Lecture 14: Azuma-Hoeffding Inequality Proof

Recall: Azuma's Inequality

Theorem (Azuma's Inequality)

Let Ω be a sample space and p be a probability distribution over Ω. Let $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration. Let $(\mathbb{F}_0, \mathbb{F}_1, \ldots, \mathbb{F}_n)$ be a martingale with respect to the filtration above. Suppose, for all $1 \leqslant i \leqslant n$, there exists c_i such that, for all $x \in \Omega$, we have

$$c_i \geqslant \max_{y \in \mathcal{F}_{i-1}(x)} \mathbb{F}_i(y) - \min_{y \in \mathcal{F}_{i-1}(x)} \mathbb{F}_i(y)$$

Then, the following bound holds

$$\mathbb{P}\left[\mathbb{F}_n - \mathbb{F}_0 \geqslant E\right] \leqslant \exp\left(-2E^2 \Big/ \sum_{i=1}^{n} c_i^2\right)$$

Proof I

- Let $(\Delta\mathbb{F}_1, \Delta\mathbb{F}_2, \ldots, \Delta\mathbb{F}_n)$ be the corresponding martingale difference sequence. That is, we define $\Delta\mathbb{F}_i = \mathbb{F}_i - \mathbb{F}_{i-1}$, for $1 \leqslant i \leqslant n$. Since, this is a martingale difference sequence, we have the following guarantee for all $1 \leqslant i \leqslant n$.

$$\mathbb{E}\left[\Delta\mathbb{F}_i | \mathcal{F}_{i-1}\right] = 0$$

- Note that the property of c_i can be written as follows (by subtracting $\mathbb{F}_{i-1}(x)$ from both the terms)

$$c_i \geqslant \max_{y \in \mathcal{F}_{i-1}(x)} \Delta\mathbb{F}_i(y) - \min_{y \in \mathcal{F}_{i-1}(x)} \Delta\mathbb{F}_i(y)$$

- Azuma's inequality is equivalent to proving

$$\mathbb{P}\left[\sum_{i=1}^{n} \Delta\mathbb{F}_i \geqslant E\right] \leqslant \exp\left(-2E^2 / \sum_{i=1}^{n} c_i^2\right)$$

Proof II

- Similar to the technique of proving Chernoff bound, we conclude that, for all $h > 0$, the following is true

$$\mathbb{P}\left[\sum_{i=1}^{n} \Delta\mathbb{F}_i \geqslant E\right] \leqslant \frac{\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\right]}{\exp(hE)}$$

- Our effort now is to upper-bound the expected value

$$\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\right]$$

Proof III

- Consider the following set of manipulations

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\Bigg|\mathcal{F}_{n-1}\right]\right]\\
=&\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n-1}\Delta\mathbb{F}_i\right)\mathbb{E}\left[\exp\left(h\Delta\mathbb{F}_n\right)\middle|\mathcal{F}_{n-1}\right]\right]
\end{aligned}
$$

The last equality is because $\exp\left(h\sum_{i=1}^{n-1}\Delta\mathbb{F}_i\right)$ is $\mathcal{F}_{n-1}$ measurable.

Proof IV

- We can apply Hoeffding's Lemma to upper bound $\mathbb{E}\left[\exp\left(h\Delta\mathbb{F}_n\right)\middle|\mathcal{F}_{n-1}\right]$ as follows

$$\mathbb{E}\left[\exp\left(h\Delta\mathbb{F}_n\right)\middle|\mathcal{F}_{n-1}\right] \leqslant \exp\left(\frac{h^2}{8}c_n^2\right)$$

- So, we obtain that

$$\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\right] \leqslant \exp\left(\frac{h^2}{8}c_n^2\right)\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n-1}\Delta\mathbb{F}_i\right)\right]$$

- Repeatedly applying the bound to the last $\Delta\mathbb{F}_i$, we get

$$\mathbb{E}\left[\exp\left(h\sum_{i=1}^{n}\Delta\mathbb{F}_i\right)\right] \leqslant \exp\left(\frac{h^2}{8}\sum_{i=1}^{n}c_i^2\right)$$

Proof V

- So, we get that

$$\mathbb{P}\left[\sum_{i=1}^{n} \Delta\mathbb{F}_i \geqslant E\right] \leqslant \exp\left(\frac{h^2}{8}\sum_{i=1}^{n} c_i^2 - hE\right)$$

- Rest of the proof is identical to the proof of the Hoeffding's Bound. The optimal choice of h that minimizes the RHS is

$$h^* = 4E/\sum_{i=1}^{n} c_i^2$$

- Substituting this value of h, we obtain

$$\mathbb{P}\left[\sum_{i=1}^{n} \Delta\mathbb{F}_i \geqslant E\right] \leqslant \exp\left(-2E^2/\sum_{i=1}^{n} c_i^2\right)$$

Concluding Note

- Students are highly recommended to use a representative example (as worked out in the class) to verify all the "equalities" and the "inequalities" used in the derivation of the proof

- The objective is to summarize all the concentration inequalities that we have studied in this course
- We shall also highlight some subtleties that differentiate the use of one concentration inequality from the others

Markov Inequality

- Let $\mathbb{X}$ be a random variable such that $\mathbb{X} \geqslant 0$ (i.e., the random variable is non-negative)
- Then the only bound that we can claim is the following for any $t \geqslant 0$
$$\mathbb{P}\left[\mathbb{X} \geqslant t\right] \leqslant \frac{\mathbb{E}\left[\mathbb{X}\right]}{t}$$
- For example, this bound implies that if $t = \alpha\mathbb{E}\left[\mathbb{X}\right]$, then $\mathbb{P}\left[\mathbb{X} \geqslant t\right] \leqslant 1/\alpha$. That is, it is unlikely that the the expectation of a random variable exceeds the expected value significantly
- **Comment:** For every t, there exists a random variable $\mathbb{X}$ such that the Markov inequality is tight. We emphasize that there is no single $\mathbb{X}$ that witnesses the tightness of the Markov inequality for all t

Chernoff Bound I

- Let $\mathbb{X}_i$, for $1 \leqslant i \leqslant n$, be independent random variables. Each $\mathbb{X}_i$ is a random variable over the sample space $[0, 1]$.
- **Comment:** Note that the $\mathbb{X}_i$s need not be identical. We can use "linear offset + linear scaling" to apply the Chernoff bound whenever there exists a, b such that $a \leqslant \mathbb{X}_i \leqslant b$, for all $1 \leqslant i \leqslant n$
- Let $p = (\mathbb{E}[\mathbb{X}_1] + \cdots + \mathbb{E}[\mathbb{X}_n])/n$
- Let $\mathbb{S}_n := \mathbb{X}_1 + \cdots + \mathbb{X}_n$
- Note that $\mathbb{E}[\mathbb{S}_n] = np$. Our objective is to prove that the probability that $\mathbb{S}_n$ significantly exceeds its expected value is very small. Chernoff bound states the following.

$$\mathbb{P}[\mathbb{S}_n \geqslant np + E] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p + E/n, p\right)\right)$$

Chernoff Bound II

- **Comment:** The Chernoff bound is true for all p and n. For example, the value of p can depend on n itself. Consider $p = 1/\sqrt{n}$. The bound continues to hold
- We can apply the Chernoff bound to the sum of $1 - \mathbb{X}_i$ random variables to conclude that

$$\mathbb{P}\left[\mathbb{S}_n \leqslant np - E\right] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left((1-p) + E/n, 1-p\right)\right)$$

- This form of the inequality, although very precise (tight lower-bound can be derived using the Stirling's Approximation and was part of your homework problems), is not easy to evaluate. This form is also unwieldy to understand the asymptotics based on n. So, we prove simpler to evaluate forms of this inequality

- Simpler Form 1.

$$\mathbb{P}\left[\mathbb{S}_n \geqslant np + E\right] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p + E/n, p\right)\right) \leqslant \exp(-2E^2/n)$$

  This bound is very easy to evaluate and highlights that most of the probability mass is concentrated around $E \approx \sqrt{n}$ radius of concentration around the mean.

- **Comment:** Note that this bound is oblivious to the value of p. So, we can bound the lower tail as follows. Let $q = 1 - p$.

$$\mathbb{P}\left[\mathbb{S}_n \leqslant np - E\right] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(q + E/n, q\right)\right) \leqslant \exp(-2E^2/n)$$

Chernoff Bound IV

- **Comment:** The property that the bound is independent of p is a significant drawback when we want to use Chernoff bound for problems where p is very small. For example, consider the case of $p = 1/\sqrt{n}$. In this case, the expected value $\mathbb{E}[\mathbb{S}_n] = \sqrt{n}$, so a radius of concentration $E \approx \sqrt{n}$ is not very meaningful. So, we go for easier to evaluate multiplicative forms of Chernoff bounds.
- **Comment:** Easier to evaluate Form 1 of the Chernoff bound is a corollary of the Hoeffding's bound (which, in turn, is a corollary of the independent bounded difference inequality; which, in turn, is a corollary of the Azuma's inequality)

Chernoff Bound V

- Simpler Form 2. Let $\varepsilon > 0$ and $\mu = \mathbb{E}[\mathbb{S}_n] = np$. Let $p' = p(1+\varepsilon)$. We have

$$\mathbb{P}\left[\mathbb{S}_n \geqslant \mu(1+\varepsilon)\right] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(p', p\right)\right) \leqslant \exp\left(-\frac{\varepsilon^2\mu}{2(1+\frac{\varepsilon}{3})}\right)$$

  This bound proves that the radius of concentration is roughly $E \approx \sqrt{\mu}$.

- Simpler Form 3. Let $1 > \varepsilon > 0$ and $\mu = \mathbb{E}[\mathbb{S}_n] = np$. Let $q = 1 - p$ and $q' = 1 - p(1-\varepsilon)$. We have

$$\mathbb{P}\left[\mathbb{S}_n \leqslant \mu(1-\varepsilon)\right] \leqslant \exp\left(-n\mathrm{D}_{\mathrm{KL}}\left(q', q\right)\right) \leqslant \exp\left(-\frac{\varepsilon^2\mu}{2}\right)$$

  This bound proves that the radius of concentration is roughly $E \approx \sqrt{\mu}$.

Hoeffding's Bound

- Let $a_i \leqslant \mathbb{X}_i \leqslant b_i$ be independent random variables for $1 \leqslant i \leqslant n$. Define $c_i = b_i - a_i$, for $1 \leqslant i \leqslant n$
- Let $\mathbb{S}_n := \mathbb{X}_1 + \cdots + \mathbb{X}_n$
- Then, the following bound holds

$$\mathbb{P}\left[\mathbb{S}_n \geqslant \mathbb{E}\left[\mathbb{S}_n\right] + E\right] \leqslant \exp\left(-2E^2 / \sum_{i=1}^{n} c_i^2\right)$$

- **Comment:** This bound is a consequence of the Bounded Independent Difference inequality
- **Comment:** The simpler to evaluate form 1 of Chernoff bound is a corollary of this bound

Independent Bounded Difference Inequality I

- Let $\mathbb{X}_1, \ldots, \mathbb{X}_n$ be <u>independent</u> random variables over the sample spaces $\Omega_1, \ldots, \Omega_n$, respectively
- Define $\Omega := \Omega_1 \times \cdots \times \Omega_n$
- Let $f : \Omega \to \mathbb{R}$ be any function that is <u>$(c_1, \ldots, c_n)$-bounded</u>
- Let $\mu = \mathbb{E}\left[f(\mathbb{X}_1, \ldots, X_n)\right]$
- Then, the following bound holds

$$\mathbb{P}\left[f(\mathbb{X}_1, \ldots, \mathbb{X}_n) \geqslant \mu + E\right] \leqslant \exp\left(-2E^2 \Big/ \sum_{i=1}^{n} c_i^2\right)$$

- **Comment:** This bound is a consequence of the Azuma-Hoeffding inequality
- **Comment:** This bound yields Hoeffding's bound as a corollary

Independent Bounded Difference Inequality II

- **Comment:** The radius of concentration is $E \approx \sqrt{\sum_{i=1}^{n} c_i^2}$. Note that the "number of variables" also factors into this bound.
- **Comment:** We emphasize that the same "quantity" can be represented in multiple ways as functions of different random variables. For example, the chromatic number of a graph whose edges are uniformly and independently included in the graph with probability p. If we define a random variable for each edge, then there are $\binom{n}{2} \approx n^2/2$ random variables. The chromatic number is a $(1, \ldots, 1)$-bounded function.

  Consequently, the radius of concentration is $\sqrt{\binom{n}{2}} \approx n$, which is not meaningful (because $\mu = \Theta(n)$, for constant p).

Independent Bounded Difference Inequality III

However, if we consider the graph where $\mathbb{X}_i$ represents how the vertex i is connected to the graph induced by $\{1, \ldots, i-1\}$, then there are n random variables. The chromatic number is a $(0, 1, \ldots, 1)$-bounded function. So, we get a $\sqrt{n}$ radius of concentration, which is meaningful.

- **Comment:** There are several applications of this bound. For example, to the chromatic number of random graphs, rank of random matrices
- **Comment:** The max-load function is also $(1, \ldots, 1)$-bounded. However, the radius of concentration is much-much larger than the expected value.

Azuma-Hoeffding Inequality I

- Let Ω be a sample space
- Consider the filtration

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$$

- Let $(\mathbb{F}_0, \mathbb{F}_1, \ldots, \mathbb{F}_n)$ be a martingale sequence with respect to the filtration mentioned above
- For $1 \leqslant i \leqslant n$, there exists c_i such that for all $x \in \Omega$

$$c_i \geqslant \max_{y \in \mathcal{F}_{i-1}(x)} \mathbb{F}_i(y) - \min_{y \in \mathcal{F}_{i-1}(x)} \mathbb{F}_i(y)$$

- Then, the following bound holds

$$\mathbb{P}\left[\mathbb{F}_n \geqslant \mathbb{F}_0 + E\right] \leqslant \exp\left(-2E^2 \Big/ \sum_{i=1}^{n} c_i^2\right)$$

Azuma-Hoeffding Inequality II

- **Comment:** Let $\mathbb{X}_1 \in \Omega_1, \dots, \mathbb{X}_n \in \Omega_n$. Define $\Omega := \Omega_1 \times \cdots \times \Omega_n$. Let $f \colon \Omega \to \mathbb{R}$ be any function. We emphasize that the distributions $\mathbb{X}_1, \dots, \mathbb{X}_n$ need not be independent. Suppose we are interested in studying the concentration of the function f around its expected value $\mu = \mathbb{E}\left[f(\mathbb{X}_1, \dots, \mathbb{X}_n)\right]$. That is, bound the quantity

$$\mathbb{P}\left[f(\mathbb{X}_1, \dots, \mathbb{X}_n) \geqslant \mu + E\right] \leqslant ?$$

1. We consider the filtration that reveals one coordinate of the functions at a time.

Azuma-Hoeffding Inequality III

2. We consider the Doob's Martingale with respect to this filtration using the following random variables. For any $0 \leqslant i \leqslant n$ we define

$$\mathbb{F}_i(x) := \mathbb{E}\left[f(\mathbb{X}_1, \ldots, \mathbb{X}_n) | \mathcal{F}_i\right](x)$$

   Note that $\mathbb{F}_0(x) = \mathbb{E}\left[f(\mathbb{X}_1, \ldots, \mathbb{X}_n)\right] = \mu$.
3. Now, we need to compute c_i, for $1 \leqslant i \leqslant n$, such that

$$c_i \geqslant \max_{y \in \mathcal{F}_{i-\mathbf{1}}(x)} \mathbb{F}_i(y) - \min_{y \in \mathcal{F}_{i-\mathbf{1}}(x)} \mathbb{F}_i(y)$$

   We emphasize that $\max_{y \in \mathcal{F}_{i-\mathbf{1}}(x)} \mathbb{F}_i(y)$ can be different from $\min_{y \in \mathcal{F}_{i-\mathbf{1}}(x)} \mathbb{F}_i(y)$ for different $x \in \Omega$. However, as long as their difference is bounded by c_i for every $x \in \Omega$, we are all set!
4. Now, Azuma-Hoeffding inequality shall imply that

$$\mathbb{P}\left[\mathbb{F}_n \geqslant \mu + E\right] \leqslant \exp\left(-2E^2 / \sum_{i=1}^{n} c_i^2\right)$$

- **Comment:** We can obtain concentration bounds for $(c_1, \dots, c_n)$-bounded functions where each coordinate is uniformly and independently chosen.
- **Comment:** We can obtain the concentration of the Hypergeometric series and in the Pólya's urn experiment

Talagrand Inequality I

- Let $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$ be independent random variables such that $\mathbb{X}_1 \in \Omega_1, \ldots, \mathbb{X}_n \in \Omega_n$.
- Define the product space $\Omega := \Omega_1 \times \cdots \times \Omega_n$
- Let $A \subseteq \Omega$ be any subset
- Talagrand's inequality states the following

$$\mathbb{P}[\mathbb{X} \in A] \cdot \mathbb{P}\left[d_T(\mathbb{X}, A) \geqslant t\right] \leqslant \exp(-t^2/4)$$

Here $d_T(\cdot, \cdot)$ represents the convex distance.

- **Comment:** Let us introduce the definition of an <u>c-configuration</u> function $f : \Omega \to \mathbb{R}$. The function f is c-configuration, if for every $x \in \Omega$, there exists α_x such that $\|\alpha_x\|_2 = 1$ and for every $y \in \Omega$ we have

$$f(y) \geqslant f(x) - \sqrt{cf(x)}d_{\alpha_x}(x, y)$$

Let m represent the median of the distribution $f(\mathbb{X})$. Then, the following bounds are corollary of the Talagrand's inequality.

$$\mathbb{P}\left[f(\mathbb{X}) \geqslant m + t\right] \leqslant \exp(-t^2/4c(m+t))$$
$$\mathbb{P}\left[f(\mathbb{X}) \leqslant m - t\right] \leqslant \exp(-t^2/4cm)$$

Talagrand Inequality III

- **Comment:** Remember that $m \leqslant 2\mathbb{E}\left[f(\mathbb{X})\right]$ by Markov inequality. This bound for c-configuration functions yields a concentration radius of $\sqrt{m} \approx \sqrt{\mu}$, which is applicable even when m is much less than n. However, note that we need $\mathbb{X}_1, \ldots, \mathbb{X}_n$ to be independent of each other. So, Talagrand and Azuma-Hoeffding inequalities are incomparable to each other (that is, they do not subsume each other).
- **Comment:** In the class we saw an application to the longest increased subsequence of n random numbers in $[0, 1)$. The mean/median of this random variable is $\Theta(\sqrt{n})$. Talagrand's inequality yields a concentration of $\sqrt{m}$ around the median.