Lecture 02: Concentration Inequalities I

Nikola Zlatanov

Innopolis University

Advanced Statistics

30-th of Jan to 6-th of Feb, 2023

Motivation

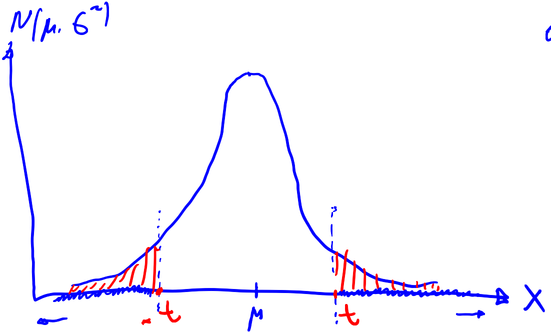
- We now begin to study probabilities in high dimensions and their applications to data science
- For that we need tools. The tools from low dimensional probabilities are not enough. We need new tools.
- The most important tool set in high dimensions is concentration inequalities.
- Concentration inequalities say the following: For many random variables (RV) X, the following holds

$$X \approx E[X], \text{ with high probability}$$

- Hence, many RVs are tightly distributed around their means.

Example: Normal Distribution

- If $X \sim N(\mu, \sigma^2)$, where
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$
then
$$|X - \mu| \leq 3\sigma \text{ with probability } 0.9973$$
- In words: 99.73% of the time, the Gaussian RV produces samples that are at most 3 standard deviations from its mean.
- Figure:



0,0017

$\underline{0,17}\%$

2

0,085

Gaussian Tails

- Fact: If $X \sim N(0,1)$, then for $t > 0$, the following holds

$\int_t^{+\infty} f_X(x)\,dx$

$$\begin{aligned}\Pr\{X > t\} = \Pr\{X < -t\} &\leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}\exp\left(-\frac{t^2}{2}\right)\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2} - \ln t\right)\end{aligned} \tag{1}$$

- Note that if $t > 1$, then $\frac{t^2}{2} + \ln t > \frac{t^2}{2}$ and as a result

$$\Pr\{X > t\} = \Pr\{X < -t\} \leq \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right), \text{ for } t > 1$$

- Note, the bound is of the same form as the distribution itself!
- This means the Gaussian tail decays exponentially fast with t^2.
- Even a simpler bound for $t > 1$: Since $1/\sqrt{2\pi} < 1$, we have

$$\Pr\{X > t\} = \Pr\{X < -t\} \leq \exp\left(-\frac{t^2}{2}\right), \text{ for } t > 1$$

Gaussian Tails

- Proof of Fact:

$$\Pr\{X > t\} = \Pr\{X < -t\} = \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Change of variables $x = t + y$, then $y = x - t$

$$\Pr\{X > t\} = \Pr\{X < -t\} = \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t+y)^2}{2}\right) dy$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \int_0^\infty \exp(-ty) \exp\left(-\frac{y^2}{2}\right) dy$$

$$\overset{(a)}{\leq} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \int_0^\infty \exp(-ty)\, dy \overset{(b)}{=} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \frac{1}{t},$$

where (a) is due to $\max_y e^{-y^2/2} = 1$, hence $e^{-y^2/2} \leq 1$, and (b) is due to $\int_0^\infty \overset{-}{e}^{ty} dy = \int_0^\infty \overset{-}{e}^{z}/t dz = (1-0)/t = 1/t$, which comes from the change of variables $z = ty$, hence $dy = dz/t$ and the integral limits stay unchanged.

Gaussian Tails

- Two sided bound

$$\Pr\{|X| \geq t\} = \Pr\{X > t \cup X < -t\} \overset{(a)}{=} \Pr\{X > t\} + \Pr\{X < -t\}$$
$$\overset{(b)}{\leq} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2} - \ln t\right),$$

  where (a) follows since $X > t$ and $X < -t$ are disjoint events and (b) is due to (1)
- If $t > 1$, then

$$\Pr\{|X| \geq t\} \leq \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \text{ for } t > 1$$

- Even simpler for $t > 1$: Since $2/\sqrt{2\pi} < 1$, we have then

$$\Pr\{|X| \geq t\} \leq \exp\left(-\frac{t^2}{2}\right), \text{ for } t > 1$$

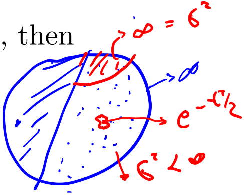
Gaussian Tails

- If instead $X \sim N(0,1)$, we have $Y \sim N(\mu, \sigma^2)$, what will be the tail bound? Note that $Y = \sigma X + \mu$, hence
$$\Pr\{Y \geq t\} = \Pr\{\sigma X + \mu > t\} = \Pr\left\{X > \frac{t-\mu}{\sigma}\right\}$$
$$\leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2} - \ln\frac{t-\mu}{\sigma}\right), \text{ but only if } t > \mu$$
- If $(t-\mu)/\sigma > 1$, which is equivalent to $t > \sigma + \mu$, we have
$$\Pr\{Y \geq t\} \leq \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right), \text{ for } t > \sigma + \mu$$
- Therefore, the two sided bound of $Y \sim N(\mu, \sigma^2)$ for $t > \mu$ is
$$\Pr\{|Y| \geq t\} \leq \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2} - \ln\frac{t-\mu}{\sigma}\right) \text{ for } t > \mu$$
- If $t > \sigma + \mu$
$$\Pr\{|Y| \geq t\} \leq \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right), \text{ for } t > \sigma + \mu$$

Tails for general distributions: Markov's bound

- For general distributions, we cannot derive such fast decaying tails as those for the Gaussian distribution.
  - Although, as we shell see, there are sub-groups of distributions for which we can derive tails bound that will decaying as fast as those for the Gaussian distribution.
- Markov's inequality: If a RV X is non-negative, then

$$\Pr\{X > t\} \leq \frac{E[X]}{t}$$



  Proof: We proved it in the last course!
- Note that the Markov's bound, since it is for general non-negative RVs, decays with linearly with t, whereas the Gaussian distribution decays exponentially with t^2.
- The above bound even works if the variance is unbounded, i.e., infinite.

Tails for dis. with bounded variance: Chebyshev bound

- For general distributions, if we know that the variance is bounded, we can use Chebyshev inequity: If X has mean μ and variance σ^2, the following bound holds

$$\Pr\{|X-\mu|>t\} \leq \frac{\sigma^2}{t^2}$$

Proof: We proved it in the last course!

- Note that the Chebyshev's bound, since it is for general RVs with bounded variance, decays with t^2, which is mush faster than Markov's bound that decays with t, but much slower than the Gaussian that decays exponentially with t^2.
- Note that if we denote the set of all distributions $\mathcal{G}$, $\mathcal{C}$, and $\mathcal{M}$ that satisfy the Markov, Chebyshev, and Gaussian tails, then $\mathcal{G} \subset \mathcal{C} \subset \mathcal{M}$

Chebyshev bound for a normalized sum

- Let X_1, X_2,..., X_n be n i.i.d. RVs with mean μ and variance σ^2 and let us define the normalized sum as

Var = 1

n

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

VAR = 1

Note, the normalized sum satisfies $E[Z] = 0$ and $\mathrm{VAR}[Z] = 1$.

- Then the Chebyshev bound is

$$\Pr\{|Z| > t\} \le \frac{1}{t^2}$$

- Even for sums, this bound seem to be very slow converging. Can we obtain Gaussian type tails for sums of i.i.d. RVs involving general distributions?
- How about if we approximate the normalized sum with the Gaussian distribution? In the following, we will see how to approximate sum of RVs from a general distributions with the Gaussian distribution

Central Limit Theorem (CLT)

- CLT: Let X_1, X_2,..., X_n be n i.i.d. RVs from some general distribution with mean μ and variance σ^2. Let Z be defined as

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \tag{2}$$

then

$$Z \sim N(0,1), \text{ when } n \to \infty$$

Proof: We state the CLT without proof.

- The CLT states that the scaled sum of i.i.d. RVs converges in distribution to the Gaussian distribution.
- This maybe be what we are looking for, since now due to the CLT, maybe a normalized sum may decay as fast as the Gaussian tail bound instead of what the Chebyshev bound showed us previously a decay with t^2. Let's check if this intuition is true.

CLT vs Chebyshev

- For Z being the normalized sums of i.i.d. RVs as in (2), we have the following bounds:
- From the Chebyshev bound:

$$\Pr\{|Z| > t\} \leq \frac{1}{t^2}$$

- From the CLT:

$$\Pr\{|Z| > t\} \leq \epsilon + e^{-\frac{t^2}{2}}, \tag{3}$$

  where ϵ is the maximum error we make due to the approximation of Z with a Gaussian RV when n is finite.
- How big is ϵ? If $\epsilon \ll e^{-\frac{t^2}{2}}$, then we can neglect it in (3) and we obtain that indeed we have a Gaussian-type tail.
- But if $\epsilon \gg e^{-\frac{t^2}{2}}$, it dominates the sum in (3) and then $e^{-\frac{t^2}{2}}$ should be one that is neglected in (3), which destroys our main goal: Gaussian-type tail via the CLT.

Berry-Esseen CLT

- How big is the approximation error in CLT ϵ, depends, intuitively, on how fast Z convergence in distribution to the Gaussian? Is it very slow (bad), or is it very fast (good)? We will look at the convergence rate in the following.
- Berry-Esseen CLT: Let X_1, X_2,..., X_n be n i.i.d. RVs from some general distribution with mean μ and variance σ^2. Let $Y \sim N(0,1)$ and let Z be the normalized sum, defined as

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

then

$$\big| \Pr\{Z \geq t\} - \Pr\{Y \geq t\} \big| \leq \frac{c}{\sqrt{n}} \triangleq \epsilon$$

where $c = E[|X_i - \mu|^3/\sigma^3]$ is a constant.

- Hence, the difference between the tails of the approximation and the actual Gaussian tail decays with rate $\sqrt{n}$, which means that $\epsilon \triangleq c/\sqrt{n} \gg e^{-t^2/2}$ for $n < e^{t^2}$.

CLT vs Chebyshev

- If we go back to CLT vs Chebyshev bounds, from the Chebyshev bound:

  $$\Pr\{|Z| > t\} \leq \frac{1}{t^2}$$

  and from the CLT:

  $$\Pr\{|Z| > t\} \leq \frac{c}{\sqrt{n}} + e^{-\frac{t^2}{2}},$$

  Hence, for $t > n^{1/4}$, we have

  $$\frac{1}{t^2} < \frac{c}{\sqrt{n}} + e^{-\frac{t^2}{2}}$$

  Hence, for $t > n^{1/4}$, in the CLT vs Chebyshev dominance, the winner is the Chebyshev bound.
- Conclusion: Gaussian-type tail bounds based on the CLT approximations are looser than the Chebyshev bound.
- Let's look at a concrete numerical examples where we compare the CLT vs Chebyshev bounds in the following.

Example of CLT vs Chebyshev Bnds for Sym Bernoulli

- Let us toss a fair coin with sides 1 and -1. Let $X \in \{-1, 1\}$ be the RV modelling the coin tosses. Then, $\Pr\{X = 1\} = \Pr\{X = -1\} = 1/2$. This distribution is called the Symmetric Bernoulli distribution.
- Problem: What is the probability that in n consecutive coin tosses of this coin, we will obtain at least 75% of the outcomes to be 1?
- Solution: Let X_i be the i-th independent toss. Then, we need to obtain the following probability

1 -1 1 1 -1 -1 1

$$\Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{3}{4}n - \frac{1}{4}n\right\} = \Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{1}{2}n\right\} \tag{4}$$

We need to bound the probability in (4)

From Chebyshev bound, we obtain the following bound

$$\Pr\left\{\left|\sum_{i=1}^{n} X_i\right| \geq \frac{1}{2}n\right\} \leq \frac{n}{\left(\frac{1}{2}n\right)^2} = \frac{4}{n} \tag{5}$$

Example of CLT vs Chebyshev Bnds for Sym Bernoulli

Now, due to symmetry of the distribution of the sum $\sum_{i=1}^{n} X_i$, we have

$$\Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{1}{2}n\right\} = \frac{1}{2}\Pr\left\{\left|\sum_{i=1}^{n} X_i\right| \geq \frac{1}{2}n\right\} \overset{(a)}{\leq} \frac{2}{n}, \tag{6}$$

where (a) comes from (5).

If we lug-in $n = 1000$ in (6), we obtain

$$\Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{1}{2}n\right\} \leq \frac{2}{1000} = \frac{1}{500} = 0.002 \tag{7}$$

Does the bound in (7) seem tight? If it is tight, then it says that we would get 1 in a 500 trials a sequence of 1000 outcomes where 75% of the outcomes will be one? To me this is unlikely to happen, intuitively, hence, the bounds seems to be very loose.

Example of CLT vs Chebyshev Bnds for Sym Bernoulli

On the other hand, from the CLT approximation, we obtain

$$\Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{1}{2}n\right\} = \Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq \frac{1}{2}\sqrt{n}\right\} \leq \frac{1}{\sqrt{n}} + e^{-\frac{1}{2}\frac{1}{4}n} \tag{8}$$

If we plug-in $n = 1000$ in (8), we obtain

$$\Pr\left\{\sum_{i=1}^{n} X_i \geq \frac{3}{4}n\right\} \leq \frac{1}{\sqrt{1000}} + e^{-\frac{1}{8}1000} \approx 0.031,$$

which is even looser bound. It says that we would get 1 in a 30 trials a series of 1000 outcomes where 75% of the outcomes will be one. Should be a much much smaller chance than that!

- Is there something more tighter we can derive for the Symmetric Bernoulli distribution?

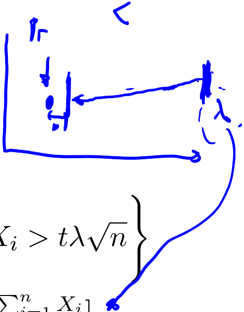
Hoeffding's Inequality For Symmetric Bernoulli

- Is there any other way to get tail bounds that decay exponentially, at least for the Symmetric Bernoulli. It turns out there is:
- Thm (Hoeffding's inequality): Let X_1, X_2,..., X_n be n i.i.d. symmetric Bernoulli RVs defined by their PMF $\Pr\{X_i = 1\} = \Pr\{X_i = -1\} = 1/2$, then a bound on the normalized sum is
$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right\} \le e^{-t^2/2} \tag{9}$$
- Note that a Symmetric Bernoulli X_i satisfies $E[X_i] = 0$ and $\mathrm{VAR}[X_i] = 1$.
- Hence, we obtained a bound exactly as the one for the Gaussian.
- Going back to the numerical example, if we plug in $n = 1000$ into (9), we obtain that the upper bound is $e^{-\frac{1}{8}1000} \approx 5^{-55}$, which is an extremely small probability. Hence, obtaining a sequence of 75% ones when the sequence length is 1000, is almost impossible!

Proof of Hoeffding's inequality

- Proof: The proof is via the Moment Generating Function (MGF)
- The MGF of RV X is defined as

$$M_X(\lambda) = E[e^{X\lambda}]$$



- Continuation of the proof:

$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right\} \overset{(a)}{=} \Pr\left\{\lambda\sum_{i=1}^{n} X_i > t\lambda\sqrt{n}\right\}$$

$$\overset{(b)}{=} \Pr\left\{e^{\lambda\sum_{i=1}^{n} X_i} > e^{t\lambda\sqrt{n}}\right\} \overset{(c)}{\leq} \frac{E[e^{\lambda\sum_{i=1}^{n} X_i}]}{e^{t\lambda\sqrt{n}}},$$

where (a) comes from multiplying both sides by $\sqrt{n}\lambda$, for $\lambda > 0$, (b) comes from taking the exponential from both sides, and (c) comes from Markov's inequality.

Proof of Hoeffding's inequality

- Note that

$$E\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = E\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] \overset{(a)}{=} \prod_{i=1}^{n} E[e^{\lambda X_i}] \overset{(b)}{=} \left(E[e^{\lambda X}]\right)^n = (M_X(\lambda))^n$$

where (a) is due to the independence of the RVs, (b) is due to the identical distribution of the RVs, where X is i.i.d. as any X_i, i.e., a symmetric Bernoulli RV.

Now, let's compute $M_X(\lambda)$

$$\begin{aligned} M_X(\lambda) &= E[e^{\lambda X}] = \Pr\{X = -1\}e^{\lambda\times(-1)} + \Pr\{X = 1\}e^{\lambda\times(1)} \\ &= \frac{1}{2}e^{-\lambda} + \frac{1}{2}e^{\lambda} \end{aligned} \tag{10}$$

Proof of Hoeffding's inequality

- Hence, we obtain

$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right\} \leq \left(\frac{e^{-\lambda}+e^{\lambda}}{2}\right)^{n}\frac{1}{e^{t\lambda\sqrt{n}}} \tag{11}$$

- Note that the above expression holds for any $\lambda > 0$.
- We would like to make the right hand-side of (11) as small as possible in order for the bound to be tight.
- To do this, we can minimize the right hand side of (11) w.r.t. $\lambda > 0$.
- However, in (11), we have a sum of two exponentials on the power n, which is a difficult expression to minimize.
- To get around this problem, we will first try to bound the sum of the two exponentials, i.e., bound (10).

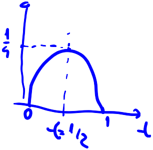
Proof of Hoeffding's inequality

- Let $a < 0 < b$ then
$$\frac{b}{b+|a|}e^{-|a|\lambda} + \frac{|a|}{b+|a|}e^{b\lambda} \le \exp\left(\lambda^2 \frac{(b+|a|)^2}{8}\right) \tag{12}$$
- Let $\gamma = \frac{|a|}{b+|a|}$, then we obtain $1-\gamma = \frac{b}{b+|a|}$, $|a|\lambda = \gamma(b+|a|)\lambda$ and $b\lambda = (1-\gamma)(b+|a|)\lambda$. If we define $u = (b+|a|)\lambda$, then we obtain $|a|\lambda = \gamma u$ and $b\lambda = (1-\gamma)u$. Inserting these in the left hand side of (12), we obtain
$$\begin{aligned}\frac{b}{b+|a|}e^{-|a|\lambda} + \frac{|a|}{b+|a|}e^{b\lambda} &= (1-\gamma)e^{-\gamma u} + \gamma e^{(1-\gamma)u}\\ &= (1-\gamma)e^{-\gamma u} + \gamma e^{u}e^{-\gamma u}\\ &= e^{-\gamma u}((1-\gamma)+\gamma e^{u})\\ &= e^{-\gamma u+\ln((1-\gamma)+\gamma e^{u})} = e^{h(u)},\end{aligned} \tag{13}$$
where
$$h(u) = -\gamma u + \ln(1-\gamma+\gamma e^{u}) \tag{14}$$

Proof of Hoeffding's inequality

- Now let's see if we can bound $h(u)$, given by (14)
- From Taylor series expansion of $h(u)$, we know that $\exists c$, $0 \leq c \leq u$, such that

$$h(u) = h(0) + uh^{'}(0) + \frac{1}{2}u^2 h^{''}(c) \tag{15}$$



- In our case, we have $h(0) = h^{'}(0) = 0$, whereas

$$h^{''}(c) = \frac{e^c\gamma}{1-\gamma+\gamma e^c}\left(1 - \frac{e^c\gamma}{1-\gamma+\gamma e^c}\right) = t(1-t) \overset{(a)}{\leq} \frac{1}{4} \tag{16}$$

where $t = \frac{e^c\gamma}{1-\gamma+\gamma e^c}$ and $0 \leq t \leq 1$, and the inequality (a) is obtains as follows. The function $t(1-t)$ is concave since $\frac{\partial^2}{\partial t^2}t(1-t) = -2 < 0$. The function $t(1-t)$ obtains its maximum for t found from $\frac{\partial}{\partial t}t(1-t) = 1 - 2t = 0$, from where we obtain $t = 1/2$. Hence, $t(1-t)$ has a maximum $1/2(1-1/2) = 1/4$, from where (a) follows.

Proof of Hoeffding's inequality

- Now inserting (16) into (15), we obtain

$$h(u) \leq \frac{1}{4}u^2 \tag{17}$$

- Inserting (17) into (13), we obtain the final bound in (12).

Proof of Hoeffding's inequality

- Using the bound in (12), we have the following bound

$$\frac{1}{2}e^{-\lambda} + \frac{1}{2}e^{\lambda} \le e^{\lambda^2/2} \tag{18}$$

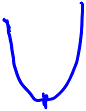
Proof: Set $a = -1$ and $b = 1$ in (12)

- Inserting (18) into (11), we obtain

$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right\} \le \frac{e^{n\lambda^2/2}}{e^{t\lambda\sqrt{n}}} = e^{n\lambda^2/2 - t\lambda\sqrt{n}} = e^{\alpha^2/2 - t\alpha}, \tag{19}$$

where $\alpha^{2} = n\lambda^2$.

- To minimize the right hand side on (18) w.r.t α, we need to minimize the expression in the exponent, leading to



$$\frac{d}{d\alpha}[\alpha^2/2 - t\alpha] = \alpha - t = 0 \tag{20}$$

Hence, the optimal α is $\alpha_* = t$, from where we obtain $e^{\alpha_*^2/2 - t\alpha_*} = e^{-t^2/2}$, Q.E.D.

Hoeffding's inequality for general symmetric Bernoulli

- We proved Gaussian-type tails for the sum of Symmetric Bernoulli RVs via Hoeffding's inequality for Symmetric Bernoulli.
- But do we obtain Gaussian-type tails for Symmetric Bernoulli with $\Pr\{X_i = 1\} = 1 - \Pr\{X_i = -1\} = p$?
- Thm (Hoeffding's inequality): Let X_1, X_2,..., X_n be n i.i.d. general Symmetric Bernoulli RVs defined by their PMF $\Pr\{X_i = 1\} = 1 - \Pr\{X_i = -1\} = p$. Then a bound on the normalized sum is given by

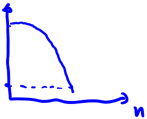
$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - E[X_i]}{\sigma} > t\right\} \leq e^{-t^2/2},$$

  where $E[X_i] = 2p - 1$ and $\sigma^2 = \mathrm{VAR}[X_i] = 4p(1-p)$
- Prove at home!

Example of Hoeffding's ineq. for general sym. Bernoulli

- Note, we will use the results in the following example later on when we propose the Medians of Means Estimator!
- Example: Let us have a biased coin with $\Pr\{X_i = 1\} = p = 1/4$ and thereby $\Pr\{X_i = -1\} = 3/4$. Then, the probability that in n consecutive tosses of this coin, we will obtain that half of the outcomes are 1, is bounded as

$$\Pr\left\{\sum_{i=1}^{n} X_i > \frac{1n}{2\cancel{n}} - \frac{1n}{2\cancel{n}}\right\} = \Pr\left\{\sum_{i=1}^{n} X_i > 0\right\}$$

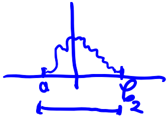


$$= \Pr\left\{\sum_{i=1}^{n}(X_i - E[X_i]) > -nE[X_i]\right\}$$

$$= \Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - E[X_i]}{\sigma} > -\sqrt{n}\frac{E[X_i]}{\sigma}\right\}$$

$$\leq e^{-\frac{n}{2}\frac{E^2[X_i]}{\sigma^2}} = e^{-\frac{n}{2}\frac{(2p-1)^2}{4p(1-p)}} = e^{-\frac{n}{2}\frac{1}{3}} = e^{-\frac{n}{6}} \qquad (21)$$

Hoeffding's inequality for sum of general bounded RVs



$b > a$

- Thm (Hoeffding's inequality): Let X_1, X_2,..., X_n be n i.i.d. general RVs that take values in the interval $[a, b]$. Then a bound on the normalized sum is

$$\Pr\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - E[X_i]}{\sigma} > t\right\} \leq e^{-t^2\frac{2\sigma^2}{(b-a)^2}}$$

- Prove at home!

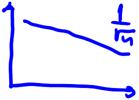
Application of HI: Median Of Means Estimator

In the rest of the lectures, we will discuss one important application of Hoeffding's inequality (HI), which is a new and very powerful estimator referred to as the "Median Of Means Estimator".

- Problem: Estimate the mean μ of a general distribution from a random sample X_1, X_2,...,X_n, where each X_i is drawn independently from the distribution.
- In the following, we discuss the classical estimator and show its performance.
- Then, we will introduce a new and very powerful estimator referred to as the "Median Of Means Estimator", and then derive its performance.

Classical Estimator

- Classical solution (learned in the previous course:) The mean estimator is $\hat{\mu}$, given by
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$



- This estimator is unbiased since $E[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$
- The mean squared error that this estimator achieves is
$$E[(\hat{\mu}-\mu)^2] = \text{VAR}[\hat{\mu}] = \text{VAR}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \overset{(a)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{VAR}[X_i] = \frac{\sigma^2}{n},$$
where (a) is due to the independence between X_i, $\forall i$.
- Now, since $E[(\hat{\mu}-\mu)^2] = O(1/n)$, we expect that the average absolute error, defined as $E[|\hat{\mu}-\mu|]$, satisfy $E[|\hat{\mu}-\mu|] = O(1/\sqrt{n})$, which means that the average absolute error is in the order of $1/\sqrt{n}$, or that it decays with rate $1/\sqrt{n}$.

Tail Bounds of the Classical Estimator

- We would like to bound $\Pr\{|\hat{\mu}-\mu| > \overset{\mu}{t}\sigma/\sqrt{n}\}$, which is the probability that the absolute error is t times larger than $\sigma/\sqrt{n}$?
- From Chebyshev inequality, we have

$$\Pr\left\{|\hat{\mu}-\mu| > t\frac{\sigma}{\sqrt{n}}\right\} \leq \frac{\text{VAR}[\hat{\mu}]}{(t\sigma/\sqrt{n})^2} = \frac{(\sigma/\sqrt{n})^2}{(t\sigma/\sqrt{n})^2} = \frac{1}{t^2}$$

- The probability that $|\hat{\mu}-\mu|$ is larger than 3 times $\sigma/\sqrt{n}$ is at most $1/9 \approx 0.1$, which means that at most it happens 1 in 10 trials, hence, very often. But is the above true for all estimators? Maybe there is some other estimator for which this happens very rarely.
- We derive such an estimator in the following.

Application of HI: Median Of Means Estimator

- Thm (Median of Means Estimator): There exists a mean estimator $\tilde{\mu}$, known as the Median of Means Estimator, such that for any $t > 0$, the following holds

$$\Pr\left\{|\tilde{\mu} - \mu| > t\frac{\sigma}{\sqrt{n}}\right\} \leq 2e^{-t^2/24} \tag{22}$$

- It is very surprising that such a powerful estimator exists.
- Proof: We partition the sample sequence $X_1, X_2, ..., X_n$ into k sets, each of size m, where $n = km$.

  Hence, we have k sets of samples $\mathcal{B}_1 = \{X_1, X_2, ..., X_m\}$, $\mathcal{B}_2 = \{X_{m+1}, X_{m+2}, ..., X_{2m}\}$,..., $\mathcal{B}_k = \{X_{n-k+1}, X_{n-k+2}, ..., X_n\}$. We now take the mean of each set and obtain $\hat{\mu}_1 = \frac{1}{m}\sum\limits_{i\in\mathcal{B}_1} X_i$, $\hat{\mu}_2 = \frac{1}{m}\sum\limits_{i\in\mathcal{B}_2} X_i$, ... , $\hat{\mu}_k = \frac{1}{m}\sum\limits_{i\in\mathcal{B}_k} X_i$.

  $[\hat{\mu}_1, \hat{\mu}_2 \dots \hat{\mu}_k]$

Application of HI: Median Of Means Estimator

- Proof Continuation: We now take the median of the sequence $\hat{\mu}_1$, $\hat{\mu}_2$, ..., $\hat{\mu}_k$, denoted by

$$[1, 2, 0.5, 3, 0.3] \xrightarrow[\text{sort}]{} [\underbrace{0.3, 0.5}, \boxed{1}, \boxed{2, 3}]$$

$$\tilde{\mu} = \text{MED}[\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_k], \tag{23}$$

and call this the Median of Means Estimator.

Note, the median of a sequence gives us a number y such that 50% of the numbers in that sequence are smaller than y and 50% of the numbers in that sequence are larger than y.

Now, to find the tail bound of the Median of Means Estimator, lets do the following. Let's check the accuracy of each $\hat{\mu}_j$, for $j = 1, 2, ..., k$, in (23). Then, we obtain that the mean is $E[\hat{\mu}_j] = \mu$, variance is $E[(\hat{\mu}_j - \mu)^2] = \sigma^2/m$, and the tail is

$$\Pr\left\{|\hat{\mu}_j - \mu| > t\frac{\sigma}{\sqrt{n}}\right\} \le \frac{E[(\hat{\mu}_j - \mu)^2]}{(t\sigma/\sqrt{n})^2} = \frac{\sigma^2/m}{t^2\sigma^2/n} = \frac{n}{t^2 m} = \frac{k}{t^2} \overset{(a)}{=} \frac{1}{4}, \tag{24}$$

where (a) comes by setting $k = t^2/4$.

Application of HI: Median Of Means Estimator

- Proof Continuation: Now note what (24) tells us: We have k very crude/inaccurate estimators, where each crude estimator, $\hat{\mu}_j$, for $j = 1, 2, ..., k$, has a tail bounded by $1/4$.

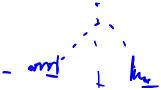
  However, if we have many such crude estimators, i.e., if k is large, then we can build from the k crude estimators one very accurate estimator.

  In fact, this one very accurate estimator, build by k crude estimators, is the Median of Means Estimator, given by (23).

Application of HI: Median Of Means Estimator

- Proof Continuation: Now, the tail of the median estimator is

$$\Pr\left\{|\tilde{\mu}-\mu|>t\frac{\sigma}{\sqrt{n}}\right\}=\Pr\left\{\tilde{\mu}-\mu>t\frac{\sigma}{\sqrt{n}}\right\}+\Pr\left\{\tilde{\mu}-\mu<t\frac{\sigma}{\sqrt{n}}\right\}$$
$$\overset{(a)}{=}2\Pr\left\{\tilde{\mu}-\mu>t\frac{\sigma}{\sqrt{n}}\right\},$$



where (a) comes since $\tilde{\mu}-\mu$ has a symmetric distribution.

On the other hand

$[\hat{\mu}_1, \hat{\mu}_2, \cdots \hat{\mu}_k]$

$\tilde{\mu}$

$$\Pr\left\{\tilde{\mu}-\mu>t\frac{\sigma}{\sqrt{n}}\right\}=\Pr\left\{\tilde{\mu}>\boxed{\mu+t\frac{\sigma}{\sqrt{n}}}\right\}$$
$$\overset{(a)}{\leq}\Pr\left\{\text{at least half of the crude estimators produce }\hat{\mu}_j>\mu+t\frac{\sigma}{\sqrt{n}}\right\}$$

where (a) comes from the definition of the median function.

Application of HI: Median Of Means Estimator

- Proof Continuation: Now we can think of each crude estimator as a biased coin. Each estimator will give as a $\hat{\mu}_j > \mu + t\frac{\sigma}{\sqrt{n}}$ with probability $1/4$ and give us $\hat{\mu}_j < \mu + t\frac{\sigma}{\sqrt{n}}$ with probability $3/4$. This equivalent to each coin will give 1 with probability $1/4$ and give us -1 with probability $3/4$, which leads to

$$\Pr\left\{\text{at least half of the crude estimators produce } \hat{\mu}_j > \mu + t\frac{\sigma}{\sqrt{n}}\right\}$$
$$= \Pr\left\{\text{from } k \text{ biased coin tosses at least } k/2 \text{ provide outcome } 1\right\}$$
$$\overset{(a)}{\leq} e^{-k/6} \overset{(b)}{=} e^{-t^2/24}$$

where (a) comes from (21) and the example on that slide and (b) comes from the fact that we have set k to $k = t^2/4$.
Hence, we have obtained (22). Q.E.D.

Application of HI: Median Of Means Estimator

- We obtained that the median mean estimator has a Gaussian type tail bound that decays exponentially with t^2, whereas the classical mean estimator has a tail that decays with $1/t^2$, hence much much slower.
- But the classical mean estimator has a variance, i.e., mean squared error of σ^2/n. How about the variance of the Median Mean Estimator? **If someone can derive the variance, i.e., mean squared error, of the Median Mean Estimator at home, I will give points!**
- We will test variance of the Median Mean Estimator with simulations in the lab.