ECE 587 / STA 563: Lecture 3 – Convergence and Typical Sets

Information Theory
Duke University, Fall 2023

Author: Galen Reeves
Last Modified: September 12, 2023

Outline of lecture:

3.1 Probability Review

3.1.1 Basic Inequalities

- **Markov's Inequality:** For any nonnegative random variable X and $t > 0$,

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

- **Proof:** We have

$$\mathbf{1}(x \geq t) \leq \frac{x}{t}, \quad \text{for all } x \geq 0$$

  Evaluating this inequality with X and the taking the expectation gives the stated result.

- **Chebyshev's Inequality:** For any random variable X with finite second moment and $t > 0$,

$$\mathbb{P}\big[\big|X - \boldsymbol{E}[X]\big| > t\big] \leq \frac{\mathsf{Var}(X)}{t^2}$$

- **Proof:** Apply Markov's inequality too $Y = (X - \mathbb{E}[X])^2$:

$$\mathbb{P}\big[\big|X - \boldsymbol{E}[X]\big| > t\big] = \mathbb{P}\big[Y > t^2\big] \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathsf{Var}(X)}{t^2}$$

- **Chernoff Bound:** For any random variable X, $t \in \mathbb{R}$, and $\lambda > 0$,

$$\mathbb{P}[X \geq t] \leq \exp(-\lambda t)\, \mathbb{E}[\exp(\lambda X)]$$

  Here $\mathbb{E}[\exp(\lambda X)]$ is the *moment generating function*

- **Proof:**

$$\begin{aligned}
\mathbb{P}[X \geq t] &= \mathbb{P}[\lambda X \geq \lambda t] && \text{Since } \lambda > 0 \\
&= \mathbb{P}\Big[e^{\lambda X} \geq e^{\lambda t}\Big] && \text{Since } \exp(\cdot) \text{ is nondecreasing} \\
&\leq e^{-\lambda t}\mathbb{E}\Big[e^{\lambda X}\Big] && \text{Markov's Inq.}
\end{aligned}$$

- **Chernoff Bound for Sums:** Let $X_1, X_2, \cdots$ be iid and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. For any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\big[\bar{X}_n \geq t\big] &\leq \exp(-n\lambda t)\,\mathbb{E}\big[\exp(n\lambda \bar{X}_n)\big] && \text{Chernoff Inq.}\\
&= \exp(-n\lambda t)\,\mathbb{E}\left[\exp\left(\sum_{i=1}^n \lambda X_i\right)\right] && \text{Definition of } \bar{X}_n\\
&= \exp(-n\lambda t)\prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)] && \text{Independence of } X_i\\
&= \Big[\exp(-\lambda t)\mathbb{E}[\exp(\lambda X_1)]\Big]^n && \text{Identically distributed}
\end{aligned}$$

3.1.2 Convergence of Random Variables

- A sequence of numbers $x_1, x_2, \cdots$ converges to a limit x if, for all $\epsilon > 0$, there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$|x_n - x| \leq \epsilon$$

  This is written as $x_n \to x$ as $n \to \infty$, or

$$\lim_{n\to\infty} x_n = x$$

- For a continuous function $g(\cdot)$,

$$x_n \to x \text{ as } n \to \infty \quad \Longrightarrow \quad g(x_n) \to g(x) \text{ as } n \to \infty$$

- There are several ways to characterize convergence for random sequences. In the following, we consider the case where the sequence of random variables $X_1, X_2, \cdots$ converges to a (non-random) limit x:

  - **convergence with probability one:** (also called *almost sure* convergence)

$$\mathbb{P}\Big[\lim_{n\to\infty} X_n = x\Big] = 1$$

  - **convergence in probability:** For every $\epsilon > 0$, $\mathbb{P}[|X_n - x| \geq \epsilon] \to 0$ as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - x| \leq \epsilon] = 1, \quad \text{for all } \epsilon > 0$$

  - **convergence in in p-th mean:**

$$\lim_{n\to\infty} \mathbb{E}[|X_n - x|^p] = 0$$

- Note:

$$\begin{aligned}
\text{convergence with probability one} \quad &\Longrightarrow \quad \text{convergence in probability}\\
\text{convergence in } p\text{-th mean } (p \geq 1) \quad &\Longrightarrow \quad \text{convergence in probability}
\end{aligned}$$

- **Example:** (Law of Large Numbers) Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}[|X|] < \infty$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. The sequence $\bar{X}_1, \bar{X}_2, \cdots$ converges to $\mathbb{E}[X]$ almost surely that is

$$\bar{X}_n \to \mathbb{E}[X] \quad \text{almost surely as } n \to \infty$$

  Thus, the sequence $\bar{X}_1, \bar{X}_2, \cdots$ also converges in probability, i.e. all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big[\big|\bar{X}_n - \mathbb{E}[X]\big| > \epsilon\big] = 0.$$

- Convergence of random sequences can be extended to functions of random sequences. If $g(\cdot)$ is a continuous function, then

$$X_n \to x \text{ in probability as } n \to \infty \quad \Longrightarrow \quad g(X_n) \to g(x) \text{ in probability as } n \to \infty$$

- **Example:** Let $X_1, X_2, \cdots$ be i.i.d. and let $g(\cdot)$ be continuous with $\mathbb{E}[|g(X_1)|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathbb{E}[g(X)] \quad \text{almost surely as } n \to \infty$$

3.2 The Typical Set and AEP

Throughout this section, let $X_1, X_2, \cdots$ be iid copies of a random variable $X \sim p(x)$ with finite support $\mathcal{X}$

3.2.1 High Probability Sets

- A length-n random sequence is denote by

$$X^n = (X_1, X_2, \cdots, X_n)$$

and a realization is denote by

$$x^n = (x_1, x_2, \cdots, x_n)$$

The joint pmf is the product measure:

$$\underbrace{p_{X^n}(x^n)}_{\text{joint pmf}} = \mathbb{P}[X^n = x^n] = \underbrace{\prod_{i=1}^{n} p(x_i)}_{\text{since iid}} = \underbrace{p(x^n)}_{\text{simplified notation}}$$

- The total number of possible sequences if length n is given by $|\mathcal{X}|^n$. This is huge!
- Our goal is to identify a subset of $A \subset \mathcal{X}^n$ which contains most of the probability, i.e. for $\epsilon \in (0,1)$, we want a set such that

$$\mathbb{P}[X^n \in A] = \sum_{x^n \in A} p(x^n) \geq 1 - \epsilon$$

The identification of a such a set with *nice* properties is a key step for many of the proofs in information theory.

- It is useful to define such a set in terms of a function $g : \mathbb{R} \to \mathbb{R}$. By the law of large numbers, for any $\epsilon > 0$ there exists N_ϵ such that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)]\right| \leq \epsilon\right] \geq 1 - \epsilon, \qquad \text{for all } n \geq N_\epsilon$$

  - This means that **almost all of the probability** is concentrated on the set of sequences $A \subset \mathcal{X}^n$ given by

$$A = \left\{x^n \in \mathcal{X}^n \ : \ \left|\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mathbb{E}[g(X)]\right| \leq \epsilon\right\}$$

- This set can be expressed equivalently as all $x^n \in \mathcal{X}^n$ such that

$$\mathbb{E}[g(X)] - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \leq \mathbb{E}[g(X)] + \epsilon$$

or equivalently

$$2^{-n(\mathbb{E}[g(X)]+\epsilon)} \leq 2^{-\sum_{i=1}^{n} g(x_i)} \leq 2^{-n(\mathbb{E}[g(X)]-\epsilon)}$$

- For which choice of $g(\cdot)$ will this set have *nice* properties? Can we characterize how large the set is?

3.2.2 The Typical Set

- For the special choice of $g(x) = -\log p(x)$, it follow that:

$$2^{-\sum_{i=1}^{n} g(x_i)} = 2^{\sum_{i=1}^{n} \log p(x_i)} = \prod_{i=1}^{n} p(x_i) = \underbrace{p_{X^n}(x^n)}_{\text{joint pmf of } X^n}$$

and

$$\mathbb{E}[g(X)] = \mathbb{E}[-\log p(X)] = \underbrace{H(X)}_{\text{entropy of } p(x)}$$

- **Definition:** The ϵ-typical set is defined by

$$A_\epsilon^{(n)} = \left\{ x^n \in \mathcal{X}^n \,:\, 2^{-n(H(X)+\epsilon)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-\epsilon)} \right\}$$

or equivalently, the set of all sequences $x^n \in \mathcal{X}^n$ obeying

$$\underbrace{H(X)}_{\text{entropy}} - \epsilon \leq \underbrace{-\frac{1}{n}\log p_{X^n}(x^n)}_{\text{empirical entropy}} \leq \underbrace{H(X)}_{\text{entropy}} + \epsilon$$

This is the set of all sequences whose probability is approximately equal to the expected probability. Unusually likely and unusually unlikely sequences are excluded.

- The typical set contains almost all of the probability. Furthermore, all of the sequences in the typical have roughly the same probability. This is know as the Asymptotic Equipartition property (AEP)

- The advantage of the this definition is that we can characterize the size of the typical set in terms of the entropy of $H(X)$:

  - By law of large numbers:

$$\mathbb{P}\left[X^n \in A_\epsilon^{(n)}\right] \geq 1 - \epsilon, \qquad \text{for all } n \geq N_\epsilon$$

  - Upper Bound:

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n\left(H(X)+\epsilon\right)}$$

  - Lower Bound:

$$\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n\left(H(X)-\epsilon\right)}, \qquad \text{for all } n \geq N_\epsilon$$

- **Proof of upper bound:**

$$1 = \sum_{x^n \in \mathcal{X}} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n\left(H(X)+\epsilon\right)} = \left|A_\epsilon^{(n)}\right| 2^{-n\left(H(X)+\epsilon\right)}$$

and so

$$2^{n\left(H(X)+\epsilon\right)} \geq \left|A_\epsilon^{(n)}\right|$$

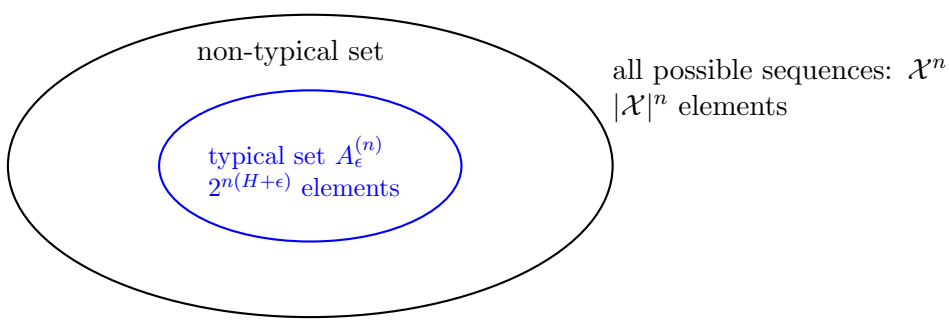
- **Proof of lower bound:** For n large enough, $\mathbb{P}\left[A_\epsilon^{(n)}\right] > 1-\epsilon$,

$$1-\epsilon < \mathbb{P}\left[A_\epsilon^{(n)}\right] = \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n\left(H(X)-\epsilon\right)} = \left|A_\epsilon^{(n)}\right| 2^{-n\left(H(X)-\epsilon\right)}$$

and so

$$(1-\epsilon)2^{n\left(H(X)-\epsilon\right)} \leq \left|A_\epsilon^{(n)}\right|$$

- Illustration of typical set with respect to $\mathcal{X}^n$. Note $|\mathcal{X}^n| = 2^{n\log(|\mathcal{X}|)}$
  - High probability sequences ($p(x^n) > 2^{-n(H(X)-\epsilon)}$) are excluded (too few to matter)
  - Low probability sequences ($p(x^n) < 2^{-n(H(X)+\epsilon)}$) are excluded (too unlikely to matter)
  - Average probability sequences ($p(x^n) \approx 2^{-n(H(X))}$) are retained



3.2.3 Examples

- **Example 1:** X_i are iid Bernoulli(3/4). ϵ is very small.

x^n	$p(x^n)$	typical?
11111111	$\left(\frac{3}{4}\right)^8$	no
11111100	$\left(\frac{3}{4}\right)^6 \times \left(\frac{1}{4}\right)^2$	yes
11101110	$\left(\frac{3}{4}\right)^6 \times \left(\frac{1}{4}\right)^2$	yes
00000000	$\left(\frac{1}{4}\right)^8$	no

- **Example 1:** X_i are iid according to

$$p(x) = \begin{cases} \frac{1}{2}, & x = 0 \\ \frac{1}{4}, & x = 1 \\ \frac{1}{4}, & x = 2 \end{cases}$$

where ϵ is very small.

x^n	$p(x^n)$	typical?
00001122	$\left(\frac{1}{2}\right)^4 \times \left(\frac{1}{4}\right)^4$	yes
00001111	$\left(\frac{1}{2}\right)^4 \times \left(\frac{1}{4}\right)^4$	yes
11002222	$\left(\frac{1}{2}\right)^2 \times \left(\frac{1}{4}\right)^6$	no