MA4261 Information and Coding Theory

AY24/25 Semester 1

by Isaac Lai

Probability

- $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- **Union bound:** In a probability space with σ-algebra $\mathscr{F}$ we have
$$\Pr\left(\bigcup_{i=1}^{k} A_i\right) \le \sum_{i=1}^{k} \Pr(A_i)$$
This holds in the infinite case too.
- $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X \mid Y]]$
- Random variables X, Y, Z form a **Markov chain** in the order $X - Y - Z$ if their joint distribution P_{XYZ} satisfies for all $(x, y, z) \in \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$
$$P_{XYZ}(x, y, z) = P_X(x) P_{Y|X}(y \mid x) P_{Z|Y}(z \mid y)$$
This is equivalent to saying X and Z are **conditionally independent given** Y.
- **Markov's Inequality:** Let X be a real-valued non-negative random variable. Then for any $a > 0$ we have $\Pr(X > a) \le \frac{\mathbb{E}[X]}{a}$.
- **Chebyshev's Inequality:** Let X be a real-valued random variable with mean μ and variance σ^2. Then for any $a > 0$
$$\Pr(|X - \mu| > a\sigma) \le \frac{1}{a^2}$$
- **Weak Law of Large Numbers:** For every $\epsilon > 0$,
$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > \epsilon\right) \le \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Information Quantities

Definition. The **entropy** $H(X)$ of a discrete random variable X is defined by
$$H(X) = -\sum_{x\in\mathscr{X}} p(x) \log p(x)$$

Properties of H

1. $H(X) \ge 0$
2. $H_b(X) = (\log_b a) H_a(X)$ (binary entropy)
3. (Conditioning does not increase entropy) For any two random variables X and Y, $H(X \mid Y) \le H(X)$ with equality iff X and Y are independent.
4. $H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i)$ with equality iff all X_i are independent.
5. $H(X) \le \log|\mathscr{X}|$ with equality iff X is distributed uniformly over $\mathscr{X}$.
6. $H(p)$ is concave in p.
7. **Han's Inequality:**
$$H(X_1, \ldots, X_n) \le \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$
8. $H(g(X)) \le H(X)$

Definition. The **relative entropy** $D(p \parallel q)$ of pmf p wrt pmf q is
$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Definition. The **mutual information** between two random variables X and Y is defined as
$$I(X;Y) = \sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Alternatively,
$$H(X) = E_p \log \frac{1}{p(X)}$$
$$H(X,Y) = E_p \log \frac{1}{p(X,Y)}$$
$$H(X \mid Y) = E_p \log \frac{1}{p(X \mid Y)}$$
$$I(X;Y) = E_p \log \frac{p(X,Y)}{p(X)p(Y)}$$
$$D(p \parallel q) = E_p \log \frac{p(X)}{q(X)}$$

Properties of D and I

1. $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y)$
2. $D(p \parallel q) \ge 0$ with equality iff $p(x) = q(x)$ for all $x \in \mathscr{X}$
3. $I(X;Y) = D(p(x,y) \parallel p(x)p(y)) \ge 0$ with equality iff $p(x,y) = p(x)p(y)$, i.e. X and Y are independent.
4. If $|\mathscr{X}| = m$ and u is the uniform distribution over $\mathscr{X}$, then $D(p \parallel q) = \log m - H(p)$.
5. $D(p \parallel q)$ is convex in the pair (p, q).

Chain rules

- Entropy: $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1)$
- Mutual information: $I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1, X_2, \ldots, X_{i-1})$
- Relative entropy: $D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y \mid x) \parallel q(y \mid x))$

Important results

- **Jensen's Inequality:** If f is a convex function, then $\mathbb{E}f(X) \ge f(\mathbb{E}X)$
- **Log sum Inequality:** For n positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \ge \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$
with equality iff $\frac{a_i}{b_i} =$ constant.
- **Data-processing Inequality:** If $X \to Y \to Z$ forms a Markov chain, $I(X;Y) \ge I(X;Z)$.
- **Sufficient statistic:** $T(X)$ is sufficient relative to $\{f_\theta(x)\}$ iff $I(\theta;X) = I(\theta;T(X))$ for all distributions on θ.
- **Fano's Inequality:** Let $P_e = \Pr\{\hat{X}(Y) \ne X\}$. Then
$$H(P_e) + P_e \log|\mathscr{X}| \ge H(X \mid Y)$$
This can be loosened to
$$P_e \ge \frac{H(X \mid Y) - 1}{\log|\mathscr{X}|}$$
- If X and X' are i.i.d., then $\Pr(X = X') \ge 2^{-H(X)}$

Asymptotic Equipartition Property

Definition. The **typical set** of X, a discrete memoryless source (DMS) is defined as
$$A_\epsilon^{(n)}(X) := \left\{x^n \in \mathscr{X}^n : \left|\frac{1}{n} \log \frac{1}{P_{X^n}(x^n)} - H(X)\right| \le \epsilon\right\}$$
where for all $x^n \in \mathscr{X}^n$
$$P_{X^n}(x^n) = \Pr(X^n = x^n) = \prod_{i=1}^{n} P_X(x_i)$$

Theorem (AEP). 1. $\Pr(X^n \in A_\epsilon^{(n)}(X)) \ge 1 - \epsilon$ for all sufficiently large n.

2. The size of the typical set satisfies $(1-\epsilon)2^{n(H(X)-\epsilon)} \le \left|A_\epsilon^{(n)}(X)\right| \le 2^{n(H(X)+\epsilon)}$.

Definition (Code). An $(n, 2^{nR})$-fixed-to-fixed-length source code consists of an encoder f and a decoder φ where

1. $f : \mathscr{X}^n \to \{1, \ldots, 2^{nR}\}$ and
2. $\varphi : \{1, \ldots, 2^{nR}\} \to \mathscr{X}^n$

n is the blocklength of the code and R is the rate of the code.

Definition (Achievable rate). $R \ge 0$ is achievable if there exists a sequence of $(n, 2^{nR})$-codes such that $\lim_{n\to\infty} \Pr(\hat{X}^n \ne X^n) = 0$ where $\hat{X}^n = \varphi(M)$ and $M = f(X^n)$ are the reconstructed source and compression index respectively.

Definition (Optimum Source Coding Rate). The optimum source coding rate for the DMS X is $R^*(X) = \inf\{R : R \text{ is achievable}\}$.

Theorem (Fixed-to-Fixed-Length Data Compression).
$$R^*(X) = H(X)$$

Theorem. If $R < H(X)$, then $P_e^{(n)} := \Pr(\hat{X}^n \ne X^n) \to 1$ as $n \to \infty$.

Theorem (Han-Verdu Lemma). Fix any $(n, 2^{nR})$-code. Then $P_e = \Pr(\hat{X}^n \ne X^n)$ satisfies
$$P_e \ge \sup_{\gamma>0} \Pr\left\{\frac{1}{n} \log \frac{1}{P_{X^n}(X^n)} \ge R + \gamma\right\} - e^{-n\gamma}$$

Theorem. Let $B_\delta^{(n)} \subset \mathscr{X}^n$ be such that if $X_1, X_2, \cdots \sim P_X$, then for every $\delta \in (0,1)$, $\Pr(X^n \in B_\delta^{(n)}) \ge 1 - \delta$ for all n sufficiently large. Then for any $\delta' > 0$,
$$\frac{1}{n} \log \left|B_\delta^{(n)}\right| \ge H(X) - \delta'$$
for n sufficiently large. Here $H(X)$ is computed wrt PMF P_X

Entropy Rates of Stochastic Processes

A **stochastic process** $\{x_i\}_{i\in\mathbb{N}}$ is an indexed sequence of random variables where i is the time.

Definition. A stochastic process is **stationary** if $\Pr(X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_{1+\ell} = x_1, \ldots, X_{n+\ell} = x_n)$ for all $n \in \mathbb{N}$ and every shift $\ell \in \mathbb{N}$, and for all $x_1, \ldots, x_n \in \mathscr{X}$

Definition. A stochastic process is a **Markov chain** if $\forall n \ge 1$, $\Pr(X_{n+1} = x_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_{n+1} = x_{n+1} \mid X_n = x_n)$ $\forall x_1, \ldots, x_{n+1} \in \mathscr{X}$

Definition. The Markov chain is **time-invariant** if $P(x_{n+1} \mid x_n)$ does not depend on n. Such a Markov chain is charactersied by a transition probability matrix (TPM) $P = [P_{ij}]$, $i, j \in \mathscr{X}$, $P_{ij} = \Pr(X_{n+1} = j \mid X_n = i)$ for all time-invariant n. In other words, we have $p_{n+1} = p_n P$

If it is possible to go from any state to any other in a finite number of steps, the Markov chain is **irreducible**. If the GCD of the lengths of different paths from a state to itself is 1, the Markov chain is **aperiodic**.

Definition (Entropy rate). Two definitions:
$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$
$$H'(X) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$

For a stationary stochastic process, $H(\mathscr{X}) = H'(\mathscr{X})$

Theorem (Cesaro mean). If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.

Theorem (Shannon-McMillan-Breiman). For a stationary, ergodic (irreducible and aperiodic) process, the AEP holds: $\lim_{n\to\infty} -\frac{1}{n} \log p(X_1, \ldots, X_n) = H(X)$

- **Entropy rate of an ergodic Markov chain:** $H(X) = H'(X) = H(X_2 \mid X_1)$
- **Functions of a Markov chain:** If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain and $Y_i = \phi(X_i)$, then

$$H(Y_n \mid Y^{n-1}, X_1) \le H(Y) \le H(Y_n \mid Y^{n-1})$$
$$\lim_{n\to\infty} H(Y_n \mid Y^{n-1}, X_1) = H(Y) = \lim_{n\to\infty} H(Y_n \mid Y^{n-1})$$

Fixed-to-Variable-Length Source Coding

Definition. A fixed-to-variable-length (F2V) source code for a random variable X is a map C for $\mathscr{X}$ to $\{0,1\}^*$. $C(x)$ is the codeword corresponding to $x \in \mathscr{X}$ and $l(w)$ is the length of the codeword corresponding to $x \in \mathscr{X}$.

Definition. The expected length $L(C)$ of a code $C : \mathscr{X} \to \{0,1\}^*$ for a random variable $X \sim p_X$ is $L(C) = \sum_{x\in\mathscr{X}} p_X(x) l(x) = \mathbb{E}_{p_X}[l(X)]$.

Definition. A code C is **non-singular** if for all $x \in \mathscr{X}$ gets mapped to a different codeword, i.e. for all $x, x' \in \mathscr{X}$ such that $x \neq x'$, we have $C(x) \neq C(x')$.

Definition. The **extension** C^* of a code C is the map from finite-length strings in $\mathscr{X}$ to finite-length strings in $\{0,1\}^*$.

Definition. A **uniquely decodable** code is one in which its extension is non-singular.

Definition. A code is called **prefix-free** or **instantaneous** if no codeword is a prefix of any other codeword.

Theorem (Kraft's Inequality). For any PF code over an alphabet of size 2, its codeword lengths $l_1, l_2, \ldots, l_m$ must satisfy $\sum_{i=1}^m 2^{-l_i} \le 1$. Conversely, if the inequality is satisfied, then there exists a PF code with those lengths.

Theorem. The expected codeword length L^* of any binary PF code for a random variable X satisfies $L^* \ge H(X)$ with equality iff $2^{-l_i} = p_i$. Moreover, $L^* < H(X) + 1$.

Definition (Shannon code). For all $i \in \mathscr{X}$, $l_i = \left\lceil \log \frac{1}{p_i} \right\rceil$

Theorem (Coding over long blocks). We have $\frac{H(X^n)}{n} \le L_n^* < \frac{H(X^n)}{n} + \frac{1}{n}$. If $\mathscr{X} = \{X_n\}_{n=1}^\infty$ is a stationary stochastic process, then $L_n^* \to H(\mathscr{X})$.

Theorem (Wrong code). For the code assignment $l(x) = \left\lfloor \log \frac{1}{q(x)} \right\rfloor$,

$$H(p) + D(p \parallel q) \le E_p l(X) < H(p) + D(p \parallel q) + 1$$

.

Huffman codes

- We expect that an optimal PF code will have the longest codeword for the two least probable symbols. Otherwise we can delete one bit from the longer one and retain the PF property while decreasing the expected codeword length.
- Algorithm: rank symbols by probability. Combine the two least probable ones, and re-rank the probabilities. Repeat this process until we only have one symbol, then generate the codewords using the binary tree by branching with 0 and 1 at each node.
- WLOG $p_1 \ge p_2 \ge \cdots \ge p_m$. The code is optimal iff $\sum p_i l_i$ is minimised.
- There exists an opitmal code with $l_1 \le l_2 \le \cdots \le l_m$ where $C(m-1)$ and $C(m)$ are siblings that differ only in their last bits.
- The Huffman procedure yields an optimal code.

Channel Capacity

Definition. A **discrete** channel is a system consisting of (1) input alphabet $\mathscr{X}$, (2) output alphabet $\mathscr{Y}$, and (3) probability transition matrix $p_{Y|X}$.

Definition. The channel is **memoryless** if the probaiblity distribution of the output at time i depends only on the input at time i, i.e. for all x^n, y^n,

$$\Pr(Y^n = y^n \mid X^n = x^n) = \prod_{i=1}^n p_{Y|X}(y_i \mid x_i)$$

Definition. The **channel capacity** of a DMC $(\mathscr{X}, \mathscr{Y}, p_{Y|X})$ is

$$C = C(p_{Y|X}) = \max_{p_X} I(X;Y)$$

Definition. The n**-th extension of a DMC** $(\mathscr{X}, p(y \mid x), \mathscr{Y})$ without feedback is the channel $(\mathscr{X}^n, p(y^n \mid x^n), \mathscr{Y}^n)$ where $p(y^n \mid x^n) = \prod_{i=1}^n p(y_i \mid x_i)$ for $(x^n, y^n) \in \mathscr{X}^n \times \mathscr{Y}^n$.

Definition. An (M,n)**-code** for the DMC $(\mathscr{X}, p(y \mid x), \mathscr{Y})$ consists of (1) the message set $\{1, \ldots, M\}$, (2) the encoder $f : \{1, \ldots, M\} \to \mathscr{X}^n$, (3) the decoder $\varphi : \mathscr{Y}^n \to \{1, \ldots, M\}$.

Definition. The **conditional probability of error** of a code (f, φ) of sending a message $w \in [M] = \{1, \ldots, M\}$ is

$$\begin{aligned}\lambda_w &= \Pr(\varphi(Y^n \neq w \mid X^n = w^n(w)) \\ &= \sum_{y^n} p_{Y^n|X^n}(y^n \mid x^n(w)) \mathbf{1}\{\varphi(y^n) \neq w\}\end{aligned}$$

Definition. The **maximal probability of error** of a code (f, φ) is $\lambda_{max}^{(n)} = \max_{w\in[M]} \lambda_w$.

Definition. The **average probability of error** of a code (f, φ) is $P_e^{(n)} = \lambda_{ave}^{(n)} = \frac{1}{M} \sum_{w=1}^M \lambda_w$.

Definition. The **rate** of an (M,n)-code is $R = \frac{1}{n} \log M$ bits per channel use, or $R = \frac{1}{n} \ln M$ nats per channel use.

Definition. A rate $R \ge 0$ is **achievable** for a DMC $p_{Y|X}$ if there exists a sequence of $(2^{nR}, n)$-codes such that $\lambda_{max}^{(n)} \to 0$ as $n \to \infty$. Note that if $R \ge 0$ is achievable, then $R' \le R$ is achievable too.

Definition. The **capacity** of a DMC $p_{Y|X}$ is

$$\tilde{C} = \tilde{C}(p_{Y|X}) = \sup\{R \ge 0 : R \text{ is achievable}\}$$

Theorem. $\tilde{C} = C(p_{Y|X}) = \max_{p_X} I(p_X, p_{Y|X})$

Examples

- Noiseless binary channel, noisy channel with non-overlapping output: $C = 1$
- Binary symmetric channel: $C = 1 - H_b(p)$ when p_X is uniform on $\{0,1\}$
- Binary erasure channel: $C = 1 - \alpha$ when p_X is uniform on $\{0,1\}$
- Symmetric channels (doubly stochastic PTM, rows and columns are permutations of one another respectively): $C = \log|\mathscr{Y}| - H(\mathbf{r})$ where $\mathbf{r}$ is the distribution on one row

Jointly Typical Sequences

Definition. The set $A_\epsilon^{(n)}(X,Y)$ of jointly typical sequences (x^n, y^n) wrt $p_{X,Y}$ is

$$\begin{aligned}A_\epsilon^{(n)} = A_\epsilon^{(n)}(X,Y) = \{&(x^n, y^n) \in \mathscr{X}^n \times \mathscr{Y}^n : \\ &\left| -\frac{1}{n} \log p_{X^n}(x^n) - H(X) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log p_{Y^n}(y^n) - H(Y) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log_{p_{X^n,Y^n}}(x^n, y^n) - H(X,Y) < \epsilon \right| \Big\}\end{aligned}$$

where $p_{X^n,Y^n}(x^n, y^n) = \prod_{i=1}^n p_{X,Y}(x_i, y_i)$.

Theorem (Joint AEP). 1. $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$

2. $\left|A_\epsilon^{(n)}\right| \le 2^{nH(X,Y)}$

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p_{X^n}(x^n) p_{Y^n}(y^n)$, then

$$\begin{aligned}(1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)} &\le \Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon(x,y)) \\ &\le 2^{-n(I(X;Y)-3\epsilon)}\end{aligned}$$

Channel Coding Theorem

Theorem (Direct/Achievability). For a DMC, all rates $R < C$ are achievable. For all $R < C$, there exists a sequence (in $n \to \infty$) of $(2^{nR}, n)_{n\in\mathbb{N}}$-codes with $\lambda_{max}^{(n)} \to 0$.

Theorem (Converse/Impossibility). Conversely, any sequence of $(2^{nR}, n)_{n\in\mathbb{N}}$-codes with $\lambda_{ave}^{(n)} \to 0$ satisfies $R \le C$. (Proof using Fano's inequality)

Proof of achievability

1. Fix $p_X(x)$. Generate the codewords in iid fashion.
2. By symmetry of codebook generation and uniformity of codewords, choose $w = 1$ WLOG. Calculate the error probability and bound it.
3. Choose a rate R such that the bound approaches 0 as $n \to \infty$.

Definition. The **feedback capacity** $C_{FB}(p_{Y|X})$ is the supremum of all achievable rates with feedback codes.

Theorem. $C_{FB} = C = \max_{p_X} I(X;Y)$

Theorem (Source-channel Separation Theorem). If $\{V_n\}$ is a finite-alphabet stationary stochastic process that satisfies AEP and $H(\mathscr{V}) < C$, then there exists a source-channel code (f_n, φ_n) with $P_e^{(n)} \to 0$. This code can be realised by a separation scheme. Conversely, for all stationary stochastic processes $\{V^n\}$ with $H(\mathscr{V}) > C$, $P_e^{(n)} \nrightarrow 0$, i.e. $\limsup_{n\to\infty} P_e^{(n)} > 0$.