

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

Information Theory
Duke University, Fall 2024

History: Written by Galen Reeves (-2023). Updated by Henry Pfister (2024-).

Last Modified: September 16, 2024

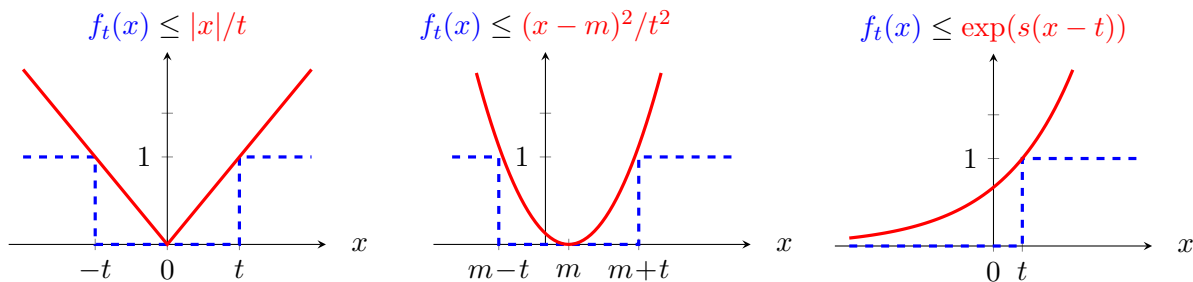
Outline of lecture:

3.1	Probability Review	1
3.1.1	Basic Inequalities	1
3.1.2	Convergence of Random Variables	2
3.2	The Typical Set and AEP	3
3.2.1	High Probability Sets	3
3.2.2	The Typical Set	4
3.2.3	Examples	6

3.1 Probability Review

3.1.1 Basic Inequalities

Let $f_t(x) = \mathbf{1}_{|x| \geq t}(x)$ satisfy $\mathbb{P}[X \geq t] = \mathbb{E}[f_t(X)]$ and



- **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

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

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

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

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

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

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

- **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:**

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

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

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

3.1.2 Convergence of Random Variables

- A sequence of real numbers x_1, x_2, \dots 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 denoted by $x_n \rightarrow x$ or

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

- For any continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ and any real sequence x_n , it holds that

$$x_n \rightarrow x \implies g(x_n) \rightarrow g(x)$$

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

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

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

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

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

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

$$\mathbb{P}\left[\lim_{n \rightarrow \infty} X_n = x\right] = \mathbb{P}\left[\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = x\right\}\right] = 1$$

- Note:

convergence in p -th mean ($p \geq 1$) \implies convergence in probability
 convergence with probability one \implies convergence in probability

- **Example:** (Law of Large Numbers) Let X_1, X_2, \dots 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 $\bar{X}_1, \bar{X}_2, \dots$ converges to $\mathbb{E}[X]$ almost surely that is

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

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

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

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

$$X_n \rightarrow x \text{ in probability as } n \rightarrow \infty \implies g(X_n) \rightarrow g(x) \text{ in probability as } n \rightarrow \infty$$

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

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

If $\text{Var}(g(X_1)) < \infty$, then there is a simple proof that follows from applying Chebyshev's inequality to the sum because, for $Z_i = g(X_i) - \mathbb{E}[g(X_i)]$, we have

$$\mathbb{E} \left[\left(\sum_{i=1}^n Z_i \right)^2 \right] = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[Z_i Z_j] = n \mathbb{E}[Z_1^2] = n \text{Var}(g(X_1)).$$

3.2 The Typical Set and AEP

Throughout this section, let X_1, X_2, \dots 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

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

and a realization is denote by

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

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 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 $g : \mathbb{R} \rightarrow \mathbb{R}$. By the law of large numbers, for any $\epsilon > 0$ there exists N_ϵ such that

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

- 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 follows 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 known as the Asymptotic Equipartition property (AEP)
- The advantage of 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:

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

- Upper Bound:

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

- Lower Bound:

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

- **Proof of upper bound:**

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

and so

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

Since this argument only uses the probability lower bound, the stated upper bound actually applies to the larger set of all sequences with probability greater than $2^{-n(H(X)+\epsilon)}$.

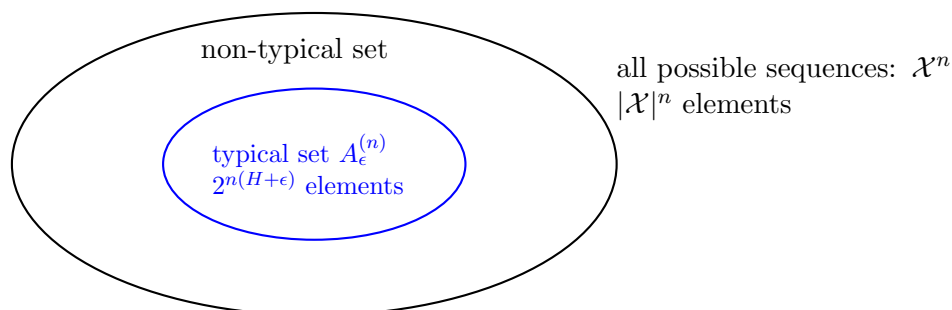
- **Proof of lower bound:** For n large enough, $\mathbb{P}[A_\epsilon^{(n)}] > 1 - \epsilon$,

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

and so

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

- 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	$(\frac{3}{4})^8$	no
11111100	$(\frac{3}{4})^6 \times (\frac{1}{4})^2$	yes
11101110	$(\frac{3}{4})^6 \times (\frac{1}{4})^2$	yes
00000000	$(\frac{1}{4})^8$	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	$(\frac{1}{2})^4 \times (\frac{1}{4})^4$	yes
00001111	$(\frac{1}{2})^4 \times (\frac{1}{4})^4$	yes
11002222	$(\frac{1}{2})^2 \times (\frac{1}{4})^6$	no