Discrete Mathematics and Its Applications 2 (CS147)

Lecture 13: Markov inequality

Fanghui Liu

Department of Computer Science, University of Warwick, UK

Logistics: Related notes and past exam

Department of Computer Science

Study with us | Teaching | Research | People | News | Events | Outreach | Welfare | Vacancies | Intranet

Online Material ▸ CS147 ▸ Lectures and Seminars

Relevant Textbooks and Lecture Notes for week 1 to 5:

- [DPV] Algorithms.
- [Aspnes] Notes on Discrete Mathematics.
- [Jeff] Algorithms.

Lectures

Week 1

Lecture 1: Introduction to the module [Lec1.pdf]

Lecture 2: Big-O notation [Lec2-updated] [Aspnes, Chapter 7]

Lecture 3: Worst-case asymptotic running time [Lec3.pdf] [DPV, Chapter 0]

Related notes: [Jeff, Chapter 0]

Week 2

Lecture 4: Bubble-sort [Lec4.pdf]

Lecture 5: Merge-sort [Lec5.pdf] [DPV, Chapter 2.1-2.3]

Lecture 6: Master theorem [Lec6.pdf] [DPV, Chapter 2.1-2.3]

https://warwick.ac.uk/fac/sci/dcs/teaching/modules/cs147/#assessment
https://warwick.ac.uk/services/exampapers/cs/2023/cs1470_a_exam_paper.pdf

Concentration inequalities

"Concentration": quantify how a random variable X deviates around its expectation μ

$$\Pr\left(\underbrace{|X-\mu| \geq t}_{\text{tail}}\right) \leq \text{small}$$

Tail probability: We wish to create an upper bound such that $|X-\mu|$ exceed t with a low probability

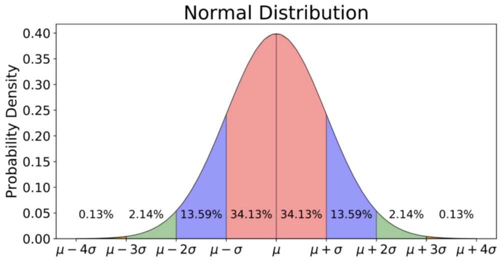
Concentration inequalities

"Concentration": quantify how a random variable X deviates around its expectation μ

$$\Pr\left(\underbrace{|X-\mu| \geq t}_{\text{tail}}\right) \leq \phi(t)$$

Intuitively, $\phi(t)$ is a decreasing function of t

*Gaussian tail[1]



$$\Pr[X - \mu \geq t] \leq \exp\left(-\frac{t^2}{\sigma^2}\right) .$$

Using moment generating function...

[1] https://vatsalsharan.github.io/lecture_notes/lec4_final.pdf if you're interested in.

Markov inequality

Theorem (Markov inequality)

For a non-negative (discrete) random variable $X \geq 0$ *and a constant* $t > 0$, *if* $\mathbb{E}(X)$ *exists, we have* $\Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}$.

Markov inequality

Theorem (Markov inequality)

For a non-negative (discrete) random variable $X \geq 0$ *and a constant* $t > 0$*, if* $\mathbb{E}(X)$ *exists, we have* $\Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}$.

Proof 1.

$$
\begin{aligned}
\mathbb{E}(X) = \sum_{x} x\Pr(X = x) &= \sum_{x:0 \leq x < t} x\Pr(X = x) + \sum_{x:x \geq t} x\Pr(X = x) \\
&\geq \sum_{x:x \geq t} x\Pr(X = x) \\
&\geq \sum_{x:x \geq t} t\Pr(X = x) \\
&= t\Pr(X \geq t)\,.
\end{aligned}
$$

□

Second way for proof

Proof 2.

Consider the following indicator random variable for any $\omega \in \Omega$

$$Y(\omega) = \begin{cases} 1 & \text{if } X(\omega) \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

$\Rightarrow Y(\omega) \leq X(\omega)/t$ for any $\omega \in \Omega$.
$\Rightarrow \mathbb{E}(Y) \leq \frac{\mathbb{E}X}{t}$

$$\Pr(X \geq t) = \Pr(Y = 1) = \mathbb{E}(Y) \leq \frac{\mathbb{E}X}{t}\,.$$

$\square$

Example (I)

Problem

Use Markov's inequality to bound the probability of obtaining at least $3n/4$ heads in a sequence of n fair coin flips.

Example (I)

Problem

Use Markov's inequality to bound the probability of obtaining at least $3n/4$ heads in a sequence of n fair coin flips.

Solution

Let X be the number of heads in a sequence of n fair coin flips. We know that $\mathbb{E}(X) = \frac{n}{2}$. Then by Markov's inequality, we have

$$\Pr[X \geq \frac{3n}{4}] \leq \frac{\mathbb{E}(X)}{\frac{3n}{4}} = \frac{2}{3}\,.$$

Application of Markov's inequality to coupon collector's problem

Problem (Recall Coupon collector's problem)

We randomly and uniformly sample one object from $\{1, 2, \cdots, n\}$, T *is the number of draws before the every* $\{1, 2, \cdots, n\}$ *is seen, we have* $\mathbb{E}(T) = nH_n$.

Application of Markov's inequality to coupon collector's problem

Problem (Recall Coupon collector's problem)

We randomly and uniformly sample one object from $\{1, 2, \cdots, n\}$, T *is the number of draws before the every* $\{1, 2, \cdots, n\}$ *is seen, we have* $\mathbb{E}(T) = nH_n$.

Now we plan to estimate the tail by Chebyshev's inequality.

Problem (Tail probability)

For coupon collector's problem, what is the probability of the event that the numbers we draw is larger than $\beta n \log n$ *with* $\beta \geq 1$*?*

Application of Markov's inequality to coupon collector's problem

Problem (Recall Coupon collector's problem)

We randomly and uniformly sample one object from $\{1, 2, \cdots, n\}$, T *is the number of draws before the every* $\{1, 2, \cdots, n\}$ *is seen, we have* $\mathbb{E}(T) = nH_n$.

Now we plan to estimate the tail by Chebyshev's inequality.

Problem (Tail probability)

For coupon collector's problem, what is the probability of the event that the numbers we draw is larger than $\beta n \log n$ *with* $\beta \geq 1$?

$$\Pr(T \geq \beta n \log n) \leq \frac{nH_n}{\beta n \log n} \leq \frac{1}{\beta} + \frac{\gamma}{\beta \log n} + \mathcal{O}(\frac{1}{n \log n})$$

Bounded random variable

Problem

Consider a random variable X such that for every $\omega \in \Omega$, $X(\omega) \geq -80$ and $\mathbb{E}(X) = -40$. Give an upper bound on $\Pr(X \geq 10)$.

Solution

Define a random variable $Y = X + 80$...

Remark: It is possible to use Markov's inequality as long as there is a lower bound on the range of the random variable under consideration.

Handle bounded random variable

Theorem

Consider a random variable X on $(\Omega, \mathcal{F}, \Pr)$, it can take values $X \geq b$ with $b < 0$. Then taking a constant $a > b$, we have

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X) - b}{a - b}.$$

Theorem

Consider a random variable X on $(\Omega, \mathcal{F}, \Pr)$, it can take values $X \leq b$. Then taking a constant $a < b$, we have

$$\Pr(X \leq a) \leq \frac{b - \mathbb{E}(X)}{b - a}.$$

Proof.

Taking $Y := b - X$... □

Summary

Statement

We can invoke Markov's inequality to bound:

- *The probability of a random variable X taking value at least c if there is a lower bound on the range of X*
- *The probability of a random variable X taking value at most c if there is a upper bound on the range of X*

Relationship between expectation and tail

Theorem

Let X be a non-negative (discrete) random variable taking values in $\{0, 1, 2, \cdots\}$, if its expectation exists, then

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} \Pr(X > i) \, .$$

Relationship between expectation and tail

Theorem

Let X be a non-negative (discrete) random variable taking values in $\{0, 1, 2, \cdots\}$, if its expectation exists, then

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} \Pr(X > i)\,.$$

Proof.

$$\sum_{i\geq 0} \Pr(X > i) = \sum_{i\geq 0} \sum_{j\geq i+1} \Pr(X = j) = \sum_{i\geq 1} i \cdot \Pr(X = i) = \sum_{i\geq 0} i \cdot \Pr(X = i) = \mathbb{E}(X).$$

□

*Continuous version

Theorem (Integral identity)

Let X be a non-negative continuous random variable, then we have

$$\mathbb{E}(X) = \int_0^{\infty} \Pr(X > t)\mathrm{d}t\,.$$

Remark: 1) The two sides of this identity are either finite or infinite simultaneously.

*Proof

Proof.

We represent any non-negative real number x via the identity

$$x = \int_0^x 1 \mathrm{d}t = \int_0^\infty 1_{\{t<x\}} \mathrm{d}t \, .$$

Substitute the random variable X for x and take expectation of both sides. This gives

$$\mathbb{E}(X) = \mathbb{E} \int_0^\infty \mathbf{1}_{\{t<X\}} \mathrm{d}t = \int_0^\infty \mathbb{E}\mathbf{1}_{\{t<X\}} \mathrm{d}t = \int_0^\infty \Pr(X > t) \mathrm{d}t \, .$$

□

Remark: To change the order of expectation and integration in the second equality, we used Fubini-Tonelli's theorem (beyond the scope of this course).

Example: expectation of Geometric distribution (proof by tail)

$X \sim \mathrm{Geo}(p)$ with the PMF

$$\Pr(X = k) = (1-p)^{k-1} p \quad \forall k \geq 1\,.$$

Statement

The expected value of a Geometric random variable is $\mathbb{E}(X) = 1/p$.

Proof.

Using the integral identity and $q := 1 - p$, we have

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=0}^{\infty} \Pr(X > i) = \sum_{i=1}^{\infty} \Pr(X \geq i) = \sum_{i=1}^{\infty}\sum_{k=i}^{\infty} (1-p)^{k-1} p := p \sum_{i=1}^{\infty}\sum_{k=i}^{\infty} q^{k-1} \\ &= p \sum_{i=1}^{\infty} \frac{q^{i-1}}{1-q} = \sum_{i=1}^{\infty} q^{i-1} = \frac{1}{1-q} = \frac{1}{p}\,.\end{aligned}$$

□

*Application in ML theory: tail and union bound [LXMZ21]

Lemma 9 (bounds of initial parameters). *Given $\delta \in (0,1)$, we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$*

$$\max_{k\in[m]} \left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\} \le \sqrt{2\log\frac{2m(d+1)}{\delta}}, \tag{45}$$

Proof If $\mathrm{X} \sim N(0,1)$, then $\mathbb{P}(|\mathrm{X}| > \varepsilon) \le 2\mathrm{e}^{-\frac{1}{2}\varepsilon^2}$ for all $\varepsilon > 0$. Since $a_k^0 \sim N(0,1)$, $(w_k^0)_\alpha \sim N(0,1)$ for $k = 1,2,\ldots,m$, $\alpha = 1,\ldots,d$ and they are all independent, by setting

$$\varepsilon = \sqrt{2\log\frac{2m(d+1)}{\delta}},$$

one can obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{k\in[m]}\left\{|a_k^0|,\ \|\boldsymbol{w}_k^0\|_\infty\right\} > \varepsilon\right) &= \mathbb{P}\left(\max_{k\in[m],\alpha\in[d]}\left\{|a_k^0|,\ |(w_k^0)_\alpha|\right\} > \varepsilon\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{m}\left(|a_k^0| > \varepsilon\right) \bigcup \left(\bigcup_{\alpha=1}^{d}\left(|(w_k^0)_\alpha| > \varepsilon\right)\right)\right) \\
&\le \sum_{k=1}^{m}\mathbb{P}\left(|a_k^0| > \varepsilon\right) + \sum_{k=1}^{m}\sum_{\alpha=1}^{d}\mathbb{P}\left(|(w_k^0)_\alpha| > \varepsilon\right) \\
&\le 2m\mathrm{e}^{-\frac{1}{2}\varepsilon^2} + 2md\mathrm{e}^{-\frac{1}{2}\varepsilon^2} \\
&= 2m(d+1)\mathrm{e}^{-\frac{1}{2}\varepsilon^2} \\
&= \delta.
\end{aligned}$$

*Application in ML theory: Markov inequality [LXMZ21]

Then

$$
\begin{aligned}
&\mathbb{E}\sum_{i,j=1}^{n}\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\quad\le \sum_{i,j=1}^{n}\frac{\kappa^2\kappa' d}{m}\sum_{k=1}^{m}\left(4\max\left\{\frac{1}{\kappa'^2},1\right\}\xi^2\mathbb{E}|D_{k,i,j}| + 6\max\left\{\frac{1}{\kappa'^2},1\right\}\xi^2 p\right) \\
&\quad\le \sum_{i,j=1}^{n}\frac{\kappa^2\kappa' d}{m}\sum_{k=1}^{m}\left(4\max\left\{\frac{1}{\kappa'^2},1\right\}\xi^2 8d\max\{\kappa',1\}\xi p + 6\max\left\{\frac{1}{\kappa'^2},1\right\}\xi^2 p\right) \\
&\quad\le \kappa^2\kappa' d n^2\left(32d\xi\max\{\kappa',\frac{1}{\kappa'^2}\} + 6\max\left\{\frac{1}{\kappa'^2},1\right\}\right)\xi^2 p \\
&\quad\le 40\kappa^2 d^2 n^2\left(2\log\frac{8m(d+1)}{\delta}\right)^{3/2}\max\{\kappa'^2,\frac{1}{\kappa'}\}p.
\end{aligned}
$$

By Markov's inequality, with probability at least $1-\delta/2$ over the choice of $\boldsymbol{\theta}^0$, we have

$$
\begin{aligned}
&\|G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\|_{\mathrm{F}} \\
&\le \sum_{i,j=1}^{n}\left|G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(t)) - G_{ij}^{[\boldsymbol{w}]}(\boldsymbol{\theta}(0))\right| \\
&\le \max\left\{\kappa'^2,\frac{1}{\kappa'}\right\}\frac{40\kappa^2 d^2 n^2\left(2\log\frac{8m(d+1)}{\delta}\right)^{3/2} p}{\delta/2}
\end{aligned}
$$

References I

[0] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang, *Phase diagram for two-layer relu neural networks at infinite-width limit*, Journal of Machine Learning Research **22** (2021), no. 71, 1–47.
(Cited on pages 22 and 23.)