# Contents

**1 Convergence** **2**
1 Convergence almost surely . . . 2
2 Convergence in probability . . . 3
3 Convergence in distribution . . . 3

**2 Information theory** **4**
1 AEP . . . 4
2 Huffman coding . . . 5
3 Entropy . . . 5
4 Mutual information . . . 5
5 Capacity . . . 6
5.1 Binary erasure channel (BEC) . . . 6

**3 Random processes** **7**
1 MC . . . 7
1.1 Big theorem . . . 9
1.2 DTMC . . . 9
2 PP . . . 10
2.1 Erlang . . . 10
2.2 Random incidence property (RIP) . . . 10

Spring 2025, taught by Professor Kannan Ramchandran.

# Chapter 1

# Convergence

Modes of convergence:

almost surely $\Rightarrow$ in probability $\Rightarrow$ in distribution

## 1 Convergence almost surely

$$X_n \xrightarrow{a.s.} X \iff \mathbb{P}(\{w \in \Omega : \lim_{n\to\infty} X_n = X\}) = 1 \iff \mathbb{P}(\lim_{n\to\infty} X_n \neq X) = 0 \tag{1.1}$$

**Theorem 1** (SLLN)**.** *If $(X_n)_{n=1}^{\infty}$ are i.i.d. with finite mean $\mathbb{E}[X_1] < \infty$, then the sample mean $\bar{X_n}$ converges almost surely to the true mean:*

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X_1]$$

**Lemma 1** (Borel-Cantelli lemma)**.** *Let $(A_n)_{n=1}^{\infty}$ be a collection of events.*

*The event that $A_n$ happens infinitely often is:*

$$A_n \ i.o. = \limsup_{n\to\infty} A_n = \bigcap_{k=1}^{\infty} \bigcup_{k\geq n} A_k \tag{1.2}$$

**Fact:** If $w \in A_n$ i.o., then $\forall n \geq 1 : \exists k \geq n$ s.t. $w \in A_k$.

Otherwise, there is a max $N$ s.t. $w \notin A_k \quad \forall k \geq N$, i.e. $w$ only appears in finitely many $A_n$.

(i) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$

(ii) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $(A_n)_{n=1}^{\infty}$ are independent, then $\mathbb{P}(A_n \text{ i.o.}) = 1$

**Fact:** If we define $A_n := \{w \in \Omega : |X_n(w) - X(w)| \geq \epsilon\}$, then we can show that $A_n$ is th event the sequence $X_n$ is not converging to $X$ (i.e. diverges and $\lim_{n\to\infty} X_n(w) \neq X(w)$).

So, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$, and $X_n \xrightarrow{a.s.} X$.

Some applications of almost sure convergence:

- In DTMC, the proportion of time spent in a state converges a.s. to the inverse of the expected time it takes to revisit that state (given a few assumptions).
- If $(X_n)_{n=1}^{\infty}$ over a finite alphabete, then the average suprise $-\frac{1}{n}\log_2 p(X_1, \dots, X_n)$ converges a.s. to the entropy $H(X)$. This is called *asymptotic equipartition property.*
- In machine learning, we can ask if the iterates of the *stochastic gradient descent* algorithm converge a.s. to the true minimizer of the given function.

# 2 Convergence in probability

$$X_n \xrightarrow{\mathbb{P}} X \iff \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \quad \forall \epsilon > 0 \tag{1.3}$$

# 3 Convergence in distribution

$$X_n \xrightarrow{d} X \iff \lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R} : \mathbb{P}(X = x) = 0 \tag{1.4}$$

Equivalently, $p_{X_n} \to p_X$ for discrete RV and $f_{X_n} \to f_X$ for continuous RV.

**Theorem 2** (Central Limit theorem)**.** *If $(X_n)_{n=1}^{\infty}$ are i.i.d. with mean $\mu$ and variance $\sigma^2$, then the standard score of the sample mean $\bar{X}_n$ converges in distirbution to the standard normal distribution.*

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{1.5}$$

# Chapter 2

# Information theory

**Shannon's separation theorem.** Source coding and channel coding can be done separately without loss of optimality.

- SC: cannot compress an i.i.d. source $X$ "on average" asymptotically below the *entropy* of the source $X$, $H(X)$.
- CC: cannot transmit reliably at rate $R$ above channel capacity $C$.

**Theorem 3** (Source Coding Theorem). *For an i.i.d. sequence $X_1, \ldots, X_n$ and an arbitrarily small $\epsilon > 0$, there is a source coding scheme for which*

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l(X_1, \ldots, X_n)\right] \leq H(X) + \epsilon \textit{ bits per symbol} \tag{2.1}$$

*s.t. the sequence $X_1, \ldots, X_n$ can be recovered from the encoding with a high probability $(1-\epsilon)$.*

## 1 AEP

**Fact:** A typical set in flipping $n$ coins with a probability of heads $p$ is when there are $np$ heads and $n(1-p)$ tails, the expected number of heads and tails.

Then, the probability of a typical sequence is:

$$p^{np}(1-p)^{n(1-p)} = 2^{\log(p^{np}(1-p)^{n(1-p)})} = 2^{n(p\log p + (1-p)\log(1-p))} = 2^{-nH(p)} = 2^{\mathbb{E}[\log p_X(X)^n]} \tag{2.2}$$

The $\epsilon$-typical set $A_\epsilon^{(n)}$ is a set of sequences s.t.

$$2^{-n(H(X)+\epsilon)} \leq p_{X^{(n)}}(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{2.3}$$

where $|A_\epsilon^{(n)}| \approx 2^{nH(X)}$.

Note that the size of the set of all possible sequence is $|\mathcal{X}|^n = 2^{n \log |X|}$.

Since $\log |\mathcal{X}|$ is the entropy of the uniform distribution $\mathcal{X}$, we have that $H(X) \leq \log |\mathcal{X}|$.

**Theorem 4** (Asymptotic Equipartition Property)**.** *If $X_1, \ldots, X_n \sim p_{X^n}$ i.i.d., then*

$$
\begin{aligned}
&-\frac{1}{n}\log_2 p_{X^n}(x_1,\ldots,x_n) \xrightarrow{i.p.} H(X) \\
&\iff \mathbb{P}\left(\left|-\frac{1}{n}\log_2 p_{X^n}(x_1,\ldots,x_n) - H(X)\right| > \epsilon\right) \to 0 \text{ as } n \to \infty \\
&\mathbb{P}(2^{-n(H(X)+\epsilon)} < p_{X^n}(x_1,\ldots,x_n) < 2^{-n(H(X)-\epsilon)}) \to 1
\end{aligned} \tag{2.4}
$$

# 2 Huffman coding

Expected Huffman code length is between $H(X)$ and $H(X) + 1$.

# 3 Entropy

Entropy of the DRV $X$:

$$H(X) = \mathbb{E}[-\log p(X)] = \mathbb{E}\left[\log_2 \frac{1}{p(X)}\right] = \sum_{x \in X} p(x) \log_2 \left(\frac{1}{p(x)}\right)$$

**Properties of entropy:**

1. $H(X) \geq 0$.
2. Joint entropy: $H(X,Y) = \mathbb{E}\left[\log_2 \frac{1}{p(X,Y)}\right] = \sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) \log \frac{1}{p_{X,Y}(x,y)}$
3. Conditional entropy: $H(Y|X) = \mathbb{E}\left[\log_2 \frac{1}{p(Y|X)}\right] = \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} \leq H(Y)$

   Note that conditioning only decreases entropy.
4. Chain rule: $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

   Note that $H(Y) = H(X,Y) - H(X|Y)$ is the remaining amount of uncertainty after observing $X$.

Entropy is maximized when the distribution is uniform.

# 4 Mutual information

Average amount of information that $X$ provides about $Y$:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

# 5 Capacity

$C(BEC(p)) = 1 - p \geq C(BSC(p)) = 1 - h(p)$.

$$C = \max_{p_X} I(X;Y) = \max_{p_X} H(X) - H(X|Y) \text{ bits per channel use} \tag{2.5}$$

Each bit we send carries $C$ bits of information.

## 5.1 Binary erasure channel (BEC)

Ensure reliability by redundancy of $(1-p)n$ unerased bits. Map one message only to one codeword.

In general, the rate is $\boxed{R := \frac{L}{n}}$.

**Fact:** We wish to send a message of length $L$ bits, and we encode to a codeword of length $n > L$.

*Shannon's random codebook argument.* We flip $n2^L$ fair coins independently, and populate a $2L \times n$ codebook accordingly ($2^L$ codewords, each with length $n$).

$\mathcal{Y}$ is a string with values $\{0, 1, e\}$.

**Theorem 5.** *The capacity of the BEC with error probability $p$ is $1 - p$.*

*Proof.* Note that we can do no better than $1-p$, since we can just resend the erased bits.

*Oracle argument.* Since the channel erases fraction $p$ of the input bits, the relaible rate of communication is $1-p$ bits per channel use.

We show that we can achieve a rate of $R := 1 - p - \epsilon$ for any $\epsilon > 0$.

Suppose the first codeword is sent.

WLOG, assume first $n(1-p)$ symbols came through.

Then, we have:

$$\begin{aligned}
\mathbb{P}(\text{error}) &= \mathbb{P}\left(\bigcup_{i=2}^{2^L}\{c_1 = c_i\}\right) \\
&\leq \sum_{i=2}^{2^L} \mathbb{P}(c_1 = c_i) \quad \mathbb{P}(c_1 = c_i) = \frac{1}{2^{n(1-p)}} \\
&= (2^L - 1) \cdot 2^{-n(1-p)} \\
&\approx 2^{L-n(1-p)} \quad L = nR \\
&= 2^{-n(1-p-R)} \to 0 \text{ as } n \to \infty \quad R < 1-p
\end{aligned} \tag{2.6}$$

□

**Fact:** In BEC, $\mathbb{P}(\text{error}) \leq 2^{-n(1-p-R)}$.

# Chapter 3

# Random processes

## 1 MC

Markov chain: $(X_n)_{n=1}^N$, where $X_n$ is the state at time $n$.

*Chapman-Kolmogorov equation* for $n$-step transition probability:

$$P_{ij}^n = \mathbb{P}(\text{going from state } i \text{ to state } j \text{ in } n \text{ steps}) = \sum_{k \in \mathcal{X}} P_{ik}^{n-1} \cdot P_{kj}$$

A MC is *irreducible* if we can reach any state from any other state.

*Periodicity*: if irreducible, gcd of all path length to return (if irreducible, same $d(i)$ for all states $i$):

$$d(i) = \gcd\{n \geq 1 | P_{ii}^n > 0\}$$

A MC is *reversible* if its stationary distribution $\pi$ and transition probability matrix $P$ satisfy the detailed balance equation:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall x, y \in \mathcal{X} \tag{3.1}$$

**Fact:** Start with a graph for an irreducible, pos. recurrent MC. Remove all arrows, multiple edges between nodes and loops. If the resulting graph is a tree, then MC is reversible and its stationary distribution satisfies DBE.

*Backwards Markov property*:

$$\mathbb{P}(X_n = x_n | X_{n+1} = x_{n+1}, \ldots, X_{n+k} = x_{n+k}) = \mathbb{P}(X_n = x_n | X_{n+1} = x_{n+1})$$

**Fact:** Given a reversible MC $(X_n)_{n \geq 0}$ with stationary distribution $\pi$.

If $X_0 \sim \pi$, then $\forall n \in \mathbb{N}$, the chain up to time $n$ is equal in distribution to its reverse:

$$\begin{aligned}
\mathbb{P}(X_{0:n} = x_{0:n}) &= \mathbb{P}(X_n = x_n) \prod_{k=0}^{n-1} \mathbb{P}(X_k = x_k | X_{k+1} = x_{k+1}) \quad \text{backwards MP} \\
&= \mathbb{P}(X_n = x_n) \prod_{k=0}^{n-1} \frac{\mathbb{P}(X_{k+1} = x_{k+1} | X_k = x_k)\, \mathbb{P}(X_k = x_k)}{\mathbb{P}(X_{k+1} = x_{k+1})} \quad \text{Bayes rule} \\
&= \pi(x_n) \prod_{k=0}^{n-1} \frac{\pi(x_k) P(x_k, x_{k+1})}{\pi(x_{k+1})} \quad \text{stationarity} \\
&= \pi(x_n) \prod_{k=0}^{n-1} P(x_{k+1}, x_k) \quad \text{reversibility} \\
&= \mathbb{P}(X_0 = x_n) \prod_{k=0}^{n-1} \mathbb{P}(X_{n-k} = x_k | X_{n-k-1} = x_{k+1}) \\
&= \mathbb{P}(X_{0:n} = x_{n:0})
\end{aligned} \tag{3.2}$$

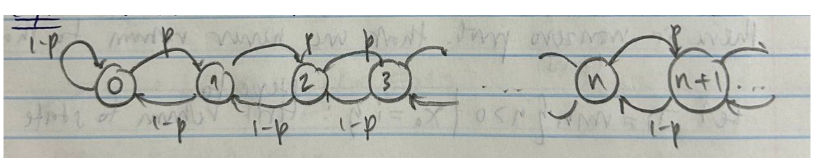
A state $i$ is *transient* if given that we start in state $i$, there is nonzero probability that we never return to that state $i$.

Let $T_x = \min\{n \geq 1 \mid X_0 = x\}$ denote number of steps to first return to state $x \in \mathcal{X}$.

- If MC is irreducible, then $\mathbb{P}(T_x < \infty | X_0 = x) = \begin{cases} 1 \text{ if recurrent} \\ < 1 \text{ if transient} \end{cases}$
- If MC is recurrent, then $\mathbb{E}[T_x \mid X_0 = x] = \mathbb{E}_x[T_x^+] = \begin{cases} < \infty \text{ postitive recurrent} \\ \infty \text{ null recurrent} \end{cases}$

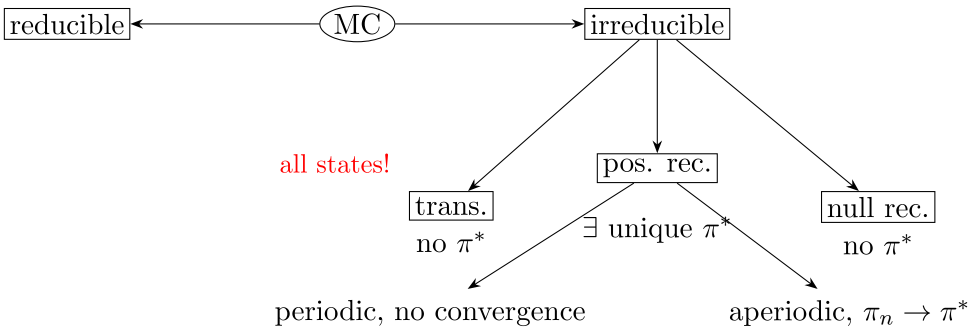
**Fact:** A random walk reflected at 0 with probability of moving to the right $p$ is:

- if $p < 1/2$, positive recurrent
- if $p = 1/2$, null recurrent
- if $p > 1/2$, transient



## 1.1 Big theorem

Big theorem for a finite state MC:




**Example:** [FA23 Q5 Ehrenfest's diffusion model]

$$\pi_i = \pi_{i-1} P_{i-1,i} + \pi_{i+1} P_{i+1,i} \quad i = 1, 2, \ldots, K-1$$

**Fact:** A finite state, irreducible MC that is *undirected* has a stationary distribution $\pi(i) = \frac{\deg(i)}{2|E|}$ and is reversible.

## 1.2 DTMC

Stationary distribution: $\pi P = \pi$, where $\sum_{i=0}^{n} \pi_i = 1$, $\pi = [\pi_0 \; \pi_1 \; \ldots \; \pi_n]$.

**Fact:** If a Markov chain starts at the stationary distribution, then every future state $X_t$ is also distributed according to $\pi$ for $t \geq 0$.

**Theorem 6.** *Suppose that the Markov chain is irreducible with a stationary disitribution $\pi$. Then, for each state $x \in \mathcal{X}$:*

$$\pi(x) = \frac{1}{\mathbb{E}[T_x^+]} \tag{3.3}$$

*Proof.*

$$\frac{t}{\sum_{i=0}^{t-1} \mathbf{1}_{X_i = x}} \to \mathbb{E}_x[T_x^+] \quad \frac{\text{total time}}{\text{number of visits to } x} \tag{3.4}$$

Then, we have:

$$\frac{1}{t} \sum_{i=0}^{t-1} \mathbf{1}_{X_i = x} \to \frac{1}{\mathbb{E}_x[T_x^+]} \tag{3.5}$$

where expectation of LHS is $\frac{1}{t} \sum_{i=0}^{t-1} \mathbb{P}(X_i = x)$.

If we start chain at the stationary distribution, then $\mathbb{P}(X_i = x) = \pi(x)$. □

# 2 PP

Poisson process: events that occur independently with some average rate $\lambda$.

Let $S_i$ be interarrival time between $(i-1)$th and $i$th arrival, where $S_i \sim \text{Exp}(\lambda)$ are i.i.d.

Poisson splitting, Poisson merging.

**Fact:** $\mathbb{P}(N(t)=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}$, where $N(t)$ is the number of arrivals on $[0,t]$.

Stationary, independent increments.

## 2.1 Erlang

Erlang$(n;\lambda)$ is a sum of $n$ i.i.d. exponential RVs with rate $\lambda$.

Then, the distribution of $i$th arrival time $T_i = S_1 + \cdots + S_i \sim \text{Erlang}(i;\lambda)$ is:

$$f_{T_i}(t) = \frac{\lambda^i t^{i-1} e^{-\lambda t}}{(i-1)!} \quad \text{for } t \geq 0 \tag{3.6}$$

*Proof.* Assume $0 < t_1 < t_2 < \cdots < t_n$.

$$\begin{aligned} f_{T_1,\ldots,T_i}(t_1,t_2,\ldots,t) &= f_{S_1,\ldots,S_i}(t_1, t_2-t_1, \ldots, t-t_{i-1}) \\ &= f_{S_1}(t_1)\cdot f_{S_2}(t_2-t_1)\cdot\cdots\cdot f_{S_i}(t-t_{i-1}) \\ &= \lambda e^{-\lambda t_1}\cdot \lambda e^{-\lambda(t_2-t_1)}\cdots \lambda e^{-\lambda(t-t_{i-1})} \\ &= \lambda^i e^{-\lambda t} \end{aligned} \tag{3.7}$$

Then, we have:

$$\begin{aligned} f_{T_i}(t) &= \int_0^t \cdots \int_0^t f_{T_1,\ldots,T_i}(t_1,t_2,\ldots,t)dt_1 dt_2\cdots dt_{i-1} \\ &= \int_0^t \cdots \int_0^t \lambda^i e^{-\lambda t} dt_1 dt_2 \cdots dt_{i-1} \\ &= \frac{\lambda^i e^{-\lambda t} t^{i-1}}{(i-1)!} \end{aligned} \tag{3.8}$$

□

**Fact:** $\mathbb{E}[T_i] = \frac{i}{\lambda}$, $\text{var}(T_i) = \frac{i}{\lambda^2}$.

## 2.2 Random incidence property (RIP)

Length of the interval with the arbitrary time point we choose will be Erlang$(2;\lambda)$

**Example:** [Disc 10 Q2 Bus arrival at Cory]