ECE595 / STAT598: Machine Learning I
Lecture 23 Probability Inequality

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

- Lecture 22 Is Learning Feasible?
- Lecture 23 Probability Inequality
- Lecture 24 Probably Approximate Correct

Today's Lecture:

- Basic Inequalities
  - Markov and Chebyshev
  - Interpreting the results
- Advance Inequalities
  - Chernoff inequality
  - Hoeffding inequality

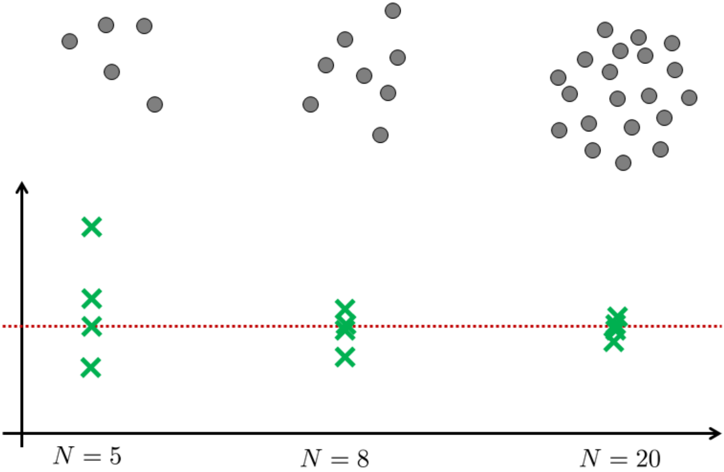
Empirical Average

- We want to take a detour to talk about probability inequalities
- These inequalities will become useful when studying learning theory

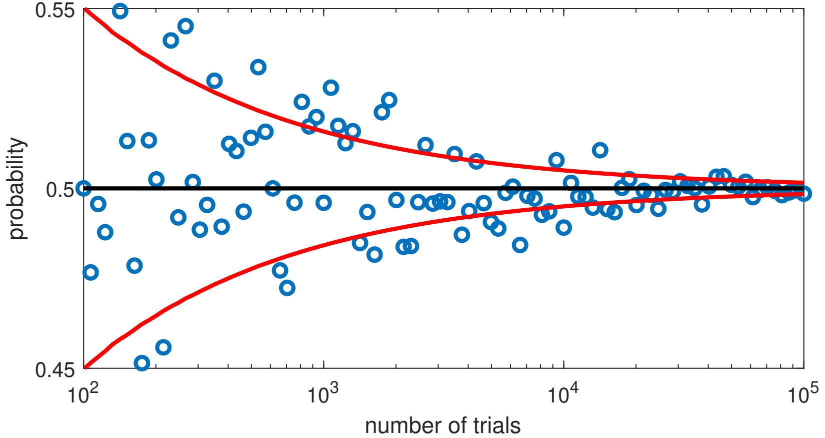
Let us look at 1D case.

- You have random variables $X_1, X_2, \ldots, X_N$.
- Assume **independently identically distributed** i.i.d.
- This implies
$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots = \mathbb{E}[X_N] = \mu$$
- You compute the **empirical average**
$$\nu = \frac{1}{N}\sum_{n=1}^{N} X_n$$
- How close is ν to μ?

As N grows ...



As N grows ...



Interpreting the Empirical Average

$$\nu = \frac{1}{N}\sum_{n=1}^{N} X_n$$

- ν is a random variable
- ν has CDF and PDF
- ν has mean
-

$$\begin{aligned}\mathbb{E}[\nu] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n] \\ &= \frac{1}{N}N\mu = \mu.\end{aligned}$$

- Note that "$\mathbb{E}[\nu] = \mu$" is not the same as "$\nu = \mu$".
- What is the probability ν deviates from μ?

Probability of Bad Event

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] = \ ?$$

- $\mathcal{B} = \{|\nu - \mu| > \epsilon\}$: The $\mathcal{B}$ad event: ν deviates from μ by at least ϵ
- $\mathbb{P}[\mathcal{B}] =$ probability that this bad event happens.
- Want $\mathbb{P}[\mathcal{B}]$ small. So upper bound it by δ.

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \ \leq \ \delta.$$

- With probability **no greater** than δ, $\mathcal{B}$ad event happens.
- Rearrange the equation:

$$\mathbb{P}\left[|\nu - \mu| \leq \epsilon\right] \ > \ 1 - \delta.$$

- With probability **at least** $1 - \delta$, the $\mathcal{B}$ad event will **not** happen.

Markov Inequality

Theorem (Markov Inequality)

For any $X > 0$ *and* $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

$$\begin{aligned}
\epsilon \mathbb{P}[X \geq \epsilon] &= \epsilon \int_{\epsilon}^{\infty} p(x) dx \\
&= \int_{\epsilon}^{\infty} \epsilon \; p(x) dx \\
&\leq \int_{\epsilon}^{\infty} x p(x) dx \\
&\leq \int_{0}^{\infty} x p(x) dx = \mathbb{E}[X].
\end{aligned}$$

Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let $X_1, \ldots, X_N$ be i.i.d. with $\mathbb{E}[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2$. Define

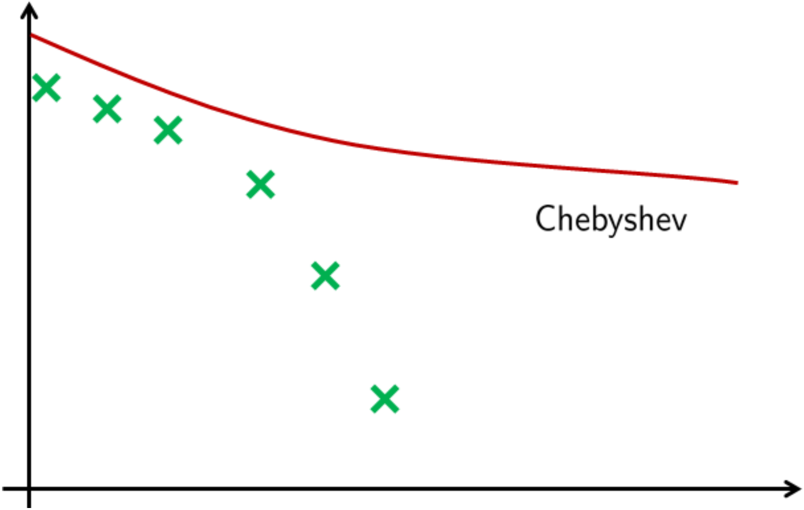
$$\nu = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

Then,

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq \frac{\sigma^2}{N\epsilon^2}$$

$$\mathbb{P}\left[|\nu - \mu|^2 > \epsilon^2\right] \underbrace{\leq \frac{\mathbb{E}[|\nu-\mu|^2]}{\epsilon^2}}_{\text{Markov}} \underbrace{= \frac{\mathrm{Var}[\nu]}{\epsilon^2}}_{\mathbb{E}[(\nu-\mu)^2]=\mathrm{var}[\nu]} \underbrace{= \frac{\sigma^2}{N\epsilon^2}}_{\mathrm{var}[\nu]=\frac{\sigma^2}{N}}.$$

How Good is Chebyshev Inequality?



Weak Law of Large Number

Theorem (WLLN)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with common mean μ. Let $M_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then, for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[|M_N - \mu| > \varepsilon\right] = 0. \tag{1}$$

Remark:

- The limit is outside the probability.
- This means that the probability of the event $|M_N - \mu| > \varepsilon$ is diminishing as $N \to$.
- But diminishing probability can still have occasions where $|M_N - \mu| > \varepsilon$.
- It just means that these occasions do not happen often.

Strong Law of Large Number

Theorem (SLLN)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with common mean μ. Let $M_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left[\lim_{N\to\infty} |M_N - \mu| > \varepsilon\right] = 0. \tag{2}$$

Remark:

- The limit is inside the probability.
- We need to analyze the limiting object $\lim_{N\to\infty} |M_N - \mu|$
- This object may or may not exist. This object is another random variable.
- The probability is measuring the event that this limiting object will deviate significantly from ε
- There is no "occasional" outliers.

Outline

- Lecture 22 Is Learning Feasible?
- Lecture 23 Probability Inequality
- Lecture 24 Probably Approximate Correct

Today's Lecture:

- Basic Inequalities
  - Markov and Chebyshev
  - Interpreting the results
- Advance Inequalities
  - Chernoff inequality
  - Hoeffding inequality

Hoeffding Inequality

Let us revisit the Bad event:

$$\begin{aligned}\mathbb{P}[|\nu-\mu|\geq\epsilon] &= \mathbb{P}[\nu-\mu\geq\epsilon \quad \textbf{or} \quad \nu-\mu\leq-\epsilon] \\ &\leq \underbrace{\mathbb{P}[\nu-\mu\geq\epsilon]}_{\leq A} + \underbrace{\mathbb{P}[\nu-\mu\leq-\epsilon]}_{\leq A}, \qquad \text{Union bound} \\ &\leq 2A, \qquad \text{(What is } A\text{? To be discussed.)}\end{aligned}$$

Theorem (Hoeffding Inequality)

Let $X_1,\ldots,X_N$ be random variables with $0\leq X_n\leq 1$, then

$$\mathbb{P}\left[|\nu-\mu|>\epsilon\right]\leq 2\underbrace{e^{-2\epsilon^2 N}}_{=A}$$

The e-trick + Markov Inequality

Let us check one side:

$$\mathbb{P}[\nu - \mu \geq \epsilon] = \mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} X_n - \mu \geq \epsilon\right] = \mathbb{P}\left[\sum_{n=1}^{N}(X_n - \mu) \geq \epsilon N\right]$$

$$= \mathbb{P}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)} \geq e^{s\epsilon N}\right], \qquad \forall s > 0$$

$$\leq \frac{\mathbb{E}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\right]}{e^{s\epsilon N}}, \qquad \text{Markov Inequality}$$

$$= \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N, \qquad \text{Independence}$$

If we let $Z_n = X_n - \mu$, then

$$\mathbb{E}[e^{s(X_n-\mu)}] = M_{Z_n}(s) = \text{MGF of } Z_n.$$

Hoeffding Lemma

So now we have

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq \left(\frac{\mathbb{E}\left[e^{s(X_n - \mu)}\right]}{e^{s\epsilon}} \right)^N$$

Lemma (Hoeffding Lemma)

If $a \leq X_n \leq b$, then

$$\mathbb{E}\left[e^{s(X_n-\mu)}\right] \leq e^{\frac{s^2(b-a)^2}{8}}$$

This leads to

$$\begin{aligned}\mathbb{P}[\nu - \mu \geq \epsilon] &= \left(\frac{\mathbb{E}\left[e^{s(X_n - \mu)}\right]}{e^{s\epsilon}} \right)^N \\ &\leq \left(\frac{e^{\frac{s^2}{8}}}{e^{s\epsilon}} \right)^N = e^{\frac{s^2 N}{8} - s\epsilon N}, \qquad \forall s > 0.\end{aligned}$$

Minimization

Finally, we arrive at:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{s^2 N}{8} - s\epsilon N}.$$

Since holds for all $s > 0$, in particular it holds for the minimizer:

$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{{s_{\min}}^2 N}{8} - s_{\min}\epsilon N} = \min_{s>0} \left\{ e^{\frac{s^2 N}{8} - s\epsilon N} \right\}$$

Minimizing the exponent gives: $\frac{d}{ds}\left\{\frac{s^2 N}{8} - s\epsilon N\right\} = \frac{sN}{4} - \epsilon N = 0$. So $s = 4\epsilon$.

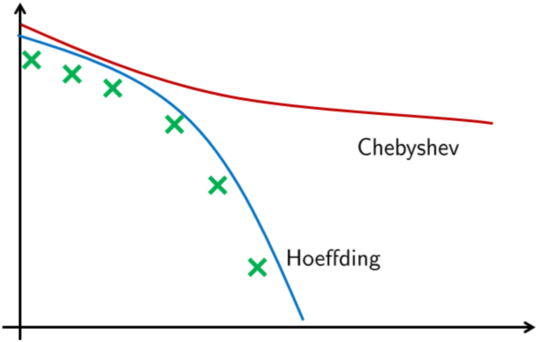
$$\mathbb{P}[\nu - \mu \geq \epsilon] \leq e^{\frac{(4\epsilon)^2 N}{8} - (4\epsilon)\epsilon N} = e^{-2\epsilon^2 N}.$$

Hoeffding Inequality

Theorem (Hoeffding Inequality)

Let $X_1, \ldots, X_N$ be random variables with $0 \leq X_n \leq 1$, then

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$



Compare Hoeffding and Chebyshev

Chebyshev:

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{N\epsilon^2}.$$

Hoeffding:

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 N}.$$

Both are in the form of

$$\mathbb{P}\left[|\nu - \mu| \geq \epsilon\right] \leq \delta.$$

Equivalent to: **For probability at least** $1 - \delta$, we have

$$\mu - \epsilon \leq \nu \leq \mu + \epsilon.$$

Error bar / **Confidence interval** of ν.

$$\delta = \frac{\sigma^2}{N\epsilon^2} \;\Rightarrow\; \epsilon = \frac{\sigma}{\sqrt{\delta N}}$$

$$\delta = 2e^{-2\epsilon^2 N} \;\Rightarrow\; \epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

Example

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \leq \nu \leq \mu + \frac{\sigma}{\sqrt{\delta N}}.$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \leq \nu \leq \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.$$

Example:

- Alex: I have data $X_1, \ldots, X_N$. I want to estimate μ. How many data points N do I need?
- Bob: How much δ can you tolerate?
- Alex: Alright. I only have limited number of data points. How good my estimate is? (ϵ)
- Bob: How many data points N do you have?

Example

Chebyshev: For probability at least $1-\delta$, we have

$$\mu - \frac{\sigma}{\sqrt{\delta N}} \leq \nu \leq \mu + \frac{\sigma}{\sqrt{\delta N}}.$$

Hoeffding: For probability at least $1-\delta$, we have

$$\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \leq \nu \leq \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.$$

Let $\delta = 0.01$, $N = 10000$, $\sigma = 1$.

$$\epsilon = \frac{\sigma}{\sqrt{\delta N}} = 0.1 \qquad\qquad \epsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = 0.016$$

Let $\delta = 0.01$, $\epsilon = 0.01$, $\sigma = 1$.

$$N \geq \frac{\sigma^2}{\epsilon^2\delta} = 1,000,000. \qquad\qquad N \geq \frac{\log\frac{2}{\delta}}{2\epsilon^2} \approx 26,500.$$

Reading List

- Abu-Mustafa, Learning from Data, Chapter 2.
- Martin Wainwright, High Dimensional Statistics, Cambridge University Press 2019. (Chapter 2)
- Cornell Note, https://www.cs.cornell.edu/~sridharan/concentration.pdf
- CMU Note, http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf
- Stanford Note, http://cs229.stanford.edu/extra-notes/hoeffding.pdf