $X = \sum_{i=1}^n X_i$, $p = \sum_{i=1}^n p_i/n$ and $q = 1 - p$. Then, for any m such that $np < m < n$ we have*

$$\text{Prob}[X \geq m] \leq e^{-n\text{RE}(m/n \| p)}. \tag{6}$$

Problem 3. Let $X_i \in [0,1]$ be independent random variables where $\text{E}[X_i] = p_i, i \in [n]$. Let $X = \sum_{i=1}^n X_i$, $p = \sum_{i=1}^n p_i/n$ and $q = 1 - p$. Prove that, for any m such that $0 < m < np$ we have

$$\text{Prob}[X \leq m] \leq e^{-n\text{RE}(m/n \| p)}. \tag{7}$$

3.2 Instantiations

There are a variety of different bounds we can get out of (6) and (7).

Theorem 3.2 (Hoeffding Bounds)**.** *Let $X_i \in [0,1]$ be independent random variables where* $\mathrm{E}[X_i] = p_i, i \in [n]$*. Let $X = \sum_{i=1}^n X_i$. Then, for any $t > 0$ we have*

$$\mathrm{Prob}[X \geq \mathrm{E}[X] + t] \leq e^{-2t^2/n}. \tag{8}$$

and

$$\mathrm{Prob}[X \leq \mathrm{E}[X] - t] \leq e^{-2t^2/n}. \tag{9}$$

Proof. We prove (8), leaving (9) as an exercise. Let $p = \sum_{i=1}^n p_i/n$ and $q = 1-p$. WLOG, we assume $0 < p < 1$. Define $m = (p+x)n$, where $0 < x < q = 1-p$, so that $np < m < n$. Also, define

$$f(x) = \mathrm{RE}\left(\frac{m}{n}\|p\right) = \mathrm{RE}\left(p+x\|p\right) = (p+x)\ln\frac{p+x}{p} + (q-x)\ln\frac{q-x}{q}. \tag{10}$$

Routine manipulations give

$$\begin{aligned} f'(x) &= \ln\frac{p+x}{p} - \ln\frac{q-x}{q} \\ f''(x) &= \frac{1}{(p+x)(q-x)} \end{aligned}$$

By Taylor's expansion, for any $x \in [0,1]$ there is some $\xi \in [0,x]$ such that

$$f(x) = f(0) + xf'(0) + \frac{1}{2}x^2 f''(\xi) = \frac{1}{2}x^2\frac{1}{(p+\xi)(q-\xi)} \geq 2x^2.$$

The last inequality follows from the fact that $(p+\xi)(q-\xi) \leq ((p+q)/2)^2 = 1/4$. Finally, set $x = t/n$. Then, $m = np + t = \mathrm{E}[X] + t$. From (6) we get

$$\mathrm{Prob}[X \geq \mathrm{E}[X] + t] \leq e^{-nf(x)} \leq e^{-2x^2n} = e^{-2t^2/n}.$$

□

Problem 4. Prove (9).

Theorem 3.3 (Chernoff Bounds)**.** *Let $X_i \in [0,1]$ be independent random variables where* $\mathrm{E}[X_i] = p_i, i \in [n]$*. Let $X = \sum_{i=1}^n X_i$. Then,*

(i) For any $0 < \delta \leq 1$,

$$\mathrm{Prob}[X \geq (1+\delta)\mathrm{E}[X]] \leq e^{-\mathrm{E}[X]\delta^2/3}. \tag{11}$$

(ii) For any $0 < \delta < 1$,

$$\mathrm{Prob}[X \leq (1-\delta)\mathrm{E}[X]] \leq e^{-\mathrm{E}[X]\delta^2/2}. \tag{12}$$

(iii) If $t > 2e\mathrm{E}[X]$, then

$$\mathrm{Prob}[X \geq t] \leq 2^{-t}. \tag{13}$$

Proof. To bound the upper tail, we apply (6) with $m = (p+\delta p)n$. Without loss of generality, we can assume $m < n$, or equivalently $\delta < q/p$. In particular, we will analyze the function

$$g(x) = \text{RE}(p + xp\|p) = (1+x)p\ln(1+x) + (q - px)\ln\frac{q-px}{q},$$

for $0 < x \le \min\{q/p, 1\}$. First, observe that

$$\ln\frac{q}{q-px} = \ln\left(1 + \frac{px}{q-px}\right) \le \frac{px}{q-px}.$$

Hence, $(q-px)\ln\frac{q-px}{q} \ge -px$, from which we can infer that

$$g(x) \ge (1+x)p\ln(1+x) - px = p\left[(1+x)\ln(1+x) - x\right].$$

Now, define

$$h(x) = (1+x)\ln(1+x) - x - x^2/3.$$

Then,

$$\begin{aligned} h'(x) &= \ln(1+x) - 2x/3 \\ h''(x) &= \frac{1}{1+x} - 2/3. \end{aligned}$$

Thus, $1/2$ is a local extremum of $h'(x)$. Note that $h'(0) = 0$, $h'(1/2) \approx 0.07 > 0$, and $h'(1) \approx 0.026 > 0$. Hence, $h'(x) \ge 0$ for all $x \in (0,1]$. The function $h(x)$ is thus non-decreasing. Hence, $h(x) \ge h(0) = 0$ for all $x \in [0,1]$. Consequently,

$$g(x) \ge p\left[(1+x)\ln(1+x) - x\right] \ge px^2/3$$

for all $x \in [0,1]$. Thus, from (6) we have

$$\text{Prob}[X \ge (1+\delta)\text{E}[X]] = \text{Prob}[X \ge (1+\delta)pn] \le e^{-n\cdot g(\delta)} \le e^{-\delta^2\text{E}[X]/3}.$$

$\square$

Problem 5. Prove (12).

Problem 6. Let $X_i \in [0,1]$ be independent random variables where $\text{E}[X_i] = p_i, i \in [n]$. Let $X = \sum_{i=1}^n X_i$, and $\mu = \text{E}[X]$. Prove the following

(i) For any $\delta, t > 0$ we have

$$\text{Prob}[X \ge (1+\delta)\text{E}[X]] \le \left(\frac{e^{e^t-1}}{e^{t(1+\delta)}}\right)^\mu$$

(**Hint**: repeat the basic structure of the proof using Bernstein's trick. Then, because $1 + x \le e^x$ we can apply $1 + p_ie^t - p_i \le e^{p_ie^t - p_i}$.)

(ii) Show that, for any $\delta > 0$ we have

$$\text{Prob}[X \ge (1+\delta)\mu] \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

(iii) Prove that, for any $t > 2e\mathrm{E}[X]$,
$$\mathrm{Prob}[X \geq t] \leq 2^{-t}.$$

Problem 7. Let $X_i \in [a_i, b_i]$ be independent random variables where a_i, b_i are real numbers. Let $X = \sum_{i=1}^{n} X_i$. Repeat the basic proof structure to show a slightly more general Hoeffding bounds:

$$\mathrm{Prob}[X - \mathrm{E}[X] \geq t] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

$$\mathrm{Prob}[X - \mathrm{E}[X] \leq -t] \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

Problem 8. Prove that, for any $0 \leq \alpha \leq n$,

$$\sum_{0 \leq k \leq \alpha n} \binom{n}{k} \leq 2^{H(\alpha)n},$$

where $H(\alpha) = -\alpha \log_2 \alpha - (1 - \alpha) \log_2(1 - \alpha)$ is the binary entropy function.

References