Information Theory

Segwang Kim

February 18, 2023

Contents

Now, we assume that **all random variables are discrete**.
For the joint pdf p of r.v.'s X, Y, denote $p(x) = \int p(x,y)dy$, $p(y) = \int p(x,y)dx$ and so on.
Denote $ran(X)$ be a range of a r.v. X.
Denote $X_i^j = (X_i, \ldots, X_j)$, its realization is $x_i^j = (x_i, \ldots, x_j)$

1 Entropy, Relative Entropy, Mutual Information

1.1 Entropy

Definition) Entropy.
X : r.v. with the pdf $p(x)$

$$H(X) = \mathbb{E}_X(\log \frac{1}{p(X)})$$

For $X = i$ w.p. p_i, $i = 1, \ldots, n$,

$$H(\{p_1, \ldots, p_n\}) := H(X)$$

Especially, for $X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$,

$$H(p) := H(X)$$

Proposition) Properties of Entropy.

(i) Shift invariant: $H(X) = H(X + a)$ for $a \in \mathbb{R}$.

(ii) Non-negativity: $H(X) \geq 0$.

(iii) $X \sim U([n])$ where $[n] = \{1, \ldots, n\}$, then $H(X) = \log(n)$.

(iv) $H(X) \leq \log |ran(X)| = H(U)$ where $|ran(X)|$ is the number of elements in the range of X, $U \sim U(ran(X))$.

(v) $H(\{p_i\})$ is concave w.r.t. $\{p_i\}$.

Proof. Consider $D(\{p_i\} \| U) = \log |ran(X)| - H(\{p_i\})$. □

Definition) Joint entropy.
X, Y : r.v.'s with the joint pdf $p(x, y)$

$$H(X, Y) = \mathbb{E}_{X,Y}(\log \frac{1}{p(X, Y)})$$

Proposition) Properties of Joint Entropy.

(i) If X, Y are independent, $H(X, Y) = H(X) + H(Y)$

1.2 Conditional entropy

Definition) Conditional entropy.

X, Y : r.v.'s with the joint pdf $p(x, y)$

$$H(Y|X) = \mathbb{E}_{X,Y}(\log \frac{1}{p(Y|X)})$$

Proposition) Properties of Conditional Entropy.

(i) Non-negativity: $H(Y|X) \geq 0$

(ii) Chain rule: $H(X, Y) = H(X|Y) + H(Y)$

(iii) Chain rule': $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1})$

(iv) $H(X, Y|Z) = H(X|Y, Z) + H(Y|Z)$

(v) $H(X|Y) \leq H(X)$. The equality holds when X, Y are indep.

(vi) For stationary process $\{X_n\}$, i.e. $p(X_i^j) = p(X_{i+1}^{j+1})$, $H(X_n|X_1^{n-1})$ is nonnegative and decreasing, thus it must have limit.

Proof. $H(X_n|X_1^{n-1}) \geq H(X_n|X_2^{n-1}) = H(X_{n-1}|X_1^{n-2}) \geq 0,$ □

(vii) For $g : ran(X) \to \mathbb{R}$, $H(g(X)) \leq H(X)$

Proof. $H(X, g(X)) = H(g(X)) + H(X|g(X)) \geq H(g(X))$, $H(X, g(X)) = H(X) + H(g(X)|X) = H(X)$ □

(viii) $H(Y|X) = 0$ iff Y is a ftn of X

(ix) A sequence of r.v.'s $\{X_i\}$ forms a Markov chain, then, $H(X_0|X_n)$ and $H(X_n|X_0)$ are non-decreasing with n.

Proof. $I(X_0; X_{n-1}) \geq I(X_0; X_n)$. Refer proposition (ii) of 1.2. □

Theorem) Fano's inequality.

Consider r.v.'s X, Y with the joint pdf. Let $P_e = \mathbb{P}(\hat{X}(Y) \neq X)$. Then,

$$P_e \geq \frac{H(X|Y) - 1}{\log |ran(X)|}$$

1.3 Relative entropy

Definition) Relative Entropy (Kullback Leibler distance).

For pdfs $p(x)$, $q(x)$,

$$D(p\|q) = \mathbb{E}_{X\sim p}(\log \frac{p(X)}{q(X)})$$

Proposition) Properties of Relative Entropy.

(i) $D(p\|q) \geq 0$. The equality holds when $p = q$ w.p. 1.

Proof. Use Jensen inequality. □

(ii) $D(p\|q)$ is convex in the pair of (p, q), i.e.
For $\lambda \in [0, 1]$, pairs of pdfs (p, q), (p', q'),

$$D(\lambda p + (1-\lambda)p' \| \lambda q + (1-\lambda)q') \leq \lambda D(p\|q) + (1-\lambda)D(p'\|q') \tag{1}$$

Proof.

$$\begin{aligned}\lambda D(p\|q) + (1-\lambda)D(p'\|q') &= \sum_x (\lambda p(x) \log(\frac{p(x)}{q(x)}) + (1-\lambda)p'(x)\log(\frac{p'(x)}{q'(x)})) \\ &= \sum_x (\lambda p(x) \log(\frac{\lambda p(x)}{\lambda q(x)}) + (1-\lambda)p'(x)\log(\frac{(1-\lambda)p'(x)}{(1-\lambda)q'(x)}))\end{aligned}$$

Note that $\sum_i^n a_i \log(\frac{a_i}{b_i}) \geq (\sum_i^n a_i) \log(\frac{\sum_i^n a_i}{\sum_i^n b_i})$ ($\because t \mapsto t\log t$ is convex). Apply this for each term of the above summation. □

Definition) Conditional Relative Entropy.

For pdfs $p(x|y)$, $q(x|y)$,

$$D(p(x|y)\|q(x|y)) = \mathbb{E}_{X,Y\sim p}(\log \frac{p(X|Y)}{q(X|Y)})$$

Proposition) Properties of Conditional Relative Entropy.

(i) $D(p(x,y)\|q(x,y)) = D(p(y)\|q(y)) + D(p(x|y)\|q(x|y))$

1.4 Mutual Information

Definition) Mutual Information.

X, Y : r.v.'s. with the joint pdf $p(x, y)$.

$$\begin{aligned} I(X;Y) &= D(p(x,y)\|p_X(x)p_y(y)) = \mathbb{E}_{X,Y\sim p}(\log(\frac{p(X,Y)}{p(X)p(Y)})) \\ &= H(X) - H(X|Y) \end{aligned}$$

Proposition) Properties of Mutual Information.

(i) $I(X;Y) \geq 0$.

(ii) $I(X;Y) = 0$ iff X, Y are indep.

(iii) $I(X;Y)$ is concave w.r.t. $p(x)$ for fixed $p(y|x)$.

Proof.

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

First, $H(Y)$ is concave w.r.t. $p(x)$ for fixed $p(y|x)$. Indeed, $H(Y)$ is concave w.r.t. $p(y) = \{p_{y,1}, \ldots, p_{y,n}\}$ and $p(y)$ is linear w.r.t. $p(x) = \{p_{x,1}, \ldots, p_{x,m}\}$ since $p_{y,i} = \sum_x p(Y = y_i|x)p(x)$. Second, $H(Y|X)$ is convex w.r.t. $p(x)$ for fixed $p(y|x)$. Indeed, $H(Y|X) = \sum_{x,y} -p(x,y)\log(p(y|x)) = \sum_x p(x)(\sum_y -p(y|x)\log(p(y|x)))$ is linear w.r.t. $p(x)$. □

(iv) $I(X;Y)$ is convex w.r.t. $p_{Y|X}(y|x)$ for fixed $p_X(x)$. i.e.,
Given $\lambda \in (0,1)$, $p_{Y|X;0}(y|x)$, $p_{Y|X;1}(y|x)$,

$$I_{(X,Y)\sim p_{X,Y;\lambda}}(X;Y) \leq \lambda I_{(X,Y)\sim p_{X,Y;0}}(X,Y) + (1-\lambda)I_{(X,Y)\sim p_{X,Y;1}}(X,Y) \tag{2}$$

where $p_{Y|X;\lambda}(y|x) = \lambda p_{Y|X;0}(y|x) + (1-\lambda)p_{Y|X;1}(y|x)$.

Proof. Note that $p_{X,Y;\lambda}(x,y) = p_X(x)p_{Y|X;\lambda}(y|x)$. Then,

$$\begin{aligned} I_{(X,Y)\sim p_{X,Y;\lambda}}(X;Y) &= \mathbb{E}_{(X,Y)\sim p_{X,Y;\lambda}} \log \frac{p_{X,Y;\lambda}(X,Y)}{p_{X;\lambda}(X)p_{Y;\lambda}(Y)} \\ &= D(p_{X,Y;\lambda}(x,y)\|p_{X;\lambda}(x)p_{Y;\lambda}(y)) \end{aligned}$$

Now, we need to compute $p_{X,Y;\lambda}(x,y)$ and $p_{X;\lambda}(x)p_{Y;\lambda}(y)$.

$$\begin{aligned} p_{X,Y;\lambda}(x,y) &= p_{X;\lambda}(x)p_{Y|X;\lambda}(y|x) \\ &= p_X(x)p_{Y|X;\lambda}(y|x) \\ &= p_X(x)(\lambda p_{Y|X;0}(y|x) + (1-\lambda)p_{Y|X;1}(y|x)) \\ &= \lambda p_{X,Y;0}(x,y) + (1-\lambda)p_{X,Y;1}(x,y) \end{aligned}$$

Also,

$$\begin{aligned}
p_{X;\lambda}(x)p_{Y;\lambda}(y) &= \int p_{X,Y;\lambda}(x,y)dy \int p_{X,Y;\lambda}(x,y)dx \\
&= \int p_{X;\lambda}(x)p_{Y|X;\lambda}(y|x)dy \int p_{X;\lambda}(x)p_{Y|X;\lambda}(y|x)dx \\
&= p_X(x) \int p_{Y|X;\lambda}(y|x)dy \int p_{X;\lambda}(x)p_{Y|X;\lambda}(y|x)dx \\
&= p_X(x) \int p_{X;\lambda}(x)(\lambda p_{Y|X;0}(y|x) + (1-\lambda)p_{Y|X;1}(y|x))dx \\
&= p_X(x)(\lambda p_{Y;0}(y) + (1-\lambda)p_{Y;1}(y)) \\
&= \lambda p_X(x)p_{Y;0}(y) + (1-\lambda)p_X(x)p_{Y;1}(y)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&I_{(X,Y)\sim p_{X,Y;\lambda}}(X;Y) = D(p_{X,Y;\lambda}(x,y)\|p_{X;\lambda}(x)p_{Y;\lambda}(y)) \\
&\quad = D(\lambda p_{X,Y;0}(x,y) + (1-\lambda)p_{X,Y;1}(x,y)\|\lambda p_X(x)p_{Y;0}(y) + (1-\lambda)p_X(x)p_{Y;1}(y)) \\
&\quad \leq \lambda D(p_{X,Y;0}(x,y)\|p_X(x)p_{Y;0}(y)) + (1-\lambda)D(p_{X,Y;1}(x,y)\|p_X(x)p_{Y;1}(y)) \quad (\because (1)) \\
&\quad \leq \lambda I_{(X,Y)\sim p_{X,Y;0}}(X,Y) + (1-\lambda)I_{(X,Y)\sim p_{X,Y;1}}(X,Y)
\end{aligned}$$

□

Definition) Conditional Mutual Information.

X, Y, Z : r.v.'s. with the joint pdf $p(x,y,z)$.

$$\begin{aligned}
I(X;Y|Z) &= \mathbb{E}_{X,Y,Z\sim p}(\log(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)})) \\
&= H(X|Z) - H(X|Y,Z)
\end{aligned}$$

Proposition) Properties of Conditional Mutual Information.

(i) $I(X;Y|Z) \geq 0$

Proof.

$$\begin{aligned}
I(X;Y|Z) &= \mathbb{E}_{X,Y,Z\sim p}(\log(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)})) \\
&= \mathbb{E}_{Z\sim p}[\mathbb{E}_{X,Y\sim p_{X,Y|Z}}(\log(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}))] \geq 0
\end{aligned}$$

□

(ii) Chain rule: $I(X_1^n;Y) = \sum_{i=1}^n I(X_i;Y|X_1^{i-1})$

Theorem) Data processing Inequality.

R.v.'s $X \to Y \to Z$ form a Markov chain. i.e. $p(z|x,y) = p(z|y)$, then,

$$I(X;Y) \geq I(X;Z)$$

This means, no clever manipulation of the data can improve the inferences that can be made from the data.

Proof. $I(X;Y) - I(X;Z) = I(X;Y|Z) \geq 0$ □

Corollary) In particular,.

(i) If $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$

(ii) If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$

Exercise) Some examples of Conditional Mutual Information.

a) $I(X;Y|Z) < I(X;Y)$ if $X \sim Ber(1/2)$, $X = Y = Z$

b) $I(X;Y|Z) > I(X;Y)$ if $X, Y \overset{i.i.d.}{\sim} Ber(1/2)$, $Z = X + Y$

2 Asymptotic Equipartition Property (AEP)

2.1 AEP

Theorem) (AEP).

X_i : i.i.d. r.v.'s with pdf p

$$-\frac{1}{n}\log p(X_1,\ldots,X_n) \to H(X) \quad \text{a.s.}$$

Definition) Typical set.

The typical set $A_\epsilon^{(n)}$ is

$$A_\epsilon^{(n)} = \{(x_1\ldots,x_n) \ : \ |-\frac{1}{n}\log p(x_1\ldots,x_n) - H(X)| < \epsilon\}$$

Proposition) Properties of Typical sets.

(i) For $x_1^n \in A_\epsilon^{(n)}$, $2^{-n(H(X)+\epsilon)} \le p(x_1^n) \le 2^{-n(H(X)-\epsilon)}$.

(ii) $\mathbb{P}(X \in A_\epsilon^{(n)}) \ge 1-\epsilon$ for sufficiently large n.

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

Proof. $1 = \sum_{x_1^n} p(x_1^n) \ge \sum_{x_1^n\in A_\epsilon^{(n)}} p(x_1^n) \ge \sum_{x_1^n\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = |A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}$ □

(iv) $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ for sufficiently large n

Proof. $1-\epsilon < \mathbb{P}(X_1^n \in A_\epsilon^{(n)}) = \sum_{x_1^n\in A_\epsilon^{(n)}} p(x_1^n) \le |A_\epsilon^{(n)}|2^{-n(H(X)-\epsilon)}$ for sufficiently large n □

Theorem) Implication of AEP to data compression.

X_i : i.i.d. r.v.'s with pdf p. There exists a data compression code (bijection) s.t. for $\epsilon > 0$

$$\mathbb{E}(\frac{1}{n}l(X_1^n)) < H(X_1^n) + \epsilon$$

where $l(X_1^n) = \sum_{X_i}$ (length of the code for X_i)$= \sum_{X_i} l(X_i)$, $X_1^n = (X_1,\ldots,X_n)$

Proof. For $X_1^n \in A_\epsilon^{(n)}$, encode it by $nH(X_1)+\epsilon+2$ bits. Otherwise, by $n\log(|ran(X_1)|)+2$ bits. It means, encode naively. (the number of possible outcome= $|ran(X_1)|^n$)

$$\begin{aligned}\mathbb{E}(l(X_1^n)) &= \sum_{x_1^n\in A_\epsilon^{(n)}} p(x_1^n)l(x_1^n) + \sum_{x_1^n\notin A_\epsilon^{(n)}} p(x_1^n)l(x_1^n)\\ &= \mathbb{P}(X_1^n \in A_\epsilon^{(n)})(nH(X_1)+\epsilon+2) + \mathbb{P}(X_1^n \notin A_\epsilon^{(n)})(n\log(|ran(X_1)|)+2)\\ &\le (nH(X_1)+\epsilon+2) + \epsilon(n\log(|ran(X_1)|)+2)\end{aligned}$$

□

3 Entropy Rates

3.1 Entropy rates

Definition) Entropy rates.

The entropy rate of a r.p. $\mathcal{X} = \{X_i\}$ is

$$H(\mathcal{X}) = \lim_n \frac{1}{n} H(X_1^n) = \lim_n \frac{1}{n} H(X_1, \ldots, X_n)$$

provided the limit exists.
Alternatively (in case of $\mathcal{X}$ is stationary),

$$H'(\mathcal{X}) = \lim_n H(X_n | X_1^{n-1})$$

provided the limit exists.

Theorem) Two definitions coincide in case of stationary distribution.

If $\mathcal{X}$ is stationary, then $H(\mathcal{X}) = H'(\mathcal{X})$, i.e.

$$\lim_n \frac{1}{n} H(X_1^n) = \lim_n H(X_n | X_1^{n-1})$$

Proof. $\frac{1}{n} H(X_1^n) = \frac{1}{n} \sum_{i=1}^n H(X_i | X_1^i) = \lim_n H(X_n | X_1^{n-1})$ by Cesaro sum. □

3.2 Markov Process

Definition) Markov Process.

A r.p. $\mathcal{X} = \{X_i\}$ is a Markov process (m.p.) if

$$\mathbb{P}(X_n = x_n | X_1^{n-1} = x_1^{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$$

for all n.

A m.p. $\mathcal{X} = \{X_i\}$ is stationary (s.m.p.) if $\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$ is indep of n. $\rightarrow$ $H(\mathcal{X}) = H(X_2 | X_1)$.

Transition matrix M for a m.p. $\mathcal{X} = \{X_i\}$ with $ran(X) = [m] = \{1, \ldots, m\}$ is

$$M = [p_{ij}]_{1 \leq i,j \leq m} \quad \text{where} \;\; p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$$

Denote $M^n = [p_{ij}^{(n)}]$.

A m.p. $\mathcal{X} = \{X_i\}$ is irreducible if there exists $m \in \mathbb{N}$ s.t. $\forall i, j \in [m]$, $\exists n \in \{0\} \cup [m]$ with $p_{i,j}^{(n)} > 0$.

A m.p. $\mathcal{X} = \{X_i\}$ is aperiodic if for given $N \in \mathbb{N}$, $\forall i, j \in [m]$, $\exists n > N$ with $p_{ij}^{(n)} > 0$.
$\rightarrow$ (Aperiodic $\subset$ Irreducible)

A stationary distribution μ for a m.p. $\mathcal{X} = \{X_i\}$ satisfies $\mu = \mu M$

Theorem) Entropy rate of s.m.p..

If $\mathcal{X}$ is s.m.p., then.

$$H(\mathcal{X}) = -\sum_{ij} \mu_i p_{ij} \log p_{ij}$$

Proof. Since it is stationary and Markov, $H(\mathcal{X}) = \lim_n H(X_n|X_1^{n-1}) = \lim_n H(X_n|X_{n-1})$. So, $\lim_n H(X_n|X_{n-1}) = H(X_2|X_1 = \mu) = \mathbb{E}_{X_1\sim\mu}(\mathbb{E}_{X_2|X_1\sim p(x_2|x_1)}(\frac{1}{\log p(X_2|X_1)}))$ where μ is a stationary distribution. □

Exercise) A few examples.

a) For a m.p. with transition matrix $M = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$,
A stationary dist. is $\mu = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$
$H(\mathcal{X}) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta) \leq H(\mu) = H(\frac{\alpha}{\alpha+\beta})$

3.3 Hidden Markov Models

Definition) Markov Process.

A r.p. $\mathcal{Y} = \{Y_i\}$ is a Hidden Markov process (h.m.p.) if $Y_i = \phi(X_i)$ for some $\phi : \mathbb{R} \to \mathbb{R}$ and a m.p. $\{X_i\}$

$\mathcal{Y}$ is stationary but not necessarily a m.p..

Lemma) Initial conditioning reduces entropy.

$\mathcal{Y} = \{Y_i\}$ is a h.m.p. associated with a m.p. $\{X_i\}$. Then,

$$H(Y_n|Y_1^{n-1}, X_1) \leq H(\mathcal{Y})$$

Proof.

$$\begin{aligned} H(Y_n|Y_1^{n-1}, X_1) &= H(Y_n|Y_1^{n-1}, X_1) \\ &= H(Y_n|Y_1^{n-1}, X_1, X_{-k}^0) \quad (\because \text{Markov property}) \\ &= H(Y_n|Y_1^{n-1}, Y_{-k}^0, X_1, X_{-k}^0) \quad (\because \mathcal{Y} = \{Y_i\} \text{ is a h.m.p.}) \\ &\leq H(Y_n|Y_{-k}^{n-1}) = H(Y_{n+k+1}|Y_1^{n+k}) \to H(\mathcal{Y}) \quad \text{as } k \to \infty \end{aligned}$$

□

Lemma) Initial conditioning approaches to the entropy rates.

$\mathcal{Y} = \{Y_i\}$ is a h.m.p. associated with a m.p.. $\{X_i\}$. $H(X_1) < \infty$. Then,

$$H(Y_n|Y_1^{n-1}) - H(Y_n|Y_1^{n-1}, X_1) \to 0 \quad \text{as } n \to \infty$$

Proof.

$$H(Y_n|Y_1^{n-1}) - H(Y_n|Y_1^{n-1}, X_1) = I(X_1; Y_n|Y_1^{n-1})$$

Since $H(X_1) \geq I(X_1; Y_1^n) = \sum_{i=1}^n I(X_1; Y_i|Y_1^{i-1})$, it follows that $I(X_1; Y_n|Y_1^{n-1}) \to 0$ as $n \to \infty$ □

Theorem) Initial conditioning approaches to the entropy rates.

$\mathcal{Y} = \{Y_i\}$ is a h.m.p. associated with a m.p.. $\{X_i\}$. $H(X_1) < \infty$. Then,

$$H(Y_n|Y_1^{n-1}, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_1^{n-1})$$
$$\lim H(Y_n|Y_1^{n-1}, X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_1^{n-1})$$

4 Data Compression

4.1 Data Compression

Denote $\mathcal{D}$ be a set of alphabets. Its size is $D = |\mathcal{D}|$
Denote $\mathcal{D}^*$ be the set of finite length strings of $\mathcal{D}$.

Definition) Codeword.

For a r.v. X, The source code is $C : ran(X) \to \mathcal{D}^*$.

The expected length $L(C)$ of a source code C is given by

$$L(C) = \mathbb{E}(l(X)) = \sum_x p(x)l(x)$$

where $l(x)$ is the length of $C(x)$

A source code is nonsingular if it is injective.

The extension of a source code $C : ran(X) \to \mathcal{D}^*$ is $C^* : ran(X)^* \to \mathcal{D}^*$ defined by concatenating codewords, i.e.

$$C^*(x_1^n) = C(x_1) \dots C(x_n)$$

for every $n \geq 0$ and $x_1^n \in ran(X)^n$

A source code $C : ran(X) \to \mathcal{D}^*$ is uniquely decodable (UD) if its extension C^* is nonsingular.

A source code is a prefix code if no codeword is a prefix of any other codeword.

Theorem) Kraft Inequality.

If C is a prefix code, then

$$\sum_i D^{-l_i} \leq 1$$

(This sum is called Kraft sum)
Conversely, given $\{l_i\}$ satisfying the above inequality, there exists a prefix code with these word lengths.

Proof. ($\Rightarrow$) Consider a D-ary full tree T with the depth $l_{\text{max}} = \max_i l_i$. Given codewords $\{C(x_i)\}$, we can find the corresponding subset nodes $\{v_i\} \subset T$ satisfying that none of nodes on the path from the root to v_i is v_j node. Therefore, v_i have $D^{l_{\text{max}} - l_i}$ descendents in T, each of those descendents is disjoint. So, $\sum_i D^{l_{\text{max}} - l_i} \leq D^{l_{\text{max}}}$.
($\Leftarrow$) Grow a D-ary full tree T with the depth $l_{\text{min}} = \min_i l_i$. □

Theorem) The expected length of a prefix code.

If C is a prefix code associated with a r.v. X on $\mathcal{D}$, then

$$L(C) \geq H_D(X) = \sum_x p(x) \log_D \frac{1}{p(x)}$$

Proof. Consider a prob. dist. $\{q_i\}$ over $ran(X)$ where $q_i = \frac{D^{-l_i}}{\sum_i D^{-l_i}}$. Then, $KL_D(\{p_i\} \| \{q_i\}) = -H_D(\{p_i\}) + L(C) + \log_D(K) \geq 0$ with log-base D where $K = \sum_i D^{-l_i}$. The conclusion follows by Kraft Inequality. Furthermore, the equality holds when $K = 1$, $p_i = q_i = D^{-l_i}$. □

4.2 Shannon Coding

Definition) D-adic. A pmf is D-adic if each of the probabilities is equal to D^{-n} for some $n \in \mathbb{N}$

Definition) Shannon Coding.
For a r.v. X, Shannon coding $C : ran(X) \to \mathcal{D}^*$ is a code satisfying $l_i = \lceil \log_D \frac{1}{p_i} \rceil$.

Proposition) Properties of Shannon Coding.

(i) Sub-optimal

(ii) prefix code ($\because$ it satisfies Kraft inequality)

(iii) $H_D(X) \le L(C) < H_D(X) + 1$ ($\because \log_D \frac{1}{p_i} \le l_i < 1 + \log_D \frac{1}{p_i}$)

Theorem) Optimal prefix codeword length.
If C^* is an optimal prefix code associated with a r.v. X on $\mathcal{D}$, then

$$H_D(X) \le L(C^*) < H_D(X) + 1$$

Proof. C^* should be better than Shannon code. Also, C^* is a prefix code. □

Theorem) The minimum average code length.
If C^* is an optimal prefix code associated with a r.v.'s $\{X_i\}$ on $\mathcal{D}$, then

$$\frac{1}{n} H_D(X_1^n) \le L_n(C^*) = \mathbb{E}(\frac{1}{n} l^*(X_1^n)) < \frac{1}{n} H_D(X_1^n) + \frac{1}{n}$$

If $\mathcal{X} = \{X_i\}$ is stationary,

$$L_n(C^*) = \mathbb{E}(\frac{1}{n} l^*(X_1^n)) \to H_D(\mathcal{X})$$

Theorem) The comparison of average code length.
If C is a prefix code associated with a r.v.' $X \sim p$ on $\mathcal{D}$ s.t. $l_C(x) = \lceil \log \frac{1}{q(x)} \rceil$ for some pmf q, then

$$H_D(p) + KL(p\|q) \le \mathbb{E}_{X\sim p}(l_C(X)) < H_D(p) + KL(p\|q) + 1$$

Proof.

$$\begin{aligned}\mathbb{E}_{X\sim p}(l_C(X)) = \sum p(x) \lceil \log \frac{1}{q(x)} \rceil < \sum p(x)(\log \frac{1}{q(x)} + 1) \\ = \sum p(x)(\log \frac{p(x)}{q(x)p(x)} + 1) = H_D(p) + KL(p\|q) + 1\end{aligned}$$

Similarly, the lower bound can be proven. □

4.3 Huffman Coding

Definition) Huffman Coding.

For a r.v. X, Huffman coding $C : ran(X) \to \mathcal{D}^*$ is a code satisfying ...

Lemma) Characterization of Huffman Coding.

For a r.v. X, there exists an optimal prefix code that satisfies

1. If $p_i > p_j$, then $l_i < l_j$.
2. The two longest codewords have the same length.
3. The two longest codewords differ only in the last bit (, and corresponds to the two least likely symbols).

Proof. Consider a corresponding tree. We can improve $\mathbb{E}(l(X))$ by swapping, rearranging and trimming. □

Proposition) Properties of Huffman Coding.

(i) Optimal

Proof. By recursion through merging the two longest codewords. □

(ii) $H_D(X) \leq L(C) < H_D(X) + 1$

4.4 Shannon-Fano-Elias Coding (Alphabetic code)

Definition) Shannon-Fano-Elias coding.

For a r.v. X with pmf p, Shannon-Fano-Elias (S.F.E) coding $C : ran(X) \to \mathcal{D}^*$ is constructed by following steps.

1. Define $\bar{F} : ran(X) \to [0,1] : x \mapsto \sum_{a<x} p(a) + \frac{1}{2}p(x)$
2. Let $l(x)$ be the integer $\left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$
3. Let $C(x)$ be the first $l(x)$ most significant bits after the decimal point of the binary expansion of $\bar{F}(x)$ i.e. $\lfloor \bar{F}(x) \rfloor_{l(x)}$.

Proposition) Properties of S.F.E Coding.

(i) Nonsingular

Proof. It is enough to show that $\lfloor \bar{F}(a_i) \rfloor_{l(a_i)}$ are distinct where $\{a_i\} = ran(X)$. Note that $F(a_i) > \bar{F}(a_i) \geq \lfloor \bar{F}(a_i) \rfloor_{l(a_i)}$. Claim that $\lfloor \bar{F}(a_i) \rfloor_{l(a_i)} > F(a_{i-1})$. Obviously, $\lfloor \bar{F}(a_i) \rfloor_{l(a_i)} \geq \bar{F}(a_i) - \frac{1}{2^{l(a_i)}}$. Also, $\bar{F}(a_i) = F(a_{i-1}) + \frac{1}{2}p(a_i) \geq F(a_{i-1}) + \frac{1}{2^{l(a_i)}}$ since $l(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$. Therefore, $F(a_i) > \lfloor \bar{F}(a_i) \rfloor_{l(a_i)} > F(a_{i-1})$ □

(ii) S.F.E coding is prefix free

(iii) $L(C) < H(X) + 2$

Proof. $L(C) = \mathbb{E}(l(C(X)) = \sum_x p(x)l(x) = \sum_x p(x)(\lceil \log_2 \frac{1}{p(x)} \rceil + 1) < H(X) + 2$ $\square$

4.5 Channel Capacity

5 Channel Capacity

Definition) Channel Capacity.

A discrete channel is a system $(X, p(Y|X), Y)$ consisting of an input r.v. X and output r.v. Y, and fixed $p(Y|X)$

Information of channel capacity is

$$C = \max_{p(X)} I(X;Y)$$

Proposition) Properties of Channel Capacity.

(i) $C \geq 0$

(ii) $C \leq \log(|ran(X)|)$, $C \leq \log(|ran(Y)|)$

(iii) C is concave w.r.t. $p(X)$

Definition) Symmetric Channel.

A channel is symmetric if the rows and the columns of the transition matrix $p(Y|X)$ are permutations with each other

Proposition) Properties of Symmetric Channel.

(i) $C = \max_{p(X)} I(X;Y) = \max_{p(X)}(H(Y) - H(r)) \leq \log|ran(Y)| - H(r)$ where r is a row of the transition matrix.

Definition) Discrete Memoryless channel.

A channel is memoryless if the prob. dist. of the output depends only on the input at the time.

The n-th extension of the discrete memoryless channel (DMC) is $(X_1^n, p(Y_1^n|x_1^n), Y_1^n)$ where $p(Y_k|x_1^k, y_1^{k-1}) = p(Y_k|x_1^k)$

Definition) Jointly typical sequences.

The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x_1^n, y_1^n)\}$ is defined as

$$A_\epsilon^{(n)} = \{(x_1^n, y_1^n) \mid \max(|-\frac{1}{n}\log p(x_1^n) - H(X)|, |-\frac{1}{n}\log p(y_1^n) - H(Y)|$$
$$, |-\frac{1}{n}\log p(x_1^n, y_1^n) - H(X,Y)|) < \epsilon\}$$

where $p(x_1^n, y_1^n) = \prod_{i=1}^n p(x_i, y_i)$

Theorem) Joint AEP.

Let (X_1^n, Y_1^n) be i.i.d. sequences from $p(x_1^n, y_1^n) = \prod_{i=1}^n p(x_i, y_i)$. Then,

1. $\mathbb{P}((X_1^n, Y_1^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
3. If $(\tilde{X_1^n}, \tilde{Y_1^n}) \sim p(x_1^n)p(y_1^n)$,
$$\mathbb{P}((\tilde{X_1^n}, \tilde{Y_1^n}) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$$
For sufficiently large n,
$$\mathbb{P}((\tilde{X_1^n}, \tilde{Y_1^n}) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Proof. 1 and 2 are obvious. For 3, $\mathbb{P}((\tilde{X_1^n}, \tilde{Y_1^n}) \in A_\epsilon^{(n)}) = \sum_{(\tilde{x}_1^n, \tilde{y}_1^n) \in A_\epsilon^{(n)}} p(\tilde{x}_1^n, \tilde{y}_1^n) = \sum_{(\tilde{x}_1^n, \tilde{y}_1^n) \in A_\epsilon^{(n)}} p(\tilde{x}_1^n)p(\tilde{y}_1^n)$. By 2, we can bound the number of terms in the summation. By definition of $A_\epsilon^{(n)}$, we can bound the each probability term. □

Definition) (M,n).

An (M, n) code consists of

1. An index set $I = \{1, \ldots, M\}$.
2. An encoding ftn $x_1^n : I \to \Omega_x^n$. This is determined by realizations of r.v. $X(w)$ n times for each $w \in I$. So, $X_1(w), \ldots, X_n(w)$ are i.i.d. r.v.'s. Denote their realization as $x_1(w), \ldots, x_n(w)$. We will determine which realizations define $x_1^n(w)$ in later.
3. A DMC $(x_1^n(w), p(Y_1^n|x_1^n(w)), Y_1^n)$. This generates a r.v. Y_1^n for given $x_1^n(w)$.
4. A decoding ftn $g : \Omega_y^n \to I$.
Since every y_1^n is always generated for given $x_1^n(w)$, a decoding ftn g can acknowledge $x_1^n(w)$. But we omit for the sake of brevity. i.e. g is a ftn of $x_1^n(w)$, as well as y_1^n.

The probability of error at input code $x_1^n(w)$ is

$$\begin{aligned}\lambda_w(x_1^n(w)) &= \mathbb{E}_{Y_1^n \sim p(\cdot|x_1^n)}(I(g(y_1^n) \neq w)) = \mathbb{P}(g(Y_1^n) \neq w | x_1^n(w)) \\ &= \sum_{y_1^n} p(y_1^n|x_1^n(w)) I(g(y_1^n) \neq w)\end{aligned}$$

The maximal probability of error at input code x_1^n is

$$\lambda^{(n)}(x_1^n) = \max_w \lambda_w(x_1^n(w))$$

The average probability of error at input code x_1^n is

$$P_e^{(n)}(x_1^n) = \mathbb{E}_{W \sim U([2^{nR}])} \lambda_W(x_1^n(W)) = \frac{1}{M} \sum_{w=1}^{M} \lambda_w(x_1^n(w))$$

The average probability of error is

$$P_e^{(n)} = \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} \lambda_W(X_1^n(W))$$

The rate R of an (M, n) code is

$$R = \frac{\log M}{n}$$

A rate R is achievable if there exists sequence of $(\lceil 2^{nR} \rceil, n)$ code s.t. $\lambda^{(n)} \to 0$ as $n \to \infty$

The capacity of a discrete memoryless channel is the supremum of all achievable rates.

Theorem) Channel Coding Theorem.

For every $\delta > 0$, $R < C$, there exist $(2^{nR}, n)$ code with $P_e^{(n)} < \delta$. Conversely, any sequence of $(2^{nR}, n)$ code with $P_e^{(n)} \to 0$ must have $R \leq C$

i.e. $(2^{nR}, n)$ code is achievable iff $R \leq C$.

Proof. First, consider i.i.d. r.v.'s $X_1(w), \ldots, X_n(w)$ for each $w \in [2^{nR}] = \{1, \ldots, 2^{nR}\}$ where $p(X_1^n(w))$ maximizes $I(X;Y)$. The number of observation n will be determined later. From the observation, we have a codebook

$$C = \begin{pmatrix} x_1(1) & x_2(1) & \ldots & x_n(1) \\ \vdots & \vdots & \ldots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \ldots & x_n(2^{nR}) \end{pmatrix} = \begin{pmatrix} x_1^n(1) \\ \vdots \\ x_1^n(2^{nR}) \end{pmatrix}$$

Fix $\epsilon > 0$ s.t. $4\epsilon < \delta$ and $R < I(X;Y) - 3\epsilon$ ($\because R < C$).
Define $E_w = \{(x_1^n(w), y_1^n) \in A_\epsilon^{(n)}\}$ for each $w \in [2^{nR}]$
Define a decoding ftn $g : ran(Y)^n \to I$ by followings.

$$g(y_1^n) = g_{x_1^n}(y_1^n) = \begin{cases} w' & \text{if } \exists!\ w' \in [2^{nR}] \text{ s.t. } (x_1^n(w'), y_1^n) \in E_{w'} \\ 2 & \text{o.w.} \end{cases}$$

Note that the second case is no matter what value you assign.
Therefore, the expected number of error (or probability of error) is

$$\begin{aligned} P_e^{(n)} &= \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} \mathbb{E}_{Y_1^n \sim p(\cdot | X_1^n(W))} (I_{g(Y_1^n) \neq W}) \\ &= \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} \mathbb{P}(g(Y_1^n) \neq W | X_1^n(W)) \\ &= \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} (\lambda_W(X_1^n(W))) \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \mathbb{E}_{X_1^n(w)} (\lambda_w(X_1^n(w))) \\ &= \mathbb{E}_{X_1^n(1)} \lambda_1(X_1^n(1)) \quad (\because \text{symmetry of code construction}) \\ &= \sum_{x_1^n(1)} \mathbb{P}(x_1^n(1)) \lambda_1(x_1^n(1)) \\ &= \sum_{x_1^n(1)} \mathbb{P}(x_1^n(1)) \cdot \mathbb{P}(g(Y_1^n) \neq 1 | x_1^n(1)) \end{aligned}$$

By the definition of g,

$$
\begin{aligned}
P_e^{(n)} &= \sum_{x_1^n(1)} \mathbb{P}(x_1^n(1)) \cdot \mathbb{P}(g(Y_1^n) \neq 1 | x_1^n(1)) \\
&= \sum_{x_1^n(1)} \mathbb{P}(x_1^n(1)) \cdot \mathbb{P}(\neg(\exists!\ 1 \in [2^{nR}] \text{ s.t. } (x_1^n(1), y_1^n) \in E_1) | x_1^n(1)) \\
&= \sum_{x_1^n(1)} \mathbb{P}(x_1^n(1)) \cdot \mathbb{P}((x_1^n(1), y_1^n) \notin E_1 \vee (x_1^n(1), y_1^n) \in E_2 \vee \cdots \vee (x_1^n(1), y_1^n) \in E_{2^{nR}} | x_1^n(1)) \\
&= \mathbb{P}((X_1^n(1), Y_1^n) \notin E_1 \vee (X_1^n(1), Y_1^n) \in E_2 \vee \cdots \vee (X_1^n(1), Y_1^n) \in E_{2^{nR}}) \\
&\leq \mathbb{P}_{X_1^n(1), Y_1^n}(E_1^c) + \mathbb{P}_{X_1^n(1), Y_1^n}(E_2) + \cdots + \mathbb{P}_{X_1^n(1), Y_1^n}(E_{2^{nR}}) \\
&\leq \epsilon + \mathbb{P}_{X_1^n(1), Y_1^n}(E_2) + \cdots + \mathbb{P}_{X_1^n(1), Y_1^n}(E_{2^{nR}}) \quad \text{for sufficiently large } n \\
&\leq \epsilon + 2^{-n(I(X;Y) - 3\epsilon - R)} \quad (\because p_{X_1^n(1)} \perp p_{Y_1^n | X_1^n(w)}\ \forall w \neq 1,\ \text{AEP 3}) \\
&\leq 2\epsilon \quad \text{for sufficiently large } n \text{ since } R < I(X;Y) - 3\epsilon
\end{aligned}
$$

Conversely, we need to show that $P_e^{(n)} \to 0$ implies $R \leq C$. First, we show Fano's inequality.

Lemma) Fano's inequality.

For a DMC, assume $W \sim U([2^{nR}])$. Let $P_e^{(n)} = \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} \lambda_W(X_1^n(W))$. Then,

$$H(X_1^n | Y_1^n) \leq 1 + P_e^{(n)} nR \tag{3}$$

or,

$$H(W | Y_1^n) \leq H(\{P_e^{(n)}, 1 - P_e^{(n)}\}) + P_e^{(n)} \log(|2^{nR}| - 1) \tag{4}$$

(Note that $H(X_1^n | Y_1^n)$ needs integration w.r.t. W, $X_1^n(W)$, Y_1^n))

Proof. Let's start from data processing inequality $H(X_1^n | Y_1^n) \leq H(W | Y_1^n)$ since $W \to X \to Y$. Define $E_{W, Y_1^n} = I(g(Y_1^n) \neq W)$ be a ftn of W and Y_1^n. Note that when we integrate E_{W, Y_1^n}, we sequentially generate $W \sim U(2^{nR})$, $X_1^n(W)$ and $Y_1^n \sim p(\cdot | X_1^n(W))$.
Consider

$$H(E_{W, Y_1^n}, W | Y_1^n) = H(W | Y_1^n) + H(E_{W, Y_1^n} | W, Y_1^n) = H(W | Y_1^n) + 0$$

Hence, $H(X_1^n(W) | Y_1^n) \leq H(W | Y_1^n) = H(E_{W, Y_1^n}, W | Y_1^n) = H(E_{W, Y_1^n} | Y_1^n) + H(W | E_{W, Y_1^n}, Y_1^n)$.
For the first term,

$H(E_{W, Y_1^n} | Y_1^n) \leq H(E_{W, Y_1^n}) \leq 1 \quad (\because E \text{ is a binary r.v..})$

For the second term,

$$
\begin{aligned}
H(W | E_{W, Y_1^n}, Y_1^n) &= \mathbb{E}_{W \sim U([2^{nR}])} (\mathbb{P}(E_{W, Y_1^n} = 0) H(W | Y_1^n, E_{W, Y_1^n} = 0) \\
&\qquad + \mathbb{P}(E_{W, Y_1^n} = 1) H(W | Y_1^n, E_{W, Y_1^n} = 1)) \\
&\qquad (\mathbb{P},\ H \text{ integrate w.r.t. } X_1^n, Y_1^n) \\
&\leq 0 + \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} (\mathbb{P}(g(Y_1^n) \neq W | X_1^n(W))) \log(|ran(W)| - 1) \\
&\qquad (\because E_{W, Y_1^n} = 0 \Leftrightarrow W \text{ is correctly determined by } g(Y_1^n)) \\
&\leq \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)} (\lambda_W(X_1^n(W))) \log(|ran(W)| - 1) \leq P_e^{(n)} nR
\end{aligned}
$$

Henceforth, $H(X_1^n(W)|Y_1^n) \leq 1 + P_e^{(n)}(x_1^n)nR$ which is (3).
For (4), note that

$$\begin{aligned} H(E_{W,Y_1^n}) &= H(\{\mathbb{P}(E_{W,Y_1^n} = 1), \mathbb{P}(E_{W,Y_1^n} = 0)\}) = H(\{\mathbb{P}(g(Y_1^n) \neq W), \mathbb{P}(g(Y_1^n) = W)\}) \\ &= H(P_e^{(n)}, 1 - P_e^{(n)}) \end{aligned}$$

□

Furthermore, we need following lemma too.

Lemma) For a DMC,.

$$I(X_1^n; Y_1^n) \leq nC \tag{5}$$

Proof.

$$\begin{aligned} I(X_1^n; Y_1^n) &= H(Y_1^n) - H(Y_1^n|X_1^n) \\ &= H(Y_1^n) - \sum_{i=1}^n H(Y_i|Y_1^{i-1}, X_1^n) \\ &= H(Y_1^n) - \sum_{i=1}^n H(Y_i|X_i) \quad (\because \text{DMC}) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) = \sum_{i=1}^n I(X_i; Y_i) \end{aligned}$$

□

Now, we can prove the converse.

$$\begin{aligned} nR = H(W) &= H(W|Y_1^n) + I(W; Y_1^n) \\ &\leq H(W|Y_1^n) + I(X_1^n(W); Y_1^n) \\ &\leq 1 + P_n^{(e)} nR + I(X_1^n; Y_1^n) \quad (\because (4),\, W \sim U([2^{nR}])) \\ &\leq 1 + P_n^{(e)} nR + nC \quad (\because (5)) \end{aligned}$$

Dividing by n, we have $R \leq \frac{1}{n} + P_e^{(n)} R + C$. Taking $n \to \infty$, we are done. □

Corollary) Bounding $\lambda^{(n)}(x_1^n)$ by specific realization.

(i) For every $\delta > 0$, $R < C$, there exist $(2^{nR}, n)$ code with $\lambda^{(n)}(x_1^n) < \delta$.

Proof. It is enough to show that we can take a codebook $(2^{n(R-1/n)}, n)$ satisfying $\lambda^{(n)}(x_1^n) < \delta$. By channel coding theorem, we have

$$P_e^{(n)} = \mathbb{E}_{W \sim U([2^{nR}])} \mathbb{E}_{X_1^n(W)}(\lambda_W(X_1^n(W))) \leq 2\epsilon.$$

Then, there exists $x_1^n(w)$ for each $w \in [2^{nR}]$ s.t. $\mathbb{E}_{W \sim U([2^{nR}])} \lambda_1(x_1^n(W)) \leq 2\epsilon$. Therefore, at least the half of w's of $[2^{nR}]$ satisfies $\lambda_w(x_1^n(w)) \leq 4\epsilon$. So we are done. □

Theorem) Zero-error codes.

$P_e^{(n)} = 0$ implies $R < C$.

Proof. $nR = H(W) = H(W|Y_1^n) + I(W;Y_1^n) = I(W;Y_1^n)$ since $P_e^{(n)} = 0$ implies W can be restored by $g(y_1^n(X_1^n(W)))$ for all $X_1^n(W)$. Data processing inequality implies that $I(W;Y_1^n) \leq I(X_1^n;Y_1^n)$. Finally, $I(X_1^n;Y_1^n) = \sum_{i=1}^n I(X_i;Y_i) \leq nC$. □

Definition) Feedback capacity.

$(2^{nR}, n)$ feedback code is a sequence of mappings $x_i(W, Y_1^{i-1})$.

The capacity with feedback, C_{FB}, of a DMC is a supremum of all rates achievable by feedback codes.

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

Proof. Clearly, $C_{FB} \geq C$. To show that $C_{FB} \leq C$, let's start from $H(W) = H(W|Y_1^n) + I(W;Y_1^n)$. Bound $I(W;Y_1^n)$ as follows.

$$\begin{aligned}
I(W;Y_1^n) &= H(Y_1^n) - H(Y_1^n|W) \\
&= H(Y_1^n) - \sum_{i=1}^n H(Y_i|Y_1^{i-1}, W) \\
&= H(Y_1^n) - \sum_{i=1}^n H(Y_i|Y_1^{i-1}, X_i, W) \quad (\because X_i \text{ is a ftn of } Y_1^{i-1},\, W) \\
&= H(Y_1^n) - \sum_{i=1}^n H(Y_i|X_i) \\
&\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) = \sum_{i=1}^n I(X_i;Y_i) \\
&\leq nC
\end{aligned}$$

Together with (3), $H(W) \leq 1 + P_e^{(n)} nR + nC$. Dividing by n and letting $n \to \infty$ give $R \leq C$. Taking supremum of R, we have $C_{FB} \leq C$. □

Theorem) Joint source-channel coding theorem.

V_1^n is a finite alphabet stochastic process $\mathcal{V}$ s.t. $V_1^n \in A_\epsilon^{(n)}$, $H(\mathcal{V}) < C$. Then there exists source-channel code s.t. $\mathbb{P}(\hat{V}_1^n \neq V_1^n) \to 0$ *a.s.*. Conversely, for any stationary stochastic process $\mathcal{V}$ with $H(\mathcal{V}) > C$, the probability of error is bounded away from zero.

Proof. Take $\epsilon > 0$ s.t. $H(\mathcal{V}) + \epsilon < C$. From AEP, we have $|A_\epsilon^{(n)}| \leq 2^{n(H(\mathcal{V})+\epsilon)}$. So, we can index them with $n(H(\mathcal{V}) + \epsilon)$ bits. From channel coding theorem, we can reliably transmit the indices since $H(\mathcal{V}) + \epsilon = R < C$ with the arbitrary small probability of error. Conversely, we need to show that $\mathbb{P}(\hat{V}_1^n \neq V_1^n) \to 0$ $a.s.$ implies $H(\mathcal{V}) < C$. Note that

$$
\begin{aligned}
H(\mathcal{V}) &\approx \frac{H(V_1^n)}{n} && (\because \text{def}) \\
&= \frac{1}{n}(H(V_1^n|\hat{V}_1^n) + I(V_1^n;\hat{V}_1^n)) \\
&\leq \frac{1}{n}(1 + \mathbb{P}(V_1^n \neq \hat{V}_1^n) n \log|\mathcal{V}| + I(V_1^n;\hat{V}_1^n)) && (\because 3, 4) \\
&\leq \frac{1}{n}(1 + \mathbb{P}(V_1^n \neq \hat{V}_1^n) n \log|\mathcal{V}| + I(X_1^n;Y_1^n)) && (\because \text{data processing inequality}) \\
&= \frac{1}{n} + \mathbb{P}(V_1^n \neq \hat{V}_1^n) \log|\mathcal{V}| + C && (\because \text{Memoryless DMC})
\end{aligned}
$$

letting $n \to \infty$, we are done. □

6 Differential Entropy

Now we are assume that all r.v.'s are continuous, i.e. $F(x) = \mathbb{P}(X \leq x)$ is continuous.

6.1 Differential Entropy, Relative Entropy, Conditional Entropy, Mutual Information

Definition) Differential Entropy.

X : r.v. with the pdf $p(x)$

$$h(X) = -\int_S p(x) \ln p(x) dx = \mathbb{E}_X(\ln \frac{1}{p(X)}; S)$$

where $S = \{x \mid p(x) > 0\}$ is the support set of X.
Comparing to discrete entropy (bits), differential entropy uses natural log (nats), i.e. ln.

Exercise) Few examples.

a) $X \sim U([a, b]) \;\Rightarrow\; h(X) = \ln(b - a)$.
Note that if $b - a < 1$, $h(X) < 0$

b) $X \sim \mathcal{N}(0, \sigma^2) \;\Rightarrow\; h(X) = \mathbb{E}_X(\frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2\sigma^2} X^2)) = \frac{1}{2} \ln 2\pi e \sigma^2$.

Proposition) Properties of Differential Entropy.

(i) Shift invariant: $h(X) = h(X + a)$ for $a \in \mathbb{R}$.

(ii) $h(aX) = h(X) + \log |a|$

Proof. $p_{aX}(y) = \frac{1}{|a|} p_x(\frac{y}{a})$ □

(iii) $h(AX) = h(X) + \log |A|$ where A is a linear map and $|A| = detA$

6.2 AEP for continuous r.v.

Theorem) (AEP).

X_i : i.i.d. r.v.'s with pdf p

$$-\frac{1}{n} \ln p(X_1, \ldots, X_n) \to h(X) = \mathbb{E}_X(-\ln p(X)) \quad \text{a.s.}$$

Definition) Typical set.

The typical set $A_\epsilon^{(n)}$ is

$$A_\epsilon^{(n)} = \{(x_1 \ldots, x_n) \in S^n \;:\; |-\frac{1}{n} \ln p(x_1 \ldots, x_n) - h(X)| < \epsilon\}$$

Define a $Vol(A)$ as

$$Vol(A) = \int_A dx_1 \cdots dx_n$$

Proposition) Properties of Typical sets.

(i) $\mathbb{P}(X \in A_\epsilon^{(n)}) \geq 1 - \epsilon$ for sufficiently large n.

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

Proof.

$$1 = \int_{S^n} p(x_1^n) dx_1^n \geq \int_{A_\epsilon^{(n)}} p(x_1^n) dx_1^n \geq \int_{A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} dx_1^n$$
$$= Vol(A_\epsilon^{(n)}) 2^{-n(H(X)+\epsilon)}$$

□

(iii) $Vol(A_\epsilon^{(n)}) \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$ for sufficiently large n

Proof. $1 - \epsilon < \mathbb{P}(X_1^n \in A_\epsilon^{(n)}) = \int_{A_\epsilon^{(n)}} p(x_1^n) dx_1^n \leq Vol(A_\epsilon^{(n)}) 2^{-n(H(X)-\epsilon)}$ for sufficiently large n □

Theorem) Relation to Discrete Entropy (Quantization).

Define $X^\Delta = \sum_i \Delta i I_{\Delta i \leq X < \Delta(i+1)}$.
If $p(x)$ is Riemann-integrable, then

$$H(X^\Delta) + \log \Delta \to h(X) \text{ as } \Delta \to 0.$$

Proof. $H(X^\Delta) = -\sum \mathbb{P}(X^\Delta = \Delta i) \log \mathbb{P}(X^\Delta = \Delta i)$. MVT implies that there exists x_i s.t. $\mathbb{P}(X^\Delta = \Delta i) = \mathbb{E}(I_{\Delta i \leq X < \Delta(i+1)}) = p(x_i)\Delta$. Therefore,

$$\begin{aligned} H(X^\Delta) &= -\sum \mathbb{P}(X^\Delta = \Delta i) \log \mathbb{P}(X^\Delta = \Delta i) \\ &= -\sum (p(x_i)\Delta) \log(p(x_i)\Delta) = -\sum (p(x_i)\Delta) \log p(x_i) - \log \Delta \sum p(x_i)\Delta \\ &= -\sum (p(x_i)\Delta) \log p(x_i) - \log \Delta \to h(X) - \log \Delta \ (bits) \end{aligned}$$

. □

Definition) Joint differential entropy.

X, Y : r.v.'s with the joint pdf $p(x, y)$

$$h(X, Y) = \mathbb{E}_{X,Y}(\ln \frac{1}{p(X, Y)})$$

Exercise) Multivariate normal distribution..

a) $X \sim \mathcal{N}(\mu, \Sigma)$

$$
\begin{aligned}
h(X) &= \mathbb{E}_X(\frac{1}{2}\ln(2\pi)^n|\Sigma| + \frac{1}{2}(X-\mu)^t\Sigma^{-1}(X-\mu)) \\
&= \frac{1}{2}\ln(2\pi)^n|\Sigma| + \frac{1}{2}tr(\mathbb{E}_X(\Sigma^{-1}(X-\mu)^t(X-\mu))) \\
&= \frac{1}{2}\ln(2\pi)^n|\Sigma| + n \ (nats)
\end{aligned}
$$

Proposition) Properties of Joint Differential Entropy.

(i) If X, Y are independent, $h(X, Y) = h(X) + h(Y)$

Definition) Conditional Differential Entropy.

X, Y : r.v.'s with the joint pdf $p(x, y)$

$$
H(Y|X) = \mathbb{E}_{X,Y}(\ln \frac{1}{p(Y|X)})
$$

Proposition) Properties of Conditional Differential Entropy.

(i) Chain rule: $h(X_1, \ldots, X_n) = \sum_{i=1}^n h(X_i|X_1^{i-1})$

(ii) Conditioning reduces entropy: $h(X_1, \ldots, X_n) \leq \sum_{i=1}^n h(X_i)$. The equality holds when $X_1, \ldots, X_n$ are indep.

Theorem) Hadamard Inequality.

K : p.s.d. matrix. Then,

$$
|K| \leq \prod_{i=1}^n K_{ii}
$$

Proof. Let $X \sim \mathcal{N}(0, K)$. From the above 2nd proposition,

$$
\frac{1}{2}\ln(2\pi e)^n|K| \leq \sum \frac{1}{2}\ln(2\pi e)K_{ii} = \frac{1}{2}\ln[(2\pi e)^n \prod_{i=1}^n K_{ii}]
$$

□

Definition) Differential Relative Entropy (Kullback Leibler distance).

For pdfs $p(x)$, $q(x)$,

$$
D(p\|q) = \mathbb{E}_{X\sim p}(\ln \frac{p(X)}{q(X)})
$$

Proposition) Properties of Differential Relative Entropy.

(i) $D(p\|q) \geq 0$. The equality holds when $p = q$ w.p. 1.

Theorem) Normal distribution maximizes entropy.
Let $X \in \mathbb{R}^n$ be a r.v. with $\mathbb{E}(X) = 0$, $\mathbb{E}(XX^t) = K$. Then,

$$h(X) \leq \frac{1}{2} \ln(2\pi e)^n |K|$$

where equality holds when $X \sim \mathcal{N}(0, K)$

Proof. Let $Y \sim \mathcal{N}(0, K)$. Then,

$$\begin{aligned} 0 \leq D(X\|Y) &= -h(X) + \mathbb{E}_X(-\log \mathcal{N}(X; 0, K)) \\ &= -h(X) + \frac{1}{2} \ln(2\pi e)^n |K| \end{aligned}$$

□

Definition) Differential Mutual Information.
X, Y : r.v.'s. with the joint pdf $p(x, y)$.

$$\begin{aligned} I(X;Y) &= D(p(x,y)\|p_X(x)p_y(y)) = \mathbb{E}_{X,Y\sim p}(\log(\frac{p(X,Y)}{p(X)p(Y)})) \\ &= h(X) - h(X|Y) \end{aligned}$$

Unlike differential entropy, the mutual information of continuous r.v. is the same as that of quantized r.v..

Proposition) Properties of Mutual Information.

(i) $I(X;Y) \geq 0$.

(ii) $I(X;Y) = 0$ iff X, Y are indep.

7 Gaussian Channel

7.1 Gaussian Channel

Definition) Gaussian channel.

$Y_i = X_i + Z_i,\ Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, N)$ where $Z_i,\ X_i$ are independent and $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

Proposition) Probability of error.

(i) Probability of error for binary transmission $X = \pm\sqrt{P}\ w.p.\frac{1}{2}$.

$$\begin{aligned} P_e &= \mathbb{E}_X(I(XY < 0)) = \frac{1}{2}(\mathbb{P}(Y < 0|X = \sqrt{P}) + \mathbb{P}(Y > 0|X = -\sqrt{P})) \\ &= \mathbb{P}(Z > \sqrt{P}) \end{aligned}$$

Definition) Information capacity.

The information capacity with power constraint is

$$C = \max_{p(x):EX^2 \le P} I(X;Y)$$

Proposition) Gaussian channel capacity.

(i) The information capacity of Gaussian Channel is

$$\frac{1}{2}\log(1 + \frac{P}{N}) \text{ where } X \sim \mathcal{N}(0, P)$$

Proof. $I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z|X) = h(Y) - h(Z)$. Note that $\mathbb{E}(Y^2) = \mathbb{E}(X^2) + \mathbb{E}(Z^2) \le P + N$. Therefore, $h(Y) \le \frac{1}{2}\log 2\pi e(P + N)$. We are done. □

Definition) (M,n) with power constraint.

An (M, n) code with power constraint consists of

1. An index set $I = \{1, \dots, M\}$.
2. An encoding ftn $x_1^n : I \to \Omega_x^n$ with power constraint of $\sum_{i=1}^{n} x_i^2(w) \le nP \quad \forall w \in I$
3. A DMC $(x_1^n(w), p(Y_1^n|x_1^n(w)), Y_1^n)$. This generates a r.v. Y_1^n for given $x_1^n(w)$.
4. A decoding ftn $g : \Omega_y^n \to I$.

Theorem) Gaussian capacity.

For every $\delta > 0$, $R < C = \frac{1}{2}\log(1 + \frac{P}{N})$, there exist $(2^{nR}, n)$ code with $P_e^{(n)} < \delta$. Conversely, any sequence of $(2^{nR}, n)$ code with $P_e^{(n)} \to 0$ must have $R \le C = \frac{1}{2}\log(1 + \frac{P}{N})$

i.e. $(2^{nR}, n)$ code is achievable iff $R \le C$.

Proof. Fix $\epsilon > 0$ s.t. $4\epsilon < \delta$ and $R < I(X;Y) - 3\epsilon$ ($\because R < C$).
Generate $X_i(w) \sim \mathcal{N}(0, P-\epsilon) \quad \forall w \in [2^{nR}]$.
Define $E_w = \{(x_1^n(w), y_1^n) \in A_\epsilon^{(n)}\}$ for each $w \in [2^{nR}]$, $F_w = \{\frac{1}{n}\sum_{i=1}^n x_i(w) > P\}$.
Define a decoding ftn $g : ran(Y)^n \to I$ by followings.

$$g(y_1^n) = g_{x_1^n}(y_1^n) = \begin{cases} w' & \text{if } \exists!\ w' \in [2^{nR}] \text{ s.t. } (x_1^n(w'), y_1^n) \in E_{w'} \wedge x_1^n(w') \in F_{w'} \\ 2 & \text{o.w.} \end{cases}$$

Note that the second case is no matter what value you assign.
Similar to channel coding theorem, the expected number of error (or probability of error) is

$$P_e^{(n)} = \mathbb{E}_{W\sim U([2^{nR}])}\mathbb{E}_{X_1^n(W)}\mathbb{E}_{Y_1^n \sim p(\ \cdot\ |x_1^n(W))}(I_{g(Y_1^n)\neq W}) = \int_{x_1^n(1)} \mathbb{P}(g(Y_1^n) \neq 1 | x_1^n(1)) d\mathbb{P}(x_1^n(1))$$

By the definition of g,

$$\begin{aligned} P_e^{(n)} &= \int_{x_1^n(1)} \mathbb{P}(g(Y_1^n) \neq 1 | x_1^n(1)) d\mathbb{P}(x_1^n(1)) \\ &\le \mathbb{P}(X_1^n(1) \in F_1) + \mathbb{P}(X_1^n(1) \in E_1^c) + \mathbb{P}(X_1^n(1) \in E_2) + \cdots + \mathbb{P}(X_1^n(1) \in E_{2^{nR}}) \\ &\le \epsilon + \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon} \quad (\because X_i(1) \sim \mathcal{N}(0, P-\epsilon)) \\ &\le 2\epsilon + 2^{-n(I(X;Y)-3\epsilon-R)} \quad \text{for sufficiently large } n \\ &\le 3\epsilon \quad \text{for sufficiently large } n \text{ since } R < I(X;Y) - 3\epsilon \end{aligned}$$

Conversely, we need to show that $P_e^{(n)} \to 0$ implies $R \le C$. Now, we can prove the converse.

$$\begin{aligned} R &= \frac{1}{n}H(W) = \frac{1}{n}(H(W|Y_1^n) + I(W;Y_1^n)) \\ &\le \frac{1}{n}(H(W|Y_1^n) + I(X_1^n(W);Y_1^n)) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{n}I(X_1^n;Y_1^n) \quad (\because (4),\ W \sim U([2^{nR}])) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{n}\sum_{i=1}^n h(Y_i) - h(Z_i) \quad (\because \text{the last line of proof of (5)},\ Y_i = X_i + Z_i) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{n}\sum_{i=1}^n [\frac{1}{2}\log(2\pi e(P_i + N)) - \frac{1}{2}\log(2\pi e N)] \quad \text{where } P_i = \mathbb{E}_{w\sim U([2^{nR}])} x_i^2(w) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\frac{P_i + N}{N} \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{2}\log(\frac{1}{n}\sum_{i=1}^n \frac{P_i + N}{N}) \quad (\because \text{Jensen's inequality}) \end{aligned}$$

Note that $\sum_{i=1}^n \frac{P_i}{n} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{w\sim U([2^{nR}])} x_i^2(w) = \mathbb{E}_{w\sim U([2^{nR}])}\frac{1}{n}\sum_{i=1}^n x_i^2(w) \le P$

$$\begin{aligned} R &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{2}\log(\frac{1}{n}\sum_{i=1}^{n}\frac{P_i+N}{N}) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{2}\log(1+\frac{1}{N}\sum_{i=1}^{n}\frac{P_i}{n}) \\ &\le \frac{1}{n} + P_n^{(e)}R + \frac{1}{2}\log(1+\frac{P}{N}) \end{aligned}$$

Taking $n \to \infty$, we are done. □

7.2 Parallel gaussian channel

Definition) Parallel Gaussian channel.

$Y_i = X_i + Z_i,\ Z_i \sim \mathcal{N}(0, N_i)$ where $Z_i,\ X_i$ are independent and $\sum_{i=1}^n x_i^2 \le P$

Proposition) Parallel gaussian channel capacity.

(i) The information capacity of parallel Gaussian Channel is

$$C = \max_{\sum EX_i^2 \le P} I(X_1^n; Y_1^n) = \sum \frac{1}{2}[\log(\frac{\nu}{N_i})]^+ \text{ where } \nu \text{ satisfies } \sum(\nu - N_i)^+ = P$$

Proof.

$$\begin{aligned} I(X_1^n;Y_1^n) &= h(Y_1^n) - h(Y_1^n|X_1^n) = h(Y_1^n) - h(Z_1^n) \\ &= h(Y_1^n) - \sum h(Z_i) \\ &= \sum h(Y_i) - h(Z_i) \\ &\le \sum \frac{1}{2}\log 2\pi e(P_i+N_i) - \frac{1}{2}\log 2\pi e(N_i) \quad \text{where } P_i = EX_i^2 \\ &= \sum \frac{1}{2}\log(1+\frac{P_i}{N_i}) \end{aligned}$$

So, we need to optimize followings

$$\begin{aligned} \text{Maximize } & \sum \frac{1}{2}\log(1+\frac{P_i}{N_i}) \\ \text{subject to } & \sum P_i \le P,\ P_i \ge 0 \end{aligned}$$

Consider $J = \sum \frac{1}{2}\log(1+\frac{P_i}{N_i}) - \frac{1}{2\nu}(\sum P_i)$. We have $\frac{\partial J}{\partial P_i} = \frac{1}{2}\frac{1}{P_i+N_i} - \frac{1}{2\nu} = 0$. Hence, $P_i = (\nu - N_i)^+ \ge 0$ must satisfy $\sum P_i = P$.
To sum up, we first find ν s.t. $\sum(\nu - N_i)^+ = P$. Then,

$$C = \sum \frac{1}{2}[\log(\frac{\nu}{N_i})]^+$$

□

7.3 Correlated gaussian noise channel

Definition) Correlated (colored) gaussian channel.

$Y_i = X_i + Z_i,\ X_1^n \sim \mathcal{N}(0, K_X),\ Z_1^n \sim \mathcal{N}(0, K_Z)$ where $Z_1^n \perp X_1^n$ and $\frac{1}{n}\sum_{i=1}^n x_i^2 \le P$

Proposition) Colored gaussian channel capacity.

(i) The information capacity of Colored Gaussian Channel is

$$C = \max_{\frac{1}{n}tr(K_X)\le P} I(X_1^n; Y_1^n) = \sum \frac{1}{2}[\log(\frac{\nu}{\lambda_i})]^+$$

$$\text{where } \lambda_i\text{'s are eigenvalues of } K_Z,\ \nu \text{ satisfies } \sum_{i=1}^n (\nu - \lambda_i)^+ = nP.$$

Proof. Note that $\frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{1}{n}tr({x_1^n}^t x_1^n)$. So, power constraint is $\frac{1}{n}tr(K_X) \le P$.

$$\begin{aligned} I(X_1^n; Y_1^n) &= h(Y_1^n) - h(Y_1^n|X_1^n) = h(Y_1^n) - h(Z_1^n) \\ &= h(Y_1^n) - \sum h(Z_i) \\ &= \frac{1}{2}\log(2\pi e)^n(|K_X + K_Z|) - \frac{1}{2}\log(2\pi e)^n|K_Z| \\ &= \sum \frac{1}{2}\log\frac{|K_X + K_Z|}{|K_Z|} \end{aligned}$$

So, we need to optimize followings

$$\text{Maximize } \sum \frac{1}{2}\log\frac{|K_X + K_Z|}{|K_Z|}$$
$$\text{subject to } K_X \ge 0,\ \frac{1}{n}tr(K_X) \le P$$

Since K_Z is p.s.d., we have $K_Z = QD_ZQ^t$ where $D_Z = diag(\text{eig}(K_Z))= diag(\lambda_1, \dots, \lambda_n)$ and Q is orthogonal. Then $\frac{1}{2}\log\frac{|K_X+K_Z|}{|K_Z|} = \frac{1}{2}\log\frac{|Q^tK_XQ+D_Z|}{|D_Z|}$. Let $A = Q^tK_XQ$. So, equivalently,

$$\text{Maximize } \sum \frac{1}{2}\log\frac{|A + D_Z|}{|D_Z|}$$
$$\text{subject to } A \ge 0,\ \frac{1}{n}tr(A) \le P$$

Hadamard inequality implies that $|A + D_Z| \le \prod_i |A_{ii} + \lambda_i|$ while equality holds when A is diagonal. From the constraint, $\frac{1}{n}tr(A) = \sum_i A_{ii} \le P$. So, it is reformulated as independent parallel channel. Therefore, we first find ν s.t. $\sum_{i=1}^n(\nu - \lambda_i)^+ = nP$. Then,

$$C = \sum \frac{1}{2}[\log(\frac{\nu}{\lambda_i})]^+$$

□

7.4 Stationary colored gaussian noise channel

Definition) Toeplitz matrix.

Toeplitz matrix or diagonal-constant matrix is a matrix in which each descending diagonal from left to right is constant.

Exercise) A few examples.

a) $\mathcal{X} = \{X_i\}$ is a stationary process, then $Var(X_1^n)$ is a Toeplitz matrix

Theorem) Toeplitz distribution theorem.

Given continuous $g : \mathbb{R} \to \mathbb{R}$, Toeplitz matrix

$$K_n = \begin{pmatrix} R(0) & R(1) & R(2) & \cdots & R(n-1) \\ R(1) & R(0) & R(1) & \cdots & R(n-2) \\ R(2) & R(1) & R(0) & \cdots & R(n-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(n-1) & R(n-2) & R(n-3) & \cdots & R(0) \end{pmatrix}$$

with eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$, let $N(f) = \sum_n R(n)e^{j2\pi fn}$ $(\theta = 2\pi f)$ where $\sqrt{-1} = j$. Then,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(\lambda_i^{(n)}) = \int_{1/2}^{1/2} g(N(f))df$$

Proof. Briefly...Check that $\nu = \begin{pmatrix} e^{j2\pi f\cdot 0} \\ \vdots \\ e^{j2\pi f\cdot(n-1)} \end{pmatrix}$ satisfies $K_n\nu = \lambda\nu$. Then, we have $\lambda_i^{(n)} \to N(f)$ as $n \to \infty$. □

Corollary) Revisit colored Gaussian channel capacity.

(i) For stationary Z, the information capacity of Colored Gaussian Channel is

$$C = \max_{\frac{1}{n}tr(K_X)\le P} I(X_1^n; Y_1^n) = \frac{1}{2}\int_{1/2}^{1/2} \log(1 + \frac{(\nu - N(f))^+}{N(f)})df$$

where λ_i's are eigenvalues of K_Z, $N(f) = \sum K_Z(n)e^{j2\pi fn}$,

ν satisfies $\sum(\nu - \lambda_i)^+ = P$.

The power constraint becomes $\int_{1/2}^{1/2}(\nu - N(f))^+ df = P$

Proof.

$$C = \max_{\frac{1}{n}tr(K_X)\le P} I(X_1^n; Y_1^n) = \sum \frac{1}{2}[\log(\frac{\nu}{\lambda_i})]^+ = \sum \frac{1}{2}\log(1 + \frac{(\nu - \lambda_i)^+}{\lambda_i})$$

where λ_i's are eigenvalues of K_Z, ν satisfies $\sum(\nu - \lambda_i)^+ = P$.

By the above theorem, $\sum \frac{1}{2}\log(1 + \frac{(\nu-\lambda_i)^+}{\lambda_i}) = \frac{1}{2}\int_{1/2}^{1/2}\log(1 + \frac{(\nu-N(f))^+}{N(f)})df$ where $N(f) = \sum_n K_Z(n)e^{j2\pi fn}$. The power constraint becomes $\int_{1/2}^{1/2}(\nu - N(f))^+ df = P$. □

7.5 Correlated gaussian channel with feedback

Definition) Correlated gaussian channel with feedback.

$Y_i = X_i + Z_i,\ X_1^n \sim \mathcal{N}(0, K_X),\ Z_1^n \sim \mathcal{N}(0, K_Z)$ where $\frac{1}{n}\sum x_i^2(w, Y_1^{i-1}) \leq P$

$(2^{nR}, n)$ feedback code for the correlated gaussian channel is a sequence of mappings $x_i(W, Y_1^{i-1})$ where $\mathbb{E}(\frac{1}{n}\sum x_i^2(w, Y_1^{i-1})) \leq P$

Proposition) Correlated gaussian channel with feedback capacity.

(i) Feedback capacity of correlated gaussian channel per transmission ($= \frac{1}{n}$) is

$$C_{FB,n} = \frac{1}{n} \max_{\frac{1}{n}tr(K_X) \leq P} I(X_1^n; Y_1^n) = \max_{\frac{1}{n}tr(K_X) \leq P} \frac{1}{2n} \log \frac{|K_{X+Z}|}{|K_Z|}$$

Proof.

$$\begin{aligned} I(X_1^n; Y_1^n) &= h(Y_1^n) - h(Y_1^n | X_1^n) = h(Y_1^n) - h(Z_1^n) \\ &= h(Y_1^n) - \sum h(Z_i) \\ &\leq \frac{1}{2}\log(2\pi e)^n(|K_{X+Z}|) - \frac{1}{2}\log(2\pi e)^n |K_Z| \\ &= \frac{1}{2}\log\frac{|K_{X+Z}|}{|K_Z|} \end{aligned}$$

where power constraint is $\frac{1}{n}tr(K_X) \leq P$. □

(ii) R with $P_e^{(n)} \to 0$ satisfies

$$R \leq \frac{1}{2n}\log\frac{|K_Y|}{|K_Z|} + \epsilon_n$$

where $\epsilon_n \to 0$

Proof. By (3), we have $H(W|Y_1^n) \leq 1 + nRP_e^{(n)} = n\epsilon_n$ where $\epsilon_n = \frac{1}{n} + RP_e^{(n)} \to 0$.

Then,

$$
\begin{aligned}
nR &= H(W) \\
&= I(W;Y_1^n) + H(W|Y_1^n) \\
&\le I(W;Y_1^n) + n\epsilon_n \\
&= \sum_i I(W;Y_i|Y_1^{i-1}) + n\epsilon_n \\
&= \sum_i (h(Y_i|Y_1^{i-1}) - h(Y_i|Y_1^{i-1}, W)) + n\epsilon_n \\
&= \sum_i (h(Y_i|Y_1^{i-1}) - h(Y_i|Y_1^{i-1}, W, X_1^i)) + n\epsilon_n \quad (\because X_1^i : \text{ ftn of } Y_1^{i-1}, W) \\
&= \sum_i (h(Y_i|Y_1^{i-1}) - h(Y_i|X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1}, W, X_i)) + n\epsilon_n \quad (\because \text{similarly}) \\
&= \sum_i (h(Y_i|Y_1^{i-1}) - h(Z_i|X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1}, W, X_i)) + n\epsilon_n \\
&= \sum_i (h(Y_i|Y_1^{i-1}) - h(Z_i|Z_1^{i-1}) + n\epsilon_n \quad (\because Z : \text{stationary}) \\
&= h(Y_1^n) - h(Z_1^n) + n\epsilon_n \\
&= \frac{1}{2}\log\frac{|K_Y|}{|K_Z|} + n\epsilon_n
\end{aligned}
$$

We are done. □

(iii) The information capacity of correlated gaussian channel with feedback per transmission ($= \frac{1}{n}$) can be bounded above as

$$C_{FB,n} \le C_n + \frac{1}{2}$$

where C_n is a correlated gaussian channel capacity per transmission.

Proof. We need a following lemma.

Lemma) Determinant preserves order on p.s.d. cone.

For $A \ge 0,\ B \ge 0,\ A - B \ge 0$, we have

$$|A| \ge |B|$$

Proof. For independent two r.v.'s $X \sim \mathcal{N}(0, B)$, $Y \sim \mathcal{N}(0, A - B)$, consider $h(X+Y)$. Then, we have $h(X+Y) \ge h(X+Y|Y) = h(X|Y)$. Hence, $\frac{1}{2}\log((2\pi e)^n|A|) \ge \frac{1}{2}\log((2\pi e)^n|B|)$. □

Now we can prove (ii). From (i), we have

$$I(X_1^n; Y_1^n) \leq \sum \frac{1}{2} \log \frac{|K_{X+Z}|}{|K_Z|}$$

Since $2(K_X + K_Z) - K_{X+Z} = K_{X-Z} \geq 0$, the above lemma implies $|K_{X+Z}| \leq |2(K_X + K_Z)| = 2^n |K_X + K_Z|$. Therefore,

$$\begin{aligned} I(X_1^n; Y_1^n) &\leq \frac{1}{2} \log \frac{|K_{X+Z}|}{|K_Z|} \\ &\leq \frac{1}{2} \log \frac{2^n |K_X + K_Z|}{|K_Z|} \\ &\leq \frac{1}{2} \log \frac{|K_X + K_Z|}{|K_Z|} + \frac{n}{2} \\ &\leq nC_n + \frac{n}{2} \end{aligned}$$

We are done. □

Definition) Causally related.

Random vector X_1^n is causally related to Z_1^n iff

$$p(x_1^n, z_1^n) = p(z_1^n) \prod_{i=1}^{n} p(x_i | x_1^{i-1}, z_1^{i-1})$$

Reflection) A few properties of causally related random vector.

(i) The most general causal dependence of X_1^n on Y_1^n is

$$X = BZ + V \quad (V \text{ depends on W})$$

where B is strictly lower triangular.

(ii) Causally related channel capacity is

$$C_{FB,n} = \max_{\frac{1}{n} tr(BK_Z B^t + K_V) \leq P} \frac{1}{2n} \log \frac{|(B+I)K_Z(B+I)^t + K_V|}{|K_Z|}$$

Proof. From the above proposition (i), □

Proposition) sharp bound for capacity.

(i) The information capacity of correlated gaussian channel with feedback per transmission can be bounded above as

$$C_{FB,n} \leq 2C_n$$

where C_n is a correlated gaussian channel capacity per transmission.

Proof. We need following lemmas.

Lemma) Determinant is log-concave on p.s.d. cone.

For $A \geq 0$, $B \geq 0, \lambda \in [0,1]$, we have

$$|\lambda A + (1-\lambda)B| \geq |A|^{\lambda}|B|^{1-\lambda} \tag{6}$$

Proof. For independent r.v.'s $X \sim \mathcal{N}(0,A)$, $Y \sim \mathcal{N}(0,B)$, $Z \sim Ber(\lambda)$, consider $W = ZX + (1-Z)Y$. Note that $Var(W) = \mathbb{E}(W^2) = \lambda A + (1-\lambda)B$. Then

$$\begin{aligned}
\frac{1}{2}\log(2\pi e)^n|\lambda A + (1-\lambda)B| &\geq h(W) \\
&\geq h(W|Z) \\
&\geq \lambda h(X) + (1-\lambda)h(Y) \\
&= \frac{1}{2}\log(2\pi e)^n|A|^{\lambda}|B|^{1-\lambda}
\end{aligned}$$

□

Lemma) Entropy and variance of causally related random process.

If X_1^n and Z_1^n re causally related, then

$$h(X_1^n - Z_1^n) \geq h(Z_1^n) \tag{7}$$

and

$$|K_{X-Z}| \geq |K_Z| \tag{8}$$

Proof.

$$\begin{aligned}
h(X_1^n - Z_1^n) &= \sum_{i=1}^{n} h(X_i - Z_i|X_1^{i-1} - Z_1^{i-1}) \\
&\geq \sum_{i=1}^{n} h(X_i - Z_i|X_1^{i}, Z_1^{i-1}) \quad (\because \text{Conditioning reduces entropy}) \\
&= \sum_{i=1}^{n} h(Z_i|X_1^{i}, Z_1^{i-1}) \\
&= \sum_{i=1}^{n} h(Z_i|Z_1^{i-1}) \\
&= h(Z_1^n)
\end{aligned}$$

First, taking a supremum w.r.t. $X_1^n - Z_1^n$ gives $\frac{1}{2}\log(2\pi e)^n|K_{X-Z}| \geq h(Z_1^n)$. Then, taking a supremum w.r.t. Z_1^n gives $|K_{X-Z}| \geq |K_Z|$. □

Now we can prove (i).

$$
\begin{aligned}
C_n = \frac{1}{2n}\log\frac{|K_X+K_Z|}{|K_Z|} &= \frac{1}{2n}\log\frac{|\frac{1}{2}K_{X+Z}+\frac{1}{2}K_{X-Z}|}{|K_Z|} \\
&\geq \frac{1}{2n}\log\frac{|K_{X+Z}|^{\frac{1}{2}}|K_{X-Z}|^{\frac{1}{2}}}{|K_Z|} \quad (\because (6)) \\
&\geq \frac{1}{2n}\log\frac{|K_{X+Z}|^{\frac{1}{2}}|K_{Z}|^{\frac{1}{2}}}{|K_Z|} \quad (\because (8)) \\
&= \frac{1}{2}\frac{1}{2n}\log\frac{|K_{X+Z}|}{|K_Z|} \\
&\geq \frac{1}{2}C_{FB,n}
\end{aligned}
$$

□

7.6 Multiple-Input Multiple-Output (MIMO)

Definition) Multiple-Input Multiple-Output (MIMO).

$$y = Hx + n$$

where $H \in \mathbb{C}^{r\times t}$, $\mathbb{E}(n) = 0$, $E(nn^*) = I_r$, with power constraint $\mathbb{E}(x^*x) = tr\mathbb{E}(x^*x) \leq P$.
Note that SNR (signal to noise ratio) is $\rho = \frac{P}{E(|n_i|^2)} = P$.

Definition) Complex gaussian.

Given $x \in \mathbb{C}^n$, define $\hat{x} = \begin{pmatrix}\mathrm{Re}(x)\\ \mathrm{Im}(x)\end{pmatrix} \in \mathbb{R}^{2n}$.

x is said to be (complex) gaussian if $\hat{x}$ is gaussian.

x is circulary symmetric if

$$\mathbb{E}((\hat{x}-\mathbb{E}(\hat{x})(\hat{x}-\mathbb{E}(\hat{x}))^*) = \frac{1}{2}\begin{pmatrix}\mathrm{Re}(Q) & -\mathrm{Im}(Q)\\ \mathrm{Im}(Q) & \mathrm{Re}(Q)\end{pmatrix} = \frac{1}{2}\hat{Q}$$

for some Hermitian p.s.d. $Q \in \mathbb{C}^{n\times n}$.
Note that $\mathbb{E}((x-\mathbb{E}(x))(x-\mathbb{E}(x)^*) = Q$.

Joint pdf is defined as

$$
\begin{aligned}
r_{\mu,Q}(x) &= det(\pi\hat{Q})^{-1/2}\exp(-(\hat{x}-\hat{\mu})^*\hat{Q}^{-1}(\hat{x}-\hat{\mu})) \\
&= det(\pi Q)^{-1/2}\exp(-(x-\mu)^*Q^{-1}(x-\mu))
\end{aligned}
$$

Reflection) Some properties.

(i) Joint entropy of complex gaussian is $H(r_Q) = \log det(\pi e Q)$.

Proposition) MIMO capacity.

(i) Let x be a circularly symmetric gaussian with zero-mean and covariance $\frac{P}{t}I_t$. The information capacity of MIMO $y = Hx + n$ is

$$C = \mathbb{E}[\log det(I_r + \frac{P}{t}HH^*)]$$

When $n \to infty$, $C \to r\log(1+P)$

Proof. For the capacity if $t \to \infty$, note that $\frac{1}{t}HH^* \to I_r$ as $t \to \infty$ by SLLN. □

7.7 MIMO Detectors

$$r = Ha + n$$

We want to find a which minimize $\|n\|$ for some sense.

7.7.1 Maximum Likelihood (ML) detector

- $\hat{a} = \arg\max_a \|r - Ha\|_F^2$ where the optimization is done by exhaustive search over $\forall a$.
- ML detection is optimal

7.7.2 Zero Forcing (ZF) detector

- $\hat{a} = G_{ZF}r = a + H^{\dagger}n$ where $G_{ZF} = H^{\dagger} = (H^*H)^{-1}H^*$.
- G_{ZF} increases noise.

7.7.3 MMSE detector

- $\hat{a} = G_{MMSE}r = a + H^{\dagger}n$ where $G_{MMSE} = (H^*H + \frac{1}{\rho}I_N)^{-1}H^*$ with SNR ρ.
- $G_{MMSE} = (H^*H + \frac{1}{\rho}I_N)^{-1}H^*$ is a solution of $\arg\min_G \epsilon\|Gr - a\|_F^2$ where
- MMSE receiver has good performance with reasonable complexity
- This is a mitigated version of ZF detector.

7.7.4 V-BLAST detector

- ?

8 Rate Distortion Theory

8.1 Lloyd algorithm

The goal of Lloyd algorithm is to find a set of reconstruction points.

1. Given t-th reconstruction points $x_1^{(t)}, \dots, x_n^{(t)}$, find optimal set of regions

$$R_i = \{x | \|x - x_i^{(n)}\| \leq \|x - x_j^{(n)}\| \ \forall j\}$$

2. Compute $x_i^{(t)} = \mathbb{E}(x|R_i) = \frac{\int_{R_i} x d\mathbb{P}(x)}{\int_{R_i} d\mathbb{P}(x)}$

3. Interate step 1 and 2.

8.2 Rate distortion code

Definition) Distortion.

A distortion measure is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_{\geq 0}$$

d is bounded iff

$$\max_{(x,\hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}} d(x, \hat{x}) < \infty$$

The distortion between sequence $x_1^n, \hat{x}_1^n$ is

$$d(x_1^n, \hat{x}_1^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$$

Definition) Rate distortion code.

A $(2^{nR}, n)$ rate distortion code consists of

1. An index set $I = \{1, \dots, 2^{nR}\}$.

2. An encoding ftn $f_n : \mathcal{X}^n \to [2^{nR}]$.

3. A decoding ftn $g_n : [2^{nR}] \to \hat{\mathcal{X}}^n$.

4. A distortion is defined by

$$\begin{aligned} D_n = \mathbb{E} d(X_1^n, \hat{X}_1^n) &= \mathbb{E} d(X_1^n, g_n(f_n(X_1^n))) \\ &= \sum_{x_1^n} p(x_1^n) d(x_1^n, g_n(f_n(x_1^n))) \end{aligned}$$

(R, D) is achievable iff $\exists (2^{nR}, n)$ codes (f_n, g_n) with $D_n \to D$ as $n \to \infty$

$R(D) = \inf_{\text{achievable } (R,D)} R$

$D(R) = \inf_{\text{achievable } (R,D)} D$
Information $R-D$ function is

$$R^{(I)}(D) = \min_{p_{\hat{X}|X}:\mathbb{E}_{(X,\hat{X})\sim p_{\hat{X}|X}p_X} d(x,\hat{x})\le D} I(X;\hat{X})$$

for given p_X

Proposition) Properties of $R^{(I)}(D)$.

(i) $R^{(I)}(D)$ is non-increasing.

Proof. Trivial from the definition. □

(ii) $R^{(I)}(D)$ is convex.

Proof. We need to consider a new distortion $D_\lambda = \lambda D_0 + (1-\lambda)D_1$ for given distortions D_0, D_1 with $\lambda \in (0,1)$. Let's assume that we achieve $(R_0^{(I)}, D_0)$, $(R_1^{(I)}, D_1)$ with distribution $p_{\hat{X},X;0}(\hat{x}|x)$, $p_{\hat{X},X}(\hat{x}|x)$. Let $p_{\hat{X}|X;\lambda}(\hat{x}|x) = \lambda p_{\hat{X}|X;0}(\hat{x}|x) + (1-\lambda)p_{\hat{X}|X;1}(\hat{x}|x)$. Then,

$$I_{p_{\hat{X}|X;\lambda}}(X;\hat{X}) \le \lambda I_{p_{\hat{X}|X;0}}(X;\hat{X}) + (1-\lambda)I_{p_{\hat{X}|X;1}}(X;\hat{X}) \quad (\because (2))$$

Therefore,

$$\begin{aligned} R^{(I)}(D_\lambda) &\le I_{p_{\hat{X}|X;\lambda}}(X;\hat{X}) \le \lambda I_{p_{\hat{X}|X;0}}(X;\hat{X}) + (1-\lambda)I_{p_{\hat{X}|X;1}}(X;\hat{X}) \\ \Rightarrow R^{(I)}(D_\lambda) &\le \lambda R^{(I)}(D_0) + (1-\lambda)R^{(I)}(D_1) \end{aligned}$$

□

Exercise) Compute $R-D$ function for a few examples.

a) Binary case.
For Hamming distance $d(x,\hat{x}) = I(x \ne \hat{x})$, $Ber(p)$ on $\mathcal{X}$,

$$R^{(I)}(D) = \begin{cases} H(p) - H(D) & 0 \le D \le \min(p, 1-p) \\ 0 & \text{o.w.} \end{cases}$$

Proof. We may assume that $p \le \frac{1}{2}$.

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &= h(\{p, 1-p\}) - h(X \oplus \hat{X}|\hat{X}) \\ &\ge h(\{p, 1-p\}) - h(X \oplus \hat{X}) \\ &= h(\{p, 1-p\}) - h(\{\mathbb{P}(X \ne \hat{X}), 1 - \mathbb{P}(X \ne \hat{X})\}) \\ &= h(\{p, 1-p\}) - h(\{\mathbb{E}d(X,\hat{X}), 1 - \mathbb{E}d(X,\hat{X})\}) \end{aligned}$$

Note that $\mathbb{E}d(X,\hat{X}) \le D$. Therefore, $h(\{\mathbb{E}d(X,\hat{X}), 1-\mathbb{E}d(X,\hat{X})\}) \le H(\{D, 1-D\})$ for $D \le \frac{1}{2}$.

$$\begin{aligned} I(X;\hat{X}) &\ge h(\{p,1-p\}) - h(\{\mathbb{E}d(X,\hat{X}), 1-\mathbb{E}d(X,\hat{X})\}) \\ &\ge h(\{p,1-p\}) - h(\{D,1-D\}) \quad \text{for } D \le \frac{1}{2} \end{aligned}$$

Consider a BSC model s.t. decode $\hat{X} \sim Ber(r)$. Distortion constraint $\mathbb{E}d(X,\hat{X}) \le D \le \frac{1}{2}$ implies $\mathbb{P}(X=1) = \mathbb{P}(X=1|\hat{X}=1)\mathbb{P}(\hat{X}=1) + \mathbb{P}(X=1|\hat{X}=0)\mathbb{P}(\hat{X}=0)$. Therefore, $r = \frac{p-D}{1-2D}$.

(a) For $D \le p \le \frac{1}{2}$, let $\mathbb{P}(\hat{X}=1) = r = \frac{p-D}{1-2D}$. Then, we have $I(X,\hat{X}) = H(p) - H(D)$.

(b) For $D > p$, let $\mathbb{P}(\hat{X}=0) = 1$. Then, we have $I(X,\hat{X}) = 0$ where $\mathbb{E}d(X,\hat{X}) = p < D$.

We are done by symmetricity for $p > \frac{1}{2}$. □

b) Gaussian case.
For L^2-distance $d(x,\hat{x}) = \|x-\hat{x}\|_2$, $X \sim \mathcal{N}(0,\sigma^2)$ on $\mathcal{X}$,

$$R^{(I)}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & 0 \le D \le \sigma^2 \\ 0 & \text{o.w.} \end{cases}$$

Proof. We may assume that $p \le \frac{1}{2}$.

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &= h(X) - h(X-\hat{X}|\hat{X}) \\ &\ge h(X) - h(X-\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e\sigma^2) - h(\mathcal{N}(0,\mathbb{E}(X-\hat{X})^2)) \\ &= \frac{1}{2}\log(\frac{\sigma^2}{\mathbb{E}(X-\hat{X})^2}) = \frac{1}{2}\log(\frac{\sigma^2}{D}) \end{aligned}$$

(a) For $D \le \sigma^2$, let $\hat{X} \sim \mathcal{N}(0,\sigma^2-D)$ and $X = \hat{X}+Z$ where $Z \sim \mathcal{N}(0,D)$, $X \perp Z$. Then, we have $I(X,\hat{X}) = \frac{1}{2}\log(\frac{\sigma^2}{D})$.

(b) For $D > \sigma^2$, let $\hat{X}=0$. Then, we have $I(X,\hat{X}) = 0$ where $\mathbb{E}d(X,\hat{X}) = \sigma^2 < D$.

□

c) Parallel gaussian case.
For L^2-distance $d(x,\hat{x}) = \|x-\hat{x}\|_2$, $X_i \sim \mathcal{N}(0,\sigma_i^2)$ on $\mathcal{X}$,

$$R(D) = \sum_{i=1}^{n} \frac{1}{2}[\log\frac{\sigma_i^2}{D_i}]^+$$

where $D_i = \min(\lambda, \sigma_i^2)$ with λ satisfying $\sum_{i=1}^{n} D_i = D$.

Proof.

$$\begin{aligned}
I(X_1^n;\hat{X}_1^n) &= h(X_1^n) - h(X_1^n|\hat{X}_1^n) \\
&= \sum_{i=1}^n h(X_i) - \sum_{i=1}^n h(X_i - \hat{X}_i|X_1^{i-1}, \hat{X}^n) \\
&\geq \sum_{i=1}^n h(X_i) - \sum_{i=1}^n h(X_i - \hat{X}_i|\hat{X}_i) \quad \text{if } f(x_1^m|\hat{x}_1^n) = \prod_{i=1}^n f(x_i|\hat{x}_i) \\
&= \sum_{i=1}^n I(X_i, \hat{X}_i) \\
&\geq \sum_{i=1}^n R(D_i) \quad \text{if } \hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i) \text{ where } D_i = \mathbb{E}((X-\hat{X})^2) \\
&= \frac{1}{2}\sum_{i=1}^n [\log \frac{\sigma_i^2}{D_i}]^+
\end{aligned}$$

So, we need to optimize followings

$$\begin{aligned}
\text{Minimize } &\sum \frac{1}{2}\log(1 + \frac{\sigma_i^2}{D_i}) \\
\text{subject to } &\sum D_i \leq D,\ D_i \geq 0
\end{aligned}$$

Therefore, we are done. □

8.3 R-D theorem

Definition) Jointly typical sequences.

The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x_1^n, \hat{x}_1^n)\}$ is defined as

$$\begin{aligned}
A_{d,\epsilon}^{(n)} = \{(x_1^n, \hat{x}_1^n) \mid \max(&|-\frac{1}{n}\log p(x_1^n) - H(X)|, \quad |-\frac{1}{n}\log p(\hat{x}_1^n) - H(\hat{X})|, \\
&|-\frac{1}{n}\log p(x_1^n, \hat{x}_1^n) - H(X, \hat{X})|, \quad |d(x_1^n, \hat{x}_1^n)) - \mathbb{E}d(X, \hat{X})|) < \epsilon\}
\end{aligned}$$

where $p(x_1^n, \hat{x}_1^n) = \prod_{i=1}^n p(x_i, \hat{x}_i)$, $d(x_1^n, \hat{x}_1^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$.

Theorem) Joint AEP.

Let $(X_1^n, \hat{X}_1^n) \stackrel{i.i.d}{\sim} p_{\hat{X}|X}p_X$. Then,

1. $\mathbb{P}((X_1^n, \hat{X}_1^n) \in A_{\epsilon,d}^{(n)}) \to 1$ as $n \to \infty$

2. $\forall (x_1^n, \hat{x}_1^n) \in A_{\epsilon,d}^{(n)})$,

$$p(\hat{x}_1^n) \geq p(\hat{x}_1^n|x_1^n)2^{-n(I(X;\hat{X})-3\epsilon)}$$

Proof. 1 is trivial. For 2,

$$p(\hat{x}_1^n) = \frac{p(x_1^n, \hat{x}_1^n)}{p(x_1^n)} = p(\hat{x}_1^n)\frac{p(x_1^n, \hat{x}_1^n)}{p(x_1^n)p(\hat{x}_1^n)}$$
$$\geq p(\hat{x}_1^n)\frac{2^{-n(H(X,\hat{X})-\epsilon)}}{2^{-n(H(X)-\epsilon)}2^{-n(H(\hat{X})-\epsilon)}}$$
$$\geq p(\hat{x}_1^n, |x_1^n)2^{-n(I(X;\hat{X})-3\epsilon)}$$

□

Theorem) R-D theorem. Assume that a distortion measure d is bounded. Then, if $R \geq R^{(I)}(D)$, then (R, D) is achievable. Conversely, any code that achieves distortion D with rate R must satisfy $R \geq R^{(I)}(D)$.

Proof. We assume that $R \geq R^{(I)}(D)$. Fix $\delta > 0$. To show that (R, D) is achievable, we need to construct encoding and decoding functions (f_n, g_n) with index set $I = [2^{nR}]$ satisfying $D_n = \mathbb{E}d(X_1^n, g_n(f_n(X_1^n))) \leq D + \delta$. First, generate $\hat{X}_i(w) \overset{i.i.d.}{\sim} p_{\hat{X}|X}$, $\forall i \in [n]$, $\forall w \in [2^{nR}]$. For $T(X_1^n) = \{w \in [2^n R] | (X_1^n, \hat{X}_1^n(w)) \in A_{d,\epsilon}^{(n)}\}$, define an encoding function $f_n : \mathcal{X}^n \to [2^{nR}]$

$$f_n(X_1^n) = \begin{cases} \min_{w \in T(X_1^n)}(w) & \text{if } T(X_1^n) \neq \emptyset \\ 1 & \text{o.w.} \end{cases}$$

Define a decoding function $g_n : [2^{nR}] \to \hat{\mathcal{X}}^n \cong \mathcal{X}^n$

$$g_n(w) = \hat{X}_1^n(w)$$

Note that $\hat{X}_1^n(X_1^n) := g_n(f_n(X_1^n))$ is a r.v. since it is a function of $\hat{X}_1^n$ and $\hat{X}_1^n$. Compute $\mathbb{E}_{(X_1^n, \hat{X}_1^n)} d(X_1^n, \hat{X}_1^n(X_1^n))$ as follows.

$$\mathbb{E}_{X \sim p_X, \hat{X} \sim p_{\hat{X}|X}} d(X_1^n, \hat{X}_1^n(X_1^n)) = \mathbb{E}_{X \sim p_X} \mathbb{E}_{\hat{X} \sim p_{\hat{X}|X}} d(X_1^n, \hat{X}_1^n(X_1^n))$$
$$= \mathbb{E}_{X \sim p_X} \mathbb{E}_{\hat{X} \sim p_{\hat{X}|X}, T(X_1^n) \neq \emptyset} d(X_1^n, \hat{X}_1^n(X_1^n)) + \mathbb{E}_{X \sim p_X} \mathbb{E}_{\hat{X} \sim p_{\hat{X}|X}, T(X_1^n) = \emptyset} d(X_1^n, \hat{X}_1^n(X_1^n))$$
$$\leq 1 \cdot (D_n + \epsilon) + \mathbb{P}((X_1^n, \hat{X}(w)_1^n) \notin A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}]) \cdot d_{\max}$$

Let's bound $\mathbb{P}((X, \hat{X}(w)) \notin A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}])$ as follows.

$$\mathbb{P}((X_1^n, \hat{X}_1^n(w)) \notin A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}]) = \sum_{x_1^n} p(x_1^n) \sum_{\hat{x}_1^n : (x_1^n, \hat{x}_1^n(w)) \notin A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}]} p(\hat{x}_1^n)$$
$$= \sum_{x_1^n} p(x_1^n) \sum_{\hat{x}_1^n} p(\hat{x}_1^n) I((x_1^n, \hat{x}_1^n(w)) \notin A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}])$$
$$= \sum_{x_1^n} p(x_1^n) [1 - \sum_{\hat{x}_1^n} p(\hat{x}_1^n) I((x_1^n, \hat{x}_1^n(w)) \in A_{d,\epsilon}^{(n)} \ \forall w \in [2^{nR}])]$$
$$= \int \prod_{w=1}^{2^{nR}} \mathbb{P}_{\hat{X} \sim p_{\hat{X}|x}}((x_1^n, \hat{X}_1^n(w)) \notin A_{d,\epsilon}^{(n)}) \ d\mathbb{P}_X(x_1^n)$$
$$= \int \prod_{w=1}^{2^{nR}} [1 - \mathbb{P}_{\hat{X} \sim p_{\hat{X}|x}}((x_1^n, \hat{X}_1^n(w)) \in A_{d,\epsilon}^{(n)})] \ d\mathbb{P}_X(x_1^n)$$

Conversely, assume that we have a code with distortion less than D. Then,

$$\begin{aligned}
nR &\geq H(\hat{X}_1^n) \\
&\geq H(\hat{X}_1^n) - H(\hat{X}_1^n|X_1^n) = I(X_1^n, \hat{X}_1^n) \quad (\because \hat{X}_1^n \text{ is a ftn of } X_1^n) \\
&\geq H(X_1^n) - H(X_1^n|\hat{X}_1^n) = \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\hat{X}_1^n, X_1^{i-1}) \quad (\because X_i \overset{i.i.d.}{\sim} p_X) \\
&\geq \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\hat{X}_i) = \sum_{i=1}^n I(X_i, \hat{X}_i) \\
&\geq \sum_{i=1}^n R^{(I)}(\mathbb{E}(d(X_i, \hat{X}_i))) = n\frac{1}{n}\sum_{i=1}^n R^{(I)}(\mathbb{E}(d(X_i, \hat{X}_i))) \\
&\geq nR^{(I)}(\frac{1}{n}\sum_{i=1}^n \mathbb{E}(d(X_i, \hat{X}_i))) \quad (\because R^{(I)} \text{ is convex , Jensen}) \\
&= nR^{(I)}(\mathbb{E}(d(X_1^n, \hat{X}_1^n))) \\
&\geq nR^{(I)}(D) \quad (\because R^{(I)} \text{ is non-increasing})
\end{aligned}$$

□

9 Variational Auto Encoder (VAE)

9.1 Problem Setting

- Given probability space $(\Omega, \mathcal{A}, \mathbb{P})$
- $\mathcal{X} = \mathbb{R}^D$: a data space
- $\mathcal{Z} = \mathbb{R}^d$: a latent space
- Data $x^{(1)}$, $x^{(2)}, \ldots$ are realizations of a r.v. $X : \Omega \to \mathcal{X}$
- Hidden states $z^{(1)}$, $z^{(2)}, \ldots$ are realization of a r.v. $Z : \Omega \to \mathcal{Z}$.
- We assume that $X, Z \sim p_{X,Z}(\ \cdot\ ,\ \cdot\ ; \theta)$ and $Z \sim p_Z(\ \cdot\ ; \theta)$ where $p_Z(\ \cdot\ ; \theta)$ is in the exponential family.
- Conventionally, we simply assume that $p_Z(\ \cdot\ ; \theta) = \mathcal{N}(0, I)$.
- $x^{(i)}$ is governed by $z^{(i)}$. Specifically,
  1. Generate $z^{(i)}$
  2. Then, $X^{(i)} \sim p_{X|Z=z^{(i)}}(\ \cdot\ |z^{(i)}; \theta^*)$

Furthermore, we assume that

1. $p_X(x; \theta) = \int p_{X|Z=z}(x|z; \theta) p_Z(z; \theta) dz$ dz is intractable (so we cannot evaluate or differentiate the marginal likelihood)
2. True posterior density $p_{Z|X=x}(z|x; \theta) = \frac{p_{X|Z=z}(x|z;\theta) p_Z(z;\theta)}{p_X(x;\theta)}$ is intractable (so the EM algorithm cannot be used), and where the required integrals for any reasonable mean-field VB algorithm are also intractable.
3. A large dataset: we have so much data that batch optimization is too costly; we would like to make parameter updates using small minibatches or even single datapoints. Sampling-based solutions, e.g. Monte Carlo EM, would in general be too slow, since it involves a typically expensive sampling loop per datapoint.

9.2 Goal

1. Infer $\hat{\theta}^*$, MAP (MLE) of θ^*
2. Given $x^{(i)}$, generate θ

9.3 The variational bound (Evidence Lower Bound, ELBO)

Recall that we want to obtain $\hat{\theta}^*$, MAP (MLE) of θ^*, that maximize log-likelihood $\log p_X(x;\theta)$. Hence, we start from:

$$\log p_X(x;\theta).$$

To estimate the log-likelihood, we introduce an alternative pdf $q_{Z|X=x}(\cdot|x;\phi)$ of Z depending on x and ϕ. We hope that this pdf would be a proxy of the true posterior $p_{Z|X=x}(\cdot|x;\theta)$. With these pdfs, we do a little trick as follows:

$$\begin{aligned}
\log p_X(x;\theta) &= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log p_X(x;\theta)] \\
&= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log(p_X(x;\theta)\cdot \frac{q_{Z|X=x}(Z|x;\phi)}{q_{Z|X=x}(Z|x;\phi)} \cdot \frac{p_{Z|X=x}(Z|x;\theta)}{p_{Z|X=x}(Z|x;\theta)})]
\end{aligned}$$

Then we extract the KL-divergence between $q_{Z|X=x}(\cdot|x;\phi)$ and $p_{Z|X=x}(\cdot|x;\theta)$:

$$\begin{aligned}
\log p_X(x;\theta) &= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log(p_X(x;\theta)\cdot \frac{p_{Z|X=x}(Z|x;\theta)}{q_{Z|X=x}(Z|x;\phi)}) + \log\frac{q_{Z|X=x}(Z|x;\phi)}{p_{Z|X=x}(Z|x;\theta)}] \\
&= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log(p_X(x;\theta)\cdot \frac{p_{Z|X=x}(Z|x;\theta)}{q_{Z|X=x}(Z|x;\phi)})] + KL(q_{Z|X=x}(Z|x;\phi)\|p_{Z|X=x}(Z|x;\theta)) \\
&= \mathcal{L}(\theta,\phi;x) + KL(q_{Z|X=x}(\cdot|x;\phi)\|p_{Z|X=x}(\cdot|x;\theta)) \\
&\geq \mathcal{L}(\theta,\phi;x)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}(\theta,\phi;x) &= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[-\log q_{Z|X=x}(Z|x;\phi) + \log p_{X,Z}(x,Z;\theta)] \\
&= \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[-\log q_{Z|X=x}(Z|x;\phi) + (\log p_Z(Z;\theta) + \log p_{X|Z}(x|Z;\theta))] \\
&= -KL(q_{Z|X=x}(\cdot|x;\phi)\|p_Z(\cdot;\theta)) + \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log p_{X|Z}(x|Z;\theta))] \\
&= \text{regularizer for } \phi + \text{negative reconstruction error} \\
&=: \text{ELBO}(\theta,\phi;x).
\end{aligned}$$

Hence, maximizing the ELBO means maximizing the log-likelihood $p_X(x;\theta)$.

9.4 The SGVB estimator

Now, we want to maximize the ELBO:

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[-\log q_{Z|X=x}(Z|x;\phi) + \log p_{X,Z}(x,Z;\theta)]$$

To this end, we need to generate $Z \sim q_{Z|X=x}(\cdot|x;\phi)$. As direct sampling from $q_{Z|X=x}(\cdot|x;\phi)$ is impossible, we *reparameterize* it by

$$\tilde{z} = g(\epsilon, x;\phi) \text{ with } \epsilon \sim r_\epsilon$$

where $g(\epsilon, x; \phi)$ is a differentiable transformation and r_ϵ is a distribution that is easy to sample. Hence, with the sampled $x = x^{(i)}$, Stochastic Gradient Variational Bayes (SGVB) estimator is defined as follows:

$$\mathcal{L}^A(\theta, \phi; x^{(i)}) \approx \frac{1}{L}\sum_{l=1}^{L}[-\log q_{Z|X=x^{(i)}}(\tilde{z}_l^{(i)}|x^{(i)}; \phi) + \log p_{X,Z}(x^{(i)}, \tilde{z}_l^{(i)}; \theta)]$$

where $\tilde{z}_l^{(i)} = g(\epsilon_l, x^{(i)}; \phi)$ with $\epsilon_l \overset{i.i.d.}{\sim} r_\epsilon$.

9.5 The AEVB estimator

The ELBO can be written in another form as well:

$$\mathcal{L}(\theta, \phi; x) = -KL(q_{Z|X=x}(\cdot|x; \phi)\|p_Z(\cdot; \theta)) + \mathbb{E}_{Z\sim q_{Z|X=x}(\cdot|x;\phi)}[\log p_{X|Z}(x|Z; \theta)).$$

Assume that we can analytically integrate $KL(q_{Z|X=x}(\cdot|x; \phi)\|p_Z(\cdot; \theta))$. Under such assumption and sampled $x = x^{(i)}$, Auto Encoding Variational Bayes (AEVB) estimator is defined as:

$$\mathcal{L}^B(\theta, \phi; x^{(i)}) \approx -KL(q_{Z|X=x^{(i)}}(\cdot|x^{(i)}; \phi)\|p_Z(\cdot; \theta)) + \frac{1}{L}\sum_{l=1}^{L}[\log p_{X|Z=\tilde{z}_l^{(i)}}(x^{(i)}|\tilde{z}_l^{(i)}; \theta))]$$

where $\tilde{z}_l^{(i)} = g(\epsilon_l, x^{(i)}; \phi)$ with $\epsilon_l \overset{i.i.d.}{\sim} r_\epsilon$.

Exercise) VAE.

a) Indeed, if we assume $\epsilon_l \overset{i.i.d.}{\sim} r_\epsilon = \mathcal{N}(0, I)$, the analytic integration becomes possible. With $\mu^{(i)} \in \mathbb{R}^d$ and $\sigma^{(i)} \in \mathbb{R}_+^d$, i.e., outputs of the encoding MLP for $x^{(i)}$ under variational parameters ϕ, we obtain $\tilde{z}_l^{(i)}$ as follows:

$$\tilde{z}_l^{(i)} = g(\epsilon_l, x^{(i)}; \phi) := \mu^{(i)} + \sigma^{(i)} \cdot \epsilon_l.$$

Hence, we have $\tilde{z}^{(i)} \sim \mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$ where $\Sigma^{(i)} = \mathrm{diag}((\sigma_1^{(i)})^2, \ldots, (\sigma_d^{(i)})^2)$. Now, we can do the analytical integration by computing the KL divergence between two normal distribution $\mathcal{N}(0, I)$ and $\mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$:

$$\begin{aligned}\mathcal{L}^B(\theta, \phi; x^{(i)}) \approx& \frac{1}{2}(d + \sum_{k=1}^{d}\log(\sigma_k^{(i)}) - \|\mu^{(i)}\|^2 - \sum_{k=1}^{d}\sigma_k^{(i)\,2}) \\ &+ \frac{1}{L}\sum_{l=1}^{L}[\log p_{X|Z=\tilde{z}_l^{(i)}}(x^{(i)}|\tilde{z}_l^{(i)}; \theta))].\end{aligned}$$

10 Parsing

10.1 CKY algorithm

We are given as follows:

- CFG (N, Σ, R, S) where N (Σ) is a of non-terminals (terminals), R is a set of rules, and $S \in N$ is a start non-terminal (NT).
- R is in CNF, i.e., $r \in R$ is either $(X \to Y_1 Y_2$ for some $X, Y_1, Y_2 \in \Sigma)$ or $(X \to \beta$ for some $X \in \Sigma$ and $\beta \in N)$.
- q is a probability over R, i.e., we have a PCFG.
- $s = w_1 \cdots w_n$ is a sentence of n tokens.

Our goal is to find the most probable derivation t of s.

Define a Chart To achieve this goal, CKY algorithm defines a $n^2|N|$-sized chart π where each cell $\pi(i, j, X)$ is the maximum probability of a tree with the root X spanning $w_i \cdots w_j$ for $i, j \in \{1, \ldots, n\}$ and $X \in N$. Then, our goal is to find the derivation that attains $\pi(1, n, S)$.

Dynamic Programming Note that we have following base cases:

$$\pi(i, i, X) = q(X \to w_i)$$

and the recursive formula:

$$\pi(i, j, X) = \max_{\substack{\forall k \in \{i, \ldots, j\}, \\ \forall (X \to YZ) \in R}} q(X \to YZ) \cdot \pi(i, k, Y) \cdot \pi(k, j, Z)$$

To fill out each cell, we need to check $n|N|^2$ candidates.

Complexity We have $n^2|N|$ cells, each of which require $n|N|^2$ checks. Hence, the computational cost is $n^3|N|^3$.

10.2 Lexicalized PCFGs

Weaknesses of PCFGS Lack of (1) sensitivity to lexical information, e.g., `workers dumped sacks into a bin`, and (2) structural frequencies, e.g., `dogs in houses and cats`. Refer to the the slide 50-55.

Lexicalization Main idea is to lexicalize rules to get lexical information at every node, i.e., define a head for every rule in the grammar (one child on RHS). This defines the head word of every phrase, which will provide a lexicaliztation of rules

The immediate question is, how to determine a head for each rule?. This is called *lexicalization rule*. Basically, a VP has a verb as a head and a NP has a noun as a head. Details about lexicalization rule is manually defined. Refer to the the slide 63- about lexicalized PCFGs.