Concentration of Measure Inequalities

Yevgeny Seldin

University of Copenhagen

What can be said about $L(h)$ based on $\hat{L}(h, S_{val})$?

- $\hat{L}(h, S_{val})$ is an unbiased estimate of $L(h)$
- But consider the case $m = 1$:
  - $\hat{L}(h, S_{val}) \in \{0,1\}$ – never close to $L(h)$!

- Being unbiased is neither sufficient, nor necessary

- We need concentration!

Relation to “coin flips” (Bernoulli random variables)

- $Z_i = \ell(h(X_i), Y_i) \in \{0,1\}$
  - Bernoulli random variable, “a coin flip”

- $\mathbb{E}[Z_i] = \mathbb{E}[\ell(h(X_i), Y_i)] = L(h) = p = \mu$
  - The bias of the coin
  - $\mathbb{E}[Z_i] = 1\,\mathbb{P}(Z_i = 1) + 0\,\mathbb{P}(Z_i = 0) = \mathbb{P}(Z_i = 1)$

- $\hat{L}(h, S_{val}) = \frac{1}{n}\sum_{i=1}^{n} Z_i = \hat{p}_n = \hat{\mu}_n$
  - An average of n “coin flips”, the empirical bias

- If $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent identically distributed (i.i.d.), then $Z_1, \dots, Z_n$ are also i.i.d.

- How far can $\hat{\mu}_n$ be from μ?

Frequentist vs. Bayesian reasoning

- Bayesian reasoning
  - Parameters (such as μ) are sampled from an unknown distribution
  - Bayesians start with a prior distribution $\mathbb{P}(\mu = x)$ on the parameters, and, given evidence $(Z_1, \dots, Z_n)$, apply the Bayes rule
    - $\mathbb{P}(\mu = x | Z_1, \dots, Z_n) = \frac{\mathbb{P}(Z_1,\dots,Z_n | \mu = x)\, \mathbb{P}(\mu = x)}{\mathbb{P}(Z_1,\dots,Z_n)}$
  - The probabilities are over observations and parameters (both are random variables)
  - If the prior $\mathbb{P}(\mu = x)$ does not match the reality, the results fall apart

- Frequentist reasoning
  - The parameters (μ) are unknown, but fixed
  - Frequentists bound the probability that the observation $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ deviates strongly from the true value
    - $\mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq \cdots$ or $\mathbb{P}(\hat{\mu}_n - \mu \geq \varepsilon) \leq \cdots$ or $\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon) \leq \cdots$
  - The random variable is $\hat{\mu}_n$, but not μ; and the probability is over $\hat{\mu}_n$, but not μ

Frequentist vs. Bayesian reasoning

ML-A follows the Frequentist Reasoning

- Frequentist reasoning
  - The parameters (μ) are unknown, but fixed
  - Frequentists bound the probability that the observation $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ deviates strongly from the true value
    - $\mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq \cdots$ or $\mathbb{P}(\hat{\mu}_n - \mu \geq \varepsilon) \leq \cdots$ or $\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon) \leq \cdots$
  - The random variable is $\hat{\mu}_n$, but not μ; and the probability is over $\hat{\mu}_n$, but not μ

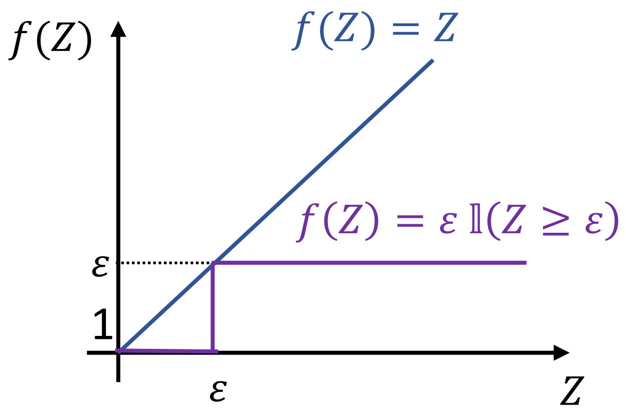
Concentration of Measure Inequalities

Markov's Inequality

- Theorem (Markov's inequality):

  For any non-negative random variable Z and $\varepsilon > 0$

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}[Z]}{\varepsilon}$$



- Proof

  Define $W = \mathbb{I}(Z \geq \varepsilon) = \begin{cases} 1, & \text{If } Z \geq \varepsilon \\ 0, & \text{Otherwise} \end{cases}$, then $W \leq \frac{Z}{\varepsilon}$

  W is a Bernoulli random variable, thus $\mathbb{P}(W = 1) = \mathbb{E}[W]$

$$\mathbb{P}(Z \geq \varepsilon) = \mathbb{P}(W = 1) = \mathbb{E}[W] \leq \mathbb{E}\left[\frac{Z}{\varepsilon}\right] = \frac{\mathbb{E}[Z]}{\varepsilon}$$

Application Example

- Our general worry is that $\hat{L}(h, S) \ll L(h)$

- We want to bound $\mathbb{P}(L(h) - \hat{L}(h, S) \geq \varepsilon)$

- Bound the probability that $\hat{L}(h, S) \leq 0.2$ when $L(h) = 0.6$, meaning that $L(h) - \hat{L}(h, S) \geq 0.4$

Markov: for $Z \geq 0$
$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}[Z]}{\varepsilon}$$

Application Example

- Let $Z_1, \dots, Z_n$ Bernoulli i.i.d.:

$$\begin{aligned}
\mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) &= \mathbb{P}(-\hat{\mu}_n \geq \varepsilon - \mu) \\
&= \mathbb{P}(1 - \hat{\mu}_n \geq \varepsilon + (1 - \mu)) \\
&\leq \frac{\mathbb{E}[1 - \hat{\mu}_n]}{\varepsilon + (1 - \mu)} \\
&= \frac{1 - \mu}{\varepsilon + (1 - \mu)} \leq \frac{1}{\varepsilon + 1}
\end{aligned}$$

- Concentration provided by Markov's inequality does not improve with n
- We used the upper bound $Z_i \leq 1$; we did not use independence

Markov: for $Z \geq 0$
$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}[Z]}{\varepsilon}$$

Chebyshev's Inequality

- Theorem (Chebyshev's inequality)

  For any $\varepsilon > 0$

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) \leq \frac{Var[Z]}{\varepsilon^2}$$

- Proof

$$\begin{aligned}\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) &= \mathbb{P}\left((Z - \mathbb{E}[Z])^2 \geq \varepsilon^2\right) \\ &\leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{\varepsilon^2} \\ &= \frac{Var[Z]}{\varepsilon^2}\end{aligned}$$

Application example

- For $Z_1, \dots, Z_n$ i.i.d.:
  $\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon)$
  $$\leq \frac{Var[\hat{\mu}_n]}{\varepsilon^2}$$
  $$= \frac{Var\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]}{\varepsilon^2}$$
  $$= \frac{Var[Z_1]}{\boldsymbol{n}\varepsilon^2}$$

- Concentration provided by Chebyshev's inequality improves at the rate of $\frac{1}{n}$

Chebyshev:

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) \leq \frac{Var[Z]}{\varepsilon^2}$$

For i.i.d. r.v. $Z_1, \dots, Z_n$ and const. c:

$$Var\left[\sum_{i=1}^{n} Z_i\right] = \sum_{i=1}^{n} Var[Z_i]$$

$$Var[cZ] = c^2 Var[Z]$$

Hoeffding's inequality

- Theorem (Hoeffding's inequality)

  Let $Z_1, \dots, Z_n$ be i.i.d., $Z_i \in [0,1]$, and $\mathbb{E}[Z_i] = \mu$, then for any $\varepsilon > 0$:

$$\left.\begin{aligned} \mathbb{P}(\hat{\mu}_n - \mu \geq \varepsilon) \leq e^{-2n\varepsilon^2} \\ \mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq e^{-2n\varepsilon^2} \end{aligned}\right\} \text{One-sided Hoeffding's inequalities}$$

- Corollary (two-sided Hoeffding's inequality)

$$\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon) \leq \mathbb{P}(\hat{\mu}_n - \mu \geq \varepsilon) + \mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

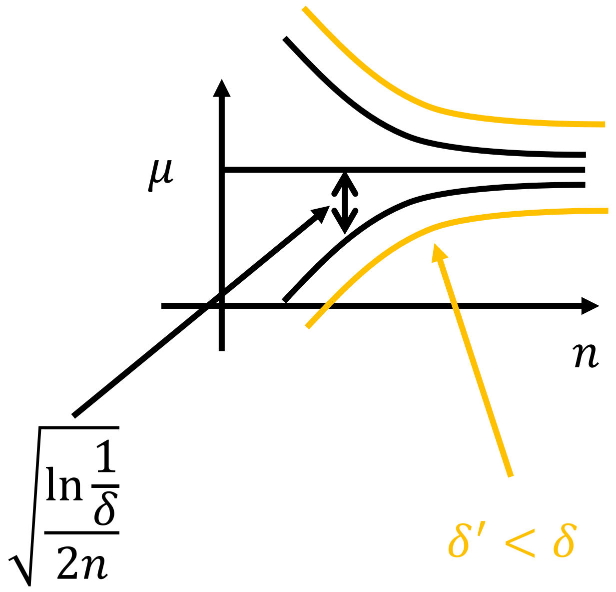
Union bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

- By Hoeffding, $\hat{\mu}_n$ converges to μ exponentially fast in n!

Understanding the bound

- $\mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq e^{-2n\varepsilon^2} = \delta$
- $\Rightarrow \varepsilon = \sqrt{\frac{\ln\frac{1}{\delta}}{n}}$ (For two-sided replace $\frac{1}{\delta}$ by $\frac{2}{\delta}$)
- $\mathbb{P}\left(\mu - \hat{\mu}_n \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$
- $\mathbb{P}\left(\mu - \hat{\mu}_n \leq \underbrace{\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}}_{\text{Precision}}\right) \geq \underbrace{1 - \delta}_{\text{Confidence}}$

- **Probably Approximately Correct (PAC) learning framework**
  - With probability at least $1 - \delta$, $\hat{\mu}_n$ is approximately equal to μ
- The probability is over $\hat{\mu}_n$ (the random variable), not over μ (deterministic)!



- $\delta = 0$
  $\mu \leq \hat{\mu}_n + \infty$
- $\delta = 1$
  $\mu \leq \hat{\mu}_n$

$\delta = e^{-2n\varepsilon^2}$ - confidence

$\varepsilon = \sqrt{\frac{\ln\frac{1}{\delta}}{n}}$ - precision

$n = \frac{\ln\frac{1}{\delta}}{2\varepsilon^2}$ - sample size

Different ways of using the bound

- $\mathbb{P}(\mu - \hat{\mu}_n \geq \varepsilon) \leq e^{-2n\varepsilon^2} = \delta$

- We can fix any two parameters and get the value for the third one
  - δ: What is the probability that $\hat{\mu}_n$ underestimates μ by more than ε given that we have n samples? (n and ε are fixed and dictate δ)
  - ε: What is the maximal underestimation of μ by $\hat{\mu}_n$ that can be guaranteed with probability at least $1 - \delta$ given a sample of size n? (n and δ are fixed and dictate ε)
  - n: How many samples do we need in order to guarantee that $\hat{\mu}_n$ does not underestimate μ by more than ε with probability at least $1 - \delta$? (ε and δ are fixed and dictate n)

Proof of Hoeffding's inequality

- The inequality: $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}$
- Hoeffing's Lemma: For r.v. $Z \in [0,1]$ and $\lambda > 0$: $\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq e^{\frac{\lambda^2}{8}}$
- Proof of Hoeffding's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \geq \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq n\varepsilon\right)$$

Chernoff's bounding technique ($\lambda > 0$) :
($x \geq y \iff e^{\lambda x} \geq e^{\lambda y}$)

$$= \mathbb{P}\left(e^{\lambda \sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])} \geq e^{\lambda n\varepsilon}\right)$$

Markov's inequality:

$$\leq \frac{\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])}\right]}{e^{\lambda n\varepsilon}}$$

$$= e^{-\lambda n\varepsilon}\,\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda(Z_i - \mathbb{E}[Z_i])}\right]$$

Independence ($\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$):

$$= e^{-\lambda n\varepsilon} \prod_{i}^{n} \mathbb{E}\left[e^{\lambda(Z_i - \mathbb{E}[Z_i])}\right]$$

Hoeffding's lemma:

$$\leq e^{-\lambda n\varepsilon} \prod_{i=1}^{n} e^{\frac{\lambda^2}{8}} = e^{-n\left(\lambda\varepsilon - \frac{\lambda^2}{8}\right)}$$

Minimize w.r.t. λ ($\lambda^* = 4\varepsilon$):

$$\leq e^{-2n\varepsilon^2}$$

λ^* is independent of $Z_1, \dots, Z_n$!

Hoeffding:
$$\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

The importance of independence

- Construct an example of **dependent** identically distributed random variables $Z_1, \dots, Z_n$, such that $Z_i \in \{0,1\}$, and $\mathbb{E}[Z_i] = \mu$, and for any n $\mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \frac{1}{2}\right) = 1$

Summary

- Means of **independent** random variables converge to their expectation

- Without independence this is not necessarily the case (home assignment)

- Hoeffding: $\mathbb{P}\left(\mu - \hat{\mu}_n \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$

- Probably Approximately Correct (PAC) learning