TA Notes

Time: 2019.7.1 Zhide Wei

1 Basic Probability Theory

1.1 Markov Inequality

Theorem 1. *Random variable (r.v.) $X \geq 0$, EX finite, then* $\forall k > 0, P(X \geq k) \leq \frac{EX}{k}$

Proof $EX \geq P(X \geq k) * k + P(0 \leq X < k) * 0$. □

In class, the following example is proved with some mistakes, it has now been corrected.

Enample 1 r.v. $X \in [0,1]$, $EX = 1 - \varepsilon$, prove $P(X \leq 1 - \sqrt{\varepsilon}) \leq \sqrt{\varepsilon}$

Proof $EX \leq P(X \leq 1 - \sqrt{\varepsilon}) * (1 - \sqrt{\varepsilon}) + P(1 - \sqrt{\varepsilon} \leq X < 1) * 1$.

So $P(X \leq 1 - \sqrt{\varepsilon}) \leq \sqrt{\varepsilon}$. □

Corollary 2 (Chebyshev Inequality)**.** *For every constant $a > 0$,*

$$P(|X - EX| \geq a) \leq \frac{varX}{a^2}.$$

Definition 3. *Denote $[r]$ to be $\{1, 2, ..., r\}$.*

Corollary 4. *r.v. $X \geq 0$, $EX, EX^2, ..., EX^r$ finite, then* $\forall k > 0, j \in [r], P(X \geq k) = P(X^j \geq k^j) \leq \frac{EX^j}{k^j}$

So we have

$$P(X \geq k) \leq \min_{j \in [r]} \frac{EX^j}{k^j}$$

1.2 Chernorff Bound

Considering the following question: i.i.d r.v. $X_1, X_2, ..., X_n \sim \mathrm{B}(1, p)$,

it's obvious that $\lim_{n \to \infty} \frac{\sum X_i}{n} = p$, but how much the average value can deviate from p?

We can give a loose bound using Chebyshev Inequality:

$$P\left(\left|\frac{1}{n}\sum_i X_i - p\right| \geq \varepsilon\right) \leq \frac{\mathrm{var}(\frac{\sum_i X_i}{n})}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}.$$

However, we know that from Probability Theory, the sum of i.i.d. random variable tend to follow the Gaussian distribution, so we hope the convergence speed to be exponential in n .

since we know the distribution of X_i, we can give a more tight bound using the following Chernoff Inequality:

Corollary 5 (Chernoff Inequality)**.** *r.v. X, $\forall t > 0, Ee^{tx}$ finite, then* $P(X \geq k) \leq \inf_{t>0} e^{-tk} Ee^{tX}$

Theorem 6 (Chernorff Bound)**.**

$$\begin{aligned} & P\left(\frac{1}{n}\sum_i X_i - p \geq \varepsilon\right) = P(\sum_i X_i \geq n(p+\varepsilon)) \\ & \leq \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)}\,\mathrm{E}(\mathrm{e}^{t\sum_i X_i}) \\ & = \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)}(\mathrm{E}(\mathrm{e}^{tX_1}))^n \\ & = \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)}(pe^t + 1 - p)^n \\ & = e^{-nD_e^{(B)}(p+\varepsilon||p)} \leq e^{-2n\varepsilon^2} \end{aligned} \tag{1}$$

here, $D_e^{(B)}(p||q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$

2 Useful Inequalities

Problem 1. $X, X_1, X_2, ..., X_n$ i.i.d. Gaussian r.v., mean 0, variance σ^2, Prove that

$$P(\frac{1}{n}\sum_{i=1}^{n} \geq \varepsilon) \leq e^{-\frac{n\varepsilon^2}{2\sigma^2}}$$

Proof L.H.S. $\leq \inf_{t>0} e^{-tn\varepsilon}(Ee^{tX})^n = e^{-\frac{n\varepsilon^2}{2\sigma^2}}$
(It's trivial that $Ee^{tX} = e^{\frac{\sigma^2 t^2}{2}}$) □

Remark 1. *Thanks for a student for providing the following solution to the problem.*

Proof By changing variable, it suffices to prove that $\forall a > 0, P(X \geq a) \leq e^{-\frac{a^2}{2\sigma^2}}$.
It's equivalent to prove that $\int_a^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{a^2-x^2}{2\sigma^2}} dx \leq 1$
Let $y = x - a$, it's equivalent to $\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-2ay-y^2}{2\sigma^2}} dx \leq 1$
Note that $ay \geq 0$, so the left hand side $\leq \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-y^2}{2\sigma^2}} dx = 1$
□

Problem 2. $X, X_1, X_2, ..., X_n, n \geq 2$ i.i.d. r.v., $\forall t > 0, E(e^{tx}) \leq e^{\frac{t^2 r^2}{2}}$ for some fixed $r > 0$, Prove that

$$E(\max_{j\in[n]} X_j) \leq r\sqrt{2\ln n}$$

Proof $e^{tE(\max_{j\in[n]} X_j)} \leq E(e^{t\max_{j\in[n]} X_j}) = E(\max_{j\in[n]} e^{tX_j}) \leq E(\sum_{j\in[n]} e^{tX_j}) \leq ne^{\frac{t^2 r^2}{2}}$
(The first inequality is due to the Jensen's inequality: for f convex, X r.v., $f(EX) \leq Ef(X)$)
take ln on both side, then $E(\max_{j\in[n]} X_j) \leq \frac{\ln n}{t} + \frac{tr^2}{2}$, take $t = \frac{\sqrt{2\ln n}}{r}$, then we get the desired inequality. □

3 Introduction to the Original Turing Machine

See WeChat group.

4 Some Application of probability in Shannon's Information Theory

We'll briefly introduce Shannon's typical sequence and typical set in this section, for more details, we'll introduce Shannon's original paper "A mathematical Theory of Communication" two weeks later.

4.1 Asymptotic Equipartition Property

According to weak law of large number, for i.i.d. r.v. $X, X_1, X_2, ...$, with finite expectation and variance, $\forall\epsilon > 0, P(|\frac{1}{n}\sum_{i=1}^{n} X_i - EX| \geq \epsilon) \xrightarrow{n\to\infty} 0$

Similarly, for any function g, we regard $g(X)$ as a r.v., suppose $Eg(X)$ and $var(g(X))$ finite, then $\forall\epsilon > 0, P(|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - Eg(X)| \geq \epsilon) \xrightarrow{n\to\infty} 0$

Now, let g be the negative logarithm of X's probability density function, namely, $g(X) = -logP(X)$.(We always assume log to be log_2)

Then

$$\begin{aligned} &P(|\frac{1}{n}\sum_{i=1}^{n} -logP(X_i) - H(X)| \geq \epsilon) \xrightarrow{n\to\infty} 0 \\ &P(2^{-n(H(X)+\epsilon)} \leq P(X_1, X_2, ..., X_n) \leq 2^{-n(H(X)-\epsilon)}) \xrightarrow{n\to\infty} 1 \end{aligned} \tag{2}$$

So roughly speaking, with high probability (w.h.p.) $P(X_1, X_2, ..., X_n) \approx 2^{-nH(X)}$

We call such sequence of $X_1, X_2, ..., X_n$ typical sequence, and the set of all typical sequence is called typical set.

So the number of elements in typical set is roughly $2^{nH(X)}$

Enample 2

i.i.d. Bernoulli r.v. $X, X_1, X_2, ..., EX = 0.3$, typical sequence consists roughly 30% 1 and 70 % 0.