Concentration of Measure: Markov's, Chebyshev's and Hoeffding's Inequalities

Mohammad Sadegh Talebi
Department of Computer Science

(Partially based on Yevgeny Seldin's Slides)

Outline

Motivation: Supervised Learning

- A finite hypothesis class $\mathcal{H}$
- A sample $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$, with elements independently drawn from a fixed (but unknown) distribution.
- $\ell(h(X), Y)$ is the loss of $h \in \mathcal{H}$ on (X, Y)
- $L(h) = \mathbb{E}[\ell(h(X), Y)]$ is unknown.
- Empirical loss of h:

$$\widehat{L}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

- $\widehat{L}(h, S)$ is an unbiased estimator of $L(h)$: $\mathbb{E}[\widehat{L}(h, S)] = L(h)$

Motivation: Supervised Learning

- A finite hypothesis class $\mathcal{H}$
- A sample $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$, with elements independently drawn from a fixed (but unknown) distribution.
- $\ell(h(X), Y)$ is the loss of $h \in \mathcal{H}$ on (X, Y)
- $L(h) = \mathbb{E}[\ell(h(X), Y)]$ is unknown.
- Empirical loss of h:

$$\widehat{L}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

- $\widehat{L}(h, S)$ is an unbiased estimator of $L(h)$: $\mathbb{E}[\widehat{L}(h, S)] = L(h)$

What can be said about $L(h)$?

Motivation: Bernoulli Trials

Assume a certain treatment is successful with probability p, and failing otherwise.

- p is unknown, but is fixed for all subject patients.
- Let $X_1, \dots, X_n$ be the realized outcomes on n patients.
- Sample mean:

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $\mathbb{E}[\widehat{p}_n] = p$

Motivation: Bernoulli Trials

Assume a certain treatment is successful with probability p, and failing otherwise.

- p is unknown, but is fixed for all subject patients.
- Let $X_1, \ldots, X_n$ be the realized outcomes on n patients.
- Sample mean:

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $\mathbb{E}[\widehat{p}_n] = p$

How close is $\widehat{p}_n$ to p for a given n?

Motivation: Bernoulli Trials

Assume a certain treatment is successful with probability p, and failing otherwise.

- p is unknown, but is fixed for all subject patients.
- Let $X_1, \dots, X_n$ be the realized outcomes on n patients.
- Sample mean:

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $\mathbb{E}[\widehat{p}_n] = p$

How close is $\widehat{p}_n$ to p for a given n?
How much patients do we need to try so that $\widehat{p}_n$ is not farther than p by some ε?

Outline

Recap: Independence

Independence

- Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Recap: Independence

Independence

- Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
- Two random variables (r.v.'s) X and Y are independent if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \qquad \forall x, y$$

Recap: Independence

Independence

- Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
- Two random variables (r.v.'s) X and Y are independent if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \qquad \forall x, y$$

- X and Y are independent $\Longrightarrow \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
- A collection of r.v.'s $X_1, \ldots, X_n$ are independent if every pair in the collection is independent.
- If $X_1, \ldots, X_n$ are independent and have identical distribution (i.e., $F_{X_1} = \ldots = F_{X_n}$), then they are called independent identically distributed (i.i.d.) r.v.'s.

Recap: Asymptotic Convergence Results

Consider i.i.d. r.v.'s $X_1, \dots, X_n$, with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$.

- Sample mean:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $\overline{X}_n$ is an unbiased estimate of μ:

$$\mathbb{E}[\overline{X}_n] = \mu$$

- By the Central Limit Theorem (CLT), $\overline{X}_n$ asymptotically converges to μ in distribution:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\text{distribution}}_{n\to\infty} \mathcal{N}(0, \sigma^2)$$

Recap: Asymptotic Convergence Results

Consider i.i.d. r.v.'s $X_1, \dots, X_n$, with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$.

- Sample mean:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $\overline{X}_n$ is an unbiased estimate of μ:

$$\mathbb{E}[\overline{X}_n] = \mu$$

- By the Central Limit Theorem (CLT), $\overline{X}_n$ asymptotically converges to μ in distribution:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\text{distribution}}_{n\to\infty} \mathcal{N}(0, \sigma^2)$$

- By the Strong Law of Large Numbers (SLLN), $\overline{X}_n$ asymptotically converges to μ almost surely:

$$\overline{X}_n \xrightarrow{\text{almost surely}}_{n\to\infty} \mu$$

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

- How close is $\overline{X}_n$ to μ when n is *finite* (and not necessarily large)?

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

- How close is $\overline{X}_n$ to μ when n is *finite* (and not necessarily large)?
- How many samples n do we need so that $\overline{X}_n$ does not exceed μ by no more than a specified ε?

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

- How close is $\overline{X}_n$ to μ when n is *finite* (and not necessarily large)?
- How many samples n do we need so that $\overline{X}_n$ does not exceed μ by no more than a specified ε?
- Can we derive a data-dependent rule to stop collecting samples for to reach a given accuracy?

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

- How close is $\overline{X}_n$ to μ when n is *finite* (and not necessarily large)?
- How many samples n do we need so that $\overline{X}_n$ does not exceed μ by no more than a specified ε?
- Can we derive a data-dependent rule to stop collecting samples for to reach a given accuracy?

Concentration inequalities (or tail inequalities) are tools from probability to study the deviation of a r.v. from its mean *non-asymptotically* (i.e., for finite n).

What About Non-asymptotic Regime?

CLT and SLLN provide nice results but asymptotically (as $n \to \infty$).

- How close is $\overline{X}_n$ to μ when n is *finite* (and not necessarily large)?
- How many samples n do we need so that $\overline{X}_n$ does not exceed μ by no more than a specified ε?
- Can we derive a data-dependent rule to stop collecting samples for to reach a given accuracy?

Concentration inequalities (or tail inequalities) are tools from probability to study the deviation of a r.v. from its mean *non-asymptotically* (i.e., for finite n).

Most often, concentration inequalities provide upper bounds on

$$\mathbb{P}\Big(f(X_1, \ldots, X_n) > \varepsilon\Big)$$

for a finite set of (conditionally) independent r.v.'s $X_1, \ldots, X_n$.

Outline

Markov's Inequality

Theorem (Markov's Inequality)

Suppose X is a non-negative r.v. Then for all $\varepsilon > 0$,

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

Markov's Inequality

Theorem (Markov's Inequality)

Suppose X is a non-negative r.v. Then for all $\varepsilon > 0$,

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

Proof. Define $Y = \mathbb{I}(X \geq \varepsilon) = \begin{cases} 1 & X \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$

Markov's Inequality

Theorem (Markov's Inequality)

Suppose X is a non-negative r.v. Then for all $\varepsilon > 0$,

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

Proof. Define $Y = \mathbb{I}(X \geq \varepsilon) = \begin{cases} 1 & X \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$. Y is a Bernoulli r.v. and $\mathbb{E}[Y] = \mathbb{P}(Y = 1)$. Thus,

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

Markov's Inequality: Examples

- $\mathbb{P}(X \geq \alpha \mathbb{E}[X]) \leq \alpha^{-1}$ valid for $X > 0$.

Markov's Inequality: Examples

- $\mathbb{P}(X \geq \alpha\mathbb{E}[X]) \leq \alpha^{-1}$ valid for $X > 0$.
- $X \in \{0, 1, \dots\}$. Then, $\mathbb{P}(X \neq 0) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$.

Markov's Inequality: Examples

- $\mathbb{P}(X \geq \alpha\mathbb{E}[X]) \leq \alpha^{-1}$ valid for $X > 0$.
- $X \in \{0, 1, \ldots\}$. Then, $\mathbb{P}(X \neq 0) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$.
- $\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^X \geq e^\varepsilon) \leq e^{-\varepsilon}\mathbb{E}[e^X]$

Markov's Inequality: Examples

- $\mathbb{P}(X \geq \alpha\mathbb{E}[X]) \leq \alpha^{-1}$ valid for $X > 0$.
- $X \in \{0, 1, \ldots\}$. Then, $\mathbb{P}(X \neq 0) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$.
- $\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^X \geq e^\varepsilon) \leq e^{-\varepsilon}\mathbb{E}[e^X]$
- For $X_1, \ldots, X_n$ i.i.d. with range $[0, 1]$ with mean μ,

$$\mathbb{P}(\mu - \overline{X}_n > \varepsilon) \leq \frac{1-\mu}{\varepsilon + 1 - \mu} \leq \frac{1}{\varepsilon + 1}$$

(Bad news: the upper bound does not decay with n)

Markov's Inequality: Examples

- $\mathbb{P}(X \geq \alpha\mathbb{E}[X]) \leq \alpha^{-1}$ valid for $X > 0$.
- $X \in \{0, 1, \ldots\}$. Then, $\mathbb{P}(X \neq 0) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$.
- $\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^X \geq e^\varepsilon) \leq e^{-\varepsilon}\mathbb{E}[e^X]$
- For $X_1, \ldots, X_n$ i.i.d. with range $[0, 1]$ with mean μ,

$$\mathbb{P}(\mu - \overline{X}_n > \varepsilon) \leq \frac{1-\mu}{\varepsilon + 1 - \mu} \leq \frac{1}{\varepsilon + 1}$$

(Bad news: the upper bound does not decay with n)

Theorem (Markov's Inequality – Extended)

Suppose X is a non-negative r.v. and f is a monotonically increasing function. Then for all $\varepsilon > 0$,

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

Outline

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

For all $\varepsilon > 0$,

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

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

For all $\varepsilon > 0$,

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

Proof.

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) &= \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq \varepsilon^2) \\ &\leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{\varepsilon^2} \qquad \text{(Markov's inequality)} \\ &= \frac{\mathrm{Var}(X)}{\varepsilon^2} \end{aligned}$$

Chebyshev's Inequality: Examples

Example 1: A fair coin is tossed 20 times. Let $X_1, \ldots, X_{20}$ be the realized outcomes. Then,

$$\mathbb{P}(|\overline{X}_{20} - \mu| \geq 0.2) \leq \frac{\mathrm{Var}(\overline{X}_{20})}{0.2^2} = \frac{\frac{1}{20}\mathrm{Var}(X_1)}{0.04} = \frac{\frac{1}{20} \cdot \frac{1}{4}}{0.04}$$

Chebyshev's Inequality: Examples

Example 1: A fair coin is tossed 20 times. Let $X_1, \dots, X_{20}$ be the realized outcomes. Then,

$$\mathbb{P}(|\overline{X}_{20} - \mu| \geq 0.2) \leq \frac{\mathrm{Var}(\overline{X}_{20})}{0.2^2} = \frac{\frac{1}{20}\mathrm{Var}(X_1)}{0.04} = \frac{\frac{1}{20} \cdot \frac{1}{4}}{0.04}$$

Example 2: A fair die is rolled 60 times. Let $X_1, \dots, X_{20}$ be the realized outcomes. Upper bound

$$\mathbb{P}(|\sum_i X_i - 210| \geq 20)$$

Chebyshev's Inequality: Examples

Example 1: A fair coin is tossed 20 times. Let $X_1, \dots, X_{20}$ be the realized outcomes. Then,

$$\mathbb{P}(|\overline{X}_{20} - \mu| \geq 0.2) \leq \frac{\mathrm{Var}(\overline{X}_{20})}{0.2^2} = \frac{\frac{1}{20}\mathrm{Var}(X_1)}{0.04} = \frac{\frac{1}{20} \cdot \frac{1}{4}}{0.04}$$

Example 2: A fair die is rolled 60 times. Let $X_1, \dots, X_{20}$ be the realized outcomes. Upper bound

$$\mathbb{P}(|\sum_i X_i - 210| \geq 20)$$

- $\mathbb{E}[X_1] = \frac{7}{2}$. Hence, $\mathbb{E}[\sum_i X_i] = 210$.
- $\mathrm{Var}(\sum_i X_i) = \frac{35}{12}$
- $\mathbb{P}(|\sum_i X_i - 210| \geq 20) \leq \frac{175}{20^2}$

Chebyshev's Inequality

Theorem (Chebyshev's Inequality for I.I.D. Variables)

Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s. Then, for all $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$$

Chebyshev's Inequality

Theorem (Chebyshev's Inequality for I.I.D. Variables)

Let $X_1, \dots, X_n$ be i.i.d. r.v.'s. Then, for all $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$$

Chebyshev's inequality provides a result that decays at a rate of $\frac{1}{n}$.

Outline

Hoeffding's Inequality

Theorem (Hoeffding's Inequality)

Let $X_1, \dots, X_n$ *be* *independent* *r.v.'s with support* $[0,1]$*. Then, for all* $\varepsilon > 0$*,*

$$\textit{(i)} \quad \mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i \right] \geq \varepsilon \right) \leq e^{-2\varepsilon^2/n}$$

$$\textit{(ii)} \quad \mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i \right] \leq -\varepsilon \right) \leq e^{-2\varepsilon^2/n}$$

Hoeffding's Inequality

Theorem (Hoeffding's Inequality)

Let $X_1, \dots, X_n$ be independent r.v.'s with support $[0,1]$. Then, for all $\varepsilon > 0$,

$$\text{(i)} \quad \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i \right] \geq \varepsilon \right) \leq e^{-2\varepsilon^2/n}$$

$$\text{(ii)} \quad \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i \right] \leq -\varepsilon \right) \leq e^{-2\varepsilon^2/n}$$

- Using a union bound:

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i \right] \right| \geq \varepsilon \right) \leq 2e^{-2\varepsilon^2/n}$$

Hoeffding's Inequality

Theorem (Hoeffding's Inequality)

Let $X_1, \dots, X_n$ be independent r.v.'s with support $[0, 1]$. Then, for all $\varepsilon > 0$,

$$\text{(i)} \quad \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geq \varepsilon\right) \leq e^{-2\varepsilon^2/n}$$

$$\text{(ii)} \quad \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \leq -\varepsilon\right) \leq e^{-2\varepsilon^2/n}$$

- Using a union bound:

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right| \geq \varepsilon\right) \leq 2e^{-2\varepsilon^2/n}$$

- Also, $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right]\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$
- If X_i's are i.i.d. with mean μ: $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \varepsilon\right) \geq 2e^{-2n\varepsilon^2}$.
  Hoeffding's bound decays exponentially fast in n.

Hoeffding's Inequality: Alternative Form

If X_i's are i.i.d. with mean μ: $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \varepsilon\right) \leq \underbrace{e^{-2n\varepsilon^2}}_{=\delta}$

Solving $\delta = e^{-2n\varepsilon^2}$ for ε yields $\varepsilon = \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$ Hence,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

or alternatively:

$$\frac{1}{n}\sum_{i=1}^n X_i - \mu \leq \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)} \qquad \text{with probability at least } 1-\delta$$

For $X_1, \ldots, X_n$ i.i.d., it holds that for all $\delta \in (0,1)$,

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)} \qquad \text{with probability at least } 1-\delta$$

Hoefdding's Inequality: Proof

The proof of Hoeffding's inequality relies on the following lemma:

Lemma (Hoeffding's Lemma)

Let X be a r.v. supported on $[a, b]$. Then,

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}, \quad \forall\lambda \in \mathbb{R}$$

For proof, see Yevgeny's lecture notes.

Hoeffding's Inequality: Proof

Proof of (i). Let $Z := \sum_{i=1}^n X_i - \mathbb{E}\Big[\sum_{i=1}^n X_i\Big]$.

Hoeffding's Inequality: Proof

Proof of (i). Let $Z := \sum_{i=1}^n X_i - \mathbb{E}\Big[\sum_{i=1}^n X_i\Big]$. Let $\lambda > 0$. Then

$$
\begin{aligned}
\mathbb{P}(Z \geq \varepsilon) &= \mathbb{P}(e^{\lambda Z} \geq e^{\lambda\varepsilon}) \\
&\leq e^{-\lambda\varepsilon}\mathbb{E}\big[e^{\lambda Z}\big] \qquad \textcolor{blue}{\text{(Markov's inequality)}} \\
&= e^{-\lambda\varepsilon}\mathbb{E}\Big[e^{\lambda\big(\sum_{i=1}^n X_i - \mathbb{E}\big[\sum_{i=1}^n X_i\big]\big)}\Big] \\
&= e^{-\lambda\varepsilon}\prod_{i=1}^n \mathbb{E}\big[e^{\lambda(X_i - \mathbb{E}[X_i])}\big] \qquad \textcolor{blue}{\text{(independence)}} \\
&\leq e^{-\lambda\varepsilon}\prod_{i=1}^n \mathbb{E}\big[e^{\lambda^2/8}\big] = e^{-\lambda\varepsilon + \frac{n\lambda^2}{8}} \qquad \textcolor{blue}{\text{(Hoeffding's lemma)}}
\end{aligned}
$$

Hoeffding's Inequality: Proof

Proof of (i). Let $Z := \sum_{i=1}^n X_i - \mathbb{E}\Big[\sum_{i=1}^n X_i\Big]$. Let $\lambda > 0$. Then

$$
\begin{aligned}
\mathbb{P}(Z \ge \varepsilon) &= \mathbb{P}(e^{\lambda Z} \ge e^{\lambda\varepsilon}) \\
&\le e^{-\lambda\varepsilon}\mathbb{E}\big[e^{\lambda Z}\big] && \text{(Markov's inequality)} \\
&= e^{-\lambda\varepsilon}\mathbb{E}\Big[e^{\lambda\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)}\Big] \\
&= e^{-\lambda\varepsilon}\prod_{i=1}^n \mathbb{E}\big[e^{\lambda(X_i - \mathbb{E}[X_i])}\big] && \text{(independence)} \\
&\le e^{-\lambda\varepsilon}\prod_{i=1}^n \mathbb{E}\big[e^{\lambda^2/8}\big] = e^{-\lambda\varepsilon + \frac{n\lambda^2}{8}} && \text{(Hoeffding's lemma)}
\end{aligned}
$$

This is valid for any $\lambda > 0$, hence

$$
\mathbb{P}(Z \ge \varepsilon) \le \min_{\lambda>0} e^{-\lambda\varepsilon + \frac{n\lambda^2}{8}} = e^{\min_{\lambda>0}\left(-\lambda\varepsilon + \frac{n\lambda^2}{8}\right)}
$$

Hence, the *smallest* (hence, best) bound, attained at $\lambda = 4\varepsilon n^{-1}$, is:

$$
\mathbb{P}(Z \ge \varepsilon) \le e^{-2\varepsilon^2/n}
$$

Hoeffding's Inequality: Generic Ranges

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ be independent r.v.'s such that $X_i \in [a_i, b_i]$ almost surely, that is $\mathbb{P}(X_i \in [a_i, b_i]) = 1$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\Big[\sum_{i=1}^n X_i\Big] \geq \varepsilon\Big) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\Big[\sum_{i=1}^n X_i\Big] \leq -\varepsilon\Big) \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Hoeffding's Inequality: Examples

Example 1 (revisited): A fair coin is tossed 20 times. Let $X_1, \ldots, X_{20}$ be the realized outcomes.

By Hoeffding's inequality,

$$\mathbb{P}\left(\left| \frac{1}{20} \sum_{i=1}^{20} X_i - \frac{1}{2} \right| \geq 0.1 \right) \leq 2e^{-2 \cdot 20 \cdot 0.1^2}$$

Hoeffding's Inequality: Examples

Example 1 (revisited): A fair coin is tossed 20 times. Let $X_1, \dots, X_{20}$ be the realized outcomes.

By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{20}\sum_{i=1}^{20} X_i - \frac{1}{2}\right| \geq 0.1\right) \leq 2e^{-2\cdot 20\cdot 0.1^2}$$

Compare it to the result from Chebyshev's inequality.

Hoeffding's Inequality: Sub-Gaussian Case

Sub-Gaussian Random Variable

A r.v. X is said to be R-sub-Gaussian if

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le e^{\frac{\lambda^2 R^2}{2}}, \quad \forall \lambda \in \mathbb{R}$$

- A r.v. with range $[a, b]$ is sub-Gaussian with $R = \frac{b-a}{2}$ (by Hoeffding's Lemma)
- A Gaussian r.v. is sub-Gaussian with $R = \sigma$ (why?).
- Intuitively, the tail of a sub-Gaussian r.v. decays at least as fast as that of a Gaussian.

Hoeffding's Inequality: Sub-Gaussian Case

Sub-Gaussian Random Variable

A r.v. X is said to be R-sub-Gaussian if

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le e^{\frac{\lambda^2 R^2}{2}}, \quad \forall \lambda \in \mathbb{R}$$

- A r.v. with range $[a, b]$ is sub-Gaussian with $R = \frac{b-a}{2}$ (by Hoeffding's Lemma)
- A Gaussian r.v. is sub-Gaussian with $R = \sigma$ (why?).
- Intuitively, the tail of a sub-Gaussian r.v. decays at least as fast as that of a Gaussian.

Theorem (Hoeffding's Inequality: Sub-Gaussian Case)

Let $X_1, \dots, X_n$ be i.i.d. R-sub-Gaussian r.v.'s with mean μ. Then, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \le R\sqrt{\frac{2}{n}\log\left(\frac{2}{\delta}\right)}$$

Outline

Confidence Interval

Definition

Consider $X_1, \ldots, X_n$ be sampled from some distribution ν, and let θ be a parameter of ν (e.g., mean, variance).

A $(1-\delta)$-confidence interval for θ is a function

$$\texttt{CI}(X_1, \ldots, X_n, \delta) \subset \mathbb{R}$$

if $\mathbb{P}(\mu \in \texttt{CI}) \geq 1 - \delta$

Confidence Interval

Definition

Consider $X_1, \ldots, X_n$ be sampled from some distribution ν, and let θ be a parameter of ν (e.g., mean, variance).

A $(1-\delta)$-confidence interval for θ is a function

$$\texttt{CI}(X_1, \ldots, X_n, \delta) \subset \mathbb{R}$$

if $\mathbb{P}(\mu \in \texttt{CI}) \geq 1 - \delta$

- $1 - \delta$ is often called the confidence level.
- $\texttt{CI}$ traps the (unknown) parameter θ with probability at least $1 - \delta$.
- $\texttt{CI}$ is a function of samples $X_1, \ldots, X_n$, hence a random interval.
- Confidence intervals act as certificate for the corresponding point estimates.

Confidence Interval

Definition

Consider $X_1, \ldots, X_n$ be sampled from some distribution ν, and let θ be a parameter of ν (e.g., mean, variance).

A $(1-\delta)$-confidence interval for θ is a function

$$\texttt{CI}(X_1, \ldots, X_n, \delta) \subset \mathbb{R}$$

if $\mathbb{P}(\mu \in \texttt{CI}) \geq 1 - \delta$

Confidence Interval

Definition

Consider $X_1, \ldots, X_n$ be sampled from some distribution ν, and let θ be a parameter of ν (e.g., mean, variance).

A $(1-\delta)$-confidence interval for θ is a function

$$\mathtt{CI}(X_1, \ldots, X_n, \delta) \subset \mathbb{R}$$

if $\mathbb{P}(\mu \in \mathtt{CI}) \geq 1 - \delta$

Example: $X_1, \ldots, X_n$ are drawn from a Bernoulli with mean μ.

- Sample mean $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
- A $(1-\delta)$-CI for $\theta = \mu$ is:

$$\mathtt{CI}(X_1, \ldots, X_n, \delta) = \left[\overline{X_n} - d, \, \overline{X_n} + d\right]$$

for some d determined by $X_1, \ldots, X_n$ and δ.

CI using Hoeffding's Inequality

How to construct confidence intervals?

CI using Hoeffding's Inequality

How to construct confidence intervals?

Let $X_1, \dots, X_n$ be i.i.d. samples from ν with mean μ and support $[0, 1]$.
Define

$$\mathtt{CI} = \left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)},\ \overline{X}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\right] .$$

By Hoeffding's inequality, $\mathbb{P}(\mu \in \mathtt{CI}) \geq 1 - \delta$

CI using Hoeffding's Inequality

How to construct confidence intervals?

Let $X_1, \ldots, X_n$ be i.i.d. samples from ν with mean μ and support $[0, 1]$.
Define

$$\mathtt{CI} = \left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)},\ \overline{X}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\right].$$

By Hoeffding's inequality, $\mathbb{P}(\mu \in \mathtt{CI}) \geq 1 - \delta \implies \mathtt{CI}$ is a $(1-\delta)$-CI for μ.

- $\mathtt{CI}$ above is a certificate for the point estimate $\overline{X}_n$ of μ.

CI using Hoeffding's Inequality

How to construct confidence intervals?

Let $X_1, \dots, X_n$ be i.i.d. samples from ν with mean μ and support $[0,1]$. Define

$$\mathtt{CI} = \left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)},\ \overline{X}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\right].$$

By Hoeffding's inequality, $\mathbb{P}(\mu \in \mathtt{CI}) \geq 1-\delta \implies \mathtt{CI}$ is a $(1-\delta)$-CI for μ.

- $\mathtt{CI}$ above is a certificate for the point estimate $\overline{X}_n$ of μ.
- Since $\mu \in [0,1]$,

$$\mathtt{CI} = \left[\max\left(\overline{X}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}, 0\right),\ \min\left(\overline{X}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}, 1\right)\right]$$

Example: CI for Bernoulli

Example: Consider samples collected in S1.csv. If they are independent samples from a Bernoulli with mean μ, then:

(a) Sample mean: $\overline{X}_n = 0.43$

Example: CI for Bernoulli

Example: Consider samples collected in S1.csv. If they are independent samples from a Bernoulli with mean μ, then:

(a) Sample mean: $\overline{X}_n = 0.43$

(b) A 0.99-CI for μ:

$$\texttt{CI}_1 = [0.267\,, 0.593]$$

Example: CI for Bernoulli

Example: Consider samples collected in S1.csv. If they are independent samples from a Bernoulli with mean μ, then:

(a) Sample mean: $\overline{X}_n = 0.43$

(b) A 0.99-$\mathtt{CI}$ for μ:

$$\mathtt{CI}_1 = [0.267\,, 0.593]$$

(c) A 0.9-$\mathtt{CI}$ for μ:

$$\mathtt{CI}_2 = [0.307\,, 0.553]$$

Example: CI for Bernoulli

Example: Consider samples collected in S1.csv. If they are independent samples from a Bernoulli with mean μ, then:

(a) Sample mean: $\overline{X}_n = 0.43$

(b) A 0.99-CI for μ:

$$\texttt{CI}_1 = [0.267\,, 0.593]$$

(c) A 0.9-CI for μ:

$$\texttt{CI}_2 = [0.307\,, 0.553]$$

(d) Repeat (a)-(c) with S2.csv and compare the results.

Example: CI for Bernoulli

Example: Consider samples collected in S1.csv. If they are independent samples from a Bernoulli with mean μ, then:

(a) Sample mean: $\overline{X}_n = 0.43$

(b) A 0.99-$\mathtt{CI}$ for μ:

$$\mathtt{CI}_1 = [0.267\,, 0.593]$$

(c) A 0.9-$\mathtt{CI}$ for μ:

$$\mathtt{CI}_2 = [0.307\,, 0.553]$$

(d) Repeat (a)-(c) with S2.csv and compare the results.

For example, $\mathtt{CI}_1$ tells us that $\mathbb{P}(0.267 \le \mu \le 0.593) \ge 0.99$. Can we obtain the exact value of μ using S1 or S2?

CI using Hoeffding's Inequality

Let $X_1, \ldots, X_n$ be independent samples with a common mean μ and common range $[b, a]$. Then, for any $\delta \in (0, 1)$,

$$\texttt{CI} = \left[\max\left(\overline{X}_n - (b-a)\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}, a\right),\ \min\left(\overline{X}_n + (b-a)\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}, b\right)\right]$$

is a $(1-\delta)$-CI for μ.

- Note that samples could have different distributions.
- Result is valid for a fixed n that does not depend on data (e.g., n is not determined by a stopping rule).
- Extension beyond fixed n using time-uniform confidence intervals.