Discrete Mathematics and Its Applications 2 (CS147)

Lecture 14: Chebyshev's inequality and application

Fanghui Liu

Department of Computer Science, University of Warwick, UK

Recall Markov inequality...

Statement

Given a non-negative random variable X, if its expectation exists, then

$$\Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}.$$

Recall Markov inequality...

Statement

Given a non-negative random variable X, if its expectation exists, then

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

Theorem (Relationship between expectation and tail)

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

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

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

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} \Pr(X > i) = \sum_{i=1}^{\infty} \Pr(X \geq i) = \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} (1-p)^{k-1} p := p \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} q^{k-1}$$

$$= p \sum_{i=1}^{\infty} \frac{q^{i-1}}{1-q} = \sum_{i=1}^{\infty} q^{i-1} = \frac{1}{1-q} = \frac{1}{p}\,.$$

□

Recall Variance...

Definition

The variance of a random variable X is defined as

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - [\mathbb{E}X]^2\,.$$

Recall Variance...

Definition

The variance of a random variable X is defined as

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - [\mathbb{E}X]^2 .$$

Property

- $\mathbb{V}(aX) = a^2\mathbb{V}(X)$ *for a constant* a.
- *If* X, Y *are independent, we have* $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y)$.

Example: Variance of Geometric distribution

$X \sim \mathrm{Geo}(p)$ with the probability mass function

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

Statement

The variance of a Geometric random variable is $\mathbb{V}(X) = \frac{1-p}{p^2}$.

Example: Variance of Geometric distribution

$X \sim \mathrm{Geo}(p)$ with the probability mass function

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

Statement

The variance of a Geometric random variable is $\mathbb{V}(X) = \frac{1-p}{p^2}$.

Proof.

We know that $\mathbb{E}(X) = 1/p$ and $\mathbb{V}[X] = \mathbb{E}X^2 - (\mathbb{E}X)^2$, then we only need to know

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p = p \sum_{k=1}^{\infty} k^2 q^{k-1} = p \sum_{k=1}^{\infty} (kq^k)' \quad \text{taking } q := 1-p \\
&= p \left(\sum_{k=1}^{\infty} kq^k \right)' = p \left(\frac{q}{(1-q)^2} \right)' = \frac{2-p}{p^2}\,.
\end{aligned}$$

[$S := \sum_{k=1}^{\infty} kq^k$, using $S - qS = \ldots$] $\square$

Chebyshev's inequality

Theorem (Chebyshev's inequality)

For a random variable X with its expectation μ and variance σ^2, then

$$\Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}.$$

Chebyshev's inequality

Theorem (Chebyshev's inequality)

For a random variable X *with its expectation* μ *and variance* σ^2*, then*

$$\Pr[|X-\mu| \geq t] \leq \frac{\sigma^2}{t^2}\,.$$

Proof.

$$\Pr[|X-\mu| \geq t] = \Pr[|X-\mu|^2 \geq t^2] \leq \frac{\mathbb{E}[|X-\mu|^2]}{t^2} = \frac{\sigma^2}{t^2}\,.$$

□

More information, better result

- Markov inequality: only use μ, convergence rate: $\mathcal{O}(1/t)$
- Chebyshev's inequality: use μ, σ^2, convergence rate: $\mathcal{O}(1/t^2)$

More information, better result

- Markov inequality: only use μ, convergence rate: $\mathcal{O}(1/t)$
- Chebyshev's inequality: use μ, σ^2, convergence rate: $\mathcal{O}(1/t^2)$

More general version: If we introduce a non-decreasing, non-negative function ϕ, then

$$\Pr(|X-\mu| \geq t) = \Pr[\phi(|X-\mu|) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}$$

More information, better result

- Markov inequality: only use μ, convergence rate: $\mathcal{O}(1/t)$
- Chebyshev's inequality: use μ, σ^2, convergence rate: $\mathcal{O}(1/t^2)$

More general version: If we introduce a non-decreasing, non-negative function ϕ, then

$$\Pr(|X-\mu| \geq t) = \Pr[\phi(|X-\mu|) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}$$

Moment (if exists and finite) by choosing ϕ as a polynomial function:

- 1st order moment: $\mathbb{E}[X]$
- 2nd order moment: $\mathbb{E}[X^2]$, $\mathbb{E}[|X-\mathbb{E}X|^2]$

  ...
- t-th order moment: $\mathbb{E}[X^t]$, $\mathbb{E}[|X-\mathbb{E}X|^t]$

Application of Chebyshev'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 Chebyshev'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 N*?*

$$\Pr(T \geq N) \leq ?$$

Solution

Solution

- T_i: *measures the number of independent trials to collect the* i*th unique coupon.*

Solution

Solution

- T_i: *measures the number of independent trials to collect the* i*th unique coupon.*
- $T_i \sim$ *Geo*(p_i) *with* $p_i = \frac{n-i+1}{n}$.

Solution

Solution

- T_i: *measures the number of independent trials to collect the* i*th unique coupon.*
- $T_i \sim$ *Geo*(p_i) *with* $p_i = \frac{n-i+1}{n}$.
- *Recall Geometric distribution:* $\mathbb{E}(X) = 1/p$ *and* $\mathbb{V}(X) = \frac{1-p}{p^2}$.

Solution

Solution

- T_i: measures the number of independent trials to collect the ith unique coupon.
- $T_i \sim$ Geo(p_i) with $p_i = \frac{n-i+1}{n}$.
- Recall Geometric distribution: $\mathbb{E}(X) = 1/p$ and $\mathbb{V}(X) = \frac{1-p}{p^2}$.
- $\mathbb{E}(T_i) = \frac{n}{n-i+1}$ and $\mathbb{V}(T_i) = \frac{n(i-1)}{(n-i+1)^2}$
- T_i and T_j are independent (each trial is independent)

We estimate the variance of T by the independence of $\{T_i\}_{i=1}^n$

Continue

Solution (To be continued)

$$
\begin{aligned}
\mathbb{V}[T] = \mathbb{V}[\sum_{i=1}^{n} T_i] = \sum_{i=1}^{n} \mathbb{V}(T_i) = \sum_{i=1}^{n} \frac{n(i-1)}{(n-i+1)^2} \\
&= 0 + \frac{n}{(n-1)^2} + \cdots + \frac{n(i-1)}{(n-i+1)^2} + \cdots + \frac{n(n-1)}{1^2} \\
&= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \cdots + \frac{i-1}{(n-i+1)^2} + \cdots + \frac{n-1}{1^2}\right) \\
&\leq n^2 \sum_{i=1}^{n} \frac{1}{i^2} = \frac{n^2\pi^2}{6}.
\end{aligned}
$$

Another way

$$
\begin{aligned}
\mathbb{V}[T] &= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \cdots + \frac{i-1}{(n-i+1)^2} + \cdots + \frac{n-1}{1^2}\right) \\
&\leq n\left(\frac{1}{(n-1)(n-2)} + \frac{2}{(n-2)(n-3)} + \frac{3}{(n-3)(n-4)} + \cdots + \frac{n-2}{2\times 1} + \frac{n-1}{1^2}\right) \\
&= n\left(\left[\frac{1}{n-2} - \frac{1}{n-1}\right] + 2\left[\frac{1}{n-3} - \frac{1}{n-2}\right] + 3\left[\frac{1}{n-4} - \frac{1}{n-3}\right] + \cdots + (n-2)\left[\frac{1}{1} - \frac{1}{2}\right] + \frac{n-1}{1}\right) \\
&\leq n\left(-\frac{1}{n-1} - \frac{1}{n-2} - \frac{1}{n-3} - \cdots - \frac{1}{2} + (n-2) + (n-1)\right) \\
&= n\left(-(H_n - 1 - 1/n) + 2n - 3\right)\,.
\end{aligned}
$$

Another way

$$
\begin{aligned}
\mathbb{V}[T] &= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \cdots + \frac{i-1}{(n-i+1)^2} + \cdots + \frac{n-1}{1^2}\right) \\
&\leq n\left(\frac{1}{(n-1)(n-2)} + \frac{2}{(n-2)(n-3)} + \frac{3}{(n-3)(n-4)} + \cdots + \frac{n-2}{2\times 1} + \frac{n-1}{1^2}\right) \\
&= n\left(\left[\frac{1}{n-2} - \frac{1}{n-1}\right] + 2\left[\frac{1}{n-3} - \frac{1}{n-2}\right] + 3\left[\frac{1}{n-4} - \frac{1}{n-3}\right] + \cdots + (n-2)\left[\frac{1}{1} - \frac{1}{2}\right] + \frac{n-1}{1}\right) \\
&\leq n\left(-\frac{1}{n-1} - \frac{1}{n-2} - \frac{1}{n-3} - \cdots - \frac{1}{2} + (n-2) + (n-1)\right) \\
&= n\left(-(H_n - 1 - 1/n) + 2n - 3\right)\,.
\end{aligned}
$$

Similarly, we have the lower bound (using $\frac{1}{n^2} \geq \frac{1}{n(n+1)}$)

$$\mathbb{V}[T] \geq n\left(-(H_n - 1) + n - 1\right)\,.$$

Continue

By Chebyshev's inequality, we have

$$\Pr(|T - \mathbb{E}(T)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2} \leq \frac{n^2\pi^2}{6t^2}.$$

That means,

$$\Pr(T \geq t + \mathbb{E}(T)) \leq \Pr(|T - \mathbb{E}(T)| \geq t) \leq \frac{n^2\pi^2}{6t^2}.$$

Continue

By Chebyshev's inequality, we have

$$\Pr(|T - \mathbb{E}(T)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2} \leq \frac{n^2\pi^2}{6t^2}\,.$$

That means,

$$\Pr(T \geq t + \mathbb{E}(T)) \leq \Pr(|T - \mathbb{E}(T)| \geq t) \leq \frac{n^2\pi^2}{6t^2}\,.$$

Recall $\mathbb{E}(T) = nH_n$, we have

$$\Pr(T \geq t + nH_n) \leq \Pr(|T - nH_n| \geq t) \leq \frac{n^2\pi^2}{6t^2}\,.$$

Taking $t := (\beta - 1)nH_n$ with $\beta > 1$, we have

$$\Pr(T \geq \beta nH_n) \leq \frac{\pi^2}{6(\beta-1)^2 H_n^2} \leq \frac{\pi^2}{6(\beta-1)^2 \log^2 n}$$

Can we do it better?

$$\Pr(T \geq N) \leq \textcolor{red}{small}$$

- strictly speaking, it should be $T \geq N + 1$...
- that means, at least one of n distinct objects has not been selected in the first N round.

Can we do it better?

$$\Pr(T \geq N) \leq \textcolor{red}{small}$$

- strictly speaking, it should be $T \geq N + 1$...
- that means, at least one of n distinct objects has not been selected in the first N round.
  - Let A_i^N denote when item i is not observed in the first N draws, i.e., $\Pr(A_i^N) = (1 - \frac{1}{n})^N$.
  - the event $\{T \geq N\} = \cup_{i=1}^n A_i^N$

Can we do it better?

$$\Pr(T \geq N) \leq \textcolor{red}{small}$$

◦ strictly speaking, it should be $T \geq N + 1$...
◦ that means, at least one of n distinct objects has not been selected in the first N round.

- Let A_i^N denote when item i is not observed in the first N draws, i.e., $\Pr(A_i^N) = (1 - \frac{1}{n})^N$.
- the event $\{T \geq N\} = \cup_{i=1}^n A_i^N$

$$\Pr(\text{not done in the first } N \text{ draws}) = \Pr(\cup_{i=1}^n A_i^N) \leq \sum_{i=1}^n \Pr(A_i^N) = \sum_{i=1}^n (1 - \frac{1}{n})^N = n(1 - \frac{1}{n})^N.$$

Can we do it better?

$$\Pr(T \geq N) \leq \textcolor{red}{small}$$

∘ strictly speaking, it should be $T \geq N + 1$...
∘ that means, at least one of n distinct objects has not been selected in the first N round.

- Let A_i^N denote when item i is not observed in the first N draws, i.e., $\Pr(A_i^N) = (1 - \frac{1}{n})^N$.
- the event $\{T \geq N\} = \cup_{i=1}^n A_i^N$

$$\Pr(\text{not done in the first } N \text{ draws}) = \Pr(\cup_{i=1}^n A_i^N) \leq \sum_{i=1}^n \Pr(A_i^N) = \sum_{i=1}^n (1 - \frac{1}{n})^N = n(1 - \frac{1}{n})^N .$$

Taking $N := \beta n \log n$, we have (using $1 + x \leq e^x$ for any $x \in \mathbb{R}$)

$$n(1 - \frac{1}{n})^N \leq n(e^{-\frac{1}{n}})^{\beta n \log n} = ne^{-\beta \log n} = n(e^{\log n})^{-\beta} = n^{-\beta+1} .$$

Can we do it better?

$$\Pr(T \geq N) \leq \textcolor{red}{small}$$

- strictly speaking, it should be $T \geq N+1$...
- that means, at least one of n distinct objects has not been selected in the first N round.
  - Let A_i^N denote when item i is not observed in the first N draws, i.e., $\Pr(A_i^N) = (1-\frac{1}{n})^N$.
  - the event $\{T \geq N\} = \cup_{i=1}^n A_i^N$

$$\Pr(\text{not done in the first } N \text{ draws}) = \Pr(\cup_{i=1}^n A_i^N) \leq \sum_{i=1}^n \Pr(A_i^N) = \sum_{i=1}^n (1-\frac{1}{n})^N = n(1-\frac{1}{n})^N .$$

Taking $N := \beta n \log n$, we have (using $1+x \leq e^x$ for any $x \in \mathbb{R}$)

$$n(1-\frac{1}{n})^N \leq n(e^{-\frac{1}{n}})^{\beta n \log n} = ne^{-\beta \log n} = n(e^{\log n})^{-\beta} = n^{-\beta+1} .$$

$\Rightarrow \Pr(T \geq N) \leq n^{-\beta+1}$.

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 31 and 32.)

*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

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