Lecture Notes for Information Theory (ECE 587/STA 563)

Jyunyi Liao

Contents

1 Measure of Information

Throughout this section, we assume that all random variables we study are discrete variables. We use capital letters like X, Y, Z to denote random variables, and their probability mass functions $p_X(x), p_Y(y), p_Z(z)$. For simplicity, we drop the subscripts and use the shorthand $p(x)$, $p(y)$, $p(z)$ instead. We use calligraphy letters like $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to denote the finite support of random variables.

1.1 Entropy and Conditional Entropy

Definition 1.1 (Entropy)**.** Let X be a random variable supported on a finite state space $\mathcal{X}$, with probability mass function $p(x)$. The *entropy* of X is a function of the distribution $p(x)$:

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

Likewise, for a collection $X_1, \cdots, X_n$ of random variables, the (joint) entropy of $X_1, \cdots, X_n$ is defined as the entropy of the random vector $(X_1, \cdots, X_n)$:

$$H(X_1, \cdots, X_n) = \sum_{x_1\in\mathcal{X}_1, \cdots, x_n\in\mathcal{X}_n} p(x_1, \cdots, x_n)\log\frac{1}{p(x_1, \cdots, x_n)}.$$

Remark I. The entropy provides a measure of uncertainty of random variables. We also frequently use the *binary entropy function* $h : [0, 1] \to \mathbb{R}_+$, which is defined as the entropy of a Bernoulli variable:

$$H(\alpha) = H(\text{Bernoulli}(\alpha)) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha), \quad \alpha \in [0, 1]$$

with the convention $0 \log 0 = 0$.

Remark II. Given any base $b > 0$, we define the entropy of X under base b to be

$$H_b(X) = \sum_{x\in\mathcal{X}} p(x)\log_b\frac{1}{p(x)} = H(X)\log_b \mathrm{e}.$$

Clearly we have $H(X) = H_{\mathrm{e}}(X)$. Another commonly used entropy is the bit entropy, in which the base $b = 2$:

$$H_2(X) = \sum_{x\in\mathcal{X}} p(x)\log_2\frac{1}{p(x)} = H(X)\log_2 \mathrm{e}.$$

Proposition 1.2. *We have the following estimate for the entropy of a random variable X:*

$$0 \le H(X) \le \log|\mathcal{X}|.$$

Proof. The lower bound follows from the definition of entropy. For the upper bound, note that

$$\begin{aligned}\sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} &= \sum_{x\in\mathcal{X}} p(x)\log\frac{|\mathcal{X}|}{p(x)|\mathcal{X}|} = \log|\mathcal{X}| + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)|\mathcal{X}|} \\ &\le \log|\mathcal{X}| + \sum_{x\in\mathcal{X}} p(x)\left(\frac{1}{p(x)|\mathcal{X}|} - 1\right) = \log|\mathcal{X}|.\end{aligned}$$

Then we complete the proof. □

Remark. If $|\mathcal{X}| = \infty$, the entropy of a random variable can be ∞. For example, let $A = \sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$, which is less than infinity. Define random variable X by

$$\mathbb{P}(X = n) = \frac{1}{An(\log n)^2}, \quad n = 2, 3, \cdots .$$

Then

$$H(X) \geq \int_2^{\infty} \frac{\log A}{x \log x}\, dx = \infty.$$

We may also wonder the uncertainty of a random variable when given potentially relevant observation.

Definition 1.3 (Conditional Entropy)**.** Let X and Y be two random variables in the same probability space. The entropy of Y conditioned on the event $X = x$ is a function of the conditional distribution $p(y|x)$:

$$H(Y|X = x) := \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)} = \mathbb{E}\left[\log \frac{1}{p(Y|x)} \middle| X = x \right].$$

The conditional entropy of Y given X is a function of the joint distribution $p(x, y)$:

$$H(Y|X) := \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} = \mathbb{E}\left[\log \frac{1}{p(Y|X)} \right].$$

Remark. Note that $H(Y|X)$ is a deterministic quantity rather than a random variable. In fact, we have

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

Next, we study the relation between joint entropy and conditional entropy.

Proposition 1.4 (Chain rule for entropy)**.** *The joint entropy of X and Y has the following decomposition:*

$$H(X, Y) = H(Y|X) + H(X). \tag{1.1}$$

More generally,

$$H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \cdots, X_1). \tag{1.2}$$

Proof. We first verify the bivariate case (1.1):

$$\begin{aligned} H(Y|X) + H(X) &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} + \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} = H(X, Y). \end{aligned}$$

The general case (1.2) follows from mathematical induction. □

Remark. The equality (1.1) also implies the chain rule for conditional entropy:

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

1.2 Mutual Information

Definition 1.5 (Mutual information)**.** Let X and Y be two discrete random variables in the same probability space. The mutual information of X and Y is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

Proposition 1.6 (Properties of mutual information)**.** *Let X and Y be two discrete random variables.*

(i) (Symmetry). $I(X;Y) = I(Y;X)$.

(ii) (Reduction). $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

(iii) (Measure of dependency). $I(X;Y) \geq 0$, and the equality holds if and only if X and Y are independent.

Proof. The assertion (i) follows from definition, and the second from direct calculation. Now we verify (iii):

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \geq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\left(1 - \frac{p(x)p(y)}{p(x,y)}\right) = 0.$$

Clearly, the equality holds if and only if $p(x,y) = p(x)p(y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. □

Remark. Combining (ii) and (iii), we see that *conditioning does not increase entropy*:

$$H(X|Y) \leq H(X), \quad \textit{and} \quad H(Y|X) \leq H(Y).$$

Based on this property, we introduce an important property of entropy as the function of distribution.

Theorem 1.7 (Concavity of entropy)**.** *Let p and q be two probability distributions that are supported in a common space $\mathcal{X}$. Then for all $0 \leq \lambda \leq 1$, we have*

$$H(\lambda p + (1-\lambda)q) \geq \lambda H(p) + (1-\lambda)H(q). \tag{1.3}$$

Proof. Let $X_1 \sim p$ and $X_2 \sim q$ be two independent random variables, and let $Z \sim \text{Bernoulli}(\lambda)$. Define

$$X_\lambda = X_1 Z + X_2(1-Z).$$

Then $X_\lambda \sim \lambda p + (1-\lambda)q$, and

$$H(X_\lambda) \geq H(X_\lambda|Z) = \lambda H(X_\lambda|Z=1) + (1-\lambda)H(X_\lambda|Z=0) = \lambda H(X_1) + (1-\lambda)H(X_2).$$

This is in fact the equality (1.3). □

Remark. Using the concavity, we can interpret why a transfer of probability that makes the distribution more uniform increases the entropy. We consider the following transformation:

$$(p_1, \cdots, p_i, \cdots, p_j, \cdots, p_m) \to \left(p_1, \cdots, \frac{p_i+p_j}{2}, \cdots, \frac{p_i+p_j}{2}, \cdots, p_m\right), \quad p_1 + \cdots + p_m = 1.$$

Let $p = (p_1, \cdots, p_i, \cdots, p_j, \cdots, p_m)$, and let $q = (p_1, \cdots, p_j, \cdots, p_i, \cdots, p_m)$ be the probability vector with i-th and j-th elements exchanged. Then

$$H\left(\frac{p+q}{2}\right) \geq \frac{1}{2}H(p) + \frac{1}{2}H(q) = H(p).$$

Mutual information as a function of distribution. If $p(x, y)$ is the joint probability mass function of random variables X and Y. The mutual information $I(X;Y)$ is in fact a function of p and does not depend on the probability space where X and Y are defined. We can write $I(X;Y) = I(p)$ with $(X, Y) \sim p$.

We consider the decomposition $p(x, y) = p(x)p(y|x)$, where $p(x)$ is the marginal distribution of X and $p(y|x)$ is the conditional distribution of Y given x. Then the mutual information between X and Y is a function of $p(x)$ and $p(y|x)$:

$$I(p(x), p(y|x)) := \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x)p(y|x) \log \frac{p(y|x)}{p(y)}, \quad \textit{where} \;\; p(y) = \sum_{x\in\mathcal{X}} p(x)p(y|x).$$

Proposition 1.8 (Marginal convexity of mutual information)**.** *Let $0 \le \lambda \le 1$. Let $p(x)$ and $q(x)$ be two distributions of X, and let $p(y|x)$ and $q(y|x)$ be two conditional distributions of Y given X. Then*

$$I(\lambda p(x) + (1-\lambda)q(x), p(y|x)) \le \lambda I(p(x), p(y|x)) + (1-\lambda) I(q(x), p(y|x)),$$

and

$$I(p(x), \lambda p(y|x) + (1-\lambda)q(y|x)) \le \lambda I(p(x), p(y|x)) + (1-\lambda) I(p(x), q(y|x)).$$

Proof. Let $Z \sim \text{Bernoulli}(\lambda)$, $X_1 \sim p$, $X_2 \sim q$, $X_\lambda = X_1 Z + X_2(1-Z)$, and $Y|X_\lambda \sim p(y|x)$. Then

$$I(\lambda p(x) + (1-\lambda)q(x), p(y|x)) = I(X_\lambda; Y).$$

Since conditioning does not increase entropy,

$$\begin{aligned} I(X_\lambda; Y) \le I(X_\lambda; Y, Z) &= I(X_\lambda; Y|Z) + I(X_\lambda; Z) \\ &= \mathbb{P}(Z=1) I(X_\lambda; Y|Z=1) + \mathbb{P}(Z=0) I(X_\lambda; Y|Z=0) + 0 \\ &= \lambda I(X_1; Y) + (1-\lambda) I(X_2; Y) \\ &= \lambda I(p(x), p(y|x)) + (1-\lambda) I(q(x), p(y|x)). \end{aligned}$$

This proves the first inequality. The second one follows in a similar approach. □

Similar to the conditional entropy, we can define the conditional mutual information.

Definition 1.9. Let X, Y and Z be discrete random variables in the same probability space. The conditional mutual information of X and Y given Z is defined as

$$I(X;Y|Z) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}, z\in\mathcal{Z}} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}.$$

Similar to Proposition 1.6, conditional mutual information has the following properties.

Proposition 1.10 (Properties of conditional mutual information)**.** *Let X, Y and Z be discrete random variables in the same probability space.*

(i) (Symmetry). $I(X;Y|Z) = I(Y;X|Z)$.

(ii) (Reduction). $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z)$.

(iii) (Measure of dependency). $I(X;Y|Z) \ge 0$, and the equality holds if and only if X and Y are conditionally independent on Z.

By direct calculation and induction, we also have the following chain rule for mutual information.

Proposition 1.11 (Chain rule for mutual information)**.** *The mutual information $I(X;Y,Z)$ has the following decomposition:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z).$$

More generally,

$$I(X;Y_1,Y_2,\cdots,Y_n) = I(X;Y_1) + I(X;Y_2|Y_1) + I(X;Y_3|Y_2,Y_1)\cdots + I(X;Y_n|Y_{n-1},\cdots,Y_1).$$

We can use this rule to derive the data processing inequality for Markov chains.

Definition 1.12 (Markov chain)**.** Random variables X, Y and Z are said to form a Markov chain, written $X \to Y \to Z$, if X and Z are conditionally independent on Y:

$$p(x,z|y) = p(x|y)p(z|y).$$

Particularly, if $Z = g(Y)$ is a function of Y, then $X \to Y \to Z$.

The following theorem asserts that no manipulation of Y can increase the mutual information.

Theorem 1.13 (Data processing inequality)**.** *If $X \to Y \to Z$, then*

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

Particularly, for any function g defined on $\mathcal{Y}$, we have

$$I(X;Y) \geq I(X;g(Y)).$$

Proof. By chain rule, we have that

$$I(X;Y) + I(X;Z|Y) = I(X;Y,Z) = I(X;Z) + I(X;Y|Z).$$

Since $X \perp\!\!\!\perp Z \,|\, Y$, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, the result follows. □

Remark. By Proposition 1.6, we also have $H(X|Z) \geq H(X|Y)$ when $X \to Y \to Z$.

Next, we introduce an alternative definition of mutual information.

Definition 1.14 (Kullback-Leibler divergence/relative entropy)**.** Let p and q be two probability distributions such that $\mathcal{X} = \operatorname{supp} q \supset \operatorname{supp} p$. The Kullback-Leibler divergence of q from p is defined as

$$D(p\|q) := \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\left[\log\frac{p(X)}{q(X)}\right].$$

This is also known as the relative entropy.

Remark. By definition, we have

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \geq \sum_{x\in\mathcal{X}} p(x)\left(1 - \frac{q(x)}{p(x)}\right) = 0.$$

Therefore, $D(p\|q) \geq 0$, and the equality holds if and only if $p = q$. Moreover, by definition, we have the following result:

$$I(X;Y) = D(p_{X,Y}\|p_X p_Y) = \mathbb{E}_{X\sim p_X}\left[D(p_{Y|X}\|p_Y)\right].$$

In other words, the mutual information of X and Y is the relative entropy of their marginal product $p_X p_Y$ from their joint distribution $p_{X,Y}$.

Application: Misclassification Rate. To end this section, we introduce a useful application of mutual information. We discuss the estimation of a discrete random variable X from an observation Y. To deal with this problem, we construct a function $\phi : \mathcal{Y} \to \mathcal{X}$. The probability of error of the estimator $\widehat{X} = \phi(Y)$ is

$$p_e = \mathbb{P}(\widehat{X} \neq X).$$

The following Fano's inequality provide a lower bound of the error rate p_e.

Theorem 1.15 (Fano's inequality)**.** *For any estimator $\widehat{X}$ of X such that $X \to Y \to \widehat{X}$, we have*

$$H(X|Y) \leq h(p_e) + p_e \log|\mathcal{X}|.$$

Particularly, we have

$$p_e \geq \frac{H(X|Y) - \log 2}{\log|\mathcal{X}|}.$$

Proof. Let $B = \mathbb{1}_{\{X=\widehat{X}\}}$, which is a Bernoulli variable with parameter p_e. By the chain rule, the conditional entropy of (B, X) given $\widehat{X}$ is

$$H(B|\widehat{X}) + H(X|B,\widehat{X}) = H(B,X|\widehat{X}) = H(X|\widehat{X}) + H(B|X,\widehat{X}).$$

Now we analyze the four terms in the equality.

(i) Since conditioning does not increase entropy, $H(B|\widehat{X}) \leq H(B) = h(p_e)$.

(ii) The conditional entropy $H(X|B,\widehat{X})$ has the following estimate:

$$\begin{aligned}
H(X|B,\widehat{X}) &= \sum_{b\in\{0,1\}}\sum_{x\in\mathcal{X}}\sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=b, X=x, \widehat{X}=\widehat{x}) \log\frac{1}{\mathbb{P}(X=x|B=b,\widehat{X}=\widehat{x})} \\
&= \sum_{x\in\mathcal{X}}\sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=0, X=x, \widehat{X}=\widehat{x}) \log\frac{1}{\mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x})} \\
&= \sum_{\widehat{x}\in\mathcal{X}} \mathbb{P}(B=0,\widehat{X}=\widehat{x}) \underbrace{\sum_{x\in\mathcal{X}} \mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x}) \log\frac{1}{\mathbb{P}(X=x|B=0,\widehat{X}=\widehat{x})}}_{\leq \log|\mathcal{X}|} \leq p_e \log|\mathcal{X}|.
\end{aligned}$$

(iii) Since $X \to Y \to \widehat{X}$, the data processing inequality implies $H(X|\widehat{X}) \geq H(X|Y)$.

(iv) Since B is a function of X and $\widehat{X}$, we have $H(B|X,\widehat{X}) = 0$.

Combining these estimates, we obtain

$$H(X|Y) \leq h(p_e) + p_e \log|\mathcal{X}| \leq \log 2 + p_e \log|\mathcal{X}|.$$

Then we complete the proof. □

1.3 Typical Sets and Asymptotic Equipartition Property

In this section, we investigate a sequence of i.i.d. copies $X_1, X_2, \cdots$ of a random variable $X \sim p(x)$ with finite support $\mathcal{X}$. We write for a random vector of length n and its realization

$$X_{1:n} = (X_1, \cdots, X_n), \quad x_{1:n} = (x_1, \cdots, x_n).$$

The joint distribution of $X_{1:n}$ is given by

$$p(x_{1:n}) = \mathbb{P}\left(X_{1:n} = x_{1:n}\right) = p(x_1)p(x_2)\cdots p(x_n).$$

In this section, we focus on finding a confidence set $A \subset \mathcal{X}^n$ that contains our observation $X_{1:n}$ with a high probability. Formally, we require $\mathbb{P}(X_{1:n} \in A) \geq 1 - \delta$, where $\delta > 0$ is an arbitrarily given small quantity.

Typical Sets. Here is an idea of constructing high probability sets. Let $g : \mathcal{X} \to \mathbb{R}$ be a function such that $\mathbb{E}|g(X)| < \infty$. By the weak law of large numbers, for each $\epsilon > 0$ and $\delta > 0$, there exists $N_{\epsilon,\delta} > 0$ such that

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

Consequently, almost all probability mass is concentrated on the following set A:

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

In the last display, the constraint can be equivalently expressed as

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

The construction of typical sets follows by plugging in $g(x) = \log_2 \frac{1}{p(x)}$.

Definition 1.16. The ϵ-typical set is defined by

$$A_\epsilon^{(n)} = \left\{x_{1:n} \in \mathcal{X}^n : 2^{-n(H_2(X)+\epsilon)} \leq p(x_{1:n}) \leq 2^{-n(H_2(X)-\epsilon)}\right\},$$

or equivalently, the set of all tuples $x_{1:n} \in \mathcal{X}^n$ obeying

$$H_2(X) - \epsilon \leq -\frac{1}{n}\log_2 p(x_{1:n}) \leq H_2(X) + \epsilon.$$

Clearly, for each $\delta > 0$, there exists a positive integer $N_{\epsilon,\delta}$ such that for all $n > N_{\epsilon,\delta}$, the typical $A_\epsilon^{(n)}$ contains $X_{1:n}$ with probability at least $1 - \delta$. In other words,

$$\lim_{n\to\infty} \mathbb{P}\left(X_{1:n} \in A_\epsilon^{(n)}\right) = 1.$$

Size of Typical Sets. When n increased, the number of possible realizations of $X_{1:n}$ would rise very quickly, which is $|\mathcal{X}|^n$. The idea of typical sets is to concentrate the probability mass of $X_{1:n}$ on a smaller set $A_\epsilon^{(n)}$:

$$A_\epsilon^{(n)} = \left\{x_{1:n} \in \mathcal{X}^n : 2^{-n(H_2(X)+\epsilon)} \leq p(x_{1:n}) \leq 2^{-n(H_2(X)-\epsilon)}\right\}.$$

In this set, all tuples have roughly the same probability mass. This is know as the *Asymptotic Equipartition property* (AEP). Here is an intuition of this typical set:

- For the low probability tuples $p(x_{1:n}) < 2^{-n(H_2(X)+\epsilon)}$, they are too unlikely to matter;
- For the high probability tuples $p(x_{1:n}) > 2^{-n(H_2(X)-\epsilon)}$, they are too few to matter;
- Therefore, we exclude those unimportant tuples and retain only the average probability tuples.

We now study the size of the reduced set.

Proposition 1.17. *Let $A_\epsilon^{(n)}$ be the ϵ-typical set for $X_{1:n}$. For each $\delta > 0$, there exists $N_{\epsilon,\delta} > 0$ such that*

$$\mathbb{P}\left(X_{1:n} \in A_\epsilon^{(n)}\right) \geq 1 - \delta, \quad \forall n \geq N_{\epsilon,\delta}.$$

Furthermore, the upper bound of the typical set is given by

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n(H_2(X)+\epsilon)}, \quad \forall n \geq 1;$$

and the lower bound of the typical set is given by

$$\left|A_\epsilon^{(n)}\right| \geq (1-\delta)2^{n(H_2(X)-\epsilon)}, \quad \forall n \geq N_{\epsilon,\delta}.$$

Proof. For the upper bound, note that

$$1 = \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n}) \geq \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) \geq \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X)+\epsilon)}.$$

For the lower bound, when $n \geq N_{\epsilon,\delta}$, we have

$$1 - \delta \leq \mathbb{P}\left(X_{1:n} \in A_\epsilon^{(n)}\right) = \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) \leq \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X)-\epsilon)}.$$

Rearranging each inequality completes the proof. □

Application: data compression. A *source code* is a mapping C from a sequence of symbols from an information source $\mathcal{X}$ to a sequence of alphabet symbols $\mathcal{D}$ (usually bits $\mathcal{D} = \{0, 1\}$) such that the source symbols can be exactly recovered from the bit sequence *(lossless source coding)* or recovered within some distortion *(lossy source coding)*. This is one approach to data compression.

We will discuss lossless coding in Chapter 2. Let us first focus on lossy source coding. Suppose the input is a sequence of i.i.d. random variables $X_1, \cdots, X_n \sim p$, and we want to compress a sequence of length n to a bit sequence. In other words, we want to find a source code $C : \mathcal{X}^n \to \{0, 1\}^*$, where $\{0, 1\}^*$ is the set of all bit sequences of finite length. The *rate* R of this code is the average length per symbol:

$$R = \frac{1}{n} \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n}) \times \text{length of } C(x_{1:n})$$

For the compression efficiency, we wish to minimize the average length per symbol. Furthermore, we also want to recover the original sequence from the code. We consider the following encoding algorithm:

- For each sequence $x_{1:n}$ in the typical set $A_\epsilon^{(n)}$, since the size of $A_\epsilon^{(n)}$ is no more than $n(H_2(X) + \epsilon)$, the encoder assigns a unique bit sequence of length $\lceil n(H_2(X) + \epsilon) \rceil$;
- Otherwise, the encoder throws an arbitrary bit sequence of length $\lceil n(H_2(X) + \epsilon) \rceil$.

For any probability of error $\delta > 0$, when n is sufficiently large, the input sequence falls in the typical set with

probability at least $1-\delta$, and the encoder does not make an error. Furthermore, the rate of this code satisfies

$$R = \frac{1}{n}\lceil n(H_2(X)+\epsilon)\rceil \leq H_2(X) + \epsilon + \frac{1}{n} \to H_2(X)+\epsilon, \quad as \;\; n \to \infty.$$

Theorem 1.18 (Shannon's source encoding theorem)**.** *The minimum rate R at which an information source can be compressed with negligible probability of error is the entropy rate $H_2(X)$ (in bits per symbol) of the source. This statement involves two aspects:*

(i) (Achievability) For each $\epsilon > 0$, there exists a source code with rate R no greater than $H_2(X)+\epsilon$ and negligible probability of error as the block length $n \to \infty$.

(ii) (Converse) Any source code with rate $R < H_2(X)$ has probability error bounded away from 0 as $n \to \infty$.

Proof. The achievability part is established by our preceding discussion. To prove the converse part, we use the following technical result:

Lemma 1.19. *Let $X_1, \cdots, X_n$ be i.i.d. variables drawn from $X \sim p$. For $0 < \delta < 1$, define*

$$S_\delta(n) = \inf\left\{|A| : A \in \mathcal{X}^n \text{ and } p(A) \geq 1-\delta\right\},$$

where we also write p for the joint distribution of $(X_1, X_2, \cdots, X_n)$ for simplicity. Then

$$\lim_{n\to\infty} \frac{\log S_\delta(n)}{n} = H(X).$$

For any $0 < \delta < 1$, to ensure that the probability of error no greater than δ, we require the source code to be one-to-one on a subset $A_n \subset \mathcal{X}^n$ with probability at least $1-\delta$. If the code has rate $R < H_2(X)$, then

$$\lim_{n\to\infty} \frac{\log_2 |A_n|}{n} = R < H_2(X),$$

which contradicts Lemma 1.19! Then we complete the proof. □

Remark. Since the number $0 < \delta < 1$ is arbitrarily chosen, we in fact prove that the probability of error for a source code with rate $R < H_2(X)$ converges to 1 as $n \to \infty$.

Proof of Lemma 1.19. □

1.4 Jointly Typical Sets

In this section, we discuss the construction of typical sets for multiple random variables.

Definition 1.20 (Jointly typical sets)**.** Let $p(x,y)$ be the joint distribution of random variables X and Y. The ϵ-typical set $A_\epsilon^{(n)}$ with respect to the joint distribution $p(x,y)$ is defined by

$$\begin{aligned} A_\epsilon^{(n)} = \big\{(x_{1:n}, y_{1:n}) \in \mathcal{X}^n \times \mathcal{Y}^n : & \, 2^{-n(H_2(X)+\epsilon)} \le p(x_{1:n}) \le 2^{-n(H_2(X)-\epsilon)}, \\ & 2^{-n(H_2(Y)+\epsilon)} \le p(y_{1:n}) \le 2^{-n(H_2(Y)-\epsilon)}, \\ & 2^{-n(H_2(X,Y)+\epsilon)} \le p(x_{1:n}, y_{1:n}) \le 2^{-n(H_2(X,Y)-\epsilon)} \big\}. \end{aligned}$$

Theorem 1.21 (Properties of jointly typical sets)**.** *Let $(X_{1:n}, Y_{1:n})$ be a sequence of length n drawn i.i.d. according to $(X,Y) \sim p(x,y)$. Let $A_n^{(\epsilon)}$ be the joint typical set with respect to $p(x,y)$. Then*

(i) High probability:

$$\lim_{n\to\infty} \mathbb{P}\left((X_{1:n}, Y_{1:n}) \in A_\epsilon^{(n)}\right) = 1.$$

(ii) Estimate of size: for all $n \in \mathbb{N}$,

$$\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X,Y)+\epsilon)};$$

Furthermore, given any $\delta > 0$, for sufficiently large n,

$$\left|A_\epsilon^{(n)}\right| \ge (1-\delta) 2^{n(H(X,Y)-\epsilon)};$$

(iii) Joint asymptotic equipartition property: If $(\widetilde{X}_{1:n}, \widetilde{Y}_{1:n}) \sim p(x_{1:n})p(y_{1:n})$, i.e. $\widetilde{X}_{1:n}, \widetilde{Y}_{1:n}$ are independent with the same marginals as $p(x^n, y^n)$, then

$$\mathbb{P}\left((\widetilde{X}_{1:n}, \widetilde{Y}_{1:n}) \in A_\epsilon^{(n)}\right) \le 2^{-n(I(X;Y)-3\epsilon)}.$$

Furthermore, given any $\delta > 0$, for sufficiently large n,

$$\mathbb{P}\left((\widetilde{X}_{1:n}, \widetilde{Y}_{1:n}) \in A_\epsilon^{(n)}\right) \ge (1-\delta) 2^{-n(I(X;Y)+3\epsilon)}.$$

Proof. By the weak law of large numbers,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\log_2 \frac{1}{p(X_{1:n})} - H_2(X)\right| > \epsilon\right) = 0, \quad \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\log_2 \frac{1}{p(Y_{1:n})} - H_2(Y)\right| > \epsilon\right) = 0, \\ &\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\log_2 \frac{1}{p(X_{1:n}, Y_{1:n})} - H_2(X,Y)\right| > \epsilon\right) = 0. \end{aligned}$$

Since the event $(X_{1:n}, Y_{1:n}) \in A_\epsilon^{(n)}$ is the complement of the union of the three events quantified above, the result (i) follows. To show the first part of (ii), just note that

$$1 \ge \sum_{x_{1:n}, y_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}, y_{1:n}) \ge \sum_{x_{1:n}, y_{1:n} \in A_\epsilon^{(n)}} 2^{-n(H_2(X,Y)+\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X,Y)+\epsilon)}.$$

It remains to show (iii). Since $p(x_{1:n}) \le 2^{-n(H_2(X)-\epsilon)}$ and $p(y_{1:n}) \le 2^{-n(H_2(Y)-\epsilon)}$ for all $(x_{1:n}, y_{1:n}) \in A_\epsilon^{(n)}$,

$$\mathbb{P}\left((\widetilde{X}_{1:n}, \widetilde{Y}_{1:n}) \in A_\epsilon^{(n)}\right) = \sum_{x_{1:n}, y_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n})p(y_{1:n}) \le \left|A_\epsilon^{(n)}\right| 2^{-n(H_2(X)+H_2(Y)-2\epsilon)} \le 2^{-n(I(X;Y)-3\epsilon)}.$$

The other part of (ii) and (iii) are similar. □

1.5 Entropy Rates

In this section, we study a discrete-time stochastic process $X = (X_t)_{t\in\mathbb{N}}$, where each X_t is a random variable in a finite range $\mathcal{X}$. These random variables do not need to be i.i.d..

Definition 1.22. Let $X = (X_t)_{t\in\mathbb{N}}$ be a stochastic process.

(i) Average entropy per symbol

$$H(X) = \lim_{n\to\infty} \frac{H(X_{1:n})}{n}$$

(ii) The k-th order entropy

$$H^k(X) = H(X_k|X_{k-1}, \cdots, X_1)$$

(iii) Rate of information innovation

$$H^\infty(X) = \lim_{k\to\infty} H^k(X) = \lim_{k\to\infty} H(X_k|X_{k-1}, \cdots, X_1)$$

Remark. If $X = (X_t)_{t\in\mathbb{N}}$ is an i.i.d. sequence, we have

$$H(X) = H^\infty(X) = H(X_1).$$

Stationarity. Recall that a stochastic process $X = (X_t)_{t\in\mathbb{N}}$ is said to be *(strongly) stationary* if

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_{k+1} = x_1, \cdots, X_{n+k} = x_n)$$

for every $n \in \mathbb{N}$, every lapse $k \in \mathbb{N}$ and all $x_1, \cdots, x_n \in \mathcal{X}$.

Theorem 1.23. *For a stationary process $X = (X_t)_{t\in\mathbb{N}}$,*

$$H(X) = H^\infty(X).$$

Proof. We first prove the existence of rate of information innovation. By stationarity,

$$H^n(X) = H(X_n|X_{n-1}, \cdots, X_2, X_1) \le H(X_n|X_{n-1}, \cdots, X_2) = H(X_{n-1}|X_{n-2}, \cdots, X_1)$$

Therefore, $H(X_n|X_{n-1}, \cdots, X_1)$ is decreasing in n. Since conditional entropy is nonnegative, the monotone sequence converges: $H^n \searrow H^\infty$. Next, by the chain rule of entropy,

$$\frac{1}{n}H(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1).$$

The right-hand side of the last display, which is a Cesàro mean, has the same limit as $H(X_n|X_{n-1}, \cdots, X_1)$, which is $H^\infty(X)$. Since the limit of the left-hand side is the average entropy per symbol, the result follows. □

Kolmogorov extension. If $(X_t)_{t\in\mathbb{N}}$ is a stationary process, then all finite-dimensional marginal distributions of this process are determined. By Kolmogorov extension theorem, we can extend the index of this process to the integer set $\mathbb{Z}$ and obtain a stationary process $(X_t)_{t\in\mathbb{Z}}$. We write for the past history

$$X_{\le 0} = (X_t)_{t\in-\mathbb{N}_0} = (X_0, X_{-1}, X_{-2}, \cdots).$$

Furthermore, we can define the conditional p.m.f. of X_1 given $X_{\le 0}$:

$$\begin{aligned} p(x_1|X_{\le 0}) = \mathbb{E}\left[\mathbb{1}_{\{X_1=x_1\}}|X_{\le 0}\right] &= \lim_{n\to\infty}\left[\mathbb{1}_{\{X_1=x_1\}}|X_0, X_{-1}, \cdots, X_{-n}\right] \\ &= \lim_{n\to\infty} p(x_1|X_0, X_{-1}, \cdots, X_{-n}). \end{aligned}$$

Here the convergence holds both in L^1 and almost surely, since the sequence we take limit of is a uniformly integrable martingale. Furthermore, by Lebesgue's dominated convergence theorem,

$$\mathbb{E}\left[-\log p(X_1|X_{\le 0})\right] = \lim_{n\to\infty} H^k(X) = H^\infty(X).$$

Ergodicity. Let $(\Omega, \mathscr{F}, P)$ be a measure space. A measurable mapping $T : (\Omega, \mathscr{F}) \to (\Omega, \mathscr{F})$ is said to be *ergodic*, if every set $A \in \mathscr{F}$ such that $TA = A$ *a.e.* satisfies $P(A) = 0$ or $P(A) = 1$. We let T play a role of time shift. The stochastic process $X = (X_t)_{t\in\mathbb{N}}$ is said to be an *ergodic* process, where $X_t(\omega) = X_0(T^t\omega)$ for all $t \in \mathbb{N}$ and $X_0 : \Omega \to \mathcal{X}$ is a random variable.

According to *Birkhoff's ergodic theorem*, the strong law of large numbers holds for a stationary ergodic process $X = (X_t)_{t\in\mathbb{N}}$:

$$\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k \to \mu = \mathbb{E}X_1, \quad a.s..$$

Lemma 1.24. *For the process $(X_t)_{t\in\mathbb{Z}}$, define the k-th order Markov approximation by*

$$p^k(X_{1:n}) = p(X_{1:k}) \prod_{j=k+1}^{n} p(X_j|X_{j-1}, \cdots, X_{j-k}).$$

If $(X_t)_{t\in\mathbb{Z}}$ is a stationary ergodic process,

$$\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} \to H^k(X) \ a.s., \quad \textit{and} \quad \frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\le 0})} \to H^\infty(X) \ a.s..$$

Proof. Since $(X_t)_{t\in\mathbb{Z}}$ is an ergodic process, so is the process $Y_t = f(X_{\le t})$, where f is any measurable function. Then both $\log p(X_n|X_{n-1}, \cdots, X_{n-k})$ and $\log p(X_n|X_{\le n-1})$ are stationary ergodic processes on $n \in \mathbb{N}$. By Birkhoff's ergodic theorem, we have

$$\begin{aligned} &\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} = \frac{1}{n}\log\frac{1}{p(X_{1:k})} + \frac{1}{n}\sum_{j=k+1}^{n}\log\frac{1}{p(X_j|X_{j-1}, \cdots, X_{j-k})} \to 0 + H^k(X), \ a.s., \\ &\frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\le 0})} = \frac{1}{n}\sum_{j=1}^{n}\log\frac{1}{p(X_j|X_{\le j-1})} \to H^\infty(X), \ a.s.. \end{aligned}$$

Then we complete the proof. □

Lemma 1.25 (Sandwich)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process. Then*

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \le 0 \ a.s., \quad \limsup_{n\to\infty}\frac{1}{n}\log\frac{p(X_{1:n})}{p(X_{1:n}|X_{\le 0})} \le 0 \ a.s..$$

Proof. Let A be the support set of $p(x_{1:n})$. Then

$$\mathbb{E}\left[\frac{p^k(X_{1:n})}{p(X_{1:n})}\right] = \sum_{x_{1:n}\in A}\frac{p^k(x_{1:n})}{p(x_{1:n})}p(x_{1:n}) = \sum_{x_{1:n}\in A} p^k(x_{1:n}) \le \sum_{x_{1:n}\in\mathcal{X}^n} p^k(x_{1:n}) = 1.$$

By Markov's inequality, we have

$$\mathbb{P}\left(\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq \frac{2\log n}{n}\right) = \mathbb{P}\left(\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq n^2\right) \leq \frac{1}{n^2}$$

By Borel-Cantelli Lemma, since $\sum_{n=1}^{\infty} n^{-2} < \infty$, the events

$$\left\{\frac{1}{n}\log\frac{p^k(X_{1:n})}{p(X_{1:n})} \geq \frac{2\log n}{n}, \quad n \in \mathbb{N}\right\}$$

happens finitely many times with probability 1, which proves the first result. On the other hand, let $B(X_{\leq 0})$ be the support set of $p(x_{1:n}|X_{\leq 0})$. Then

$$\mathbb{E}\left[\frac{p(X_{1:n})}{p(X_{1:n}|X_{\leq 0})}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{p(X_{1:n})}{p(X_{1:n}|X_{\leq 0})}\middle| X_{\leq 0}\right]\right] = \mathbb{E}\left[\sum_{x_{1:n}\in B(X_{\leq 0})} p(X_{1:n})\right] \leq 1.$$

The second result then follows from a similar procedure. □

Now we point out that, the Asymptotic Equilibrium property holds not only for i.i.d. sequences, but also for stationary ergodic processes.

Theorem 1.26 (Shannon-McMillan-Breiman)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process. Then*

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} = H^{\infty}(X).$$

Proof. By Lemmas 1.24 and 1.25, almost surely,

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \leq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p^k(X_{1:n})} = H^k(X),$$
$$\liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \geq \limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n}|X_{\leq 0})} = H^{\infty}(X).$$

Therefore, for all $k \in \mathbb{N}$, we have

$$H^{\infty}(X) \leq \liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \leq \limsup_{n\to\infty}\frac{1}{n}\log\frac{1}{p(X_{1:n})} \leq H^k(X).$$

Since X is stationary, $H^k(X) \searrow H^{\infty}(X)$ as $k \to \infty$. Hence $\frac{1}{n}\log\frac{1}{p(X_{1:n})} \overset{a.s.}{\to} H^{\infty}(X)$. □

Remark. An example for stationary ergodic process is the irreducible and aperiodic Markov chain.

2 Lossless Compression

In this section, we study the problem of lossless coding. To begin with, we have a source alphabet $\mathcal{X}$ and a D-ary alphabet $\{0, 1, \cdots, D-1\}$. Our key goal is to transform a string of $\mathcal{X}$ to a string of $\mathcal{D}$.

- A *source code* is a mapping $C : \mathcal{X} \to \mathcal{D}^*$, where $\mathcal{D}$ is a D-ary alphabet $\{0, 1, \cdots, D-1\}$, and

$$\mathcal{D}^* = \bigcup_{n=1}^{\infty} \mathcal{D}^n.$$

  The elements of $C(\mathcal{X})$ are called *codewords*. For every symbol $x \in \mathcal{X}$, we denote by $\ell(x)$ the length of the codeword $C(x)$ associated with x.
- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *nonsingular* if it is injective.
- The *extension* $C^* : \mathcal{X}^* \to \mathcal{D}^*$ of a source code C is the mapping from finite length strings of $\mathcal{X}$ to finite length strings of $\mathcal{D}$:

$$C^*(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n).$$

- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *uniquely decodable* if its extension C^* is injective.
- A source code $C : \mathcal{X} \to \mathcal{D}^*$ is said to be *instantaneous* (or *prefix-free*) if no codeword of C is prefixed by any other codeword.
- We have the inclusions: *nonsingular codes* $\supset$ *uniquely decodable codes* $\supset$ *instantaneous codes*.

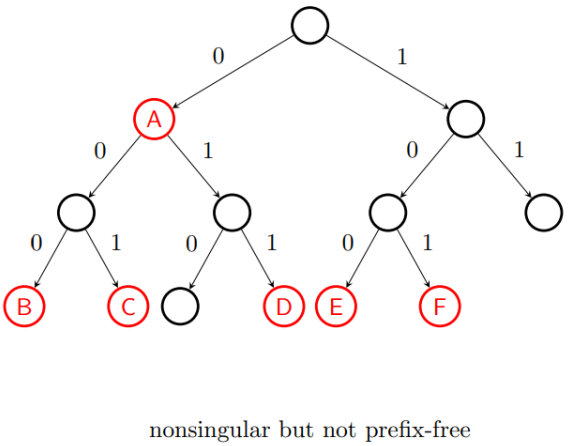
In general, some nice properties of a code are wanted:

- it is uniquely decodable;
- it is prefix free, so one can decode a string instantaneously while reading;
- it is efficient, i.e. given the distribution p of letters $\mathcal{X}$ in a string, we would like to minimize the average codeword length:

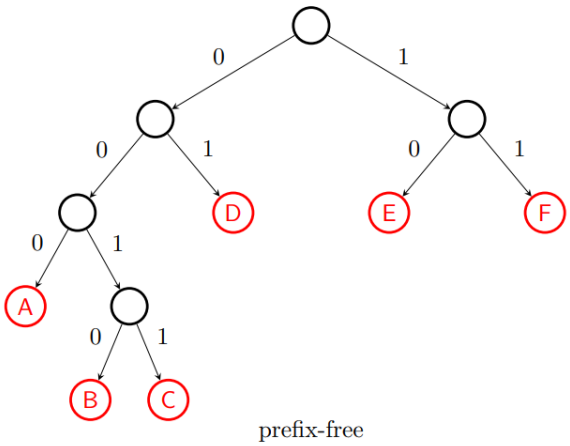
$$\mathbb{E}\left[\ell(X)\right] = \sum_{x \in \mathcal{X}} p(x)\ell(x).$$

2.1 Kraft-McMillan Inequality

Tree representation. A D-ary code $C : \mathcal{X} \to \mathcal{D}$ can be represented as a D-ary tree that consists of a root with branches, nodes and leaves. The root and every node has exactly D children, with each branch labeled by a letter in $\mathcal{D}$. Starting from the root, each vertex is uniquely associated with a string $d \in \mathcal{D}^*$, specified by the path from the root to itself. Some examples of binary trees are given below.



nonsingular but not prefix-free



prefix-free

We can determine whether a code is instantaneous right away by looking at its tree.

Proposition 2.1. *A code $C : \mathcal{X} \to \mathcal{D}^*$ is instantaneous if and only if all its codeword are leaves.*

Proof. If $C : \mathcal{X} \to \mathcal{D}^*$ is an instantaneous code, then each of its codeword has no descendant in the tree, which is a leaf; conversely, if each codeword of C is a leaf in the tree, it has no ancestor which is also a codeword, and C is instantaneous. □

Using the tree representation, we can show a property which characterizes the instantaneous codes.

Theorem 2.2 (Kraft's inequality)**.** *Let $\ell : \mathcal{X} \to \mathbb{N}$ be a length function. Then ℓ is the length function of an instantaneous code if and only if it satisfies Kraft's inequality:*

$$\sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1. \tag{2.1}$$

Proof. We first prove necessity. Let ℓ is the length function of an instantaneous code C, and let L be the depth of the tree. Then every codeword $C(x)$ at depth $\ell(x)$ prunes away $D^{L-\ell(x)}$ leaves from the complete tree of depth L. Since there are no more than D^L leaves in the complete tree, we have

$$\sum_{x \in \mathcal{X}} D^{L-\ell(x)} \leq D^L \quad \Rightarrow \quad \sum_{x \in \mathcal{X}} D^{-\ell(x)} \leq 1.$$

Now we prove the sufficiency. To this end, we prove the following argument: at every step $k \in \mathbb{N}$, after all codewords of length $\ell(x) < k$ have been assigned, there is enough room left at the depth k for the codewords of length $\ell(x) = k$. More explicitly, we want to show

$$D^k - \sum_{x \in \mathcal{X} : \ell(x) < k} D^{k-\ell(x)} \geq \left| C^{-1}(\mathcal{D}^k) \right|, \quad \forall 1 \leq k \leq L.$$

Note that

$$\left| C^{-1}(\mathcal{D}^k) \right| = \sum_{x \in \mathcal{X} : \ell(x) = k} D^{k-\ell(x)}.$$

Then our conclusion holds if

$$\sum_{x \in \mathcal{X} : \ell(x) \leq k} D^{-\ell(x)} \leq 1, \quad \forall k \in \mathbb{N}.$$

Clearly this is valid by Kraft's inequality (2.1). □

The Kraft's inequality is also a necessary condition for a code to be uniquely decodable.

Theorem 2.3 (McMillan)**.** *Every uniquely decodable code $C : \mathcal{X} \to \mathcal{D}^*$ satisfies Kraft's inequality (2.1).*

Proof. Let $C : \mathcal{X} \to \mathcal{D}^*$ be a uniquely decodable code, and let $L = \max_{x \in \mathcal{X}} \ell(x)$, where ℓ is the length function of C. Then for a source string $x_{1:n}$, the length of the extended codeword $C^*(x_{1:n})$ is given by

$$\ell^*(x_{1:n}) = \sum_{i=1}^{n} \ell(x_i) \leq nL.$$

Let N_k be the number of source strings of length n with $\ell^*(x_{1:n}) = k$. Since C is uniquely decodable, the source strings with codewords of length k are no more than D-ary strings of length k, i.e. $N_k \leq D^k$. Then

$$\sum_{x_{1:n} \in \mathcal{X}^n} D^{-\ell^*(x_{1:n})} = \sum_{k=1}^{nL} N_k D^{-k} \leq \sum_{k=1}^{nL} D^k D^{-k} \leq nL.$$

On the other hand,

$$\sum_{x_{1:n}\in\mathcal{X}^n} D^{-\ell^*(x_{1:n})} = \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_n\in\mathcal{X}} D^{-\ell(x_1)}D^{-\ell(x_2)}\cdots D^{-\ell(x_n)}$$
$$= \sum_{x_1\in\mathcal{X}} D^{-\ell(x_1)} \sum_{x_2\in\mathcal{X}} D^{-\ell(x_2)}\cdots\sum_{x_n\in\mathcal{X}} D^{-\ell(x_n)} = \left(\sum_{x\in\mathcal{X}} D^{-\ell(x)}\right)^n.$$

Therefore, we have

$$\sum_{x\in\mathcal{X}} D^{-\ell(x)} \le \inf_{n\in\mathbb{N}} \sqrt[n]{nL} = 1.$$

Then we complete the proof. □

Remark. To summarize, the Kraft's inequality (2.1) is a

- sufficient condition for the existence of an instantaneous code;
- necessary condition for a code to be uniquelt decodable.

2.2 Fundamental Limits of Compression

In this section, we study the limits of lossless compression. Given a source distribution p on $\mathcal{X}$, we want to minimize the average codeword length of our code. By Kraft-McMillan inequality, the search for optimal code can be expressed as the following optimization problem:

$$\min_{l:\mathcal{X}\to\mathbb{N}} \sum_{x\in\mathcal{X}} p(x)\ell(x) \quad \textit{subject to} \quad \sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1.$$

Following is a fundamental result of lossless compression.

Theorem 2.4. *For any source distribution $X \sim p$ on $\mathcal{X}$, the expected length $\mathbb{E}[\ell(X)]$ of an optimal uniquely decodable D-ary code satisfies*

$$\frac{H(X)}{\log D} \le \mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 1. \tag{2.2}$$

Proof. Upper bound. By Theorem 2.2, it suffices to construct a length function $\ell : \mathcal{X} \to \mathbb{N}$ that satisfies both the Kraft's inequality and the second (strict) inequality given in (2.2). Consider Shannon's length function:

$$\ell(x) = \left\lceil \log_D \frac{1}{p(x)} \right\rceil, \quad x \in \mathcal{X}, \tag{2.3}$$

Since

$$\sum_{x\in\mathcal{X}} D^{-\ell(x)} \le \sum_{x\in\mathcal{X}} D^{\log_D p(x)} = \sum_{x\in\mathcal{X}} p(x) = 1,$$

there exists an instantaneous code $C : \mathcal{X} \to \mathcal{D}^*$ whose length function is ℓ. On the other hand,

$$\mathbb{E}[\ell(X)] = \sum_{x\in\mathcal{X}} p(x)\ell(x) < \sum_{x\in\mathcal{X}} p(x)\left(\log_D \frac{1}{p(x)} + 1\right) = \frac{H(X)}{\log D} + 1.$$

Hence the upper bound holds.

LOWER BOUND. We consider the following relaxed optimization problem:

$$\min_{l:\mathcal{X}\to\mathbb{R}} \sum_{x\in\mathcal{X}} p(x)\ell(x) \quad \textit{subject to} \quad \sum_{x\in\mathcal{X}} D^{-\ell(x)} \le 1.$$

Note that the range of ℓ is $\mathbb{R}_+$. The Lagrange function is

$$L(l,\lambda) = \sum_{x\in\mathcal{X}} p(x)\ell(x) + \lambda\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1\right),$$

with KKT conditions

$$\begin{cases} \frac{\partial L}{\partial l(x)} = p(x) - \lambda D^{-l(x)} \log D = 0, \\ \lambda \ge 0,\ \sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1 \le 0, \\ \lambda\left(\sum_{x\in\mathcal{X}} D^{-\ell(x)} - 1\right) = 0. \end{cases}$$

The optimal solution is given by

$$\lambda = \frac{1}{\log D}, \quad l(x) = \log_D \frac{\lambda \log D}{p(x)} = \log_D \frac{1}{p(x)},\ x\in\mathcal{X},$$

and the optimal value is

$$\sum_{x\in\mathcal{X}} p(x)\ell(x) = \sum_{x\in\mathcal{X}} p(x)\log_D \frac{1}{p(x)} = \frac{H(X)}{\log D}. \tag{2.4}$$

Since our problem is relaxed, the primal problem (2.3) has optimal value no less than (2.4). Hence the lower bound holds for all uniquely decodable codes. □

Remark. In fact, we proved the existence of an *instantaneous* code with

$$\mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 1.$$

Coding over blocks. Using integer codeword lengths may lead to waste of memory. To overcome this effect, we consider coding over blocks of input symbols. If the input data $X_1, X_2, \cdots$ is an i.i.d. sequence of symbols, we partition it into blocks of size n and create a new source $\widetilde{X}_1, \widetilde{X}_2, \cdots$, where

$$\widetilde{X}_1 = (X_1, \cdots, X_n),\ \widetilde{X}_2 = (X_{n+1}, \cdots, X_{2n}),\ \cdots,\ \widetilde{X}_k = (X_{(k-1)n+1}, \cdots, X_{kn}),\ \cdots.$$

Consequently, every vector $\widetilde{X}_k$ can be viewed as a symbol from the alphabet $\widetilde{\mathcal{X}} = \mathcal{X}^n$, and we can find an optimal code $\widetilde{C} : \widetilde{X} \to \mathcal{D}$, whose length function ℓ satisfies

$$\frac{H(\widetilde{X})}{\log D} \le \mathbb{E}\left[\ell(\widetilde{X})\right] \le \frac{H(\widetilde{X})}{\log D} + 1.$$

Note that $H(\widetilde{X}) = nH(X)$, the average codeword length per symbol (in $\mathcal{X}$) satisfies

$$\frac{H(X)}{\log D} \le \frac{1}{n}\mathbb{E}\left[\ell(\widetilde{X})\right] < \frac{H(X)}{\log D} + \frac{1}{n}.$$

As the block size n increases, the integer effect becomes negligible. However, we also introduce delay in our system and increase the complexity of our code.

2.3 Shannon-Fano-Elias Coding

In this section, we introduce a specific coding approach that is near-optimal.

Midpoints of CDF. Without loss of generality, we assume that the source alphabet is $\mathcal{X} = \{1, 2, \cdots, m\}$, and $p(1) \geq p(2) \geq \cdots \geq p(m)$. The cumulative distribution function of p is

$$F(r) = \sum_{j=1}^{m} \mathbb{1}_{\{j \leq r\}} p(j), \quad r \in \mathbb{R}.$$

We define $\overline{F}(x)$ to be the midpoint of the interval $[F(x-1), F(x))$:

$$\overline{F}(x) = \sum_{j=1}^{x-1} p(j) + \frac{p(x)}{2}, \quad x = 1, \cdots, m.$$

Then $\overline{F}(x)$ is a real number in $(0, 1)$ that uniquely identifies $x \in \mathcal{X}$.

D-ary expansion and truncation. The D-ary expansion of a real number $\overline{F}(x) \in (0, 1)$ is given by

$$\overline{F}(x) = (0.z_1 z_2 \cdots)_D = \sum_{k=1}^{\infty} z_k D^{-k} = z_1 D^{-1} + z_2 D^{-2} + \cdots, \quad z_1, z_2, \cdots \in \{0, 1, \cdots, D-1\}.$$

Given a positive integer $\ell \in \mathbb{N}$, one have the ℓ-truncation of the D-ary expansion of $\overline{F}(x)$:

$$C(x) = (0.z_1 z_2 \cdots z_\ell)_D = \sum_{k=1}^{\ell} z_k D^{-k}$$

To ensure that the codeword of x is unique, we let $\overline{F}(x) - C(x) < \frac{p(x)}{2}$, so that

$$C(x-1) \leq \overline{F}(x-1) < F(x-1) < C(x).$$

To this end, we set

$$\ell = \left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1,$$

then

$$\overline{F}(x) - C(x) < D^{-\ell} \leq D^{-\log_D \frac{1}{p(x)} - 1} \leq \frac{p(x)}{D} \leq \frac{p(x)}{2}.$$

Construction of the Shannon-Fano-Elias code. For each $x \in \mathcal{X}$:

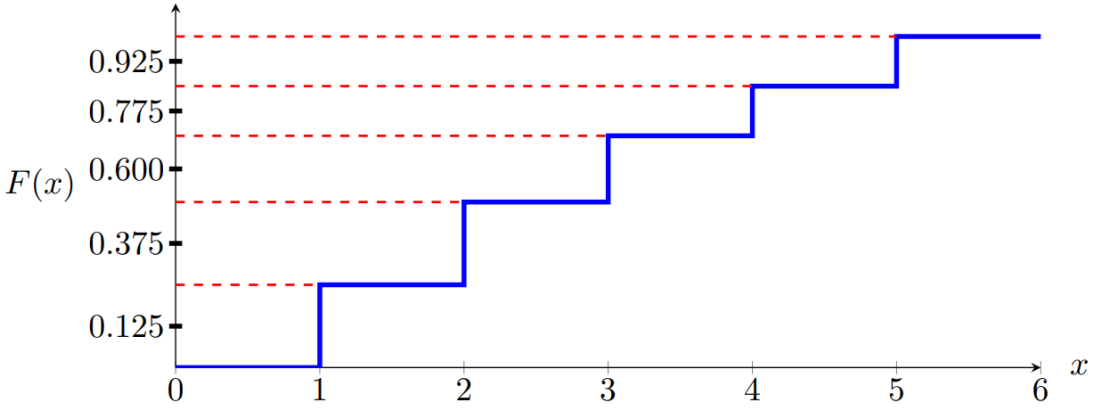
- Let z be the D-ary expansion of x;
- Choose the length of the codeword of x:

$$\ell(x) = \left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1;$$

- Choose the codeword of x to be the first most significant D-ary digits:

$$z = 0.\underbrace{z_1 z_2 \cdots z_{\ell(x)}}_{C(x)} z_{\ell(x)+1} \cdots.$$

An example of binary Shannon-Fano-Elias code. Here we let $\mathcal{X} = \{1, 2, 3, 4, 5\}$, and $D = 2$.



x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ in binary	$\ell(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$	codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.25	0.5	0.375	0.011	3	011
3	0.2	0.7	0.6	0.10011	4	1001
4	0.15	0.85	0.775	0.1100011	4	1100
5	0.15	1.0	0.925	0.1110110	4	1110

Shannon-Fano-Elias code is instantaneous. If the codeword $C(x) = (0.z_1 \cdots z_{\ell(x)})_D$ is a prefix of another codeword, this codeword lies in the half-open interval

$$\left[(0.z_1 \cdots z_{\ell(x)})_D, (0.z_1 \cdots z_{\ell(x)})_D + \frac{1}{D^{l(x)}} \right).$$

However, a contradiction rises because

$$C(x+1) - C(x) > F(x) - \overline{F}(x) = \frac{p(x)}{2} \geq D^{-l(x)}.$$

Average codeword length. The average codeword length of Shannon-Fano-Elias code is given by

$$\mathbb{E}[\ell(X)] = \sum_{x \in \mathcal{X}} p(x) \left(\left\lceil \log_D \frac{1}{p(x)} \right\rceil + 1 \right),$$

which satisfies

$$\frac{H(X)}{\log D} + 1 \leq \mathbb{E}\left[\ell(X)\right] < \frac{H(X)}{\log D} + 2.$$

It is revealed that the Shannon code is sub-optimal.

2.4 Shannon Code

Improvement of Shannon-Fano-Elias code: Shannon code. We consider

$$F(x) = \sum_{j=1}^{x-1} p(x), \quad \ell(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil.$$

We choose the codeword $c(x)$ to be the $\ell(x)$-truncation of the D-ary expansion of $F(x)$.

Shannon code is instantaneous. For every symbol i, the number $F_i = \sum_{k=0}^{i-1} p_i$ has the binary expansion

$$F_i = (0.z_1 z_2 \cdots)_2 = \sum_{k=1}^{\infty} z_k 2^{-k}, \quad z_1, z_2 \cdots \in \{0, 1\}.$$

The round off to ℓ_i is obtained by truncating the bits after ℓ_i:

$$c_i = (0.z_1 z_2 \cdots z_{\ell_i})_2 = \sum_{k=1}^{\ell_i} z_k 2^{-k}.$$

Fix $i \in \{1, \cdots, m-1\}$. Since $\ell_i = \left\lceil \log_2 \frac{1}{p_i} \right\rceil \geq \log_2 \frac{1}{p_i}$, we have

$$F_{i+1} - F_i = p_i = 2^{\log_2 p_i} \geq 2^{-\ell_i}.$$

For any $j > i$, we have $F_j - F_i \geq 2^{-\ell_i}$. If c_j is prefixed by c_i, then F_j and F_i share the first ℓ_i bits, which implies $F_j - F_i < 2^{-\ell_i}$, a contradiction! Hence the Shannon code is prefix-free.

Average length of Shannon code. The average length of this code $L = \sum_{i=1}^{m} p_i \ell_i$ satisfies

$$H(X) = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} \leq \sum_{i=1}^{m} p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil < \sum_{i=1}^{m} p_i \left(\log_2 \frac{1}{p_i} + 1 \right) = H(X) + 1.$$

Hence $H(X) \leq L < H(X) + 1$.

Example. We construct the Shannon code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$ for example. The code is shown below.

i	p_i	$F(x)$	$\ell_i = \left\lceil \log_2 \frac{1}{p_i} \right\rceil$	Codeword
1	0.5	$0 = (0.0)_2$	1	0
2	0.25	$0.5 = (0.1)_2$	2	10
3	0.125	$0.75 = (0.11)_2$	3	110
4	0.125	$0.875 = (0.1110)_2$	3	111

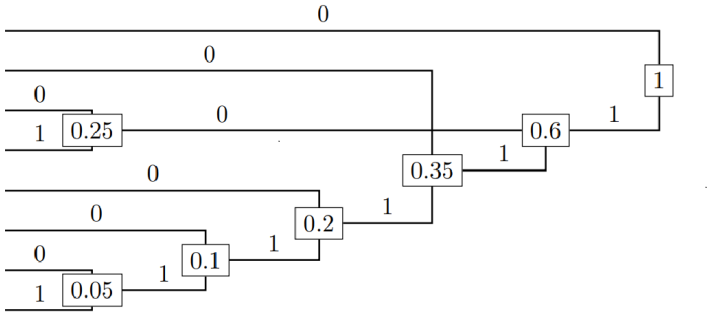
2.5 Huffman Coding

The search for binary optimal code was discovered by David Huffman (1952).

Construction of Huffman tree. The construction procedure is greedy.

- Take the two least probable symbols, which will be assigned the longest codewords having equal lengths and differing only at the last digit;
- Merge these two symbols into a new symbol with combined probability mass and repeat.

Codeword	x	$p(x)$
0	1	0.40
110	2	0.15
100	3	0.15
101	4	0.10
1110	5	0.10
11110	6	0.05
111110	7	0.04
111111	8	0.01



Optimality of Huffman code. Let $\mathcal{X} = \{1, 2, \cdots, m\}$. Without loss of generality, assume probabilities are in descending order $p(1) \geq p(2) \geq \cdots \geq p(m)$. We prove the optimality of Huffman code through three step.

Lemma 2.5. *In an optimal code, shorter codewords are assigned larger probabilities, i.e.* $p(i) > p(j)$ *implies* $\ell(i) \leq \ell(j)$.

Proof. Argue by contradiction. If there exists $i, j \in \mathcal{X}$ with $\ell(i) \leq \ell(j)$ and $p(i) > p(j)$, then we can exchange these codewords and reduce the expected length. Hence the code is not optimal. □

Lemma 2.6. *There exists an optimal code for which the codewords assigned to the smallest probabilities are siblings, i.e., they have the same length and differ only in the last symbol.*

Proof. Consider any optimal code. By Lemma 2.5, the codeword $C(m)$ has the longest length. Assume for the sake of contradiction, its sibling is not a codeword. Then the expected length can be decreased by moving $C(m)$ to its parent. Thus, the code is not optimal and a contradiction is reached.

Now, we know the sibling of $C(m)$ is a codeword. If it is $C(m-1)$, we are done. If it is some $C(i)$ for $i \neq m-1$ and the code is optimal, by Lemma 2.5, we have $p(i) = p(m-1)$. Therefore, $C(i)$ and $C(m-1)$ can be exchanged without changing expected length. □

Theorem 2.7 (Optimality of Huffman coding)**.** *Huffman's coding algorithm produces an optimal code tree.*

Proof. Let ℓ be the length function of the optimal code. By Lemmas 2.5 and 2.6, $C(m)$ and $C(m-1)$ are siblings and the longest codewords. Then we merge the two symbols and let $\widetilde{p}_1 \geq \cdots \geq \widetilde{p}_{m-1}$ denote the reordered probabilities after merging $p(m)$ and $p(m-1)$, and denote by $\widetilde{C}_1, \cdots, \widetilde{C}_{m-1}$ the corresponding codewords. The reduced length function $\widetilde{\ell}$ satisfies

$$\mathbb{E}\left[\ell(X)\right] = \mathbb{E}\left[\widetilde{\ell}(\widetilde{X})\right] + \mathbb{P}\left(\ell(X) \neq \widetilde{\ell}(\widetilde{X})\right) = \mathbb{E}\left[\widetilde{\ell}(\widetilde{X})\right] + p(m-1) + p(m).$$

Hence ℓ is the length function of an optimal code if and only if $\widetilde{\ell}$ is the length function of an optimal code for the reduced alphabet. The problem then is reduced to finding an optimal code tree for $\widetilde{p}_1 \geq \cdots \geq \widetilde{p}_{m-1}$. Repeat the merging procedure above for m times, and the result follows. □

2.6 Coding with Unknown Distributions

Given a distribution $X \sim p$, it is possible to construct a code that achieves the optimal expected length. However, we do not know what to do when the distribution p is unknown. In this section, we suppose that X is drawn from some distribution p_θ parameterized by an unknown parameter $\theta \in \Theta$.

Definition 2.8 (Redundancy)**.** The redundancy of coding a distribution p with respect to the optimal code for a distribution q, i.e. $\ell(x) = -\log q(x)$, is given by

$$R(p, q) = \sum_{x\in\mathcal{X}} p(x)\ell(x) - H(p) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = D(p||q).$$

Given a family of distributions $\{p_\theta\}_{\theta\in\Theta}$, the minimax redundancy is

$$R^* = \min_q \max_{\theta\in\Theta} R(p_\theta, q).$$

Remark. Intuitively, the distribution q leading to a code that minimizes the maximum redundancy is the distribution at the center of the "information ball" of radius R^*. Therefore, by constructing an optimal code based on q, we can reduce the redundancy in the worst case.

Lemma 2.9. *We impose a prior distribution π on Θ. Then*

$$\max_{\theta\in\Theta} R(p_\theta, q) = \max_\pi \sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q).$$

Proof. On the one hand,

$$\max_{\theta\in\Theta} R(p_\theta, q) = \max_{\theta_0\in\Theta} \sum_{\theta\in\Theta} \delta_{\theta_0}(\theta) R(p_\theta, q) \le \max_\pi \sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q).$$

On the other hand, if $\theta^* \in \Theta$ maximizes $R(p_\theta, q)$, one have

$$\sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q) \le \sum_{\theta\in\Theta} \pi(\theta) R(p_{\theta^*}, q) = R(p_{\theta^*}, q) = \max_{\theta\in\Theta} R(p_\theta, q), \quad \forall \pi \in \Delta(\Theta).$$

Then we complete the proof. □

We also introduce another technical theorem.

Theorem 2.10 (Minimax theorem)**.** *If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function that is convex in the first variable and concave in the second variable. If both $\mathcal{X}$ and $\mathcal{Y}$ are convex compact sets, then*

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x, y).$$

Remark. To develop the following theorem, we use the joint convexity of Kullback-Leibler divergence:

$$D((1-\lambda)p_0 + \lambda p_1||(1-\lambda)q_0 + \lambda q_1) \le (1-\lambda)D(p_0||q_0) + \lambda D(p_1||q_1).$$

Theorem 2.11. *The minimax redundancy is the maximum mutual information between θ and X:*

$$R^* = \max_\pi I(\theta; X),$$

where $\pi(\theta)$ is the prior distribution of the parameter θ, and $X|\theta \sim p_\theta(x)$.

Proof. Using Lemma 2.9 and Theorem 2.10, we reformulate the optimization problem:

$$R^* = \min_q \max_{\theta\in\Theta} R(p_\theta, q) = \min_q \max_\pi \sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q) = \max_\pi \min_q \sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q). \tag{2.5}$$

We write

$$q_\pi(x) = \sum_{\theta\in\Theta} \pi(\theta) p_\theta(x).$$

Then

$$\begin{aligned}
\sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q) &= \sum_{\theta\in\Theta} \pi(\theta) D(p_\theta||q) - D(q_\pi||q) + D(q_\pi||q) \\
&= \sum_{\theta\in\Theta}\sum_{x\in\mathcal{X}} \pi(\theta) p_\theta(x) \log \frac{p_\theta(x)}{q(x)} - \sum_{x\in\mathcal{X}}\sum_{\theta\in\Theta} \pi(\theta) p_\theta(x) \log \frac{q_\pi(x)}{q(x)} + D(q_\pi||q) \\
&= \sum_{\theta\in\Theta}\sum_{x\in\mathcal{X}} \pi(\theta) p_\theta(x) \log \frac{p_\theta(x)}{q_\pi(x)} + D(q_\pi||q)
\end{aligned}$$

Since the first term does not depends on q, the last display reaches its minimum if and only if $q = q_\pi$:

$$\begin{aligned}
\min_q \sum_{\theta\in\Theta} \pi(\theta) R(p_\theta, q) &= \sum_{\theta\in\Theta}\sum_{x\in\mathcal{X}} \pi(\theta) p_\theta(x) \log \frac{p_\theta(x)}{q_\pi(x)} \\
&= \sum_{\theta\in\Theta}\sum_{x\in\mathcal{X}} \pi(\theta) p_\theta(x) \log \frac{\pi(\theta) p_\theta(x)}{\pi(\theta) q_\pi(x)} = I(\theta; X),
\end{aligned}$$

where $\pi(\theta)p_\theta(x)$ is the joint distribution of θ and X, and $q_\pi(x)$ is the marginal distribution of X. Plugging in this expression to (2.5) completes the proof. □

3 Channel Coding

Motivation. In a communication situation, we often have two primary goals:

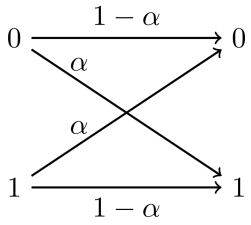
- *Reliability.* The received message should be equal to the transmitted message in most cases. In other words, we wish to reduce the error probability:

$$P_e = \mathbb{P}\left(\textit{received message} \neq \textit{transmitted message}\right).$$

- *Efficiency.* The message should be transmitted as quickly as possible. In other words, we wish to send as much information as possible in a unit time:

$$R = \textit{average number of information bits transmitted per unit time.}$$

However, these two goals often conflict with each other. We use the Binary Symmetric Channel (BSC) to interpret this. Suppose that we want to send a bit $W \in \{0,1\}$. A binary symmetric channel has a binary input $X \in \{0,1\}$ and a binary output $Y \in \{0,1\}$. While sending a bit, it flips the bit with probability α:



To reduce the error probability, we use the channel multiple times. Assume that each use of the channel consumes a unit time, and the channel is memoryless, i.e., given the input, the outputs of the channel are conditionally independent. We encode the bit using a repetition code:

$$W = 0 \quad \Rightarrow \quad X_{1:n} = \underbrace{00\cdots0}_{n}, \qquad W = 1 \quad \Rightarrow \quad X_{1:n} = \underbrace{11\cdots1}_{n}.$$

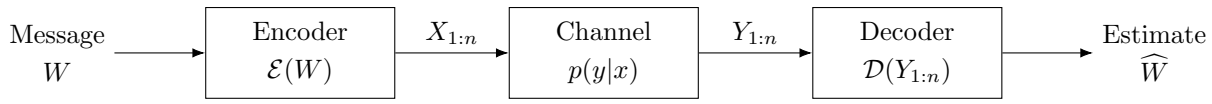
Given the output $Y_{1:n}$, we decode the bit using the maximum likelihood rule:

$$\widehat{W} = \begin{cases} 0, & \textit{if there are more 0's observed in } Y_{1:n} \textit{ than 1's}, \\ 1, & \textit{otherwise}. \end{cases}$$

As the uses n of channel increases, the error probability decreases, but the bit the channel transmitted every unit time $R = 1/n$ also decreases. Hence a tradeoff between reliability and efficiency is required.

3.1 Set-up of Channel Encoding

In this section, we study the problem of channel coding. Consider the communication over a random channel:



- The message $W \in \{1, \cdots, M\}$ is one of the possible M numbers that we want to send. We always assume W to be uniformly distributed over all possibilities.
- An (M, n)-coding scheme is an encoder $\mathcal{E} : \{1, \cdots, M\} \to \mathcal{X}^n$ that maps the message M to an n-length string of channel inputs X^n;

- The channel specifies the probabilistic transformation from inputs to outputs:

$$p(y_{1:n}|x_{1:n}) = \mathbb{P}\left(Y_1 = y_1, \cdots, Y_n = y_n | X_1 = x_1, \cdots, X_n = x_n\right).$$

  We are particularly interested in the *discrete memoryless channel (DMC)*, which is specified by

  (i) an input alphabet $\mathcal{X}$,

  (ii) an output alphabet $\mathcal{Y}$, and

  (iii) a conditional probability distribution $p_{Y|X}(y|x)$ such that the outputs between channel uses are conditionally independent given the inputs:

$$p(y_{1:n}|x_{1:n}) = p_{Y|X}(y_1|x_1) \cdots p_{Y|X}(y_n|x_n).$$

- A decoder $\mathcal{D} : \mathcal{Y}^n \to \{1, \cdots, M\}$ maps an n-length string of channel outputs $Y_{1:n}$ to an estimate $\widehat{W}$ of the transmitted message.

Now recall our two primary goals in communication:

- *Reliability.* Assuming that the message W is uniformly distributed over all possibilities, the *conditional error probability* and the *average error probability* are

$$P_e^{(n)}(w) = \mathbb{P}(\widehat{W} \neq w | W = w), \qquad P_e^{(n)} = \mathbb{P}(\widehat{W} \neq W) = \frac{1}{M} \sum_{w=1}^{M} P_e^{(n)}(W).$$

  The maximum error probability is

$$P_{\text{e,max}}^{(n)}(w) = \max_{w \in \{1, \cdots, M\}} P_e^{(n)}(w) = \max_{w \in \{1, \cdots, M\}} \mathbb{P}(\widehat{W} \neq w | W = w).$$

- *Efficiency.* The *rate* R of an (M, n) encoding scheme is

$$R = \frac{\log_2 M}{n} \quad \textit{bits/transmission.}$$

  Alternatively, the number of messages for a given rate R and block-length n is given by $M = 2^{nR}$. To specify a rate R code, we write $(2^{nR}, n)$ instead of (M, n). Particularly, are interested in the case that the error probability becomes negligible as the coding length n goes infinity.

Definition 3.1 (Operational Capacity)**.** A rate R is achievable for given discrete memoryless channel $p(y|x)$, if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ coding schemes such that maximum error probability

$$\lim_{n \to \infty} P_{\text{e,max}}^{(n)} = 0.$$

The operational capacity C_{op} is the supremum over all achievable rates:

$$C_{\text{op}} = \sup \left\{ R : R \textit{ is achievable} \right\}.$$

Definition 3.2 (Information Capacity)**.** The information capacity of a discrete memoryless channel is

$$C = \sup_{p_X} I(X;Y) = \sup_{p_X} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_X(x) p_{Y|X}(y|x) \log_2 \frac{p_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} p_{Y|X}(y|x') p_X(x')}$$

Remark. Since the map $p_X, p_{Y|X} \mapsto I(X,Y)$ is concave about p_X, we can always find a maximizer p_X^* that reaches the supremum: $C = \max_{p_X} I(X;Y)$.

3.2 Shannon's Channel Coding Theorem: Achievability

In the next two sections, we will establish Shannon's channel coding theorem.

Theorem 3.3 (Shannon's channel coding theorem)**.** *The operational capacity of a discrete memoryless channel is equal to the information capacity:*

$$C_{\text{op}} = \sup_{p_X} I(X;Y).$$

Remark. In fact, the channel coding theorem consists of two statements:

- *Achievability.* Every rate $R < C$ is achievable, i.e. there exists a sequence of $(2^{nR}, n)$ coding schemes such that the maximum error probability $P_{\text{e,max}}^{(n)} \to 0$ as $n \to \infty$:

$$R < C \quad \Rightarrow \quad R \text{ is achievable.}$$

- *Converse.* Any sequence of $(2^{nR}, n)$ coding schemes with the maximum error probability $P_{\text{e,max}}^{(n)} \to 0$ as $n \to \infty$ must satisfy $R \leq C$.

$$R \text{ is achievable} \quad \Rightarrow \quad R \leq C.$$

In this section, we are going to establish the achievability part of channel encoding theorem.

Construction of encoder $\mathcal{E}$. A $(2^{nR}, n)$ encoder $\mathcal{E}$ can be represented by a codebook:

$$\mathcal{E} = \begin{pmatrix} x_{1:n}(1) \\ x_{1:n}(2) \\ \vdots \\ x_{1:n}(2^{nR}) \end{pmatrix} = \begin{pmatrix} x_1(1) & x_1(2) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{pmatrix} \in \mathcal{X}^{2^{nR} \times n}. \tag{3.1}$$

To transmit a message w, the encoder assigns

$$\mathcal{E}(w) = x_{1:n}(w), \quad w \in \left\{1, 2, \cdots, 2^{nR}\right\}.$$

We consider the construction of random encoder. To proceed, we first choose a input distribution p_X. We let each entry in the codebook $\mathcal{E}$ to be drawn from i.i.d. p_X. The probability of generating any particular random codebook (3.1) is then given by

$$p(\mathcal{E}) = \prod_{w}^{2^{nR}} \prod_{i=1}^{n} p_X(x_n(w)).$$

With the codebook $\mathcal{E}$ specified, the conditional distribution of input string $X_{1:n}$ is the

$$p_{X_{1:n}|\mathcal{E}}(x_{1:n}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \mathbb{1}_{\{x_{1:n} = \mathcal{E}(w)\}}, \quad x_{1:n} \in \mathcal{X}^n,$$

and

$$p_{Y_{1:n}|\mathcal{E}}(y_{1:n}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} p_{Y_{1:n}|X_{1:n}}(y_{1:n}|\mathcal{E}(w)), \quad y_{1:n} \in \mathcal{Y}^n.$$

To find the unconditional distribution, note that each row of the codebook has the same distribution:

$$\begin{aligned}
p_{X_{1:n}}(x_{1:n}) &= \prod_{i=1}^{n} p_X(x_i); \\
p_{Y_{1:n}}(y_{1:n}) &= \sum_{x_{1:n}\in\mathcal{X}^n} p_{X_{1:n}}(x_{1:n}) p_{Y_{1:n}|X_{1:n}}(y_{1:n}|x_{1:n}) \\
&= \sum_{x_1\in\mathcal{X}} \sum_{x_2\in\mathcal{X}} \cdots \sum_{x_n\in\mathcal{X}} \prod_{i=1}^{n} p_X(x_i) p_{Y|X}(y_i|x_i) \\
&= \prod_{i=1}^{n} \underbrace{\left(\sum_{x_i\in\mathcal{X}} p_X(x_i) p_{Y|X}(y_i|x_i) \right)}_{p_Y(y_i)} = \prod_{i=1}^{n} p_Y(y_i).
\end{aligned}$$

Since the channel is memoryless, the *information density* of $(X_{1:n}, Y_{1:n})$ can be factorized:

$$\begin{aligned}
i(x_{1:n}; y_{1:n}) &= \log_2 \frac{p_{X_{1:n},Y_{1:n}}(x_{1:n}, y_{1:n})}{p_{X_{1:n}}(x_{1:n}) p_{Y_{1:n}}(y_{1:n})} = \log_2 \frac{p_{Y_{1:n}|X_{1:n}}(y_{1:n}|x_{1:n})}{p_{Y_{1:n}}(y_{1:n})} \\
&= \sum_{k=1}^{n} \log_2 \frac{p_{Y|X}(y_k|x_k)}{p_Y(y_k)} = \sum_{k=1}^{n} i(x_k; y_k).
\end{aligned}$$

These distributions arise from the randomness in both the codebook and the message.

Construction of decoder $\mathcal{D}$. To finish the construction of a coding scheme, we need to find an optimal decoder. To minimize the probability of error, we use a *maximum a posteriori* (MAP) decoder:

$$\begin{aligned}
\mathcal{D}^*(y_{1:n}) &= \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} p_{W|Y_{1:n}}(w|y_{1:n}) \\
&= \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} p_W(w) p_{Y_{1:n}|W}(y_{1:n}|w).
\end{aligned}$$

Since the message W is uniform, the MAP decoder is equivalent to the maximum likelihood decoder:

$$\mathcal{D}^*(y_{1:n}) = \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} p_{Y_{1:n}|W}(y_{1:n}|w).$$

Using the information density, we have

$$\begin{aligned}
\mathcal{D}^*(y_{1:n}) &= \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} p_{Y_{1:n}|X_{1:n}}(y_{1:n}|x_{1:n}(w)) \\
&= \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} \frac{p_{Y_{1:n}|X_{1:n}}(y_{1:n}|x_{1:n}(w))}{p_{Y_{1:n}}(y_{1:n})} \\
&= \operatorname*{argmax}_{w\in\{1,\cdots,2^{nR}\}} i(x_{1:n}(w); y_{1:n}).
\end{aligned}$$

To simplify the analysis, we study a sub-optimal thresholding decoder: For a given threshold T_n, we define the decoding rule as follows:

$$\mathcal{D}(y_{1:n}) = \begin{cases} \widehat{w}, & \text{if } i(x_{1:n}(\widehat{w}); y_{1:n}) > T_n \text{ and } i(x_{1:n}(w); y_{1:n}) \le T_n \text{ for all } w \neq \widehat{w}, \\ 0, & \text{otherwise.} \end{cases}$$

Decoding error is uniform. We now analyze the decoding error of our coding scheme. By uniformity of our construction of codebook and the message W,

$$\begin{aligned}
\mathbb{P}(\widehat{W} \neq W) &= \sum_{\mathcal{E}} p(\mathcal{E})\, \mathbb{P}(\widehat{W} \neq W|\mathcal{E}) \\
&= \sum_{\mathcal{E}} p(\mathcal{E}) \sum_{w=1}^{2^{nR}} \frac{1}{2^{nR}} \mathbb{P}(\widehat{W} \neq W|\mathcal{E}, W = w) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{E}} p(\mathcal{E}) \mathbb{P}(\widehat{W} \neq W|\mathcal{E}, W = w) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{E}} p(\mathcal{E}) \mathbb{P}(\widehat{W} \neq W|\mathcal{E}, W = 1) \\
&= \sum_{\mathcal{E}} p(\mathcal{E}) \mathbb{P}(\widehat{W} \neq W|\mathcal{E}, W = 1) \\
&= \mathbb{P}(\widehat{W} \neq W|W = 1)
\end{aligned}$$

Therefore, it suffices to control the decoding error conditioned on the event $W = 1$.

Proof of Theorem 3.3 (Achievability). Define events A and B as follows:

$$A_n = \{i(X_{1:n}(1); Y_{1:n}) > T_n\}, \quad B_n = \bigcap_{w=2}^{2^{nR}} \{i(X_{1:n}(w); Y_{1:n}) \leq T_n\}.$$

Consider the following bound:

$$P(\widehat{W} \neq W|W = 1) = P(A_n^c \cup B_n^c) \leq \mathbb{P}(A_n^c) + \mathbb{P}(B_n^c).$$

Analysis of $\mathbb{P}(A_n^c)$. By construction, the input $X_{1:n}(1)$ and output $Y_{1:n}$ satisfies

$$(X_k(1), Y_k) \overset{\text{i.i.d.}}{\sim} p_X p_{Y|X}.$$

Meanwhile,

$$\mathbb{E}\left[i(X_k(1), Y_k)\right] = \mathbb{E}\left[\log_2 \frac{p_{Y|X}(Y_k|X_k(1))}{p_Y(Y_k)}\right] = I(X;Y), \quad \textit{where} \;\; (X,Y) \sim p_X p_{Y|X}.$$

By strong law of large numbers,

$$\frac{i(X_{1:n}(1); Y_{1:n})}{n} = \frac{1}{n} \sum_{k=1}^{n} i(X_k(1), Y_k) \overset{a.s.}{\to} I(X;Y) \;\; \textit{as} \;\; n \to \infty.$$

Fix any $\epsilon > 0$, and set $T_n = n(I(X;Y) - \epsilon)$. Hence

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}(A_n^c) &= \limsup_{n\to\infty} \mathbb{P}\left(\frac{i(X_{1:n}(1); Y_{1:n})}{n} \leq I(X;Y) - \epsilon\right) \\
&\leq \mathbb{P}\left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \left\{\frac{i(X_{1:n}(1); Y_{1:n})}{n} \leq I(X;Y) - \epsilon\right\}\right) \\
&= \mathbb{P}\left(\limsup_{n\to\infty} \frac{i(X_{1:n}(1); Y_{1:n})}{n} \leq I(X;Y) - \epsilon\right) = 0.
\end{aligned}$$

Analysis of $\mathbb{P}(B_n^c)$. By construction, for all $w \neq 1$, $X_{1:n}(w)$ is independent of $X_{1:n}(1)$. Since the output $Y_{1:n}$ is generated from $X_{1:n}(1)$ and $p_{Y|X}$, it is independent of $X_{1:n}(w)$:

$$(X_k(w), Y_k) \overset{\text{i.i.d.}}{\sim} p_X p_Y.$$

Using the Chernoff bound, we have

$$\begin{aligned}
\mathbb{P}\left(i(X_{1:n}(w), Y_{1:n}) > T_n\right) &\leq 2^{-T_n} \mathbb{E}\left[2^{i(X_{1:n}(w);Y_{1:n})}\right] \\
&= 2^{-T_n} \mathbb{E}\left[\frac{p_{X_{1:n},Y_{1:n}}(X_{1:n}(w), Y_{1:n})}{p_{X_{1:n}}(X_{1:n}(w)) p_{Y_{1:n}}(Y_{1:n})}\right] \\
&= 2^{-T_n} \sum_{x_{1:n} \in \mathcal{X}^n} \sum_{y_{1:n} \in \mathcal{Y}^n} p_{X_{1:n}}(x_{1:n}) p_{Y_{1:n}}(y_{1:n}) \frac{p_{X_{1:n},Y_{1:n}}(x_{1:n}, y_{1:n})}{p_{X_{1:n}}(x_{1:n}) p_{Y_{1:n}}(y_{1:n})} \\
&= 2^{-T_n}.
\end{aligned}$$

We then employ a union bound:

$$\begin{aligned}
\mathbb{P}(B_n^c) &= \mathbb{P}\left(\bigcup_{w=2}^{2^{nR}} \{i(X_{1:n}(w), Y_{1:n}) > T_n\}\right) \\
&\leq \sum_{w=2}^{2^{nR}} \mathbb{P}\left(i(X_{1:n}(w), Y_{1:n}) > T_n\right) \\
&\leq 2^{nR - T_n} \\
&= 2^{n(R - I(X;Y) + \epsilon)}.
\end{aligned}$$

Choice of ϵ and p_X. Since $R < C = \sup_{p_X} I(X;Y)$, we choose $\epsilon = \frac{1}{3}(C - R)$, and choose p_X such that

$$\begin{aligned}
I(X;Y) &\geq R + 2\epsilon \\
&= C - \frac{1}{3}(C - R).
\end{aligned}$$

Then we have

$$\begin{aligned}
\lim_{n \to \infty} \mathbb{P}(\widehat{W} \neq W | W = 1) &\leq \lim_{n \to \infty} \mathbb{P}(A_n^c) + \lim_{n \to \infty} \mathbb{P}(B_n^c) \\
&\leq \lim_{n \to \infty} 2^{-n\epsilon} = 0.
\end{aligned}$$

Based on our previous discussion, the result follows.

Strengthening the proof. Yet we have not find a deterministic codebook with small error of probability. To finish the proof, we will strengthen this conclusion by getting rid of the average over codebooks. Note that the average probability of error over codebooks is small:

$$\mathbb{P}(\widehat{W} \neq W) = \sum_{\mathcal{E}} \mathbb{P}(\widehat{W} \neq W \,|\, \mathcal{E}) \mathbb{P}(\mathcal{E}) < \epsilon,$$

where $\epsilon > 0$ is an arbitrarily fixed quantity. Hence, over the set of possible codebooks, there exists at least one codebook $\mathcal{E}^*$ with a small probability of error:

$$\mathbb{P}(\widehat{W} \neq W \,|\, \mathcal{E} = \mathcal{E}^*) \leq \epsilon.$$

At this point, it is still possible that the codebook $\mathcal{E}^*$ contains some codewords with bad conditional error probabilities. Define

$$\lambda(w) = \mathbb{P}(\widehat{W} \neq W \,|\, \mathcal{E} = \mathcal{E}^*, W = w).$$

Since W is uniformly distributed over $\{1, 2, \cdots, 2^{nR}\}$, the number of "bad" codewords satisfies

$$\sum_{w=1}^{2^{nR}} \mathbb{1}_{\{\lambda(w) \geq 2\epsilon\}} \leq \sum_{w=1}^{nR} \frac{\lambda(w)}{2\epsilon} = \frac{1}{2\epsilon} 2^{nR} \mathbb{P}(\widehat{W} \neq W | \mathcal{E} = \mathcal{E}^*) \leq 2^{nR\left(1-\frac{1}{n}\right)}.$$

Therefore, if we expunge the worst half of the codewords, the maximum conditional error of the remaining codewords is $P_{e,\max}^{(n)} \leq 2\epsilon$, and the rate of the new codebook is $R - \frac{1}{n}$. Since this difference goes to zero as $n \to \infty$, we can conclude that $P_{e,\max}^{(n)}$ converges to 0 as $n \to \infty$. □

Remark. Although the theorem shows that there exist good codes with arbitrarily small error probability for long block lengths, it does not provide an approach to construct the optimal codebooks. Without some structure in the code, the simple decoding scheme of table lookup requires an exponentially large table.

3.3 Shannon's Channel Coding Theorem: Weak Converse

In this section, we prove the converse part of Shannon's channel coding theorem.

Lemma 3.4. *Let $C = \sup_{p_X}(X;Y)$ be the information capacity of a discrete memoryless channel $p_{Y|X}$. For any input distribution $p_{X_{1:n}}(x_{1:n})$, it holds*

$$I(X_{1:n}; Y_{1:n}) \leq nC.$$

Proof. We decompose the mutual information $I(X_{1:n}; Y_{1:n})$ by chain rule:

$$\begin{aligned}
I(X_{1:n}; Y_{1:n}) &= H(Y_{1:n}) - H(Y_{1:n}|X_1, \cdots, X_n) \\
&= \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \cdots, Y_1) - \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \cdots, Y_1, X_1, \cdots, X_n) \\
&= \sum_{i=1}^{n} H(Y_i|Y_{i-1}, \cdots, Y_1) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i; Y_i) \leq nC.
\end{aligned}$$

Hence we conclude the proof. □

Proof of Theorem 3.3 (Converse). By Fano's inequality [Theorem 1.15],

$$P_e^{(n)} = \mathbb{P}(\widehat{W} \neq W) \geq \frac{H(W|\widehat{W}) - 1}{\log_2 2^{nR}} = \frac{H(W|\widehat{W}) - 1}{nR}.$$

Since W is uniform over all possibilities,

$$\begin{aligned}
nR = H(W) &= H(W|\widehat{W}) + I(W; \widehat{W}) = nRP_e^{(n)} + 1 + I(W; \widehat{W}) \\
&\leq nRP_e^{(n)} + 1 + I(X_{1:n}; Y_{1:n}) && \text{(By data processing inequality)} \\
&\leq nRP_e^{(n)} + 1 + nC.
\end{aligned}$$

Therefore, we have

$$P_e^{(n)} \geq \frac{n(R-C)-1}{nR} \geq 1 - \frac{C}{R}, \quad \forall n \in \mathbb{N}.$$

If $R > C$, the error probability $P_e^{(n)}$ does not converge to 0, and R is not achievable. □

Further discussion about random coding: Privacy. We will provide more analysis about the privacy of this random coding scheme. Suppose that an eavesdropper observes the channel output $Y_{1:n}$ but does not know the codebook $\mathcal{E}$. We are worried that the eavesdropper might figure out the codebook.

Since the codebook $\mathcal{E}$ is randomly chosen, the difficulty of recovering the codebook $\mathcal{E}$ from the outputs $Y_{1:n}$ depends on the their mutual information. We will prove the following bound:

$$I(\mathcal{E}; Y_{1:n}) \leq n(C-R) + H_b(P_e^{(n)}) + P_e^{(n)} nR.$$

Using the chain rule, we have the decomposition

$$I(\mathcal{E}; Y_{1:n}) = I(Y_{1:n}; \mathcal{E}, W) - I(Y_{1:n}; W \,|\, \mathcal{E}).$$

- We first bound $I(Y_{1:n}; \mathcal{E}, W)$. Since $X_{1:n}$ is a function of W and $\mathcal{E}$, and $Y_{1:n}$ is conditionally independent of $\mathcal{E}, W$ given $X_{1:n}$,

  $$I(Y_{1:n}; \mathcal{E}, W) = I(Y^n; \mathcal{E}, W, X_{1:n}) = I(Y_{1:n}; X_{1:n}) \leq nC.$$

  where the last inequality follows from Lemma 3.4.
- Now we bound $I(Y_{1:n}; W \,|\, \mathcal{E})$. Since the message W and the codebook $\mathcal{E}$ are independent, we have

  $$I(W; Y_{1:n} \,|\, \mathcal{E}) = I(W; Y_{1:n}, \mathcal{E}) - I(W; \mathcal{E}) = I(W; Y_{1:n}, \mathcal{E}).$$

  Since W is conditionally independent of $\widehat{W}$ given Y^n and $\mathcal{E}$, we have

  $$\begin{aligned} I(W; Y_{1:n} \,|\, \mathcal{E}) = I(W; Y_{1:n}, \mathcal{E}) &\geq I(W; \widehat{W}, \mathcal{E}) && \text{(data processing inequality)} \\ &= H(W) - H(W|\widehat{W}, \mathcal{E}) && \text{(chain rule)} \\ &\geq H(W) - H(W|\widehat{W}). && \text{(Conditioning does not increase entropy)} \end{aligned}$$

  By Fano's inequality,

  $$H(W \,|\, \widehat{W}) \leq H_b(P_e^{(n)}) + P_e^{(n)} \log |\mathcal{W}| \leq nRP_e^{(n)} + H_b(P_e^{(n)}).$$

  Note that $W \sim \text{Unif}(1, 2, \cdots, 2^{nR})$, we have

  $$I(W; Y_{1:n} \,|\, \mathcal{E}) \geq H(W) - H(W|\widehat{W}) = (1 - P_e^{(n)})nR - H_b(P_e^{(n)}).$$

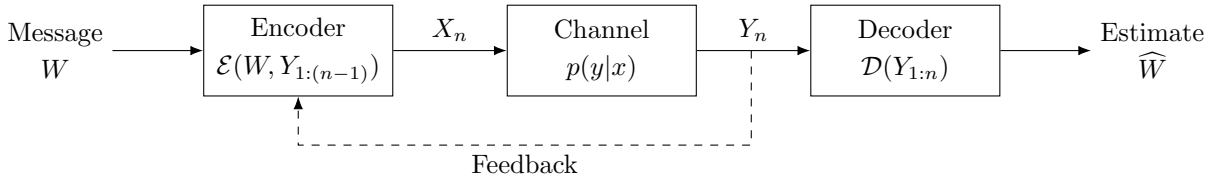
According to the two bounds, we have

$$I(\mathcal{E}; Y_{1:n}) = I(Y_{1:n}; \mathcal{E}, W) - I(Y_{1:n}; W \,|\, \mathcal{E}) \leq nC - (1 - P_e^{(n)})nR + H_b(P_e^{(n)}).$$

This proves the result. As long as the error probability $P_e^{(n)}$ is sufficiently small, increasing the rate R leads to better privacy. An interpretation is that a coding scheme with higher rate R produces less redundancy while transmitting a message. In this case, there is less information about the codebook $\mathcal{E}$ in the output $Y_{1:n}$.

3.4 Feedback Capacity

We turn to another setting of channel coding, where we allow our encoder to use previous outputs. That is, at the n-th step, our encoder assigns a channel input X_n according to not only the message W to be transmitted, but also the previous outputs $Y_{1:(n-1)}$. This setting is called the channel coding with *feedback*.



Theorem 3.5. *Feedback cannot increase capacity. For a discrete memoryless channel, the capacity with feedback, C_{FB}, is the same as the capacity without feedback:*

$$C_{\mathrm{FB}} = C.$$

Proof. Like the proof of the weak converse, since W is uniform over all possibilities,

$$\begin{aligned} nR = H(W) &= H(W|\widehat{W}) + I(W;\widehat{W}) \\ &= nRP_e^{(n)} + 1 + I(W;\widehat{W}) && \text{(By Fano's inequality)} \\ &\le nRP_e^{(n)} + 1 + I(W;Y_{1:n}). && \text{(By data processing inequality)} \end{aligned}$$

Then it remains to bound the mutual information $I(W;Y_{1:n})$. Since X_i is a function of W and $(Y_{i-1},\cdots,Y_i)$, and Y_i is conditionally independent of W and $(Y_{i-1},\cdots,Y_i)$ given X_i, we have

$$H(Y_i|Y_{i-1},\cdots,Y_1,W) = H(Y_i|Y_{i-1},\cdots,Y_1,W,X_i) = H(Y_i|X_i)$$

Then

$$\begin{aligned} I(W;Y_{1:n}) &= H(Y_{1:n}) - H(Y_{1:n}|W) \\ &= \sum_{i=1}^n H(Y_i|Y_{i-1},\cdots,Y_1) - \sum_{i=1}^n H(Y_i|Y_{i-1},\cdots,Y_1,W) \\ &= \sum_{i=1}^n H(Y_i|Y_{i-1},\cdots,Y_1) - \sum_{i=1}^n H(Y_i|X_i) \\ &\le \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) \\ &= \sum_{i=1}^n I(X_i;Y_i) \le nC. \end{aligned}$$

Therefore,

$$P_e^{(n)} \ge \frac{n(R-C)-1}{nR} \ge 1 - \frac{C}{R}, \quad \forall n \in \mathbb{N}.$$

If $R > C$, the error probability $P_e^{(n)}$ does not converge to 0, and R is not achievable. Hence $R \le C$. □

Remark. This surprising fact stems from the memorylessness of the channel. Of course, feedback can help simplify our encoding and decoding schemes in terms of complexity.

3.5 Hamming Code

Motivation. The object of coding is to introduce *redundancy* so that even if some of the information is lost or corrupted, it is still possible to recover the message at the receiver.

A simplest coding scheme is to repeat information. For example, consider sending a bit $W \in \{0,1\}$ with a binary symmetric channel. One repeat the bit over n channel uses, i.e. send $\underbrace{11\cdots1}_{n}$ for 1 and $\underbrace{00\cdots0}_{n}$ for 0. This code can correct up to $\frac{n-1}{2}$ flips, and the error probability converges to 0 as $n \to \infty$. However, the rate $R = 1/n$ of this code also goes to 0, which is not very useful.

Parity check code. Instead of simply repeating the bits, we can introduce each extra bit to check whether there is an error in some subset of the information bits. This is called an *error-detecting code.*

A *single parity check* code is a $(2^{n-1}, n)$ coding scheme for a binary symmetric channel which sends $n-1$ information bits, and the n-th bit encodes the *parity* of the entire block, i.e. whether the number of 1's in the information bits is even or odd. Then if there is an odd number of errors during transmission, the receiver will notice that the parity has changed and detect the error. This code does not detect an even number of errors and does not give any information about how to correct the errors that occur.

Hamming Code. To illustrate the idea of Hamming codes, we begin with an $m \times (2^m - 1)$ binary matrix formed by arranging the $2^m - 1$ nonzero binary column vectors of length m in ascending order. The matrix H is called a *parity check matrix*. For example, when $m = 3$, the parity check matrix $H \in \{0,1\}^{3\times 7}$ is given by

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

From now on, all operations will be done modulo 2. Under this setting, the set $\{0,1\}$ becomes a field:

$$0 \pm 0 = 0, \quad 0 \pm 1 = 1, \quad 1 \pm 1 = 0, \qquad 0 \cdot 0 = 0, \quad 0 \cdot 1 = 0, \quad 1 \cdot 1 = 1, \qquad \frac{0}{1} = 0, \quad \frac{1}{1} = 1.$$

The *Hamming codewords* correspond to the *null space of the parity check matrix*. In other words, each Hamming codeword c is a solution of the linear system

$$Hc = 0,$$

where $c \in \{0,1\}^{2^m-1}$ is a binary vector. For the case $m = 3$, there are 16 Hamming codewords:

$$\begin{array}{cccc} 0000000 & 0100101 & 1000011 & 1100110 \\ 0001111 & 0101010 & 1001100 & 1101001 \\ 0010110 & 0110011 & 1010101 & 1110000 \\ 0011001 & 0111100 & 1011010 & 1111111 \end{array} \tag{3.2}$$

We call this a $(7,4)$ *Hamming code*, and the rate is

$$R = \frac{\log_2 16}{7} = \frac{4}{7}.$$

Furthermore, since the null space $\ker(H)$ is a subspace of the vector space $\{0,1\}^{2^m-1}$, the sum of any two codewords is also a codeword.

Rate of the Hamming code. According the rank-nullity theorem, for a parity matrix $H \in \{0,1\}^{m\times(2^m-1)}$,

$$\operatorname{rank}(H) + \dim\ker(H) = 2^m - 1.$$

Since we can always pick the m distinct one-hot vectors from the columns of H, we have $\operatorname{rank}(H) = m$, and $\dim\ker(H) = 2^m - m - 1$. Therefore, the null space of H has dimension $k = 2^m - m - 1$, and over the binary field there are 2^k Hamming codewords. This is called a (N, k) *Hamming code*, which carries $k = 2^m - m - 1$ information bits via $N = 2^m - 1$ channel uses. The rate of this code is

$$R = \frac{k}{N} = 1 - \frac{m+1}{2^m - 1}.$$

As we can see, the rate R of the Hamming code converges to 1 as $m \to \infty$.

Minimum weight and minimum distance. Since the columns of H are distinct, the sum of any two columns of H must not be the all-0 vector. Hence the minimum number of 1's in any nonzero codeword is 3. This is called the *minimum weight* of the Hamming code.

If $c_1 \neq c_2$ are two distinct Hamming codewords, we have $H(c_1 - c_2) = 0$, and $c_1 - c_2$ has minimum weight 3. Hence c_1 and c_2 differ at no less than 3 bits. This is called the *minimum distance* of the Hamming code.

Covering property of the Hamming codewords. We can show that the Hamming words are widely dispersed in the space of bit words. Let $c \in \{0,1\}^{2^m-1}$ be a Hamming codeword, and write by $[c]$ the ball centered at c of radius 1 in $\{0,1\}^{2^m-1}$, i.e. $[c]$ is set of all bit words of length $2^m - 1$ whose distance to c is not greater than 1. For example, when $m = 3$ and $c = 0100101$,

$$[0100101] = \{0100101, 1100101, 0000101, 0110101, 0101101, 0100001, 0100111, 0100100\}$$

Generally, the ball $[c]$ contains 2^m words, which are c it self and the $2^m - 1$ words obtained by flipping exactly one bit of c. Since the minimum distance of the Hamming code is 3, we have $[c] \cap [\widetilde{c}] = \emptyset$ for any codewords $c \neq \widetilde{c}$. As a result, there are $2^k \cdot 2^m = 2^{2^m-1}$ distinct bit words in the union of the unit balls centered the Hamming codewords $c_1, c_2, \cdots, c_{2^k}$. Since there are in total 2^{2^m-1} bit words of length $2^m - 1$,

$$\{0,1\}^{2^m-1} = [c_1] \cup [c_2] \cup \cdots \cup [c_{2^k}]$$

Thus we obtain a cover of the space of all bit words generated by the Hamming codewords. In this sense, every bit word of length $2^m - 1$ either is a codeword or differs from a unique codeword in exactly 1 bit.

Hamming code corrects up to 1 flip. If a codeword c is corrupted in only one bit, it will differ from any other codeword in at least two bits. Hence c is the unique closest codeword.

In fact, we can identify the closest codeword without a brutal search of all codewords. We assume that e_i is the one-hot vector whose i^{th} bit is 1. If the i^{th} bit of the codeword c is flipped, the received vector is then given by $r = c + e_i$, which satisfies

$$Hr = H(c + e_i) = Hc + He_i = He_i.$$

This is simply the i^{th} column of the parity check matrix H.

Thus, assuming that only one bit was flipped, the vector Hr is the binary representation of index of the flipped bit. By flipping this bit in the received vector r, we recover the original codeword c.

Application: the hat game. We see an application of the Hamming code in game theory. In a *hat game* of N players, each player is independently assigned a hat. Each hat is colored 0 or 1 with probability 1/2. Here are the rules of the game:

- Players act a team – everyone wins or everyone loses.
- A player can observe the hats of all other players, but cannot observe the color of her own hat.
- Once hats have been distributed, there no communication between team members.
- When asked the color of their hats, all players must answer simultaneously.
- Each person is allowed to pass rather than guess a color.
- Team wins if at least one player guesses correctly and none guess incorrectly. Otherwise, the team loses.

We focus on finding an optimal strategy that maximizes the winning rate. Before we proceed, let us take a look at the best result the players can make. We let x_i be the color of the i^{th} player's hat.

- In this game, each player's decision making process is independent of the color of their own hat.
- If the j^{th} player gives a correct guess in the case $(x_1, \cdots, x_{j-1}, 0, x_{j+1}, \cdots, x_N)$, she must give a wrong guess in the case $(x_1, \cdots, x_{j-1}, 1, x_{j+1}, \cdots, x_N)$, and vice versa. Therefore, no matter what strategy the players take, there must be an equal number of correct and wrong guesses among all possible outcomes.
- However, this fact does not mean that our overall strategy has to lose as much as it wins! According to the rule, we require each win to have at least one correct guess and no wrong guess. To increase our overall winning rate, we would like that there are less correct guesses in each win and more wrong guesses in each loss. In the optimal case, we would have exactly one correct guess in every win.
- Among all 2^N outcomes, we assume that there are G wins. According to the constraint we discussed previously, to maximize G, we assume that each win has only a single correct guess. Since each loss has up to N wrong guesses, we have

$$G \leq N(2^N - G).$$

This gives an upper bound of the winning rate, and we cannot do any better:

$$\mathbb{P}(win) = \frac{G}{2^N} \leq \frac{N}{N+1}.$$

The optimal strategy. In the hat game, when the number of the players is of the form $N = 2^m - 1$, we consider the following strategy: Player j forms the bit word $(x_1, \cdots, x_{j-1}, *, x_{j+1}, \cdots, x_N)$, where x_i is the color of the i^{th} hat.

- If $(x_1, \cdots, x_{j-1}, 0, x_{j+1}, \cdots, x_N)$ forms a Hamming codeword, the player j guesses 1;
- If $(x_1, \cdots, x_{j-1}, 1, x_{j+1}, \cdots, x_N)$ forms a Hamming codeword, the player j guesses 0;
- Otherwise, the player j passes.

Using this strategy, there are only two possible outcomes:

- If $(x_1, \cdots, x_n)$ is not a Hamming codeword, then it differs from a unique Hamming codeword in exactly one bit, denoted by x_j. In this case, all players except j pass and the player j gives a correct guess.
- If $(x_1, \cdots, x_n)$ is a Hamming codeword, then each player gives a wrong guess.

Then the winning rate is one minus the proportion of Hamming codewords to all bit words:

$$\mathbb{P}(win) = 1 - \frac{2^k}{2^N} = \frac{2^{2^m - m - 1}}{2^{2^m - 1}} = 1 - 2^{-m} = \frac{N}{N+1}.$$

Hence this strategy reaches the optimal winning rate. Furthermore, the winning rate converges to 1 as $m \to \infty$.

4 Differential Entropy and Gaussian Channels

4.1 Differential Entropy of Continuous Random Variables

Motivation: Entropy of continuous random variables. We let X be a continuous real-valued random variable supported on $[a, b]$. Assume that the density function f of X is a continuous function. Then

$$\mathbb{P}(X \le x) = \int_a^x f(t)\, dt, \quad a \le x \le b.$$

We divide the range of X into bins of width $\delta > 0$:

$$a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < b < t_n, \quad t_i - t_{i-1} = \delta.$$

By mean-value theorem, there exists $x_i \in [t_{i-1}, t_i]$ such that

$$f(x_i)\delta = \int_{t_{i-1}}^{t_i} f(x)\, dx.$$

We then quantize X by defining

$$X^\delta = x_i, \quad if \;\; t_{i-1} \le X < t_i.$$

Then X^δ is a discrete random variable, and its probability mass function is given by

$$\mathbb{P}(X^\delta = x_i) = \int_{t_{i-1}}^{t_i} f(x)\, dx = f(x_i)\delta.$$

The entropy of X^δ is

$$\begin{aligned} H(X^\delta) = \sum_{i=1}^n f(x_i)\delta \log \frac{1}{f(x_i)\delta} &= \sum_{i=1}^n f(x_i)\delta \frac{1}{f(x_i)} + \sum_{i=1}^n f(x_i)\delta \log \frac{1}{\delta} \\ &= \sum_{i=1}^n (t_i - t_{i-1}) f(x_i) \log \frac{1}{f(x_i)} + \log \frac{1}{\delta} \sum_{i=1}^n \int_{t_{i-1}}^{t_i} f(x)\, dx \\ &= \sum_{i=1}^n (t_i - t_{i-1}) f(x_i) \log \frac{1}{f(x_i)} + \log \frac{1}{\delta}. \end{aligned}$$

This entropy blows up as $\delta \to \infty$. Therefore, the entropy of a continuous random variable is infinite. However, since $f : [a, b] \to \mathbb{R}_+$ is Riemann integrable, we have

$$\begin{aligned} \lim_{\delta \downarrow 0} \left(H(X^\delta) - \log \frac{1}{\delta} \right) &= \int_a^b f(x) \log \frac{1}{f(x)}\, dx \\ &= \mathbb{E}\left[\log \frac{1}{f(X)} \right]. \end{aligned}$$

We can extend this definition to multidimensional spaces.

Definition 4.1 (Differential entropy)**.** Let $X \sim f$ be a continuous random variable, and the range of X is $\mathcal{X} \subset \mathbb{R}^p$. If the function $x \mapsto f(x) \log f(x)$ is integrable, define the *differential entropy of* X to be

$$h(X) = \int_{\mathcal{X}} f(x) \log \frac{1}{f(x)}\, dx = -\mathbb{E}\left[\log f(X) \right].$$

Example 4.2. *Here are some examples of differential entropy.*

(i) Let X be a uniform random variable on $[0,a]$. Then $h(X) = \int_0^a \frac{1}{a}\log a\,dx = \log a$. When $0 < a < 1$, we have $h(X) < 0$. It is seen that the differential entropy can be negative!

(ii) Let $X \sim N(0,\sigma^2)$ be a Gaussian random variable. Then

$$h(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \left(\log\left(\sqrt{2\pi}\sigma\right) + \frac{x^2}{2\sigma^2} \right) dx = \frac{1}{2} + \frac{1}{2}\log\left(2\pi\sigma^2\right).$$

(iii) Let $X \sim N(0,\Sigma)$ be a p-dimensional Gaussian random vector, where the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ is nonsingular. Then

$$\begin{aligned} h(X) &= \int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}x^\top\Sigma^{-1}x} \left(\log\left((2\pi)^{p/2}\det(\Sigma)^{1/2}\right) + \frac{1}{2}x^\top\Sigma^{-1}x \right) dx \\ &= \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \underbrace{\frac{1}{2}\mathbb{E}\left[X^\top\Sigma^{-1}X\right]}_{=\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\mathbb{E}[XX^\top])} \\ &= \frac{p}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}\log\det(\Sigma). \end{aligned}$$

The definition of conditional differential entropy, mutual information and relative entropy then follows from the differential entropy.

Definition 4.3. Let $X, Y, Z \sim f$ be three continuous random variables. For brevity, we also write $f(x)$ and $f(y)$ for the marginal density function of X and Y, respectively.

(i) The *joint differential entropy* between X and Y is the differential entropy of the random vector (X,Y);

(ii) The *conditional differential entropy* of Y given X is

$$h(Y|X) = -\int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\log f(y|x)\,dx\,dy.$$

(iii) The *mutual information* between X and Y is

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}\,dx\,dy.$$

(iv) The *conditional mutual information* between X and Y given Z is

$$I(X;Y|Z) = \int_{\mathcal{X}\times\mathcal{Y}\times Z} f(x,y,z)\log\frac{f(x,y|z)}{f(x|z)f(y|z)}\,dx\,dy\,dz.$$

(v) Given two density functions f and g defined in the same space $\mathcal{X} \subset \mathbb{R}^p$ such that $g \ll f$, i.e. $g(x) = 0$ for all $x \in U$ with $f(x) = 0$. Then the *Kullback-Leibler divergence* of g from f is

$$D(f \,\|\, g) := \int_{\mathcal{X}} f(x)\log\frac{f(x)}{g(x)}\,dx = \mathbb{E}_{X\sim f}\left[\log\frac{f(X)}{g(X)}\right].$$

Remark. Many identities and inequalities in the discrete case also applies to the continuous case:

- $h(X,Y) = h(X) + h(Y|X)$.
- $I(X;Y) = h(Y) - h(Y|X) = h(X) - h(X|Y)$.
- $I(X;Y) = D(f_{XY} \,\|\, f_X f_Y)$.
- $I(X;Y|Z) = h(X|Z) - h(X|Y,Z) = h(Y|Z) - h(Y|X,Z)$.
- $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$.

Example 4.4. *We aim to compute the mutual information between two jointly Gaussian variables.*

(a) Let X and Y be two jointly Gaussian random vectors:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

where $\Sigma_{11} \in \mathbb{R}^{p\times p}$ and $\Sigma_{22} \in \mathbb{R}^{q\times q}$ are both nonsingular, and the covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ is also nonsingular. Then

$$\begin{aligned} h(Y|X) &= \int_{\mathbb{R}^p} f(x) \int_{\mathbb{R}^q} f(y|x) \log \frac{1}{f(y|x)}\, dy\, dx \\ &= \int_{\mathbb{R}^p} f(x) \left(\frac{p}{2} \log(2\pi e) + \frac{1}{2} \log\det(\Sigma_{22.1}) \right) dx \\ &= \frac{q}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log\det(\Sigma_{22.1}) \end{aligned}$$

where the conditional covariance matrix is $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. By Schur complement,

$$\det(\Sigma) = \det(\Sigma_{11}) \det(\Sigma_{22.1}) \quad \Rightarrow \quad \log\det(\Sigma) = \log\det(\Sigma_{11}) + \log\det(\Sigma_{22.1}).$$

Therefore,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= \frac{q}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log\det(\Sigma_{11}) - \frac{q}{2} \log(2\pi \mathrm{e}) - \frac{1}{2} \log\det(\Sigma_{22.1}) \\ &= \frac{1}{2} \log \frac{\det(\Sigma_{11}) \det(\Sigma_{22})}{\det(\Sigma)}. \end{aligned}$$

To summarize,

$$h(Y|X) = \frac{q}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log \frac{\det(\Sigma)}{\det(\Sigma_{11})}, \quad h(X|Y) = \frac{p}{2} \log(2\pi \mathrm{e}) + \frac{1}{2} \log \frac{\det(\Sigma)}{\det(\Sigma_{22})},$$

and

$$I(X;Y) = \frac{1}{2} \log \frac{\det(\Sigma_{11}) \det(\Sigma_{22})}{\det(\Sigma)}.$$

In particular, if X and Y are independent, the covariance matrix is $\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$, and $I(X;Y) = 0$.

(b) We consider the bivariate Gaussian distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} \right),$$

where $\rho \in (-1, 1)$ is the correlation coefficient between X and Y. Then

$$I(X;Y) = \frac{1}{2} \log \frac{1}{1-\rho^2}.$$

In particular, if $\rho = 0$, the mutual information between X and Y is 0; and if $\rho = \pm 1$, the mutual information between X and Y is infinity.

(c) Let $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$, where $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p\times p}$. Then

$$\begin{aligned}
D(X\|Y) &= \mathbb{E}\left[\log \frac{f_X(X)}{f_Y(X)}\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathbb{E}\left[(X-\mu_1)^\top \Sigma_1^{-1}(X-\mu_1)\right] + \frac{1}{2}\mathbb{E}\left[(X-\mu_2)^\top \Sigma_2^{-1}(X-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{p}{2} + \frac{1}{2}\operatorname{tr}\left(\Sigma_2^{-1}\mathbb{E}\left[(X-\mu_2)(X-\mu_2)^\top\right]\right) \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{p}{2} + \frac{1}{2}\left((\mu_1-\mu_2)^\top\Sigma_2(\mu_1-\mu_2) + \operatorname{tr}(\Sigma_2^{-1}\Sigma_1)\right).
\end{aligned}$$

Theorem 4.5 (Linear transformation)**.** *Let $A \in \mathbb{R}^{p\times p}$ be a nonsingular matrix, and $b \in \mathbb{R}^p$. Let X be a continuous p-dimensional random vector. Then*

$$h(AX+b) = h(X) + \log|\det(A)|.$$

Proof. Let $Y = AX$. If X has density function f, the density of Y is given by

$$g(y) = \frac{f(A^{-1}y)}{|\det(A)|}, \quad y \in \mathbb{R}^p.$$

Then the differential entropy of Y is

$$\begin{aligned}
h(Y) &= -\int_{\mathbb{R}^p} g(y)\log g(y)\,dy \\
&= -\int_{\mathbb{R}^p} \frac{f(A^{-1}y)}{|\det(A)|}\log\frac{f(A^{-1}y)}{|\det(A)|}\,dy \\
&= -\int_{\mathbb{R}^p} \frac{f(x)}{|\det(A)|}\log\frac{f(x)}{|\det(A)|}\,|\det(A)|\,dx \qquad \text{(change the variable } x = A^{-1}y\text{)} \\
&= -\int_{\mathbb{R}^p} f(x)\log f(x)\,dx + \int_{\mathbb{R}^n} f(x)\log|\det(A)|\,dx \\
&= h(X) + \log|\det(A)|.
\end{aligned}$$

By change the variable $Z = Y + b = Ax + b$, we know that $h(Z) = h(Y)$. This is the desired result. □

Remark. This transformation formula also holds for conditional differential entropy. Analogous to this formula, we have the transformation invariance for mutual information and KL-divergence:

$$\begin{aligned}
h(Ax+b|Y) &= h(X|Y) + \log|\det(A)|, \\
I(AX+b;Y) &= I(X;Y), \\
D(f_{Ax+b}\|f_{AY+b}) &= D(f_X\|f_Y).
\end{aligned}$$

We have the following estimate for the differential entropy of a random vector.

Theorem 4.6 (Upper bound of the differential entropy)**.** *If X is a p-dimensional random vector with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$,*

$$h(X) \le \frac{p}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}\log\det(\Sigma)$$

The inequality holds if and only if $X \sim N(\mu, \Sigma)$. In other words, the Gaussian distribution maximizes the differential entropy under second moment constraints.

Proof. We may assume $\mu = \mathbb{E}[X] = 0$ without loss of generality. Let $Z \sim f_Z$ be the Gaussian random variable with $\mathbb{E}[Z] = \mathbb{E}[X] = 0$ and $\mathrm{Cov}(Z) = \mathrm{Cov}(X) = \Sigma$. Then

$$\begin{aligned} 0 \le D(f_X \,\|\, f_Z) &= \mathbb{E}\left[\log \frac{f_X(X)}{f_Z(X)}\right] \\ &= -h(X) + \int_{\mathbb{R}^p} f_X(x) \left(\frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}x^\top \Sigma^{-1} x\right) dx \\ &= -h(X) + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\int_{\mathbb{R}^p} f_X(x)\,\mathrm{tr}\left(\Sigma^{-1}xx^\top\right) dx \\ &= -h(X) + \frac{p}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}\log\det(\Sigma). \end{aligned}$$

Therefore,

$$h(X) \le \frac{p}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}\log\det(\Sigma) = h(Z).$$

The equality holds if and only if $D(f_X \,\|f_Z) = 0$, which is equivalent to $X \overset{d}{=} Z$. □

Theorem 4.7 (Estimation error and differential entropy)**.** *Let X be a p-dimensional random vector, and let $\widehat{X}$ be an estimate of X. If $X \to Y \to \widehat{X}$ form a Markov chain,*

$$\mathbb{E}\left[\left|X - \widehat{X}\right|^2\right] \ge \frac{p\,\mathrm{e}^{\frac{2}{p}h(X|Y)}}{2\pi\mathrm{e}},$$

where $|\cdot|$ denotes the Euclidean norm.

Proof. Conditioning on the event $\{Y = y\}$, the variables X and $\widehat{X}$ are independent. We assume $\Sigma \in \mathbb{R}^{p\times p}$ is the conditional covariance matrix of X given $Y = y$. Since the expectation $\mu = \mathbb{E}[X \,|\, Y = y]$ minimizes the mean square error $\mathbb{E}[|X - \mu|^2 \,|\, Y = y]$, we have

$$\mathbb{E}\left[\left|X - \widehat{X}\right|^2 \,\middle|\, Y = y\right] \ge \mathbb{E}\left[(X-\mu)^\top (X-\mu) \,\middle|\, Y = y\right] = \mathrm{tr}(\Sigma).$$

We let $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ be the eigenvalues of Σ. Then

$$\begin{aligned} \log\mathrm{tr}(\Sigma) = \log p + \log\left(\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_p}{p}\right) &\ge \log p + \frac{1}{p}\log\lambda_1 + \frac{1}{p}\log\lambda_2 + \cdots + \frac{1}{p}\log\lambda_p \\ &= \log p + \frac{1}{p}\log\det(\Sigma). \end{aligned}$$

By Theorem 4.6, we have

$$\frac{1}{2}\log\det(\Sigma) \ge h(X \,|\, Y = y) - \frac{p}{2}\log(2\pi\mathrm{e}).$$

Hence

$$\mathbb{E}\left[\left|X - \widehat{X}\right|^2 \,\middle|\, Y = y\right] = \mathrm{tr}(\Sigma) \ge \exp\left(\log p + \frac{2}{p}h(X \,|\, Y = y) - \log(2\pi\mathrm{e})\right) = \frac{p}{2\pi\mathrm{e}}\mathrm{e}^{\frac{2}{p}h(X|Y=y)}.$$

Take expectation on both sides. The result follows then from Jensen's inequality. □

4.2 Capacity of Gaussian Channels

Motivation. In many scenarios, the error between the sent message X and the received message Y can be modeled as additive white Gaussian noise (AWGN). A discrete-time Gaussian channel is given by

$$Y_i = X_i + Z_i, \quad \textit{where} \;\; Z_i \sim N(0, N) \textit{ is independent of } X_i.$$

If there is no constraint on the input, we can choose an infinite subset of inputs arbitrarily far apart to separate the output with arbitrarily small probability of error. To mode real-world constraints, we impose average power constraint on codewords $(x_1, \cdots, x_n)$:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P.$$

Communication of one bit. We provide a simple strategy for communication on the AWGN channel. To transmit a single bit, we send $X = -\sqrt{P}$ for 0 and send $X = \sqrt{P}$ for 1. Then the received signal

$$Y = \pm\sqrt{P} + Z$$

is symmetric. For the decoder, we can simply choose $\sqrt{P}$ when $Y \ge 0$ and $-\sqrt{P}$ when $Y < 0$. Then the probability of error is

$$\begin{aligned} P_e &= \frac{1}{2}\mathbb{P}\left(Y \ge 0 \,|\, X = -\sqrt{P}\right) + \frac{1}{2}\mathbb{P}\left(Y < 0 \,|\, X = \sqrt{P}\right) \\ &= \frac{1}{2}\mathbb{P}\left(Z \ge \sqrt{P}\right) + \frac{1}{2}\mathbb{P}\left(Z < -\sqrt{P}\right) \\ &= \mathbb{P}\left(Z > \sqrt{P}\right) = 1 - \Phi\left(\sqrt{P/N}\right), \end{aligned}$$

where Φ is the cumulative distribution function of $N(0,1)$ distribution. It is seen that the probability of error is small when the signal-noise ratio (SNR) P/N is large.

Theorem 4.8. *The information capacity of the Gaussian channel with additive noise power B and power constraint P is*

$$C := \max_{f_X : \mathbb{E}[X^2] \le P} I(X;Y) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right).$$

Proof. The mutual information between X and Y is

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X) = h(Y) - h(Z) = h(Y) - \frac{1}{2}\log(2\pi\mathrm{e}N).$$

Since X and Z are independent, the variance of $Y = X + Z$ is less than or equal to $P + N$, and the differential entropy of Y is maximized when Y is Gaussian:

$$\max_{\mathbb{E}[Y^2] \le P+N} h(Y) = \frac{1}{2}\log(2\pi\mathrm{e}(P+N)).$$

Then

$$\max_{f_X : \mathbb{E}[X^2] \le P} I(X;Y) = \frac{1}{2}\log(2\pi\mathrm{e}(P+N)) - \frac{1}{2}\log(2\pi\mathrm{e}N) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right).$$

The equality holds when $X \sim N(0, P)$. □

Definition 4.9. A rate R is *achievable* for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $P_{e,\max}^{(n)}$ converges to zero. The *capacity* of the channel is the supremum of the achievable rates:

$$C_{\text{op}} = \sup\{R : R \textit{ is achievable}\}$$

Theorem 4.10. *The capacity of the Gaussian channel with additive noise power N and power constraint P is equal to the information capacity:*

$$C_{\text{op}} = \frac{1}{2}\log\left(1 + \frac{P}{N}\right).$$

Remark. This theorem also has two parts:

- (Achievability) If $R < \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$, then R is achievable.
- (Converse) If R is achievable, then $R \leq \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$.

Proof of Theorem 4.10 (Achievability part). Similar to our proof of the availability part of Theorem 3.3 in the case of discrete channels, we employ a random coding approach as follows:

- *Construction of a random codebook.* For each message $w \in \{1, 2, \cdots, 2^{nR}\}$, independently generate

  $$X_1(w), X_2(w), \cdots, X_n(w) \overset{i.i.d.}{\sim} N(0, P - \epsilon).$$

  Then we get a codebook $\mathcal{E} : \mathcal{W} \to \mathbb{R}^n$, and it is revealed to both the encoder and the decoder. When the encoder receives a message w, it sends $X_{1:n}(w)$ to the Gaussian channel $Y = X + Z$.
- *Decoding.* When receiving the output $Y_{1:n}$, the decoder looks down the list of codewords $X_{1:n}(w)$, and searches for a codeword that is *jointly typical* with $Y_{1:n}$. If there exists a unique such codeword $X_{1:n}(w)$, the decoder declares $\widehat{W} = w$; otherwise, it declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint $\frac{1}{n}\sum_{i=1}^{n} X_i(w)^2 \leq P$.
- *Probability of error.* Without loss of generality, assume the message 1 is transmitted. Then the output is $Y_{1:n} = X_{1:n}(1) + Z_{1:n}$. Define the following events:

  $$E_{n,0} = \left\{\frac{1}{n}\sum_{i=1}^{n} X_i(w)^2 > P\right\}, \quad E_{n,i} = \left\{(X_{1:n}(i), Y_{1:n}) \in A_\epsilon^{(n)}\right\}, \quad i = 1, 2, \cdots, 2^{nR}.$$

We fix $\epsilon > 0$. By the weak law of large numbers,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\left(X_1(1)^2 + X_2(1)^2 + \cdots + X_n(1)^2\right) > P\right) = 0.$$

Since $X_{1:n}(1)$ and $Y_{1:n}$ are jointly typical,

$$\lim_{n\to\infty} \mathbb{P}(E_{n,1}^c) = 0.$$

Furthermore, by joint asymptotic equipartition property, since $X_{1:n}(w), Y_{1:n}$ have the same marginal as $X_{1:n}(1), Y_{1:n}$ and are independent for all $i = 2, 3, \cdots, 2^{nR}$,

$$\mathbb{P}(E_{n,i}) \leq 2^{-n(I(X;Y)-3\epsilon)}, \quad i = 2, 3, \cdots, 2^{nR}.$$

We choose $N_\epsilon > 0$ great enough such that

$$\mathbb{P}(E_{n,0}) < \epsilon \quad \textit{and} \quad \mathbb{P}(E_{n,1}^c) < \epsilon \quad \textit{for all } n \geq N_\epsilon.$$

Similar to the analysis in the discrete case, the probability of error is uniform over the events $W = 1, 2, \cdots, 2^{nR}$. Then for all $n \geq N_\epsilon$,

$$\begin{aligned}\mathbb{P}(\widehat{W} \neq W) = \mathbb{P}(\widehat{W} \neq W | W = 1) &= \mathbb{P}\left(E_{n,0} \cup E_{n,1}^c \cup E_{n,2} \cup \cdots \cap E_{n,2^{nR}}\right) \\ &\leq 2\epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}.\end{aligned}$$

Note that

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X) = h(Y) - h(Z) = \frac{1}{2}\log\left(1 + \frac{P-\epsilon}{N}\right)$$

If the rate $R < \frac{1}{2}\log(1 + \frac{P}{N})$, we can find a sufficiently small $\epsilon > 0$ such that

$$I(X;Y) - R - 3\epsilon = \frac{1}{2}\log\left(1 + \frac{P-\epsilon}{N}\right) - R - 3\epsilon > 0.$$

Then the error probability tends to 0 as $n \to \infty$ and $\epsilon \to 0$.

Since this error probability is the average over all codebooks and all messages, we reapply our trick in the proof of discrete memory loss channel: choose a good codebook $\mathcal{E}^*$ and expunge the worst half of the codewords. Then the maximal conditional probability of error is small. In particular, each of the remaining codewords must satisfy the power constraint, otherwise it has conditional probability of error 1 and must belong to the worst half. The new code has rate $R - \frac{1}{n}$, which can be arbitrarily close to the capacity C. Thus we proved the availability part of the theorem. □

Proof of Theorem 4.10 (Converse part). Consider any $(2^{nR}, n)$ code that satisfies the power constraint:

$$\sum_{i=1}^{n} x_i(w)^2 \leq P, \quad w = 1, 2, \cdots, 2^{nR}.$$

Let $W \sim \text{Unif}\{1, 2, \cdots, 2^{nR}\}$. We then consider the Markov chain $W \to X_{1:n}(W) \to Y_{1:n} \to \widehat{W}$. By Fano's inequality, if $\mathbb{P}(\widehat{W} \neq W) = P_e^{(n)}$,

$$H(W|\widehat{W}) \leq 1 + nRP_e^{(n)}.$$

Let $X_{1:n}(W) = X_{1:n}$. Then

$$\begin{aligned}nR = H(W) &= I(W;\widehat{W}) + H(W|\widehat{W}) \\ &\leq I(X_{1:n};Y_{1:n}) + 1 + nRP_e^{(n)} && \text{(by data processing inequality)} \\ &= h(Y_{1:n}) - h(Y_{1:n}|X_{1:n}) + 1 + nRP_e^{(n)} \\ &= h(Y_{1:n}) - h(Z_{1:n}) + 1 + nRP_e^{(n)} \\ &\leq \sum_{i=1}^{n} h(Y_i) - h(Z_{1:n}) + 1 + nRP_e^{(n)} && \text{(conditioning does not increase entropy)} \\ &= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + 1 + nRP_e^{(n)}. && (4.1)\end{aligned}$$

Assume that the average power of the i-th column of the codebook:

$$\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} x_i^2(w) = P_i, \quad i = 1, 2, \cdots, n.$$

Since $Y_i = X_i + Z_i$, and since X_i and Z_i are independent, the average power $\mathbb{E}[Y_i^2] = P_i + N$. The differential entropy is maximized by the Gaussian distribution:

$$h(Y_i) \leq \frac{1}{2}\log(2\pi \mathrm{e}(P_i + N)).$$

Plugging in this to (4.1), we have

$$\begin{aligned} nR &\leq \sum_{i=1}^{n} \frac{1}{2}\log\left(1 + \frac{P_i}{N}\right) + 1 + nRP_e^{(n)} \\ &\leq \frac{n}{2}\log\left(1 + \sum_{i=1}^{n}\frac{P_i}{nN}\right) + 1 + nRP_e^{(n)} && \text{(by Jensen's inequality)} \\ &\leq \frac{n}{2}\log\left(1 + \frac{P}{N}\right) + 1 + nRP_e^{(n)} \end{aligned}$$

Therefore

$$P_e^{(n)} \geq 1 - \frac{1}{2R}\log\left(1 + \frac{P}{N}\right) - \frac{1}{nR}, \quad \forall n \in \mathbb{N}.$$

Since $P_e^{(n)} \to 0$ as $n \to 0$, we require $R \leq \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$. □

4.3 Parallel Gaussian Channels

Problem Setting. We consider k independent Gaussian channels with a common power constraint:

$$\begin{aligned} &\textit{Channel}: \ Y_i = X_i + Z_i, \quad i = 1, 2, \cdots, k, \\ &\textit{Power constraint}: \sum_{i=1}^{k}\mathbb{E}[X_i^2] := \sum_{i=1}^{k} P_i \leq P, \\ &\textit{Independent additive Gaussian noises}: Z_i \sim N(0, N_i), \quad i = 1, 2, \cdots, k \end{aligned}$$

Our goal is to distribute the power amongst the channels to maximize the total capacity:

$$C = \max\left\{ I(X_{1:k}; Y_{1:k}) \,\middle|\, X_1, X_2, \cdots, X_k : \sum_{i=1}^{k}\mathbb{E}[X_i^2] \leq P \right\}$$

An upper bound. As usual, we decompose and estimate the mutual information as follows:

$$\begin{aligned} I(X_{1:k}; Y_{1:k}) &= h(Y_{1:k}) - h(Y_{1:k}|X_{1:k}) \\ &= h(Y_{1:k}) - h(X_{1:k} + Z_{1:k}|X_{1:k}) \\ &= h(Y_{1:k}) - h(Z_{1:k}) \\ &= \sum_{i=1}^{k} h(Y_i|Y_{i-1}, \cdots, Y_1) - \sum_{i=1}^{k} h(Z_i) \\ &\leq \sum_{i=1}^{k}(h(Y_i) - h(Z_i)) \leq \frac{1}{2}\sum_{i=1}^{k}\log\left(1 + \frac{P_i}{N_i}\right). \end{aligned}$$

This upper bound can be reached when $X_1, X_2, \cdots, X_k$ are independent with

$$X_i \sim N(0, P_i), \quad i = 1, 2, \cdots, k.$$

Solution. To solve the capacity, we consider the following optimization problem:

$$\max_{P_1,\cdots,P_k} \sum_{i=1}^{k} \log\left(1+\frac{P_i}{N_i}\right), \quad \textit{subject to} \quad P_1,\cdots,P_k \geq 0,\ \sum_{i=1}^{k} P_i \leq P.$$

Since the objective function is concave about $P_1,\cdots,P_k$, define the Lagrangian function:

$$L(P_1,\cdots,P_k,\lambda) = \sum_{i=1}^{k} \log\left(1+\frac{P_i}{N_i}\right) - \sum_{i=1}^{k} \mu_i P_i - \lambda\left(\sum_{i=1}^{k} P_i - P\right), \quad P_1,\cdots,P_k,\mu_1,\cdots,\mu_k,\lambda \geq 0.$$

Apply the KKT conditions to solve the problem:

$$\begin{cases} \dfrac{\partial L}{\partial P_i} = \dfrac{1}{P_i+N_i} - \mu_i - \lambda = 0, \\ \sum_{i=1}^{k} P_i - P = 0, \\ \sum_{i=1}^{k} \mu_i P_i = 0, \\ P_1,\cdots,P_k,\mu_1,\cdots,\mu_k,\lambda \geq 0. \end{cases}$$

Then for each $i = 1,2,\cdots,k$, the optimal solution satisfies

$$P_i^* = \frac{1}{\mu_i^*+\lambda^*} - N_i,$$

and at least one of P_i and μ_i^* is zero. This implies

$$P_i^*(\lambda^*) = \begin{cases} \dfrac{1}{\lambda^*} - N_i, & \frac{1}{\lambda^*} - N_i \geq 0 \\ 0, & \frac{1}{\lambda^*} - N_i < 0 \end{cases} = \max\left(\frac{1}{\lambda^*} - N_i, 0\right)$$

Furthermore, we require

$$\sum_{i=1}^{k} P_i^*(\lambda^*) - P = 0. \tag{4.2}$$

Since the function $\lambda \mapsto \sum_{i=1}^{n} \max\left(\frac{1}{\lambda} - N_i, 0\right)$ is strictly monotone decreasing from ∞ to 0 on the interval $(0, \frac{1}{\min_{1\leq i\leq k} N_i})$, one can solve $\lambda^* > 0$ uniquely from (4.2). By construction,the power allocation $P_i^*(\lambda^*)$ satisfies the constraints of our problem and thus is a feasible solution. Hence

$$C = \frac{1}{2}\sum_{i=1}^{k} \log\left(1+\frac{P_i^*(\lambda^*)}{N_i}\right) = \frac{1}{2}\sum_{i=1}^{k} \log\left(1+\frac{1}{N_i}\max\left(\frac{1}{\lambda^*} - N_i, 0\right)\right),$$

where $\lambda^* > 0$ is the unique solution to the equation

$$\sum_{i=1}^{k} \max\left(\frac{1}{\lambda^*} - N_i, 0\right) = P.$$

This is also known as the *water filling* solution.

4.4 I-MMSE Relationship

Setting. Given a random variable X with finite variance $\sigma^2 > 0$, let

$$Y = \sqrt{s}X + W, \quad W \sim \mathcal{N}(0,1) \text{ is independent of } X.$$

This is a variant of the Gaussian channel, and the constant $s \geq 0$ is called the *signal-to-noise ratio (SNR)* of the channel. The capacity of this channel is measured by the mutual information $I(X;Y)$.

Estimation error. The *minimum mean-squared error (MMSE)* in estimating X from Y is defined as

$$\mathsf{mmse}(X|Y) = \min_{g:\mathcal{Y}\to\mathbb{R}} \mathbb{E}\left[(g(Y) - X)^2\right]$$

It is easy to verify that the optimal estimation function $g(Y)$ is given by the conditional expectation $\mathbb{E}[X|Y]$. Consequently, the MMSE can be written as

$$\mathsf{mmse}(X|Y) = \mathbb{E}\left[|X - \mathbb{E}[X|Y]|^2\right] = \mathbb{E}\left[\mathrm{Var}(X|Y)\right].$$

This is in fact the squared distance from X to its projection onto the subspace spanned by Y. The MMSE measures the uncertainty of X given an observation Y.

The I-MMSE relationship states that for any distribution on X, the derivative of the mutual information with respect to the signal-to-noise ratio is equal to one half the MMSE:

$$\frac{d}{ds} I(X;Y) = \frac{1}{2}\mathsf{mmse}(X|Y).$$

In the remainder of this section, we aim to establish this result.

Lemma 4.11 (Almost Gaussian variable)**.** *Let X be a random variable with mean μ and finite variance $\sigma^2 > 0$, and let $W \sim \mathcal{N}(0,1)$ be a noise independent of X. When $Y = \sqrt{s}X + W$, and $Y' \sim \mathcal{N}(\sqrt{s}\mu, 1 + s\sigma^2)$ is a Gaussian variable that has the same mean and variance as Y, then*

$$\lim_{s\to 0} \frac{D(Y\|Y')}{s} = 0.$$

Proof. We may assume $\mu = 0$ without loss of generality, otherwise we replace X with $X - \mu$. By definition,

$$\begin{aligned} D(Y\|Y') = \int_{\mathbb{R}} f_Y(y) \log \frac{f_Y(y)}{f_{Y'}(y)}\, dy &= \int_{\mathbb{R}} f_Y(y) \left(\frac{1}{2}\log(2\pi(1+s\sigma^2)) + \frac{y^2}{2(1+s\sigma^2)}\right) dy - h(Y) \\ &= \left(\frac{1}{2}\log\left(2\pi(1+s\sigma^2)\right) + \frac{\mathbb{E}[Y^2]}{2(1+s\sigma^2)}\right) dy - h(Y) \\ &= \frac{1}{2}\log\left(2\pi\mathrm{e}(1+s\sigma^2)\right) - h(Y). \end{aligned}$$

We fix $M > 0$, and define $B = \mathbb{1}_{\{|X|\leq M\}}$. Then

$$\begin{aligned} h(Y) &= \mathbb{P}(B=1)h(Y|B=1) + \mathbb{P}(B=0)h(Y|B=0) \\ &= \mathbb{P}(B=1)h(Y|B=1) + \mathbb{P}(B=0)h(\sqrt{s}X + W|B=0) \\ &\geq \mathbb{P}(B=1)h(Y|B=1) + \mathbb{P}(B=0)h(W), \end{aligned}$$

where the last inequality holds because W is independent of B and X, and B is a function of X:

$$h(\sqrt{s}X + W|B=0) \leq h(\sqrt{s}X + W|X, B=0) = h(\sqrt{s}X + W|X) = h(W).$$

Since $|X| \le M$ conditioning on the event $B = 1$, using Taylor's expansion, we have

$$\begin{aligned} f_Y(y|B=1) &= \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(y-\sqrt{s}X)^2/2}\middle| B=1\right] \\ &= \frac{\mathrm{e}^{-y^2/2}}{\sqrt{2\pi}}\mathbb{E}\left[1+\sqrt{s}yX+\frac{s}{2}(y^2-1)X^2+o(s)\middle| B=1\right] \\ &= \frac{\mathrm{e}^{-y^2/2}}{\sqrt{2\pi}}\left(1+\sqrt{s}y\mathbb{E}[X|B=1]+\frac{s}{2}(y^2-1)\mathbb{E}[X^2|B=1]+o(s)\right), \end{aligned}$$

where $o(s)$ is a quantity smaller than s in the sense that $\lim_{s\to 0}\frac{o(s)}{s} = 0$. Hence

$$\begin{aligned} &h(Y|B=1) = -\mathbb{E}\left[\log f_Y(Y|B=1)|B=1\right] \\ &= \frac{1}{2}\mathbb{E}[Y^2|B=1] + \frac{1}{2}\log(2\pi) - \sqrt{s}\mathbb{E}[Y|B=1]\mathbb{E}[X|B=1] + \frac{s}{2}\mathbb{E}[Y^2|B=1]\mathbb{E}[X|B=1]^2 \\ &\quad - \frac{s}{2}\mathbb{E}[Y^2-1|B=1]\mathbb{E}[X^2|B=1] + o(s) \\ &= \frac{s}{2}\operatorname{Var}(X|B=1) + \frac{1}{2}\log(2\pi\mathrm{e}) - s\mathbb{E}[X|B=1]^2 + o(s) \end{aligned}$$

where the last equality holds because

$$\begin{aligned} &\mathbb{E}[Y|B=1] = \mathbb{E}[\sqrt{s}X+W|B=1] = \sqrt{s}\mathbb{E}[X|B=1], \\ &\mathbb{E}[Y^2|B=1] = \mathbb{E}\left[(\sqrt{s}X+W)^2|B=1\right] = s\mathbb{E}[X^2|B=1]+1. \end{aligned}$$

For any $\delta > 0$, by Lebesgue's dominated convergence theorem, we can choose $M = M_\delta > 0$ such that $\mathbb{P}(B=1) \ge 1-\delta$, $|\mathbb{E}[X|B=1]| \le \delta$ and $\left|\operatorname{Var}(X|B=1) - \sigma^2\right| \le \delta$. Therefore

$$h(Y|B=1) \ge \frac{1}{2}\log(2\pi\mathrm{e}) + \frac{s\sigma^2}{2} - \frac{3}{2}\delta + o(s).$$

Then for sufficiently small s, we have

$$\begin{aligned} h(Y) &= \mathbb{P}(B=1)h(Y|B=1) + \mathbb{P}(B=0)h(W) \\ &= \mathbb{P}(B=1)\left(\frac{1}{2}\log(2\pi\mathrm{e}) + \frac{s\sigma^2}{2} - \frac{3}{2}\delta + o(s)\right) + \mathbb{P}(B=0)\frac{1}{2}\log(2\pi\mathrm{e}) \\ &\ge \frac{1}{2}\log(2\pi\mathrm{e}) + (1-\delta)\frac{s\sigma^2}{2} - \frac{3}{2}\delta + o(s), \end{aligned}$$

and

$$\begin{aligned} D(Y\|Y') &= \frac{1}{2}\log(2\pi\mathrm{e}(1+s\sigma^2)) - h(Y) \\ &\le \frac{1}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}s\sigma^2 + o(s) - h(Y) \\ &\le \left(\frac{s\sigma^2}{2} + \frac{3}{2}\right)\delta + o(s). \end{aligned}$$

Since the choice $\delta > 0$ is arbitrary and does not depend on s, we have $D(Y\|Y') \le o(s)$. □

Remark. We have an intuitive interpretation for this lemma. When $s > 0$ is sufficiently small, the random variable $Y = \sqrt{s}X + W$ is almost Gaussian. In fact, the density of Y is the *convolution* of the density of Gaussian variable W and a "pulse" near 0. Hence $Y = \sqrt{s}X + W$ is "close" to the Gaussian variable $Y' \sim \mathcal{N}(0, 1+s\sigma^2)$, which has the same mean and variance as Y.

Lemma 4.12. *Under the assumption of Lemma 4.11, one have*

$$\lim_{s\to 0}\frac{I(X;Y)}{s}=\frac{\sigma^2}{2}.$$

Proof. We may assume $\mathbb{E}[X]=0$. Let $Y'\sim\mathcal{N}(0,1+s\sigma^2)$. Then

$$\begin{aligned}
I(X;Y)&=\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X,Y}(x,y)\frac{f_{Y|X}(y|x)}{f_Y(y)}\,dx\,dy\\
&=\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X,Y}(x,y)\frac{f_{Y|X}(y|x)}{f_{Y'}(y)}\,dx\,dy-\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X,Y}(x,y)\frac{f_Y(y)}{f_{Y'}(y)}\,dx\,dy\\
&=\int_{\mathbb{R}} f_X(x)\int_{\mathbb{R}} f_{Y|X}(y|x)\frac{f_{Y|X}(y|x)}{f_{Y'}(y)}\,dy\,dx-\int_{\mathbb{R}} f_Y(y)\frac{f_Y(y)}{f_{Y'}(y)}\,dy\\
&=\int_{\mathbb{R}} f_X(x)D(\sqrt{s}x+W\|Y')\,dx-D(Y\|Y').
\end{aligned}$$

We analyze the first term. Since $\sqrt{s}x+W\sim\mathcal{N}(\sqrt{s}x,1)$ and $Y'\sim\mathcal{N}(0,1+s\sigma^2)$ are both Gaussian,

$$D(\sqrt{s}x+W\|Y')=\frac{1}{2}\log(1+s\sigma^2)+\frac{1}{2}\frac{s(x^2-\sigma^2)}{1+s\sigma^2},$$

and

$$\int_{\mathbb{R}} f_X(x)D(\sqrt{s}x+W\|Y')\,dx=\mathbb{E}\left[\frac{1}{2}\log(1+s\sigma^2)+\frac{1}{2}\frac{s(X^2-\sigma^2)}{1+s\sigma^2}\right]=\frac{1}{2}\log(1+s\sigma^2).$$

According to Lemma 4.11, the second term is controlled by $o(s)$, and

$$I(X;Y)=\frac{1}{2}\log(1+s\sigma^2)+o(s)=\frac{s\sigma^2}{2}+o(s).$$

Thus we finish the proof. □

Now we are prepared to prove the main result.

Theorem 4.13. *Let X be a random variable with finite variance, and let $W\sim\mathcal{N}(0,1)$ be a noise independent of X. Then*

$$\frac{d}{ds}I(X;\sqrt{s}X+W)=\frac{1}{2}\mathsf{mmse}(X\mid\sqrt{s}X+W).$$

Proof. Let $Y=\sqrt{s}X+W$. We compute the derivative of $I(X;Y)$. We write

$$I(s)=I(X;Y)=I(X;\sqrt{s}X+W)=I\left(X;X+\frac{1}{\sqrt{s}}W\right),\quad s>0.$$

Define

$$Z_1=X+\frac{1}{\sqrt{s+h}}W_1,\quad Z_2=Z_1+\sqrt{\frac{h}{s(s+h)}}W_2,$$

where W_1 and W_2 are independent $\mathcal{N}(0,1)$ variables that are also independent of X. Then $X\to Y_1\to Y_2$ is a Markov chain, and

$$I(s+h)-I(s)=I(X;Z_1)-I(X;Z_2)=I(X;Z_1,Z_2)-I(X;Z_2)=I(X;Z_1|Z_2).$$

We define

$$W := \sqrt{\frac{s}{s+h}}W_1 + \sqrt{\frac{h}{s+h}}W_2, \quad U := \sqrt{\frac{h}{s+h}}W_1 - \sqrt{\frac{s}{s+h}}W_2$$

Clearly, $U, W \sim \mathcal{N}(0,1)$ are two independent Gaussian variables. Since U is a function of W_1 and W_2, it is independent of X. Hence U is independent of $Z_2 = X + W/\sqrt{s}$. Moreover, we decompose Z_1 as

$$\begin{aligned} Z_1 &= \frac{s}{s+h}\left(Z_2 - \sqrt{\frac{h}{s(s+h)}}W_2\right) + \frac{h}{s+h}\left(X + \frac{1}{\sqrt{s+h}}W_1\right) \\ &= \frac{sZ_2}{s+h} + \frac{hX}{s+h} + \frac{\sqrt{h}}{s+h}U. \end{aligned}$$

We fix the event $\{Z_2 = z_2\}$, where $z_2 \in \mathbb{R}$. Under this event,

$$I(X; Z_1|Z_2 = z_2) = I\left(X; \frac{sZ_2}{s+h} + \frac{hX}{s+h} + \frac{\sqrt{h}}{s+h}U \middle| Z_2 = z_2\right) = I(X; \sqrt{h}X + U|Z_2 = z_2).$$

According to Lemma 4.12,

$$I(X; \sqrt{h}X + U|Z_2 = z_2) = \frac{h}{2}\operatorname{Var}(X|Z_2 = z_2) + o(h).$$

Note that $Y = \sqrt{s}X + W = \sqrt{s}Z_2$. Hence

$$I(X; Z_1|Z_2) = \frac{h}{2}\mathbb{E}[\operatorname{Var}(X|Z_2)] + o(h) = \frac{h}{2}\mathbb{E}[\operatorname{Var}(X|Y)] + o(h) = \frac{h}{2}\mathsf{mmse}(X|Y) + o(h),$$

and

$$\lim_{h\downarrow 0}\frac{I(s+h) - I(s)}{h} = \frac{1}{2}\mathsf{mmse}(X|Y).$$

The case $h \uparrow 0$ follows from a similar approach. Thus we finish the proof. □

Remark. We can also write this theorem to an integral form:

$$I(X; \sqrt{s}X + W) = \frac{1}{2}\int_0^s \mathsf{mmse}\left(X \mid \sqrt{\gamma}X + W\right) d\gamma.$$

Now we use this result to derive a new representation of differential entropy.

Lemma 4.14. *Under the assumption of Lemma 4.11, one have*

$$\lim_{s\to\infty} D(Y\|Y') = D(X\|X'),$$

where $X' \sim \mathcal{N}(\mu, \sigma^2)$ is a Gaussian variable with the same mean and variance as X.

Proof. Let $W_1, W_2 \sim \mathcal{N}(0,1)$ be independent Gaussian variables that are also independent of X. If $t_1 < t_2$, by data processing inequality for KL-divergence,

$$\begin{aligned} D(X + \sqrt{t_2}W \,\|\, X' + \sqrt{t_2}W) &= D(X + \sqrt{t_1}W_1 + \sqrt{t_2 - t_1}W_2 \,\|X' + \sqrt{t_1}W_1 + \sqrt{t_2 - t_1}W_2) \\ &\le D(X + \sqrt{t_1}W_1 \,\|X' + \sqrt{t_1}W_1) \\ &= D(X + \sqrt{t_1}W \,\|X' + \sqrt{t_1}W). \end{aligned}$$

By rescaling by $\sqrt{s}$, one have $D(Y\|Y') = D(X + W/\sqrt{s} \,\|\, X' + W/\sqrt{s})$, which is monotone increasing with respect to $s > 0$. Furthermore, it is bounded by $D(X\|X')$ from above. Hence

$$\lim_{s\to\infty} D(Y\|Y') \le D(X\|X').$$

On the other hand, by Fatou's lemma,

$$D(X\|X') \le \liminf_{s\to\infty} D\left(X + \frac{W}{\sqrt{s}} \,\middle\|\, X' + \frac{W}{\sqrt{s}}\right) = \lim_{s\to\infty} D(Y\|Y').$$

Thus we complete the proof. □

Theorem 4.15. *Let X be a random variable with finite variance $\sigma^2 > 0$, and let $W \sim \mathcal{N}(0,1)$ be a noise independent of X. Then*

$$h(X) = \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\int_0^\infty \left(\frac{\sigma^2}{1+\gamma\sigma^2} - \mathsf{mmse}(X \mid \sqrt{\gamma}X + W)\right) d\gamma.$$

Proof. Let X' be a Gaussian variable with the same mean and variance as X. Define $Y = \sqrt{s}X + W$ and $Y' = \sqrt{s}X' + W$. In the proof of Lemma 4.12, we obtained

$$I(X;Y) = \frac{1}{2}\log(1+s\sigma^2) - D(Y\|Y').$$

By Lemma 4.14 and Theorem 4.13,

$$\begin{aligned} D(X\|X') &= \lim_{s\to\infty} D(Y\|Y') \\ &= -\lim_{s\to\infty}\left(\frac{1}{2}\log(1+s\sigma^2) - I(X;Y)\right) \\ &= \frac{1}{2}\int_0^\infty \left(\frac{\sigma^2}{1+\gamma\sigma^2} - \mathsf{mmse}(X \mid \sqrt{\gamma}X + W)\right) d\gamma. \end{aligned}$$

Note that $h(X) = h(X') - D(X\|X')$, the result follows. □

Remark. This result can be extended to multi-dimensional vectors. Let X be a p-dimensional random vector with covariance matrix $\Sigma \succ 0$, and let $W \sim \mathcal{N}(0, \mathrm{Id}_p)$ be a noise independent of X. Then

$$h(X) = \frac{1}{2}\log\left((2\pi e)^p \det(\Sigma)\right) - \frac{1}{2}\int_0^\infty \left(\mathrm{tr}\left(\gamma\,\mathrm{Id} + \Sigma^{-1}\right)^{-1} - \mathsf{mmse}(X \mid \sqrt{\gamma}X + W)\right) d\gamma.$$

4.5 Entropy Power Inequality

Lemma 4.16. *Let X and Y be independent random variables with finite variance, and $\alpha \in [0, 2\pi)$. Then*

$$h(X\cos(\alpha) + Y\sin(\alpha)) \ge h(X)\cos^2(\alpha) + h(Y)\sin^2(\alpha).$$

Proof. Let $Z = X\cos(\alpha) + Y\sin(\alpha)$. According to Theorem 4.15,

$$\begin{aligned} &h(Z) - h(X)\cos^2(\alpha) - h(Y)\sin^2(\alpha) \\ &= \frac{1}{2}\int_0^\infty \left(\mathsf{mmse}(Z \mid \sqrt{\gamma}Z + W) - \mathsf{mmse}(X \mid \sqrt{\gamma}X + W)\cos^2(\alpha) - \mathsf{mmse}(Y \mid \sqrt{\gamma}Y + W)\sin^2(\alpha)\right) d\gamma \end{aligned} \tag{4.3}$$

Let $W_1, W_2 \sim \mathcal{N}(0,1)$ be independent Gaussian variables, and define

$$U = \sqrt{\gamma}X + W_1, \quad V = \sqrt{\gamma}Y + W_2.$$

Let $W = W_1\cos(\alpha) + W_2\sin(\alpha)$. Then $\sqrt{\gamma}Z + W = U\cos(\alpha) + V\sin(\alpha)$.

$$\mathsf{mmse}(Z \,|\, \sqrt{\gamma}Z + W) \geq \mathsf{mmse}(Z \,|\, U, V) = \mathsf{mmse}(X \,|\, U)\cos^2(\alpha) + \mathsf{mmse}(Y \,|\, V)\sin^2(\alpha).$$

Hence the integrand in (4.3) is nonnegative, and the result follows. □

Theorem 4.17. *Let X and Y be independent one-dimensional random variables such that $h(X)$, $h(Y)$ and $h(X+Y)$ exists. Then*

$$\mathrm{e}^{2h(X+Y)} \geq \mathrm{e}^{2h(X)} + \mathrm{e}^{2h(Y)}. \tag{4.4}$$

Proof. We choose $\alpha \in [0, \pi/2)$ such that

$$\tan(\alpha) = \mathrm{e}^{h(Y)-h(X)}.$$

We define $U = X/\cos(\alpha)$, and $V = Y/\sin(\alpha)$. By Lemma 4.16,

$$\begin{aligned} h(X+Y) = h(U\cos(\alpha) + V\sin(\alpha)) &\geq h(U)\cos^2(\alpha) + h(V)\sin^2(\alpha) \\ &= \cos^2(\alpha)\log\frac{\mathrm{e}^{h(X)}}{\cos(\alpha)} + \sin^2(\alpha)\log\frac{\mathrm{e}^{h(Y)}}{\sin(\alpha)} \\ &= \frac{1}{2}\log\left(\mathrm{e}^{2h(X)} + \mathrm{e}^{2h(Y)}\right). \end{aligned}$$

Then we complete the proof of (4.4). □

Remark. This conclusion can be generalized to multi-dimensional cases. Let X and Y be independent p-dimensional random vectors such that $h(X)$, $h(Y)$ and $h(X+Y)$ exists. Then

$$\mathrm{e}^{\frac{2}{p}h(X+Y)} \geq \mathrm{e}^{\frac{2}{p}h(X)} + \mathrm{e}^{\frac{2}{p}h(Y)}.$$

4.6 Entropic Central Limit Theorem

5 Rate Distortion Theory

A continuous source contains an infinite amount of information and cannot be represented exactly using a finite number of bits. In lossy source coding, we seek instead a representation that is close to the source (with respect to some fidelity criterion) and can be represented using a finite number of bits.

5.1 Quantization

Setting. Let X be a continuous random variable. For every value $x \in \mathcal{X}$, we would like to find a representation $\hat{x}(x)$ where $\hat{x}$ can take on only 2^R different values for a given rate R (measured in bits).

Example: Quantizing a Gaussian variable with squared error distortion. Let $X \sim \mathcal{N}(0, \sigma^2)$ be a Gaussian random variable. We wish to minimize the mean-squared error distortion $\mathbb{E}\left[(X - \hat{x}(X))^2\right]$.

If we use $R = 1$ bit to represent X (i.e., we can chose only $2^R = 2$ different reconstruction symbols), then we should use the bit to indicate whether X is positive or negative. To minimize the square error distortion, the reconstruction symbols should be the conditional mean given the sign:

$$\mathbb{E}[X|X > 0] = \int_0^\infty \frac{2x}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\, dx = \sqrt{\frac{2\sigma^2}{\pi}}.$$

Hence

$$\hat{x}(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\sigma, & x \geq 0, \\ -\sqrt{\frac{2}{\pi}}\sigma, & x < 0. \end{cases}$$

The average distortion is

$$\mathbb{E}\left[(X - \hat{x}(X))^2\right] = \left(1 - \frac{2}{\pi}\right)\sigma^2.$$

General quantization. A quantization scheme is characterized by a partition (V_i) of the metric space $(\mathcal{X}, d)$ and the corresponding reconstruction points $(\hat{x}_i) \subset \mathcal{X}$:

$$\hat{x}(x) = \hat{x}_i, \quad \textit{if } \ x \in V_i.$$

The regions and reconstruction points should satisfy:

- Given a set of reconstruction points (x_i), the regions should be chosen to minimize the distortion. This occurs if the regions are the Voronoi cells:

$$V_i = \{x \in \mathcal{X} : d(x, x_i) < d(x, x_j), \quad \forall j \neq i\}$$

- Given a set of regions, the reconstruction points should be chosen to minimize the distortion. Under squared error distortion, this is given by the conditional expectation:

$$\hat{x}_i = \mathbb{E}\left[X | X \in V_i\right].$$

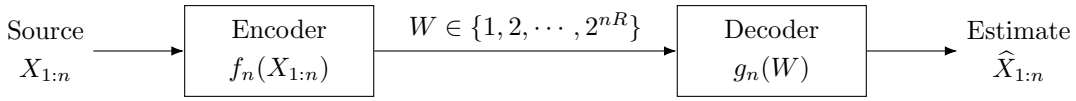
Lloyd's Algorithm is an iterative algorithm for constructing a quantization function. Starting with an initial set of reconstruction points, the algorithm repeats the following two steps:

- Given reconstruction points (x_i), find the optimal partition (V_i);
- Given a partition (V_i), find optimal set of reconstruction points (x_i).

This algorithm will converge to a local optimum (but not necessarily the global optimum).

5.2 Lossy Source Coding

Vector quantization. Let $X_{1:n}$ be an length-n random vector with i.i.d. entries. For every realization $X_{1:n} = x_{1:n}$, we would like to find a representation $\hat{x}^n(x_{1:n})$ where $\hat{x}^n$ can take on only 2^{nR} different values for a given rate R. One option is to use the rate R scalar quantization strategy we discussed in the previous section. However, it turns out that quantizing jointly can be much better than quantizing separately.



- The source produces a sequence $X_1, X_2, \cdots$ of i.i.d. random variable with distribution $p(x)$ supported on a possibly infinite alphabet $\mathcal{X}$.
- The encoder is a mapping $f_n : \mathcal{X}^n \to \{1, 2, \cdots, 2^{nR}\}$ that describes every source sequence by an index w. The rate is given by

$$R = \frac{1}{n} \log_2 |\mathcal{W}| .$$

- The decoder $g_n : \{1, 2, \cdots, 2^{nR}\} \to \widehat{\mathcal{X}}^n$ maps each index w to an estimate $\hat{x}^n \in \widehat{\mathcal{X}}$, where $\widehat{\mathcal{X}}$ is a finite reconstruction alphabet.

Definition 5.1 (Rate measure)**.** A *per-letter distortion measure* is a mapping $\mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}_+$ from the set of source alphabet-reconstruction pairs into the nonnegative real numbers. The distortion measure is *bounded* if the maximum value of the distortion is finite:

$$\sup_{x \in \in \mathcal{X}, \widehat{x} \in \widehat{\mathcal{X}}} d(x, \hat{x}) < \infty.$$

The distortion between two sequences $x_{1:n}$ and $\widehat{x}_{1:n}$ is given by the average per-letter distortion:

$$d(x_{1:n}, \widehat{x}_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \widehat{x}_i).$$

Example 5.2. *Here are two examples of distortion.*

(i) (Hamming distortion). $d(x, \widehat{x}) = \mathbb{1}_{\{x \neq \widehat{x}\}}$. *This is often used for discrete alphabets.*

(ii) (Square-error distortion). $d(x, \widehat{x}) = |x - \widehat{x}|^2$. *This is one of the most popular distortion measures used for continuous alphabets.*

Definition 5.3. A $(2^{nR}, n)$ rate distortion coding scheme consists of

- a source alphabet $\mathcal{X}$ and a reconstruction alphabet $\widehat{\mathcal{X}}$,
- a encoder $f_n : \mathcal{X}^n \to \{1, 2, \cdots, 2^{nR}\}$ and a decoder $g_n : \{1, 2, \cdots, 2^{nR}\} \to \widehat{\mathcal{X}}^n$, and
- a distortion measure $d : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}_+$.

The (expected) distortion associated with this coding scheme is defined as

$$D = \mathbb{E}\left[d(X_{1:n}, \widehat{X}_{1:n})\right] = \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n})\, d(x_{1:n}, g_n(f_n(x_{1:n})))$$

The collection of n-tuples $g_n(1), g_n(2), \cdots, g_n(2^{nR})$, denoted by $\widehat{X}_{1:n}(1), \widehat{X}_{1:n}(2), \cdots, \widehat{X}_n(2^{nR})$, constitutes the *codebook*, and $f_n^-(1), f_n^{-1}(2), \cdots, f_n^{-1}(2^{nR})$ are the associated *assignment regions.*

Remark. $\widehat{X}_{1:n}$ is referred to as the *vector quantization, reconstruction, or estimate* of $X_{1:n}$.

Definition 5.4. We call (R, D) a *rate-distortion pair*.

(i) A rate-distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$ rate distortion coding schemes (f_n, g_n) with

$$\limsup_{n\to\infty} \mathbb{E}\left[d(X_{1:n}, \widehat{X}_{1:n})\right] \leq D.$$

(ii) The *rate distortion region* for a source is the closure of the set of achievable rate distortion pairs (R, D).

(iii) The *rate distortion function* $R(D)$ is the infimum of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.

(iv) The *distortion rate function* $D(R)$ is the infimum of all distortions D such that (R, D) is in the rate distortion region of the source for a given rate R.

Remark. If the distortion rate is $D = 0$, the coding scheme is accurate. According to Shannon's source coding theorem, we require $R = H(X)$. This is not feasible for continuous variable X.

5.3 Information Rate Distortion Function

Definition 5.5. Let X be a source from a distribution $p(x)$ on $\mathcal{X}$. The *information rate distortion function* $R^{(I)}(D)$ for a source X with distortion measure d is defined as

$$R^{(I)}(D) = \inf_{p(\widehat{x}|x):\, \mathbb{E}[d(X,\widehat{X})]\leq D} I(X;\widehat{X}) = \inf_{p(\widehat{x}|x):\, \mathbb{E}[d(X,\widehat{X})]\leq D} I(p(x), p(\widehat{x}|x)).$$

Here the infimum is taken over all conditional distributions of $\widehat{X}$ given X such that the expected distortion constraint $\mathbb{E}[d(X, \widehat{X})] \leq D$ is satisfied.

We first introduce an important property of rate distortion function $R(D)$, then calculate the information rate distortion function for some sources.

Theorem 5.6. *The information rate distortion function $R^{(I)}(D)$ is a non-increasing convex function of D.*

Proof. When the distortion D increases, the set of feasible conditional distributions $p(\widehat{x}|x)$ is also increasing. Since $R^{(I)}(D)$ is the infimum taken over this set, it is non-increasing.

To show the convexity of $R^{(I)}$, take $D_1, D_2 > 0$ and $\epsilon > 0$. Let $\widehat{X}_1|X \sim p_1(\widehat{x}|x)$ and $\widehat{X}_2|X \sim p_2(\widehat{x}|x)$ be the conditional distributions such that

$$I(X;\widehat{X}_1) \leq R^{(I)}(D_1) + \epsilon, \quad I(X;\widehat{X}_2) \leq R^{(I)}(D_2) + \epsilon.$$

For any $\lambda \in [0, 1]$, consider the distribution $p_\lambda(\widehat{x}|x) = \lambda p_1(\widehat{x}|x) + (1-\lambda)p_2(\widehat{x}|x)$. The distortion associated to the distribution $p_\lambda(x, \widehat{x}) = p_\lambda(\widehat{x}|x)p(x)$, by linearity of expectation, satisfies

$$\mathbb{E}\left[d(X, \widehat{X})\right] \leq D_\lambda = \lambda D_1 + (1-\lambda)D_2.$$

By convexity of mutual information,

$$I(p(x), p_\lambda(\widehat{x}|x)) \leq \lambda I(p(x), p_1(\widehat{x}|x)) + (1-\lambda)I(p(x), p_2(\widehat{x}|x)) \leq \lambda R^{(I)}(D_1) + (1-\lambda)R^{(I)}(D_2) + \epsilon.$$

Since $R^{(I)}(D_\lambda)$ is no greater than the last display, and $\epsilon > 0$ is arbitrarily taken, we have

$$R^{(I)}(D_\lambda) \leq \lambda R^{(I)}(D_1) + (1-\lambda)R^{(I)}(D_2)$$

Thus we complete the proof. □

Example 5.7 (Binary source with Hamming distortion)**.** *The information rate distortion function for an i.i.d.* $X \sim \text{Bernoulli}(p)$ *source with Hamming distortion is given by*

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

Proof. Let $\widehat{\mathcal{X}} = \{0, 1\}$, and let $p(\widehat{x}|x)$ be any distribution satisfying the expected distortion constraint, That is, $\mathbb{P}(\widehat{X} \neq X) \le D$. Consider $B = \mathbb{1}_{\{\widehat{X} \neq X\}} = X + \widehat{X} \mod 2$. Then

$$I(X; \widehat{X}) = H(X) - H(X|\widehat{X}) = H(X) - H(B|\widehat{X}) \ge H(X) - H(B).$$

When $D \le 1/2$, we have $H(B) \le H(D)$, and $I(X; \widehat{X}) \ge H(p) - H(D)$; otherwise, we have $D > \min\{p, 1-p\}$, and by definition $I(X; \widehat{X}) \ge 0$. Hence

$$R^{(I)}(D) \ge \begin{cases} H(p) - H(D), & 0 \le D \le \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\}. \end{cases}$$

It remains to show the opposite inequality. If $p < \frac{1}{2}$ and $D \ge p$, we let $\widehat{X} = 0$ with probability 1. Then $\mathbb{P}(\widehat{X} \neq X) = p \le D$, and $I(X; \widehat{X}) = 0$. A similar conclusion applies for $p > \frac{1}{2}$ and $D \ge 1 - p$. Hence

$$R^{(I)}(D) = 0, \quad D > \min\{p, 1-p\}.$$

Now we show the case $0 \le D < \min\{p, 1-p\}$. Without loss of generality assume $0 \le D < p \le \frac{1}{2}$. We consider the joint distribution

$$\begin{aligned} \mathbb{P}(X = 0, \widehat{X} = 0) &= \frac{(1-D)(1-p-D)}{1-2D}, \quad \mathbb{P}(X = 0, \widehat{X} = 1) = \frac{D(p-D)}{1-2D}, \\ \mathbb{P}(X = 1, \widehat{X} = 0) &= \frac{D(1-p-D)}{1-2D}, \quad \mathbb{P}(X = 1, \widehat{X} = 1) = \frac{(1-D)(p-D)}{1-2D}. \end{aligned}$$

This distribution satisfies the expected distortion constraint $\mathbb{P}(\widehat{X} \neq X) = D$, and

$$R^{(I)}(D) \le I(X; \widehat{X}) = H(X) - H(X|\widehat{X}) = H(p) - H(D).$$

Thus we complete the proof. □

Example 5.8 (Gaussian source with square-error distortion)**.** *The information rate distortion function for an i.i.d.* $X \sim \mathcal{N}(0, \sigma^2)$ *source with square-error distortion is given by*

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

Proof. Let $(X, \widehat{X})$ be distributed such that $X \sim \mathcal{N}(0, \sigma^2)$ and $\mathbb{E}\left[(X - \widehat{X})^2\right] \le D$. Then

$$I(X; \widehat{X}) = h(X) - h(X|\widehat{X}) = \log(2\pi \mathrm{e}\sigma^2) - h(X - \widehat{X}|\widehat{X}) \ge \log(2\pi \mathrm{e}\sigma^2) - h(X - \widehat{X}).$$

By Theorem 4.6,

$$h(X - \widehat{X}) \le \log\left(2\pi \mathrm{e}\mathbb{E}\left[(X - \widehat{X})^2\right]\right) \le \log(2\pi \mathrm{e}D).$$

Hence

$$R^{(I)}(D) \geq \inf_{p(\widehat{x}|x):\, \mathbb{E}[(X-\widehat{X})^2] \leq D} I(X;\widehat{X}) \geq \max\left\{\log \frac{\sigma^2}{D}, 0\right\}.$$

Now we prove the other side. If $D \geq \sigma^2$, we can simply set $\widehat{X} = 0$, which satisfies the expected distortion constraint $\mathbb{E}[(X-\widehat{X})^2] \leq D$. If $D < \sigma^2$, we choose distribution given by the Gaussian kernel

$$X = \widehat{X} + Z, \quad \widehat{X} \sim \mathcal{N}(0, \sigma^2 - D) \quad \textit{and} \quad Z \sim \mathcal{N}(0, D).$$

which also satisfies the expected distortion constraint $\mathbb{E}[(X-\widehat{X})^2] = D$ Then

$$R^{(I)}(D) \geq I(X;\widehat{X}) = h(X) - h(X|\widehat{X}) = \log \frac{\sigma^2}{D}.$$

Thus we complete the proof. □

5.4 Rate Distortion Theorem

The main theorem of rate distortion theory can now be stated as follows:

Theorem 5.9. *The rate distortion function for an i.i.d. source $X \sim p$ and a bounded distortion measure $d : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}_+$ is equal to the associated information rate distortion function, i.e.*

$$R(D) = R^{(I)}(D).$$

This theorem includes two parts:

- *(Achievability). If $R > R^{(I)}(D)$, the rate-distortion pair (R, D) is achievable.*
- *(Converse). If the rate-distortion pair (R, D) is achievable, then $R \geq R^{(I)}(D)$.*

Proof of Theorem 5.9 (Converse). For any sequence of $(2^{nR}, n)$ coding schemes such that

$$\lim_{n\to\infty} \mathbb{E}[d(X_{1:n}; \widehat{X}_{1:n})] \leq D,$$

we want to show that $R \geq R^{(I)}(D)$. We take $\epsilon > 0$, then there exists N such that $\mathbb{E}[d(X_{1:n}; \widehat{X}_{1:n})] \leq D + \epsilon$ for all $n \geq N$. Since there are 2^{nR} values in the range of f_n,

$$nR \geq H(f_n(X_{1:n})) = H(f_n(X_{1:n})) - H(f_n(X_{1:n})|X_{1:n}) = I(X_{1:n}; f(X_{1:n})) \tag{5.1}$$

By data processing inequality,

$$\begin{aligned}
I(X_{1:n}; f(X_{1:n})) &\geq I(X_{1:n}; \widehat{X}_{1:n}) = H(X_{1:n}) - H(X_{1:n}|\widehat{X}_{1:n}) \\
&= \sum_{i=1}^{n} H(X_i) - H(X_{1:n}|\widehat{X}_{1:n}) && \text{(By independence of } X_{1:n}) \\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1, \widehat{X}_{1:n}) && \text{(By the chain rule)} \\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\widehat{X}_i) && \text{(Conditioning does not increase entropy)} \\
&= \sum_{i=1}^{n} I(X_i; \widehat{X}_i). && (5.2)
\end{aligned}$$

By definition of the information distortion function $R^{(I)}(D)$,

$$I(X_i; \widehat{X}_i) \geq R^{(I)}\left(\mathbb{E}\left[d(X_i; \widehat{X}_i)\right]\right).$$

By convexity of $R^{(I)}(D)$ and Jensen's inequality, we get

$$\sum_{i=1}^{n} I(X_i; \widehat{X}_i) \geq \sum_{i=1}^{n} R^{(I)}\left(\mathbb{E}\left[d(X_i; \widehat{X}_i)\right]\right) \geq nR^{(I)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d(X_i; \widehat{X}_i)\right]\right) = nR^{(I)}\left(\mathbb{E}\left[d(X_{1:n}; \widehat{X}_{1:n})\right]\right). \tag{5.3}$$

Combining (5.1), (5.2) and (5.3), we obtain

$$R \geq R^{(I)}\left(\mathbb{E}\left[d(X_{1:n}; \widehat{X}_{1:n})\right]\right) \geq R^{(I)}(D+\epsilon).$$

This inequality holds for all $\epsilon > 0$. Since the function $R^{(I)}(D)$ is convex, it is continuous, and

$$R \geq \lim_{\epsilon \downarrow 0} R^{(I)}(D+\epsilon) = R^{(I)}(D).$$

Thus we complete the proof. □

Definition 5.10 (Distortion ϵ-typical set)**.**

Proof of Theorem 5.9 (Achievability). □

6 f-Divergences

6.1 Definition and Properties

Definition 6.1 (f-divergence)**.** Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \mathscr{F})$. For any convex function $f : [0, \infty) \to (-\infty, \infty]$ such that (i) $f(1) = 0$, (ii) f is strictly convex at 1,[1] and (iii) f is finite except possibly at 0, the f-divergence of Q with respect to P is defined as follows:

(i) If $Q \ll P$,

$$D_f(Q\|P) := \int f\left(\frac{dQ}{dP}\right) dP,$$

where the notation $\frac{dQ}{dP}$ stands for the Radon-Nikodym derivative of Q with respect to P.

(ii) More generally, let μ be any dominating measure of P and Q, i.e. $P \ll \mu$ and $Q \ll \mu$. Assume $dP = p\, d\mu$ and $dQ = q\, d\mu$. Then

$$D_f(Q\|P) := \int_{\{p>0\}} f\left(\frac{q}{p}\right) p\, d\mu + f'(\infty)\, Q\{p = 0\}, \tag{6.1}$$

where $f'(\infty) = \lim_{x\to 0^+} xf(\frac{1}{x})$.

Remark. In fact, the definition (6.1) comes from the following division:

$$D_f(Q\|P) = \int_{\{p>0\}} f\left(\frac{q}{p}\right) dP + \int_{\{p=0\}} \frac{p}{q} f\left(\frac{q}{p}\right) dQ = \int_{\{p>0\}} f\left(\frac{q}{p}\right) p\, d\mu + \lim_{x\to 0^+} xf\left(\frac{1}{x}\right) Q\{p = 0\}.$$

In practice, we often use the following two forms of f-divergence:

- When $\mathcal{X}$ is discrete, P and Q are probability mass functions:

$$D_f(Q\|P) = \sum_{x\in\mathcal{X}} f\left(\frac{Q(x)}{P(x)}\right) P(x) = \mathbb{E}_P\left[f\left(\frac{Q}{P}\right)\right].$$

- When P and Q are characterized by density functions p and q (i.e. their Radon-Nikodym derivatives with respect to the Lebesgue measure), respectively, then

$$D_f(q\|p) = \int f\left(\frac{q(x)}{p(x)}\right) p(x)\, dx = \mathbb{E}_{X\sim p}\left[f\left(\frac{p(X)}{q(X)}\right)\right].$$

We use the convention that

- $f(0) = f(0^+)$,
- $0f\left(\frac{0}{0}\right) = 0$, and
- $0f\left(\frac{a}{0}\right) = \lim_{x\to 0^+} xf\left(\frac{a}{x}\right) = af'(\infty)$.

Furthermore, by definition, if $P \perp Q$,

$$D_f(Q\|P) = f(0) + f'(\infty) = \lim_{x\to 0^+}\left[(1-x)f\left(\frac{x}{1-x}\right) + xf\left(\frac{1}{x}\right)\right] > \lim_{x\to 0^+} f(x+1) = f(1) = 0.$$

An f-divergence provide an evaluation of the difference between two probability distributions.

[1] By strict convexity at 1, we mean that for all $x, y \in (0, \infty)$ and $0 < \lambda < 1$ such that $\lambda x + (1-\lambda)y = 1$, we have

$$\lambda f(x) + (1-\lambda) f(y) > f(1).$$

For a random variable X with $\mathbb{E}[X] = 1$, the Jensen's inequality $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ is strict if X is not a constant.

Proposition 6.2 (Positive-definiteness of f-divergence)**.** *Let D_f be an f-divergence. Then*

$$D_f(Q\|P) \geq 0,$$

and the equality holds if and only if $P = Q$.

Proof. From Jensen's inequality, the convexity of f implies:

$$D_f(Q\|P) = \mathbb{E}_P\left[f\left(\frac{Q}{P}\right)\right] \geq f\left(\mathbb{E}_P\left[\frac{Q}{P}\right]\right) = f(1) = 0.$$

By the strict convexity of f at 1, the equality holds if and only if $P = Q$. □

Definition 6.3 (Examples of f-divergence)**.** The following are some commonly used f-divergences:

(i) Total variation distance. $f(x) = \frac{1}{2}|x-1|$:

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\mathbb{E}_P\left[\left|\frac{Q}{P} - 1\right|\right] = \frac{1}{2}\int |dQ - dP|.$$

Clearly, we have $d_{\mathrm{TV}}(P,Q) = d_{\mathrm{TV}}(Q,P)$. Furthermore, the triangle inequality follows from definition:

$$d_{\mathrm{TV}}(P,Q) \leq d_{\mathrm{TV}}(P,R) + d_{\mathrm{TV}}(R,Q).$$

Therefore, the total variation distance is a metric on the space of all probability measures on $(\mathcal{X}, \Sigma)$.

(ii) Kullback-Leibler divergence. $f(x) = x\log x$:

$$D(Q\|P) = \mathbb{E}_P\left[\frac{Q}{P}\log\left(\frac{Q}{P}\right)\right] = \mathbb{E}_Q\left[\log\left(\frac{Q}{P}\right)\right].$$

(iii) Pearson χ^2-divergence. $f(x) = x^2 - 1$:

$$\chi^2(Q\|P) = \int \frac{Q^2}{P} - 1 = \int \frac{(P-Q)^2}{P}.$$

(iv) Squared Hellinger distance. $f(x) = \frac{1}{2}(1-\sqrt{x})^2$:

$$H^2(P,Q) = \frac{1}{2}\mathbb{E}_P\left[\left(1 - \sqrt{\frac{Q}{P}}\right)^2\right] = \frac{1}{2}\int \left(\sqrt{P} - \sqrt{Q}\right)^2.$$

Clearly, we have $H^2(P,Q) = H^2(Q,P)$. We further define the Hellinger distance $H(P,Q) = \sqrt{H^2(P,Q)}$. Then the triangle inequality $H(P,Q) \leq H(P,R) + H(R,Q)$ follows from the case for L^2-norm. Therefore, the Hellinger distance $H(\cdot,\cdot)$ is a metric on the space of all probability measures on $(\mathcal{X}, \Sigma)$.

(v) Jensen-Shannon divergence. $f(x) = \frac{x}{2}\log x - \frac{1+x}{2}\log\left(\frac{1+x}{2}\right)$:

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M),$$

where $M = \frac{1}{2}P + \frac{1}{2}Q$. This is also known as the symmetrized Kullback-Leibler divergence.

(vi) Le Cam distance. $f(x) = \frac{(1-x)^2}{2(1+x)}$:

$$\mathrm{Le}(P,Q) = \frac{1}{2}\int \frac{(P-Q)^2}{P+Q}.$$

Now we discuss more properties of f-divergences.

Proposition 6.4 (Properties of f-divergences)**.** *Let D_f be an f-divergence.*

(i) (Monotonicity). Let $P_{X,Y}$ and $Q_{X,Y}$ be two joint distributions of random variables X and Y. Then

$$\max\{D_f(Q_X\|P_X), D_f(Q_Y\|P_Y)\} \le D_f(Q_{XY}\|P_{XY}).$$

(ii) (Data processing inequality). Fix the conditional distribution $P_{Y|X}$ of Y given X. Let $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$. Then

$$D_f(Q_Y\|P_Y) \le D_f(Q_X\|P_X).$$

(iii) (Joint convexity). The mapping $(Q,P) \mapsto D_f(Q\|P)$ is jointly convex. That is, for any distributions P_1, P_2, Q_1, Q_2 and any $0 \le \lambda \le 1$,

$$D_f\left(\lambda Q_1 + (1-\lambda)Q_2 \| \lambda P_1 + (1-\lambda)P_2\right) \le \lambda D_f(Q_1\|P_1) + (1-\lambda) D_f(Q_2\|P_2).$$

(iv) (Conditional increment). Given two conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ and a marginal distribution P_X, define the conditional f-divergence:

$$D_f(Q_{Y|X}\|P_{Y|X}|P_X) := \int_{\mathcal{X}} D_f(Q_{Y|X=x}\|P_{Y|X=x})\, dP(x) = \mathbb{E}_{X\sim P_X}\left[D_f(Q_{Y|X}\|P_{Y|X})\right].$$

Let $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$. Then

$$D_f(Q_Y\|P_Y) \le D_f(Q_{Y|X}\|P_{Y|X}|P_X).$$

Proof. (i) Using Jensen's inequality:

$$\begin{aligned}
D_f(Q_{X,Y}\|P_{X,Y}) &= \mathbb{E}_{P_{X,Y}}\left[f\left(\frac{Q_{X,Y}}{P_{X,Y}}\right)\right] = \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[f\left(\frac{Q_{X,Y}}{P_{X,Y}}\right)\right]\right] \\
&\ge \mathbb{E}_{P_X}\left[f\left(\mathbb{E}_{P_{Y|X}}\left[\frac{Q_X}{P_X}\frac{Q_{Y|X}}{P_{Y|X}}\right]\right)\right] \\
&= \mathbb{E}_{P_X}\left[f\left(\frac{Q_X}{P_X}\mathbb{E}_{P_{Y|X}}\left[\frac{Q_{Y|X}}{P_{Y|X}}\right]\right)\right] \\
&= \mathbb{E}_{P_X}\left[f\left(\frac{Q_X}{P_X}\right)\right] \\
&= D_f(Q_X\|P_X).
\end{aligned}$$

Switching X and Y yields $D_f(Q_{X,Y}\|P_{X,Y}) \ge D_f(Q_Y\|P_Y)$.

(ii) Following that (i), it suffices to show $D_f(Q_X\|P_X) = D_f(Q_{X,Y}\|P_{X,Y})$. This is true since the conditional distribution $P_{Y|X}$ is fixed:

$$\begin{aligned}
D_f(Q_{X,Y}\|P_{X,Y}) &= \mathbb{E}_{P_{X,Y}}\left[f\left(\frac{Q_{X,Y}}{P_{X,Y}}\right)\right] = \mathbb{E}_{P_{X,Y}}\left[f\left(\frac{Q_X}{P_X}\right)\right] \\
&= \mathbb{E}_{P_X}\left[f\left(\frac{Q_X}{P_X}\right)\right] = D_f(Q_X\|P_X).
\end{aligned}$$

(iii) Fix $\lambda \in [0,1]$, and let $B \sim \text{Bernoulli}(\lambda)$. We set $P_{X|B=1} = P_1$, $P_{X|B=0} = P_2$, and $Q_{X|B_1} = Q_1$ and $Q_{X|B=0} = Q_2$. Since the distribution P_X of X is fixed,

$$\frac{Q_{X,B}}{P_{X,B}} = \frac{Q_B}{P_B}\frac{Q_{X|B}}{P_{X|B}} = \frac{Q_{X|B}}{P_{X|B}}.$$

Then

$$D_f(Q_{X,B}\|P_{X,B}) = \mathbb{E}_{P_B}\left[\mathbb{E}_{P_{X|B}}\left[f\left(\frac{Q_{X|B}}{P_{X|B}}\right)\right]\right] = \lambda D_f(Q_1\|P_1) + (1-\lambda)D_f(Q_2\|P_2).$$

On the other hand, the monotonicity implies

$$D_f(Q_{X,B}\|P_{X,B}) \geq D_f(Q_X\|P_X) = D_f\left(\lambda Q_1 + (1-\lambda)Q_2\|\lambda P_1 + (1-\lambda)P_2\right).$$

Combining the last two displays gives the wanted result.

(iv) By calculus, the marginal distributions of Y are given by

$$Q_Y = \int_{\mathcal{X}} Q_{Y|X=x}\, dP_X(x) = \mathbb{E}_{X\sim P_X}[Q_{Y|X}], \quad P_Y = \int_{\mathcal{X}} P_{Y|X=x}\, dP_X(x) = \mathbb{E}_{X\sim P_X}[P_{Y|X}].$$

Then by joint convexity of $D_f(\cdot\|\cdot)$ and Jensen's inequality,

$$D_f(Q_{Y|X}\|P_{Y|X}|P_X) = \mathbb{E}_{X\sim P_X}\left[D_f(Q_{Y|X}\|P_{Y|X})\right] \geq D_f\left(\mathbb{E}_{X\sim P_X}[Q_{Y|X}]\,\|\,\mathbb{E}_{X\sim P_X}[P_{Y|X}]\right) = D_f(Q_Y\|P_Y).$$

Thus we complete the proof. □

The data processing inequality for f-divergence has many applications. Here are some basic examples.

Proposition 6.5. *Let P and Q be two probability measures on $(\mathcal{X}, \mathscr{F})$, and let $A \in \mathscr{F}$.*
(i) $|P(A) - Q(A)| \leq d_{\mathrm{TV}}(P, Q)$. *(ii)* $|\sqrt{P(A)} - \sqrt{Q(A)}| \leq \sqrt{2}H(P, Q)$.
(iii) $|P(A) - Q(A)|^2 \leq \chi^2(Q\|P)P(A)(1 - P(A))$.
(iv) $Q(A)\log\frac{1}{P(A)} \leq D(Q\|P) + \log 2$.

Proof. We fix $X \sim P_X = P$ or $X \sim Q_X = Q$, and define $Y = \mathbb{1}_{\{X\in A\}}$. Then $P_Y = \mathrm{Bernoulli}(P(A))$, and $Q_Y = \mathrm{Bernoulli}(Q(A))$. Use the data processing inequality:

$$D_f(Q_Y\|P_Y) \leq D_f(Q_X\|P_X) = D_f(Q\|P).$$

(i) Since $d_{\mathrm{TV}}(P_Y, Q_Y) = \frac{1}{2}|Q(A) - P(A)| + \frac{1}{2}|1 - Q(A) - (1 - P(A))| = |Q(A) - P(A)|$, we have

$$|P(A) - Q(A)| \leq d_{\mathrm{TV}}(P, Q).$$

(ii) By definition,

$$H^2(P_Y, Q_Y) = \frac{1}{2}\left(\sqrt{P(A)} - \sqrt{Q(A)}\right)^2 + \frac{1}{2}\left(\sqrt{1 - P(A)} - \sqrt{1 - Q(A)}\right)^2.$$

Hence

$$\frac{1}{2}\left(\sqrt{P(A)} - \sqrt{Q(A)}\right)^2 \leq H^2(P_Y, Q_Y) \leq H^2(P, Q).$$

(iii) By definition,

$$\chi^2(Q_Y\|P_Y) = \frac{|Q(A) - P(A)|^2}{P(A)} + \frac{|1 - Q(A) - (1 - P(A))|^2}{1 - P(A)} = \frac{|Q(A) - P(A)|^2}{P(A)(1 - P(A))}$$

Hence

$$|Q(A) - P(A)| \leq \chi^2(Q\|P)P(A)(1 - P(A)).$$

(iv) By definition,

$$\begin{aligned} D(Q_Y \| P_Y) &= Q(A) \log \frac{Q(A)}{P(A)} + (1 - Q(A)) \log \frac{1 - Q(A)}{1 - P(A)} \\ &\geq Q(A) \log \frac{1}{P(A)} + (1 - Q(A)) \log \frac{1}{1 - P(A)} - \log 2. \end{aligned}$$

Hence

$$Q(A) \log \frac{1}{P(A)} \leq D(Q\|P) + \log 2.$$

Thus we conclude the proof.

□

In fact, the first inequality in the above proposition can become equality.

Proposition 6.6. *Let P and Q be two probability measures on $(\mathcal{X}, \mathscr{F})$. Then*

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathscr{F}} P(A) - Q(A).$$

Proof. We consider the signed measure $\mu = P - Q$. By Hahn decomposition theorem, there exists a partition $\mathcal{X} = \mathcal{P} \cap \mathcal{N}$ such that

(i) $\mathcal{P}, \mathcal{N} \in \mathscr{F}$ and $\mathcal{P} \cap \mathcal{N} = \emptyset$,

(ii) $P(A \cap \mathcal{P}) - Q(A \cap \mathcal{P}) \geq 0$ for all $A \in \mathscr{F}$, and

(iii) $P(A \cap \mathcal{P}) - Q(A \cap \mathcal{P}) \leq 0$ for all $A \in \mathscr{F}$.

We take $A = \mathcal{P}$. Then

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= \frac{1}{2} \int_{\mathcal{P}} [dP - dQ] + \frac{1}{2} \int_{\mathcal{N}} [dQ - dP] \\ &= \int_{\mathcal{P}} [dP - dQ] = P(A) - Q(A). \end{aligned}$$

The conclusion then follows from Proposition 6.5(i). □

6.2 Variational Representation

Definition 6.7 (Fenchel conjugate)**.** Let $(\mathcal{X}, \langle \cdot, \cdot \rangle)$ be a real Hilbert space, and let $f : \mathcal{X} \to (-\infty, +\infty]$ be a proper function, that is, $\mathrm{dom}(f) := \{x \in \mathcal{X} : f(x) \in \mathbb{R}\} \neq \emptyset$. The Fenchel conjugate of f is defined as

$$f^*(t) = \sup_{x \in \mathcal{X}} \{\langle x, t \rangle - f(x)\}, \ t \in \mathcal{X}. \tag{6.2}$$

Remark. It can be seen that f^* is the pointwise supremum of a collection of affine functions, hence f^* is convex, regardless of f is convex or not. Moreover, it can be shown that the duality $(f^*)^* = f$ holds if f is convex and lower semicontinuous. Below is an immediate consequence of this definition.

Proposition 6.8 (Fenchel-Young inequality)**.** *For all $x, t \in \mathcal{X}$,*

$$f(x) + f^*(t) \geq \langle x, t \rangle.$$

Remark. Recall that in Definition 6.1, f is defined on $[0, +\infty)$. We complete f by redefining $f(x) = \infty$ for $x < 0$, which preserves the convexity of f. Moreover, the Fenchel conjugate of $f : \mathbb{R} \to (-\infty, +\infty]$ is well defined: $f^*(t) = \sup_{x \in \mathbb{R}} \{tx - f(x)\}, \ t \in \mathbb{R}$. The f-divergence admits the following variational representation.

Lemma 6.9 (Variational representation of f-divergence)**.** *Denote $\mathcal{M}$ by the class of measurable functions on $(\mathcal{X}, \Sigma)$. Then the f-divergence can be represented as*

$$D_f(Q\|P) = \sup_{g\in\mathcal{M}} \left\{ \int g\, dQ - \int (f^* \circ g)\, dP \right\} = \sup_{g\in\mathcal{M}} \{\mathbb{E}_Q[g(X)] - \mathbb{E}_P[f^*(g(X))]\}. \tag{6.3}$$

Where f^* is the Fenchel conjugate of f. If f is differentiable, the supremum is reached at $g = f'(\frac{dQ}{dP})$.

Proof. We fix the measurable function $g \in \mathcal{M}$. By Fenchel's duality, we have

$$g(x)\frac{Q(x)}{P(x)} - f\left(\frac{Q(x)}{P(x)}\right) \le f^*(g(x)).$$

Take integration with respect to P on both sides of the equation above, we have

$$\int g(x)\, dQ(x) - D_f(Q\|P) \le \int f^*(g(x))\, dP(x).$$

Since g is arbitrarily chosen, we immediately conclude the equality (6.3). The supremum can be found when the derivative of (6.2) vanishes. □

Proposition 6.10. *We provide the variational form of f-divergences in Definition 6.3.*

- *Total variation distance.* $f^*(t) = \begin{cases} t, & |t| \le 1/2 \\ \infty, & |t| > 1/2 \end{cases}$:

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sup_{\|g\|_\infty \le 1} \left(\int g\, dP - \int g\, dQ \right)$$

- *Kullback-Leibler divergence.* $f^*(t) = \mathrm{e}^{t-1}$:

$$D(Q\|P) = \sup_{g\in\mathcal{M}} \left\{ \int g(x)\, dQ(x) - \int \mathrm{e}^{g(x)-1}\, dP(x) \right\}.$$

- *Squared Hellinger distance.* $f^*(t) = \begin{cases} \frac{t}{1-2t}, & t < 1/2 \\ \infty, & t \ge 1/2 \end{cases}$:

$$H^2(P, Q) = \inf_{g<\frac{1}{2}} \left\{ \int g\, dQ - \int \frac{g}{1-2g}\, dP \right\} = \inf_{h<1} \frac{1}{2} \left(\int h\, dQ - \int \frac{h}{1-h}\, dP \right).$$

- *Jensen-Shannon divergence.* $f^*(t) = \begin{cases} -\frac{1}{2}\log(2 - \mathrm{e}^{2t}), & t < \frac{1}{2}\log 2 \\ \infty, & t \ge \frac{1}{2}\log 2. \end{cases}$:

$$\begin{aligned} D_{\mathrm{JS}}(P, Q) &= \sup_{g \le \frac{1}{2}\log 2} \left\{ \int g\, dQ + \frac{1}{2} \int \log(2 - \mathrm{e}^{2g})\, dP \right\} \\ &= \frac{1}{2} \sup_{\|h\|_\infty < 1} \left\{ \int \log(1+h)\, dQ + \int \log(1-h)\, dP \right\}. \qquad (h = \mathrm{e}^{2g} - 1) \end{aligned}$$

- *Pearson χ^2-divergence.* $f^*(t) = \frac{1}{4}t^2 + 1$*:*

$$\chi^2(Q\|P) = \sup_{g\in\mathcal{M}} \left\{ \int g\, dQ - \frac{1}{4} \int g^2\, dP - 1 \right\}. \tag{6.4}$$

Let $g = a + bh$, and solve (6.4) with respect to a, b, we obtain a more symmetric version which is directly related to the bias-variance tradeoff:

$$\begin{aligned}
\chi^2(Q\|P) &= \sup_{h\in\mathcal{M}} \sup_{a,b\in\mathbb{R}} \left\{ a - \frac{a^2}{4} + b\int h\,dQ - \frac{ab}{2}\int h\,dP - \frac{b^2}{4}\int h^2\,dP - 1 \right\} \\
&= \sup_{h\in\mathcal{M}} \sup_{a\in\mathbb{R}} \left\{ \frac{\left(a\int h\,dP - 2\int h\,dQ\right)^2}{4\int h^2\,dP} - \left(1 - \frac{a}{2}\right)^2 \right\} \qquad \left(\text{take } b = \frac{a\int h\,dP - 2\int h\,dQ}{\int h^2\,dP}\right) \\
&= \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{\left(\int h\,dQ - \int h\,dP\right)^2}{\int h^2\,dP - \left(\int h\,dP\right)^2}. \qquad \left(\text{take } a = \frac{2\int h^2\,dP - 2\int h\,dP\int h\,dQ}{\int h^2\,dP - (\int h\,dP)^2}\right)
\end{aligned}$$

We can write this equality to the expectation form:

$$\chi^2(Q\|P) = \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{\left(\mathbb{E}_Q[h(X)] - \mathbb{E}_P[h(X)]\right)^2}{\mathrm{Var}_P(h(X))} \tag{6.5}$$

This bound extremely useful later.

Theorem 6.11 (Donsker-Varadhan)**.**

6.3 Inequality between f-Divergences and Joint Range

Theorem 6.12. *For two probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{F})$,*

$$D(Q\|P) \le \log\left(1 + \chi^2(Q\|P)\right). \tag{6.6}$$

Proof. By Jensen's inequality,

$$\log\left(1 + \chi^2(Q\|P)\right) = \log\left(\int \frac{Q^2}{P}\right) \ge \int Q\log\frac{Q}{P} = D(Q\|P).$$

Thus we complete the proof. □

Theorem 6.13 (Pinsker's inequality)**.** *For two probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{F})$,*

$$d_{\mathrm{TV}}(P, Q) \le \sqrt{\frac{1}{2}D(Q\|P)}.$$

Proof. We first consider the case $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$ with $p \le q$. Then

$$\begin{aligned}
D(Q\|P) &= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \\
&= q\int_p^q \frac{dt}{t} - (1-q)\int_p^q \frac{dt}{1-t} \\
&= \int_p^q \frac{q-t}{t(1-t)}\,dt \ge 4\int_p^q (q-t)\,dt = 2(q-p)^2.
\end{aligned}$$

Since $d_{\mathrm{TV}}(P, Q) = q - p$, the inequality follows. For the general case, we take $A \in \mathscr{F}$. Let $P_A = \text{Bernoulli}(P(A))$ be the distribution of the variable $\mathbb{1}_{\{X\in A\}}$ under $X \sim P$, and we define $Q_A = \text{Bernoulli}(Q(A))$ similarly. Using the conclusion above and the data processing inequality,

$$|P(A) - Q(A)| = d_{\mathrm{TV}}(P_A, Q_A) \le \sqrt{\frac{1}{2}D(Q_A\|P_A)} \le \sqrt{\frac{1}{2}D(Q\|P)}.$$

Taking the supremum $d_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathscr{F}} |P(A) - Q(A)|$, we have the desired result. □

A refinement of Pinsker's inequality is presented below.

Theorem 6.14 (Bretagnolle-Huber)**.** *For two probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{F})$,*

$$d_{\mathrm{TV}}(P,Q) \le \sqrt{1 - \mathrm{e}^{-D(Q\|P)}} \le 1 - \frac{1}{2}\mathrm{e}^{-D(Q\|P)}.$$

Proof. Similar to the proof of Pinsker's inequality, it suffices to show the Bernoulli case. Let $P = \mathrm{Bernoulli}(p)$ and $Q = \mathrm{Bernoulli}(q)$ with $p \le q$. Then

$$\begin{aligned} D(Q\|P) &= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \\ &= -2q\log\sqrt{p}q - 2(1-q)\log\frac{1-p}{1-q} \\ &\ge -2\log\left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right). \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{e}^{-D(Q\|P)} &\le \left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right)^2 \\ &\le \left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right)^2 + \left(\sqrt{p(1-p)} - \sqrt{(1-q)q}\right)^2 \\ &= 1 - (q-p)^2 = 1 - d_{\mathrm{TV}}(P,Q)^2. \end{aligned}$$

The desired bound then follows. □

The downside of ad hoc approaches is that it is hard to tell whether those inequalities can be improved or not. However, the key step when we proved the Pinksers inequality, reduction to the case for Bernoulli random variables, is inspiring: is it possible to reduce inequalities between any two f-divergences to the binary case?

The joint range of f-divergences provides a systematic approach to find inequalities between f-divergences.

Definition 6.15 (Joint range)**.** Consider two f-divergences D_f and D_g. The joint range between D_f and D_g is a subset of $\mathbb{R}^2_+$ defined by

$$\begin{aligned} \mathcal{R} &= \{(D_f(Q\|P), D_g(Q\|P)) : P \text{ and } Q \text{ are probability measures on some measure space}\}, \\ \mathcal{R}_k &= \{(D_f(Q\|P), D_g(Q\|P)) : P \text{ and } Q \text{ are probability measures on } \{1,2,\cdots,k\}\}, \quad k = 2,3,\cdots. \end{aligned}$$

If we know the region $\mathcal{R}$, we can find a tight inequality between D_f and D_g:

$$D_g(Q\|P) \ge F(D_f(Q\|P)),$$

where F is the lower boundary of $\mathcal{R}$:

$$F(t) := \inf\{x \ge 0 : (t,x) \in \mathcal{R}\} = \inf_{(P,Q):D_f(P\|Q)=t} D_g(P\|Q), \quad t \ge 0.$$

The region $\mathcal{R}$ seems difficult to characterize since we need to consider probability measures P and Q over all measurable spaces. On the other hand, the region R_k for small k is easy to obtain. The main theorem we will prove is the following, which provides a simple characterization of $\mathcal{R}$.

Theorem 6.16 (Harremoës-Vajda)**.** *Given two f-divergences D_f and D_g, their joint range satisfies*

$$\mathcal{R} = \mathrm{Conv}(\mathcal{R}_2).$$

The proof of this theorem requires some technical lemmata.

Lemma 6.17. *Given two f-divergences D_f and D_g, their joint range is*

$$\mathcal{R} = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + f'(\infty)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + g'(\infty)(1-\mathbb{E}[X]) \end{pmatrix} : X \text{ is a random variable with } X \geq 0 \text{ and } \mathbb{E}[X] \leq 1 \right\}. \tag{6.7}$$

Furthermore, for any integer k greater than 1,

$$\mathcal{R}_k = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + f'(\infty)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + g'(\infty)(1-\mathbb{E}[X]) \end{pmatrix} : \begin{array}{c} X \text{ takes at most } k-1 \text{ values, } X \geq 0 \text{ and } \mathbb{E}[X] \leq 1 \\ \text{or } X \text{ takes at most } k \text{ values, } X \geq 0 \text{ and } \mathbb{E}[X] = 1 \end{array} \right\}. \tag{6.8}$$

Proof. Given any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, let p, q denote the densities of P, Q under some dominating measure μ, respectively. Take

$$X = \mathbb{1}_{\{p>0\}} \frac{q}{p}, \quad \mu_X = P. \tag{6.9}$$

Then $X \geq 0$ and $\mathbb{E}[X] = Q(\{p > 0\}) \leq 1$. Moreover,

$$\begin{aligned} D_f(Q\|P) &= \int_{\{p>0\}} f\left(\frac{q}{p}\right) p \, d\mu + f'(\infty) Q(\{p = 0\}) = \mathbb{E}[f(X)] + f'(\infty)(1-\mathbb{E}[X]), \\ D_g(Q\|P) &= \int_{\{p>0\}} g\left(\frac{q}{p}\right) p \, d\mu + g'(\infty) Q(\{p = 0\}) = \mathbb{E}[g(X)] + g'(\infty)(1-\mathbb{E}[X]). \end{aligned}$$

On the other hand, for any random variable X with $X \geq 0$ and $\mathbb{E}[X] \leq 1$ with $X \sim \mu$, let

$$dP = d\mu, \quad dQ = X \, d\mu + (1 - \mathbb{E}[X])\delta_{-\infty}, \tag{6.10}$$

where $-\infty$ is an arbitrary symbol outside the support of X. Then

$$D_f(Q\|P) = \mathbb{E}[f(X)] + f'(\infty)(1-\mathbb{E}[X]), \quad D_g(Q\|P) = \mathbb{E}[g(X)] + g'(\infty)(1-\mathbb{E}[X]).$$

Now we consider $\mathcal{R}_k$. Consider any two probability measures P and Q on $\{1, 2, \cdots, k\}$. If $P \ll Q$, the likelihood ratio X defined in (6.9) takes at most k values and $\mathbb{E}[X] = 1$; otherwise, X takes at most $k-1$ values and $\mathbb{E}[X] \leq 1$. On the other hand, for any variable X taking at most k values with $\mathbb{E}[X] \geq 0$ and $\mathbb{E}[X] = 1$, the construction of P and Q in (6.10) are on the same support of size k; for any variable X taking at most $k-1$ values with $\mathbb{E}[X] \geq 0$ and $\mathbb{E}[X] \leq 1$, the support of Q increases at most by 1. □

Theorem 6.18 (Carathéodory)**.** *Let S be a nonempty subset of $\mathbb{R}^n$. For each $x \in \mathrm{Conv}(S)$, there exist $n+1$ points $x_1, x_2, \cdots, x_{n+1} \in S$ such that $x \in \mathrm{Conv}\{x_1, x_2, \cdots, x_{n+1}\}$.*

Proof. We first prove that *each point $x \in \mathrm{Cone}(S)$ can be represented as a positive combination of linearly independent vectors from S*, where $\mathrm{Cone}(S)$ is the minimum convex cone containing S, i.e.

$$\mathrm{Cone}(S) = \left\{ \sum_{i=1}^{N} \alpha_i x_i : N \in \mathbb{N}, \ x_1, \cdots, x_N \in S, \alpha_1, \cdots, \alpha_N \geq 0 \right\}.$$

Take $x \in \mathrm{Cone}(S)$ with $x \neq 0$. Let m be the minimum integer such that there exist $x_1, \cdots, x_m \in S$ and $\alpha_1, \cdots, \alpha_m > 0$ satisfying $x = \sum_{i=1}^{m} \alpha_i x_i$. If the vectors $x_1, \cdots, x_m$ are not linearly independent, there exist $\lambda_1, \cdots, \lambda_m \in \mathbb{R}$ with at least one $\lambda_i > 0$ such that $\sum_{i=1}^{n} \lambda_i x_i = 0$. Consider the greatest $\gamma^* \in \mathbb{R}$ such that

$\alpha_i - \gamma\lambda_i \geq 0$ for all $i = 1, \cdots, m$. Then

$$x = \sum_{i=1}^{m} (\alpha_i - \gamma^* \lambda_i) x_i,$$

which is a positive combination of at most $m-1$ vectors in S, contradicting the minimality of m!

Now we consider the set $U = \{(x, 1) : x \in S\}$. If $x \in \mathrm{Conv}(S)$, the extended vector $(x, 1) \in \mathrm{Cone}(U)$. By our conclusion above, one can find linearly independent vectors $(x_1, 1), \cdots, (x_m, 1) \in U \subset \mathbb{R}^{n+1}$, with $m \leq n+1$, and corresponding weights $\alpha_1, \cdots, \alpha_m > 0$ such that

$$(x, 1) = \sum_{i=1}^{m} \alpha_i (x_i, 1).$$

The last coordinate implies that $\sum_{i=1}^{n} \alpha_i = 1$. Therefore, any $x \in \mathrm{Conv}(S)$ is the convex combination of no more than $n+1$ points of S, which finishes the proof. □

Lemma 6.19. *Given two f-divergences D_f and D_g, their joint range satisfies*

$$\mathcal{R} = \mathcal{R}_5.$$

Proof. It suffices to show that $\mathcal{R} \subset \mathcal{R}_5$. We define the set $S = \{(x, f(x), g(x)) : x \geq 0\} \subset \mathbb{R}^3$. For any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, consider the likelihood ratio X defined in (6.9). Then $(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) \in \mathrm{Conv}(S)$. By Carathéodory theorem, there exist points $x_1, x_2, x_3, x_4 \geq 0$ and the corresponding weights $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$ with $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ such that

$$\sum_{i=1}^{4} \alpha_i (x_i, f(x_i), g(x_i)) = (\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)])$$

Consider the random variable Y supported on $\{x_1, x_2, x_3, x_4\}$ and taking value x_i with probability α_i. Then

$$(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) = (\mathbb{E}[Y], \mathbb{E}[f(Y)], \mathbb{E}[g(Y)]).$$

Since Y takes at most 4 values, $Y \geq 0$ and $\mathbb{E}[Y] = \mathbb{E}[X] \leq 1$, by Lemma 6.17,

$$\begin{pmatrix} D_f(Q\|P) \\ D_g(Q\|P) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X)] + f'(\infty)(1 - \mathbb{E}[X]) \\ \mathbb{E}[g(X)] + g'(\infty)(1 - \mathbb{E}[X]) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(Y)] + f'(\infty)(1 - \mathbb{E}[Y]) \\ \mathbb{E}[g(Y)] + g'(\infty)(1 - \mathbb{E}[Y]) \end{pmatrix} \in \mathcal{R}_5.$$

Thus we conclude that $\mathcal{R} \subset \mathcal{R}_5$. □

Lemma 6.20. *Given two f-divergences D_f and D_g, their joint range $\mathcal{R}$ is a convex set in $\mathbb{R}^2$.*

Proof. Given any two pairs of distributions (P_0, Q_0) and (P_1, Q_1) on some measurable space $(\mathcal{X}, \mathscr{F})$ and given any $0 \leq \lambda \leq 1$, we construct a random variable $Z = (X, B)$ such that $B \sim \mathrm{Bernoulli}(\lambda)$, $P_{X|B} \sim P_i$ and $Q_{X|B=i} = Q_i$, where $i = 0, 1$. Then we can verify that

$$D_f(Q_{X,B}\|P_{X,B}) = (1 - \lambda) D_f(Q_0\|P_0) + \lambda D_f(Q_1\|P_1).$$

The same conclusion holds for D_g. Hence $\mathcal{R}$ is convex. □

Remark. If we further assume that $\mathcal{X} = \{1, 2, \cdots, k\}$ in our proof, where $k = 2, 3, \cdots$, it turns out that

$$\mathrm{Conv}(\mathcal{R}_k) \subset \mathcal{R}_{2k}.$$

Lemma 6.21. *Given two f-divergences D_f and D_g, their joint range satisfies*

$$\mathcal{R}_{k+1} \subset \mathrm{Conv}(\mathcal{R}_2 \cup \mathcal{R}_k), \quad k = 2, 3, \cdots . \tag{6.11}$$

Proof. Given any pair of distributions (P, Q) on $\{1, 2, \cdots, k+1\}$ that produces a point of $\mathcal{R}_{k+1}$, we take the likelihood ratio X as in (6.9) that takes at most $k+1$ values.

- If $\mathbb{E}[X] = Q(\{p > 0\}) < 1$, and P is supported on at most k values. Denote by x the smallest possible value of X and then $x < 1$. Assume $\mu_X(x) = \lambda$, then

  $$\mu_X = \lambda\delta_x + (1-\lambda)\mu',$$

  where μ' is supported on at most $k-1$ values of X other than x. Let $\mu_2 = \delta_x$. To prove (6.11), we aim to find a probability measure μ_1 and $0 \le \alpha \le 1$ such that

  $$\mu_X = \alpha\mu_1 + (1-\alpha)\mu_2,$$

  where $Y \sim \mu_1$ takes at most $k-1$ values and $\mathbb{E}[Y] \le 1$, or $Y \sim \mu_1$ takes at most k values and $\mathbb{E}[Y] = 1$.
  - If $\mathbb{E}_{\mu'}[X] \le 1$, we let $\mu_1 = \mu'$ and $\alpha = 1 - \lambda$;
  - If $\mathbb{E}_{\mu'}[X] > 1$, we consider $\mu_1 = \beta\delta_x + (1-\beta)\mu'$, where we take $\beta = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ so that $\mathbb{E}[Y] = 1$. In this setting, we let $\alpha = \frac{\mathbb{E}[X]-x}{1-x}$.
- If $\mathbb{E}[X] = Q(\{p > 0\}) = 1$, we have $Q \ll P$. Denote the smallest value of X by x and the largest value by y, respectively, and then $x \le 1$, $y \ge 1$. Assume $\mu_X(x) = r$ and $\mu_X(y) = s$. Then

  $$\mu_X = r\delta_x + s\delta_y + (1-r-s)\mu',$$

  where μ' is supported on at most $k-1$ values of X other than x and y.
  Let $\mu_2 = \frac{y-1}{y-x}\delta_x + \frac{1-x}{y-x}\delta_y$, so $Z \sim \mu_2$ takes at most 2 values and $\mathbb{E}[Z] = 1$. To prove (6.11), we aim to find a probability measure μ_1 and $0 \le \alpha \le 1$ such that

  $$\mu_X = \alpha\mu_1 + (1-\alpha)\mu_2,$$

  where $Y \sim \mu_1$ takes at most $k-1$ values and $\mathbb{E}[Y] \le 1$, or $Y \sim \mu_1$ takes at most k values and $\mathbb{E}[Y] = 1$.
  - If $\mathbb{E}_{\mu'}[X] \le 1$, we consider $\mu_1 = \beta\delta_y + (1-\beta)\mu'$, where we take $\beta = \frac{1-\mathbb{E}_{\mu'}[X]}{y-\mathbb{E}_{\mu'}[X]}$ so that $\mathbb{E}[Y] = 1$. In this setting, we let $\alpha = 1 - \frac{r(y-x)}{y-1}$;
  - If $\mathbb{E}_{\mu'}[X] > 1$, we consider $\mu_1 = \beta\delta_x + (1-\beta)\mu'$, where we take $\beta = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ so that $\mathbb{E}[Y] = 1$. In this setting, we let $\alpha = 1 - \frac{s(y-x)}{1-x}$.

Let $Y \sim \mu_1$ and $Z \sim \mu_2$. Applying the construction in (6.10) with μ_1 and μ_2, we obtain two pairs of measures (P_1, Q_1) supported on k values and (P_2, Q_2) supported on two values, respectively. Then

$$\begin{aligned}
\begin{pmatrix} D_f(Q\|P) \\ D_g(Q\|P) \end{pmatrix} &= \begin{pmatrix} \mathbb{E}[f(X)] + f'(\infty)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + g'(\infty)(1-\mathbb{E}[X]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} \mathbb{E}[f(Y)] + f'(\infty)(1-\mathbb{E}[Y]) \\ \mathbb{E}[g(Y)] + g'(\infty)(1-\mathbb{E}[Y]) \end{pmatrix} + (1-\alpha) \begin{pmatrix} \mathbb{E}[f(Z)] + f'(\infty)(1-\mathbb{E}[Z]) \\ \mathbb{E}[g(Z)] + g'(\infty)(1-\mathbb{E}[Z]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} D_f(Q_1\|P_1) \\ D_g(Q_1\|P_1) \end{pmatrix} + (1-\alpha) \begin{pmatrix} D_f(Q_2\|P_2) \\ D_g(Q_2\|P_2) \end{pmatrix} \in \mathrm{Conv}(\mathcal{R}_2 \cup \mathcal{R}_k)
\end{aligned}$$

Therefore, $\mathcal{R}_{k+1} \subset \mathrm{Conv}(\mathcal{R}_2 \cup \mathcal{R}_k)$. □

Now we are prepared to prove the main result.

Proof of Theorem 6.16. According to Lemma 6.21, we have $\mathcal{R}_3 \subset \mathrm{Conv}(\mathcal{R}_2)$. By induction, we conclude that

$$\mathcal{R}_k \subset \mathrm{Conv}(\mathcal{R}_2 \cup \mathcal{R}_{k-1}) = \mathrm{Conv}(\mathcal{R}_2), \quad k = 4, 5, \cdots .$$

Particularly, we have $\mathcal{R}_5 \subset \mathrm{Conv}(\mathcal{R}_2)$. On the other hand, by Lemma 6.20 and the definition of $\mathcal{R}_k$, we have

$$\mathrm{Conv}(\mathcal{R}_2) \subset \mathcal{R}_4 \subset \mathcal{R}_5 \subset \cdots .$$

Finally, using Lemma 6.19, we obtain $\mathcal{R} = \mathcal{R}_5 = \mathrm{Conv}(\mathcal{R}_2)$. □

Remark. To summarize, we have shown that

$$\mathcal{R}_2 \subset \mathcal{R}_3 \subset \mathcal{R}_4 = \mathcal{R}_5 = \cdots = \mathcal{R} = \mathrm{Conv}(\mathcal{R}_2).$$

Every point the joint range $\mathcal{R}_2$ can be parameterized as $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$, where $p, q \in [0, 1]$. Note that $D_f(P\|Q) = D_f(\overline{P}\|\overline{Q})$, where $\overline{P} = \text{Bernoulli}(1-p)$ and $\overline{Q} = \text{Bernoulli}(1-q)$. Therefore, to determine the region $\mathcal{R}_2$, it suffices to consider the image of the triangle

$$\mathcal{S} = \{(p, q) : 0 \le p \le q \le 1\}$$

under the transformation $(p, q) \mapsto (D_f, D_g)$. Then, to determine the joint range $\mathcal{R}$, we simply take the convex hull of the image of the triangle $\mathcal{S}$.

Theorem 6.22 (Sandwich bound)**.** *Let P and Q be two probability measure on some measurable space. Then*

$$H^2(P, Q) \le d_{\mathrm{TV}}(P, Q) \le H(P, Q)\sqrt{2 - H^2(P, Q)},$$

Proof. We consider the distributions $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$, where $0 \le p, q \le 1$. Clearly,

$$d_{\mathrm{TV}}(P, Q) = |q - p|, \quad H^2(P, Q) = 1 - \sqrt{pq} - \sqrt{(1-p)(1-q)}.$$

The joint range of d_{TV} and H^2 is

$$\mathcal{R}_2 = \left\{ \begin{pmatrix} |q-p| \\ 1 - \sqrt{pq} - \sqrt{(1-p)(1-q)} \end{pmatrix} : 0 \le p, q \le 1 \right\}$$

Since both d_{TV} and H^2 are symmetric, it suffices to consider the case $p \le q$. We fix $t = d_{\mathrm{TV}}(P, Q) \ge 0$. Consider the function

$$\varphi(p) = H^2(P, Q) = 1 - \sqrt{p(p+t)} - \sqrt{(1-p)(1-p-t)}, \quad 0 \le p \le 1 - t.$$

This function attains minimum at $p = \frac{1-t}{2}$ and maximum at both $p = 0$ and $p = 1 - t$. Hence

$$1 - \sqrt{1 - t^2} \le \varphi(p) \le 1 - \sqrt{1 - t},$$

and $\mathcal{R}_2$ is the region between the two curves given in the start and end of the last display. Since $\mathcal{R} = \mathrm{Conv}(\mathcal{R}_2)$, we take the convex hull of $\mathcal{R}$ and get

$$\mathcal{R} = \left\{ (x, y) : 0 \le x \le 1, \text{ and } 1 - \sqrt{1 - x^2} \le y \le x \right\}.$$

According to this joint range, for all probability measures P and Q on some measurable space, we have

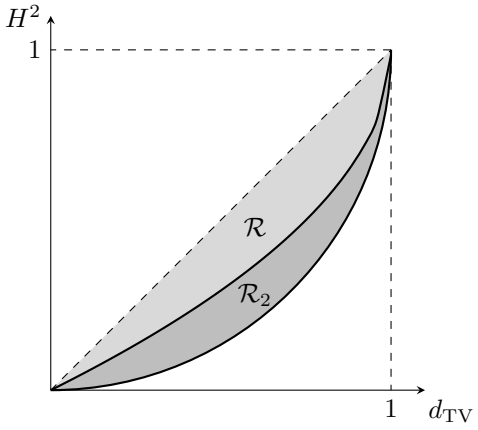
$$1 - \sqrt{1 - d_{\mathrm{TV}}(P,Q)^2} \le H^2(P,Q) \le d_{\mathrm{TV}}(P,Q),$$

or equivalently,

$$H^2(P,Q) \le d_{\mathrm{TV}}(P,Q) \le H(P,Q)\sqrt{2 - H^2(P,Q)},$$

which is the desired bound. □

Remark. We visualize the joint range of d_{TV} and H^2 in the following figure.



The sandwich bound is described by the diagonal line and the lower arc. According to our discussion, this sandwich bound is non-improvable. Under the constraint $d_{\mathrm{TV}}(P,Q) = t$, the upper bound is attained when

$$P = \text{Bernoulli}\left(\frac{1-t}{2}\right), Q = \text{Bernoulli}\left(\frac{1+t}{2}\right),$$

and the lower bound is attained when

$$P = (1-t, t, 0), Q = (1-t, 0, t).$$

Theorem 6.23 (Total variation versus chi-square divergence)**.** *Let P and Q be two probability measure on some measurable space. Then*

$$\chi^2(P\|Q) \ge f(d_{\mathrm{TV}}(P,Q)), \quad \textit{where} \quad f(t) = \begin{cases} 4t^2, & 0 \le t \le \frac{1}{2}, \\ \frac{t}{1-t}, & \frac{1}{2} < t \le 1. \end{cases}$$

Proof. We consider the distributions $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$, where $0 \le p, q \le 1$. Clearly,

$$d_{\mathrm{TV}}(P,Q) = |q-p|, \quad \chi^2(P\|Q) = \frac{p^2}{q} + \frac{(1-p)^2}{1-q} - 1 = \frac{(q-p)^2}{q(1-q)}.$$

The joint range of d_{TV} and H^2 is

$$\mathcal{R}_2 = \left\{ \left(|q-p|, \frac{(q-p)^2}{q(1-q)} \right) : 0 \le p, q \le 1 \right\}$$

We fix $d_{\mathrm{TV}}(P,Q) = |q-p| = t \in (0,1)$. It turns out that

$$\chi^2(P\|Q) = \frac{t^2}{q(1-q)} = \frac{t^2}{q(1-q)}.$$

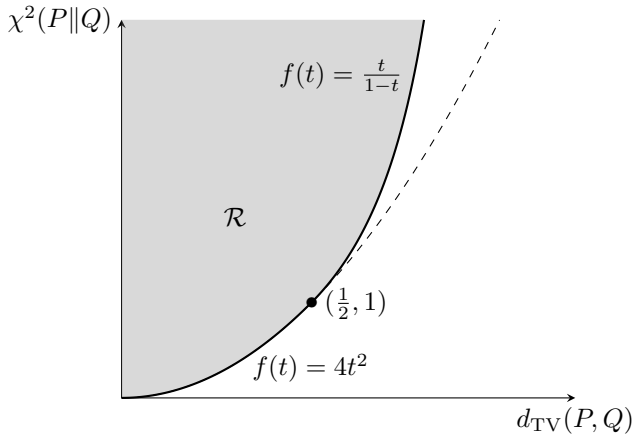
- If $t \le \frac{1}{2}$, the minimum is attained when $q = \frac{1}{2}$ and $p = \frac{1}{2} - t$, and $\chi^2(P\|Q) = 4t^2$;
- If $t > \frac{1}{2}$, the minimum is attained when $q = t$ and $p = 0$, and $\chi^2(P\|Q) = \frac{t}{1-t}$.

We consider the function $f : [0,1] \to \mathbb{R}_+$ defined as follows:

$$f(t) = \begin{cases} 4t^2, & 0 \le t \le \frac{1}{2}, \\ \frac{t}{1-t}, & \frac{1}{2} < t \le 1. \end{cases}$$

Then f is a convex function on $[0,1]$, and $\mathcal{R}_2$ is the epigraph $\{(t,x) : 0 \le t \le 1,\ x \ge f(t)\}$. Since $\mathcal{R}_2$ is convex, we have $\mathcal{R} = \mathrm{Conv}(\mathcal{R}_2) = \mathcal{R}_2$. The desired bound follows this range. □

Remark. We visualize the joint range of d_{TV} and χ^2 in the following figure.



A direct corollary of this range is that as $d_{\mathrm{TV}}(P,Q) \to 1$, we have $\chi^2(P\|Q) \to \infty$.

Theorem 6.24 (Total variation versus Le Cam divergence)**.** *Let P and Q be two probability measure on some measurable space. Then*

$$d_{\mathrm{TV}}(P,Q)^2 \le \mathrm{Le}(P\|Q) \le d_{\mathrm{TV}}(P,Q).$$

Proof. We consider the distributions $P = \mathrm{Bernoulli}(p)$ and $Q = \mathrm{Bernoulli}(q)$, where $0 \le p, q \le 1$. Clearly,

$$d_{\mathrm{TV}}(P,Q) = |q-p|, \quad \mathrm{Le}(P,Q) = \frac{(p-q)^2}{2(p+q)} + \frac{(p-q)^2}{2(2-p-q)}.$$

The joint range of d_{TV} and H^2 is

$$\mathcal{R}_2 = \left\{ \left(|q-p|, \frac{(p-q)^2}{2(p+q)} + \frac{(p-q)^2}{2(2-p-q)} \right) : 0 \le p, q \le 1 \right\}$$

We fix $d_{\mathrm{TV}}(P,Q) = |q-p| = t \in (0,1)$. It turns out that

$$\mathrm{Le}(P,Q) = \frac{t^2}{(2p+t)(2-2p-t)}.$$

Over the interval $0 \le p \le 1-t$, we have

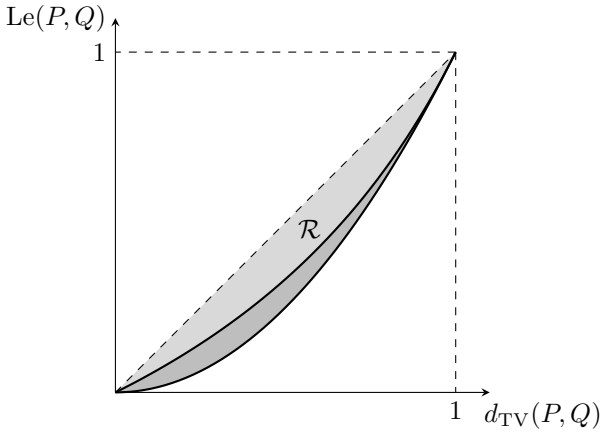
$$t^2 \le \mathrm{Le}(P,Q) \le \frac{t}{2-t}.$$

Therefore, the joint range is

$$\mathcal{R}_2 = \left\{(t,x) : 0 \le t \le 1,\ t^2 \le x \le \frac{t}{2-t}\right\}, \quad \mathcal{R} = \mathrm{Conv}(\mathcal{R}_2) = \left\{(t,x) : 0 \le t \le 1,\ t^2 \le x \le t\right\}.$$

The desired inequality follows from this joint range. □

Remark. We visualize the joint range of $d_{\mathrm{TV}}(P,Q)$ and $\mathrm{Le}(P,Q)$ in the following figure.



Example 6.25 (Total variation versus Jensen-Shannon divergence)**.** *Let P and Q be two probability measure on some measurable space. Then*

$$\frac{1-d_{\mathrm{TV}}(P,Q)}{2}\log\left(1-d_{\mathrm{TV}}(P,Q)\right) + \frac{1+d_{\mathrm{TV}}(P,Q)}{2}\log(1+d_{\mathrm{TV}}(P,Q)) \le d_{\mathrm{JS}}(P\|Q) \le \log 2 \cdot d_{\mathrm{TV}}(P,Q).$$

Proof. We consider the distributions $P = \mathrm{Bernoulli}(p)$ and $Q = \mathrm{Bernoulli}(q)$, where $0 \le p, q \le 1$. Then

$$d_{\mathrm{TV}}(P,Q) = |q-p|, \quad d_{\mathrm{JS}}(P,Q) = H\left(\frac{p+q}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(q).$$

The joint range of d_{TV} and d_{JS} is

$$\mathcal{R}_2 = \left\{\left(|q-p|, H\left(\frac{p+q}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(q)\right) : 0 \le p \le q \le 1\right\}$$

We fix $p = \alpha - \frac{t}{2}, q = \alpha + \frac{t}{2}$. It turns out that

$$d_{\mathrm{JS}}(P,Q) = H(\alpha) - \frac{1}{2}H\left(\alpha - \frac{t}{2}\right) - \frac{1}{2}H\left(\alpha + \frac{t}{2}\right).$$

Over the interval $\frac{t}{2} \le \alpha \le 1 - \frac{t}{2}$, we have

$$\frac{\partial}{\partial\alpha} d_{\mathrm{JS}}(P,Q) = \log\frac{1-\alpha}{\alpha} - \frac{1}{2}\left(\log\frac{1-\alpha-\frac{t}{2}}{\alpha+\frac{t}{2}} + \log\frac{1-\alpha+\frac{t}{2}}{\alpha+\frac{t}{2}}\right),$$
$$\frac{\partial^2}{\partial\alpha^2} d_{\mathrm{JS}}(P,Q) = \frac{t^2}{\alpha(4\alpha^2-t^2)} + \frac{t^2}{4(1-\alpha)((1-\alpha)^2-t^2)} > 0.$$

Hence $d_{\mathrm{JS}}(P,Q)$ attains minimum when $\alpha = \frac{t}{2}$ or $1-\frac{t}{2}$, and attains maximum when $\alpha = \frac{1}{3}$.

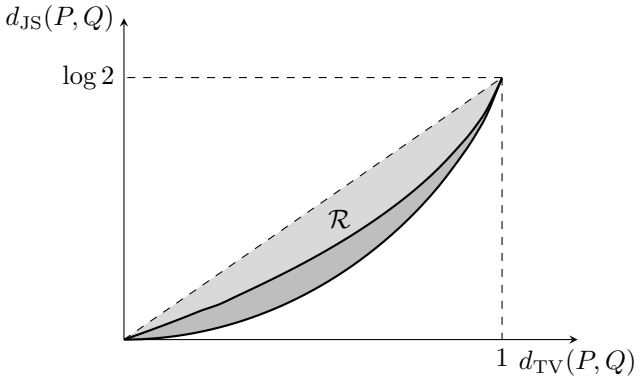
$$H\left(\frac{t}{2}\right) - \frac{H(t)}{2} \le d_{\mathrm{JS}}(P,Q) \le H\left(\frac{1}{2}\right) - \frac{1}{2}H\left(\frac{1-t}{2}\right) - \frac{1}{2}H\left(\frac{1+t}{2}\right).$$

Therefore

$$\mathcal{R}_2 = \left\{(t,x) : \frac{1-t}{2}\log(1-t) + \frac{1+t}{2}\log(1+t) \le x \le H\left(\frac{t}{2}\right) - \frac{H(t)}{2}\right\},$$
$$\mathcal{R} = \mathrm{Conv}(\mathcal{R}_2) = \left\{(t,x) : \frac{1-t}{2}\log(1-t) + \frac{1+t}{2}\log(1+t) \le x \le t\log 2\right\}.$$

The desired bound follows from the joint range. □

Remark. We visualize the joint range of $d_{\mathrm{TV}}(P,Q)$ and $d_{\mathrm{JS}}(P,Q)$ in the following figure.

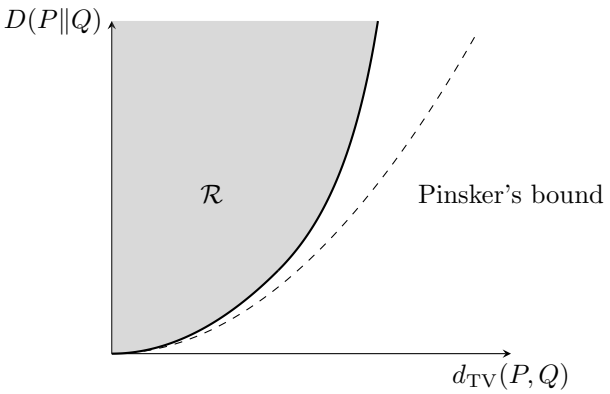


Example 6.26 (Total variation versus Kullback-Leibler divergence)**.** *The joint range between KL and TV is shown in the following figure. Although there is no known close-form expression, the following parametric formula of the lower boundary is known:*

$$\begin{cases} \mathrm{TV}_t = \frac{t}{2}\left(1 - \left(\coth(t) - \frac{1}{t}\right)^2\right) \\ \mathrm{KL}_t = t\coth(t) + \log\left(t\,\mathrm{csch}(t)\right) - t^2\mathrm{csch}^2(t) \end{cases}, \quad t \ge 0.$$

A direct corollary of this formula is Vajda's lower bound:

$$D(P\|Q) \ge \log\frac{1+d_{\mathrm{TV}}(P,Q)}{1-d_{\mathrm{TV}}(P,Q)} - \frac{2d_{\mathrm{TV}}(P,Q)}{1+d_{\mathrm{TV}}(P,Q)}.$$



6.4 Pearson χ^2-Divergence and Information Bounds

The Pearson χ^2-divergence is special because most f-divergences are locally χ^2-like.

Theorem 6.27. *Let D_f be an f-divergence such that $f \in C^2(\mathbb{R}_+)$ and $\limsup_{x\to\infty} f''(x) < \infty$. Then*

(i) If $\chi^2(Q\|P) < \infty$, then for any $0 < \lambda < 1$,

$$D_f(\lambda Q + (1-\lambda)P\|P) < \infty;$$

(ii) If $\chi^2(Q\|P) < \infty$,

$$\lim_{\lambda\to 0^+} \frac{1}{\lambda^2} D_f(\lambda Q + (1-\lambda)P\|P) = \frac{f''(1)}{2}\chi^2(Q\|P). \tag{6.12}$$

Proof. We will use the integral remainder of Taylor's expansion:

$$f(x) = f(a) + f'(a)(x-a) + (x-a)^2 \int_0^1 (1-\theta) f''(a+\theta(x-a))\, d\theta.$$

For any $0 < \lambda < 1$,

$$\begin{aligned}
D_f(\lambda Q + (1-\lambda)P\|P) &= \int f\left(1 + \lambda\frac{dQ-dP}{dP}\right) dP \\
&= \int \left(f(1) + f'(1)\left(\lambda\frac{dQ-dP}{dP}\right) + \left(\lambda\frac{dQ-dP}{dP}\right)^2 \int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \right) dP \\
&= \lambda^2 \int \left(\frac{dQ}{dP} - 1\right)^2 \left(\int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \right) dP.
\end{aligned}$$

Since $1 + \theta\lambda\frac{dQ-dP}{dP} \geq 1-\lambda$ and $\limsup_{x\to\infty} f''(x) < \infty$, the function f'' is bounded on $[1-\lambda, \infty)$. Hence

$$\int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \leq \frac{1}{2} \sup_{x\in[1-\lambda,\infty)} f''(x) := C_\lambda, \tag{6.13}$$

and therefore $D_f(\lambda Q + (1-\lambda)P\|P) \leq \lambda^2 C_\lambda \chi^2(Q\|P) < \infty$. To prove (ii), it remains to determine

$$\lim_{\lambda\to 0^+} \frac{1}{\lambda^2} D_f(\lambda Q + (1-\lambda)P\|P) = \lim_{\lambda\to 0^+} \int \left(\frac{dQ}{dP} - 1\right)^2 \left(\int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \right) dP.$$

According to the bound (6.13), for all $0 < \lambda < \frac{1}{2}$,

$$\left(\frac{dQ}{dP} - 1\right)^2 \int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \leq C_{\frac{1}{2}} \left(\frac{dQ}{dP} - 1\right)^2 \in L^1(\mathcal{X}, \mathscr{F}, P).$$

By dominated convergence theorem,

$$\begin{aligned}
&\lim_{\lambda\to 0^+} \int \left(\frac{dQ}{dP} - 1\right)^2 \left(\int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \right) dP \\
&= \int \left(\frac{dQ}{dP} - 1\right)^2 \left(\lim_{\lambda\to 0^+} \int_0^1 (1-\theta) f''\left(1+\theta\lambda\frac{dQ-dP}{dP}\right) d\theta \right) dP \\
&= \int \left(\frac{dQ}{dP} - 1\right)^2 \frac{f''(1)}{2}\, dP = \frac{f''(1)}{2}\chi^2(Q\|P).
\end{aligned} \tag{6.14}$$

Thus we complete the proof. □

Remark. In fact, the identity (6.12) remains correct even if $\chi^2(Q\|P) = \infty$ and $f''(1) > 0$, i.e.

$$D_f(\lambda Q + (1-\lambda)P\|P) = \omega(\lambda^2), \quad \lambda \to 0.$$

To see this, under the condition $Q \ll P$, we replace the equality (6.14) with "$\geq$" by Fatou's lemma:

$$\lim_{\lambda\to 0^+} \int \left(\frac{dQ}{dP} - 1\right)^2 \left(\int_0^1 (1-\theta) f'' \left(1 + \theta\lambda \frac{dQ - dP}{dP}\right) d\theta\right) dP \geq \frac{f''(1)}{2}\chi^2(Q\|P). \tag{6.15}$$

Since $\chi^2(Q\|P) = \infty$ and $f''(1) > 0$, the RHS of the last display is ∞, and the equality holds naturally.

Parameter estimation. Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions parameterized by $\theta \in \Theta$. The estimation of parameter θ can be described by the following Markov chain:

$$\theta^* \to X \to \widehat{\theta},$$

where θ^* is the true parameter, the sample $X \sim P_{\theta^*}$ is drawn from the distribution P_{θ^*}, and the estimator $\widehat{\theta} = \widehat{\theta}(X)$ is a (possibly random) mapping from the sample space $\mathcal{X}$ to the parameter space Θ. We use the *quadratic loss* to evaluate the difference between the real and the predicted parameter, i.e. $\ell(\widehat{\theta}) = |\widehat{\theta} - \theta^*|^2$. The *mean-squared error/risk* of estimator $\widehat{\theta}$ when the real parameter is given by θ^* is

$$R_{\theta^*}(\widehat{\theta}) = \mathbb{E}_{\theta^*}[|\widehat{\theta} - \theta^*|^2] = \mathbb{E}_{X\sim P_{\theta^*}}\left[\left|\widehat{\theta}(X) - \theta^*\right|^2\right].$$

For an estimator $\widehat{\theta}$ of θ, we have the following Hammersley-Chapman-Robbins (HCR) lower bound of risk.

Theorem 6.28 (Univariate Hammersley-Chapman-Robbins bound)**.** *If $\Theta \subset \mathbb{R}$, any estimator $\widehat{\theta}$ satisfies*

$$\mathrm{Var}_{\theta^*}(\widehat{\theta}) \geq \sup_{\theta\in\Theta:\theta\neq\theta^*} \frac{\left(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\right)^2}{\chi^2(P_\theta\|P_{\theta^*})}, \quad \forall \theta^* \in \Theta.$$

Proof. For all $\theta \in \Theta$, the distribution P_θ of X and the mapping rule $\widehat{\theta} = \widehat{\theta}(X)$ together induce a distribution Q_θ of the estimator $\widehat{\theta}$. We fix $\theta^*, \theta \in \Theta$ with $\theta \neq \theta^*$. By data processing inequality,

$$\chi^2(P_\theta\|P_{\theta^*}) \geq \chi^2(Q_\theta\|Q_{\theta^*}). \tag{6.16}$$

We take $h(x) = x$ in the variational representation (6.5) of χ^2-divergence, hence

$$\chi^2(Q_\theta\|Q_{\theta^*}) \geq \frac{\left(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\right)^2}{\mathrm{Var}_{\theta^*}(\widehat{\theta})}. \tag{6.17}$$

Combining (6.16) and (6.17), we get

$$\mathrm{Var}_{\theta^*}(\widehat{\theta}) \geq \frac{\left(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\right)^2}{\chi^2(P_\theta\|P_{\theta^*})}.$$

Since this inequality holds for all $\theta \neq \theta^*$, we take supremum on both sides to get the desired bound. □

Remark. Define the *bias function* of $\widehat{\theta}$ by $b(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$. According to the bias-variance decomposition,

$$R_{\theta^*}(\widehat{\theta}) = \mathbb{E}_{\theta^*}[(\widehat{\theta} - \theta^*)^2] \geq \sup_{\theta\in\Theta:\theta\neq\theta^*} \frac{\left(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\right)^2}{\chi^2(P_\theta\|P_{\theta^*})} + b(\theta^*)^2.$$

Theorem 6.29 (Multivariate Hammersley-Chapman-Robbins bound)**.** *If $\Theta \subset \mathbb{R}^n$, any estimator $\widehat{\theta}$ satisfies*

$$\chi^2(P_\theta \| P_{\theta^*}) \geq \big(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\big)^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta})^{-1} \big(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\big), \quad \forall \theta^*, \theta \in \Theta.$$

Proof. Based on the same procedure establishing (6.16), we take $h(x) = \xi^\top x$ in the variational representation (6.5) of χ^2-divergence, where $\xi \in \mathbb{R}^n \backslash \{0\}$. Then

$$\chi^2(Q_\theta \| Q_{\theta^*}) \geq \frac{\big(\xi^\top \mathbb{E}_{\theta^*}[\widehat{\theta}] - \xi^\top \mathbb{E}_\theta[\widehat{\theta}]\big)^2}{\xi^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta}) \xi}. \tag{6.18}$$

Combining (6.16), (6.17) and taking the supremum with respect to $\xi \in \mathbb{R}^n \backslash \{0\}$, we get

$$\chi^2(P_\theta \| P_{\theta^*}) \geq \sup_{\xi \in \mathbb{R}^n \backslash \{0\}} \frac{\xi^\top \big(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\big) \big(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\big)^\top \xi}{\xi^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta}) \xi}.$$

Using the fact $\sup_{\xi \in \mathbb{R}^n \backslash \{0\}} \frac{\langle w, \xi \rangle^2}{\xi^\top M \xi} = w^\top M^{-1} w$ from linear algebra, we obtain the desired bound. □

Score function. In many cases, we can express P_θ in form of the likelihood function:

$$\mathbb{P}_\theta(X \in A) = \int_A P_\theta(x)\, \mu(dx), \quad A \in \mathscr{F},$$

where μ is a dominating measure. For example, μ is the counting measure in the discrete case or the Lebesgue measure in the continuous case. The partial derivative of the log-likelihood $\log P_\theta(x)$ with respect to $\theta \in \Theta$ is called the *score*. Under regularity conditions on $P_\theta(x)$[2], the expectation of the score function evaluated at the true parameter θ is zero:

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log P_\theta(x) \right] = \int_{\mathcal{X}} \frac{\frac{\partial}{\partial \theta} P_\theta(x)}{P_\theta(x)} P_\theta(x)\, dx = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} P_\theta(x)\, dx = \nabla_\theta \int_{\mathcal{X}} P_\theta(x)\, dx = 0.$$

[2]To be specific, we assume Θ is an open subset of $\mathbb{R}^n$, and assume $\theta \mapsto P_\theta(x)$ to be a continuously differentiable function. For fixed parameter value $\theta_0 \in \Theta$ and $\epsilon > 0$, we consider the following conditions:

(i) $\int_{\mathcal{X}} P_\theta(x)\, dx < \infty$ for all $\theta \in B(\theta_0, \epsilon) := \{\theta \in \Theta : |\theta - \theta^*| < \epsilon\}$;

(ii) $\int_{\mathcal{X}} \frac{\partial}{\partial \theta} P_\theta(x)\, dx$ is continuous at $\theta = \theta_0$;

(ii') $\int_{\mathcal{X}} \sup_{\theta \in B(\theta_0, \epsilon)} \left| \frac{\partial}{\partial \theta} P_\theta(x) \right| dx < \infty$;

(iii) $\int_{\mathcal{X}} \int_0^\epsilon \left| \frac{\partial}{\partial \theta} P_{\theta_0 + tu}(x) \right| dt\, dx < \infty$ for all unit vectors $|u| = 1$.

Under conditions (i), (ii) and (iii), we use Fubini's theorem:

$$\frac{1}{h} \int_{\mathcal{X}} \big(P_{\theta_0 + hu}(x) - P_{\theta_0}(x)\big)\, dx = \frac{1}{h} \int_{\mathcal{X}} \int_0^h u^\top \frac{\partial}{\partial \theta} P_{\theta_0 + tu}(x)\, dt\, dx = \frac{1}{h} u^\top \int_0^h \int_{\mathcal{X}} \frac{\partial}{\partial \theta} P_\theta(x)\, dx\, dt, \quad \forall |u| = 1.$$

Letting $h \to 0$ on the both sides, we obtain

$$\nabla_\theta \int_{\mathcal{X}} P_\theta(x)\, dx \Big|_{\theta = \theta_0} = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} P_{\theta_0}(x)\, dx$$

Under conditions (i) and (ii'), for all $|h| < \epsilon$ and $|u| = 1$, the difference quotients are dominated by an integrable function:

$$\left| \frac{P_{\theta_0 + hu}(x) - P_{\theta_0}(x)}{h} \right| = \left| \frac{1}{h} \int_0^h \frac{\partial}{\partial \theta} P_{\theta_0 + tu}(x) d\theta \right| \leq \sup_{\theta \in B(\theta_0, \epsilon)} \left| \frac{\partial}{\partial \theta} P_\theta(x) \right|.$$

By Dominated Convergence Theorem,

$$\lim_{h \to 0} \int_{\mathcal{X}} \frac{P_{\theta_0 + hu}(x) - P_{\theta_0}(x)}{h}\, dx = \int_{\mathcal{X}} \lim_{h \to 0} \frac{P_{\theta_0 + hu}(x) - P_{\theta_0}(x)}{h}\, dx, \ \forall |u| = 1 \quad \Leftrightarrow \quad \nabla_\theta \int_{\mathcal{X}} P_\theta(x)\, dx \Big|_{\theta = \theta_0} = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} P_{\theta_0}(x)\, dx.$$

Fisher information. The *Fisher information* $\mathcal{I}(\theta)$ is defined to be the covariance of the score when the true parameter is θ:

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log P_\theta(X)\right)\left(\frac{\partial}{\partial\theta}\log P_\theta(X)\right)^\top\right] = \int \frac{1}{P_\theta}\frac{\partial P_\theta}{\partial\theta}\frac{\partial P_\theta}{\partial\theta^\top}.$$

Using the Taylor's expansion of P_θ at θ^*, we have

$$P_\theta - P_{\theta^*} = (\theta-\theta^*)^\top\frac{\partial P_{\theta^*}}{\partial\theta} + o(|\theta-\theta^*|).$$

Under regularity conditions [3] on $P_\theta(x)$,

$$\begin{aligned}\chi^2(P_\theta\|P_{\theta^*}) &= \int\frac{(P_\theta-P_{\theta^*})^2}{P_{\theta^*}}\\ &= (\theta-\theta^*)^\top\mathcal{I}(\theta^*)(\theta-\theta^*) + o\left(|\theta-\theta^*|^2\right).\end{aligned} \tag{6.19}$$

Therefore, the χ^2-divergence is "locally Fisher information". Using this property, we can derive a universal bound for estimation error in terms of Fisher information.

Theorem 6.30 (Cramér-Rao bound)**.** *Let $\Theta\subset\mathbb{R}$. Under regularity conditions, for the quadratic loss, any unbiased estimator $\widehat{\theta}$ satisfies*

$$R_{\theta^*}(\widehat{\theta}) = \mathbb{E}_{\theta^*}[(\widehat{\theta}-\theta^*)^2] \geq \frac{1}{\mathcal{I}(\theta^*)},\quad \forall\theta^*\in\Theta.$$

where $\mathcal{I}(\theta^*)$ is the Fisher information evaluated at the true parameter $\theta^*\in\Theta$.

Proof. Using the Hammersley-Chapman-Robbins bound and the unbiasedness of $\widehat{\theta}$,

$$\begin{aligned}\mathbb{E}_{\theta^*}[(\widehat{\theta}-\theta^*)^2] &\geq \lim_{\theta\to\theta^*}\frac{\left(\mathbb{E}_{\theta^*}[\widehat{\theta}]-\mathbb{E}_\theta[\widehat{\theta}]\right)^2}{\chi^2(P_\theta\|P_{\theta^*})}\\ &= \lim_{\theta\to\theta^*}\frac{\left(\theta^*-\theta\right)^2}{(\theta^*-\theta)^2\mathcal{I}(\theta^*)+o\left((\theta^*-\theta)^2\right)} = \frac{1}{\mathcal{I}(\theta^*)}.\end{aligned}$$

Thus we complete the proof. □

For biased estimators, we also have a similar bound.

---

[3] Define the remainder of the first-order Taylor expansion as follows:

$$r_\theta = P_\theta - P_{\theta^*} - (\theta-\theta^*)^\top\frac{\partial P_{\theta^*}}{\partial\theta} = o(|\theta-\theta^*|).$$

By Cauchy-Schwarz inequality, (6.19) is valid if r_θ vanishes under the weighted inner product:

$$\int\frac{r_\theta^2(x)}{P_{\theta^*}(x)}\,dx = o(|\theta-\theta^*|^2).$$

By Dominated Convergence Theorem, if the mapping $x\mapsto\frac{r_\theta^2(x)}{P_{\theta^*}(x)|\theta-\theta^*|^2}$ is dominated by some L^1-function on $\mathcal{X}$ within some deleted neighborhood $0<|\theta-\theta^*|<\epsilon$, we have

$$\lim_{\theta\to\theta^*}\frac{1}{|\theta-\theta^*|^2}\int\frac{r_\theta^2(x)}{P_{\theta^*}(x)}\,dx = \int\lim_{\theta\to\theta^*}\frac{r_\theta^2(x)}{P_{\theta^*}(x)|\theta-\theta^*|^2}\,dx = 0.$$

We can require that for some $\epsilon>0$,

$$\int\sup_{\theta:0<|\theta-\theta^*|<\epsilon}\frac{r_\theta^2(x)}{P_{\theta^*}(x)|\theta-\theta^*|^2}\,dx<\infty.$$

Theorem 6.31 (Biased Cramér-Rao bound)**.** *Let $\Theta \subset \mathbb{R}$. Given an estimator $\widehat{\theta}$, assume that the function $\theta \mapsto \mathbb{E}_\theta[\widehat{\theta}]$ is continuously differentiable. Under regularity conditions, for the quadratic loss, the estimator $\widehat{\theta}$ satisfies*

$$R_{\theta^*}(\widehat{\theta}) = \mathbb{E}_{\theta^*}[(\widehat{\theta} - \theta^*)^2] \geq \frac{(1 + b'(\theta^*))^2}{\mathcal{I}(\theta^*)} + b(\theta^*)^2, \quad \forall \theta^* \in \Theta,$$

where $\mathcal{I}(\theta^*)$ is the Fisher information evaluated at $\theta^* \in \Theta$, and $b(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$ is the bias of $\widehat{\theta}$.

Proof. Using the Hammersley-Chapman-Robbins bound, we have

$$\mathrm{Var}_{\theta^*}(\widehat{\theta}) \geq \lim_{\theta \to \theta^*} \frac{\big(\mathbb{E}_{\theta^*}[\widehat{\theta}] - \mathbb{E}_\theta[\widehat{\theta}]\big)^2}{\chi^2(P_\theta \| P_{\theta^*})} = \lim_{\theta \to \theta^*} \frac{\big(\theta - \theta^* + b(\theta) - b(\theta^*)\big)^2}{(\theta^* - \theta)^2 \mathcal{I}(\theta^*) + o\big((\theta^* - \theta)^2\big)} = \frac{(1 + b'(\theta^*))^2}{\mathcal{I}(\theta^*)}.$$

The result then follows from bias-variance decomposition. □

Theorem 6.32 (General Cramér-Rao bound)**.** *Let $\Theta \subset \mathbb{R}^n$. Given an estimator $\widehat{\theta}$, assume that the function $\theta \mapsto \mathbb{E}_\theta[\widehat{\theta}]$ is continuously differentiable. Let $\phi(\theta) = \frac{\partial}{\partial \theta} \mathbb{E}_\theta[\widehat{\theta}]$. Then estimator $\widehat{\theta}$ satisfies*

$$\mathcal{I}(\theta^*) \succeq \phi(\theta^*) \operatorname{Cov}_{\theta^*}(\widehat{\theta})^{-1} \phi(\theta^*)^\top, \quad \textit{and} \quad \operatorname{Cov}_{\theta^*} \succeq \phi(\theta^*)(\widehat{\theta}) \mathcal{I}(\theta^*)^{-1} \phi(\theta^*)^\top,$$

where $A \succeq B$ denotes that $\xi^\top (A - B) \xi \geq 0$ for all $\xi \in \mathbb{R}^n$.

Proof. Fix any $\xi \in \mathbb{R}^n$. Since $\phi \mapsto \mathbb{E}_\theta[\widehat{\theta}]$ is continuously differentiable, we have

$$\mathbb{E}_{\theta^* + h\xi}[\widehat{\theta}] - \mathbb{E}_{\theta^*}[\widehat{\theta}] = h \xi^\top \phi(\theta^*) + o(h).$$

Plugging in this equation and (6.19) into the multivariate Hammersley-Chapman-Robbins bound, we have

$$\xi^\top \mathcal{I}(\theta^*) \xi + \frac{o(h^2)}{h^2} \geq \xi^\top \phi(\theta^*) \operatorname{Cov}_{\theta^*}(\widehat{\theta})^{-1} \phi(\theta^*)^\top \xi + \frac{o(h^2)}{h^2}$$

Letting $h \to 0$, we obtain the first bound. Furthermore, according to the univariate HCR bound,

$$\xi^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta}) \xi = \mathrm{Var}_{\theta^*}(\xi^\top \widehat{\theta}) \geq \frac{\big(\xi^\top \mathbb{E}_{\theta^*}[\widehat{\theta}] - \xi^\top \mathbb{E}_\theta[\widehat{\theta}]\big)^2}{\chi^2(P_\theta \| P_{\theta^*})}.$$

For any $\eta \in \mathbb{R}^n$, letting $\theta = \theta^* + h\eta$, we have

$$\xi^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta}) \xi \geq \frac{h^2 \big(\xi^\top \phi(\theta^*) \eta\big)^2 + o(h^2)}{h^2 \eta^\top \mathcal{I}(\theta^*) \eta + o(h^2)} \to \frac{\big(\xi^\top \phi(\theta^*) \eta\big)^2}{\eta^\top \mathcal{I}(\theta^*) \eta}, \quad \forall \eta \in \mathbb{R}^n.$$

Using the fact $\sup_{\eta \in \mathbb{R}^n \setminus \{0\}} \frac{\langle w, \eta \rangle}{\eta^\top M \eta} = w^\top M^{-1} w$ from linear algebra, we have

$$\xi^\top \operatorname{Cov}_{\theta^*}(\widehat{\theta}) \xi \geq \xi^\top \phi(\theta^*) \mathcal{I}(\theta^*)^{-1} \phi(\theta^*)^\top \xi.$$

Since $\xi \in \mathbb{R}^n$ is arbitrary, we obtain the second bound, concluding the proof. □

Remark. If $\widehat{\theta}$ is an unbiased estimator, $\phi(\theta) = \mathrm{Id}_n$. We have the unbiased Cramér-Rao bound:

$$\operatorname{Cov}_{\theta^*}(\widehat{\theta}) \succeq \mathcal{I}(\theta^*)^{-1}.$$

When the dimension $n = 1$, the Theorem reduces to the case in Theorem 6.31.

Bayesian case. From a Bayesian perspective, in each experiment, the true parameter θ is subject to a prior distribution π on Θ. The average risk of $\widehat{\theta}$ over the prior π is

$$R_\pi(\widehat{\theta}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\widehat{\theta}) = \mathbb{E}_{\theta\sim\pi}\left[(\widehat{\theta}-\theta)^2\right].$$

Given a prior distribution π, the Bayesian risk for π is defined as the infimum of the average risk:

$$R_\pi^* = \inf_{\widehat{\theta}} R_\pi(\widehat{\theta}) = \inf_{\widehat{\theta}} \mathbb{E}_{\theta\sim\pi}\left[(\widehat{\theta}-\theta)^2\right]$$

Theorem 6.33 (Bayesian Cramér-Rao bound)**.**

$$R_\pi^* = \inf_{\widehat{\theta}} R_\pi(\widehat{\theta}) \geq \frac{1}{\mathbb{E}_{\theta\sim\pi}[\mathcal{I}(\theta)] + \mathcal{I}(\pi)},$$

where $\mathcal{I}(\pi) = \int_\Theta \frac{\pi'}{\pi}$ *is the Fisher information of the prior.*

Proof. Consider the following comparison of experiments:

$$p : \pi \to \theta \xrightarrow{X\sim p_{X|\theta}} X \to \widehat{\theta}, \qquad q : \widetilde{\pi} \to \theta \xrightarrow{X\sim q_{X|\theta}} X \to \widehat{\theta}.$$

By the data processing inequality and the variational representation,

$$\chi^2(q_{\theta,X}\|p_{\theta,X}) \geq \chi^2(q_{\theta,\widehat{\theta}}\|p_{\theta,\widehat{\theta}}) \geq \chi^2(q_{\theta-\widehat{\theta}}\|p_{\theta-\widehat{\theta}}) \geq \frac{\left(\mathbb{E}_p[\theta-\widehat{\theta}] - \mathbb{E}_q[\theta-\widehat{\theta}]\right)^2}{\mathrm{Var}_p(\theta-\widehat{\theta})}$$

Let $\widetilde{\pi}$ be the prior obtained by shifting π by δ, i.e. $\widetilde{\pi}(\theta) = \pi(\theta-\delta)$. We choose $p_{X|\theta} = P_\theta$ and $q_{X|\theta} = P_{\theta-\delta}$. Then $p_X = q_X$, and $p_{\widehat{\theta}} = q_{\widehat{\theta}}$. Hence $\mathbb{E}_q[\theta-\widehat{\theta}] - \mathbb{E}_p[\theta-\widehat{\theta}] = \delta$, and

$$\mathrm{Var}_p(\theta-\widehat{\theta}) \geq \frac{\delta^2}{\chi^2(q_{\theta,X}\|p_{\theta,X})}. \tag{6.20}$$

On the other hand,

$$\begin{aligned}
\chi^2(q_{\theta,X}\|p_{\theta,X}) &= \int\int \frac{(q_{\theta,X}-p_{\theta,X})^2}{p_{\theta,X}}\,dx\,d\theta = \int\int \frac{\left[q_\theta(q_{X|\theta}-p_{X|\theta}) + (q_\theta-p_\theta)p_{X|\theta}\right]^2}{p_{\theta,X}}\,d\theta\,dx \\
&= \int \frac{q_\theta^2}{p_\theta}\,d\theta \int \frac{(q_{X|\theta}-p_{X|\theta})^2}{p_{X|\theta}}\,dx + \int \frac{(q_\theta-p_\theta)^2}{p_\theta}\,d\theta \int p_{X|\theta}\,dx + \int \frac{q_\theta(q_\theta-p_\theta)}{p_\theta}\,d\theta \int (q_{X|\theta}-p_{X|\theta})\,dx \\
&= \int \chi^2(q_{X|\theta}\|p_{X|\theta})\frac{q_\theta^2}{p_\theta}\,d\theta + \chi^2(q_\theta\|p_\theta).
\end{aligned}$$

According to Taylor's expansion, we have $\chi^2(q_{X|\theta}\|p_{X|\theta}) = \chi^2(P_{\theta-\delta}\|P_\theta) = \delta^2\mathcal{I}(\theta) + o(\delta^2)$, and $\chi^2(q_\theta\|p_\theta) = \chi^2(\pi(\cdot-\delta)\|\pi) = \delta^2\mathcal{I}(\pi) + o(\delta^2)$. Hence

$$\chi^2(q_{\theta,X}\|p_{\theta,X}) = \delta^2\mathbb{E}_{p_\theta}[\mathcal{I}(\theta)] + \delta^2\mathcal{I}(\pi) + o(\delta^2). \tag{6.21}$$

Combining (6.20) and (6.21), and letting $\delta \to 0$, we have

$$\mathbb{E}[(\theta-\widehat{\theta})^2] \geq \mathrm{Var}_{\theta\sim\pi}(\theta-\widehat{\theta}) \geq \frac{1}{\mathbb{E}_{\theta\sim\pi}[\mathcal{I}(\theta)] + \mathcal{I}(\pi)}.$$

Since $\widehat{\theta}$ is arbitrary, we obtain the desired bound for Bayesian risk. □

6.5 Application: Kernel Density Estimator

Setting. In this section, we consider the estimation of the density of a channel output Y. Assume that we are given a known channel $P_{Y|X}$, and let P_X be any unknown input distribution. Given observations $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_X$, the empirical distribution of X is

$$\widehat{P}_n = \sum_{j=1}^{n} \delta_{X_j}.$$

A natural estimate of the distribution of Y is $P_{Y|X} \circ \widehat{P}_n$. In many practical cases, the conditional density of Y given $X = x$ is of the form $P_{Y|X=x} = \phi(\cdot - x)$, where ϕ is a fixed density. For example, the additional channel $Y = X + Z$ satisfies this formula when $Z \sim \phi$ is independent of X. In this case, we can think of $P_{Y|X} \circ \widehat{P}_n$ as a *kernel density estimator* (KDE), whose density is

$$\widehat{p}_n(y) = (\phi * \widehat{P}_n)(y) = \frac{1}{n} \sum_{j=1}^{n} \phi(y - X_n).$$

Using the fact that $\mathbb{E}[\widehat{p}_n(y)] = p_Y(y)$ for all $y \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\left[D(P_{Y|X} \circ \widehat{P}_n \| P_X)\right] &= \mathbb{E}\left[\int \widehat{p}_n(y) \log \frac{\widehat{p}_n(y)}{p_X(y)}\, dy\right] \\
&= \mathbb{E}\left[\int \widehat{p}_n(y) \log \frac{\widehat{p}_n(y)}{p_Y(y)}\, dy\right] + \mathbb{E}\left[\int \widehat{p}_n(y) \log \frac{p_Y(y)}{p_X(y)}\, dy\right] \\
&= \mathbb{E}\left[\int \widehat{p}_n(y) \log \frac{\widehat{p}_n(y)}{p_Y(y)}\, dy\right] + \int \mathbb{E}\left[\widehat{p}_n(y)\right] \log \frac{p_Y(y)}{p_X(y)}\, dy \\
&= \mathbb{E}\left[D(P_{Y|X} \circ \widehat{P}_n \| P_Y)\right] + D(P_Y \| P_X)
\end{aligned}$$

In this section, we determine the convergence rate of the expected estimation error $\mathbb{E}[D(P_{Y|X} \circ \widehat{P}_n \| P_Y)]$.

Mutual χ^2-information. Consider two random variables $X, Y \sim P_{X,Y}$. Let P_X and P_Y be the marginal distributions of X and Y, respectively. The mutual information between X and Y is defined as

$$I(X;Y) = D\left(P_{XY} \| P_X P_Y\right) = \int P_{XY} \log \frac{P_{XY}}{P_X P_Y}.$$

Similarly, we define the mutual χ^2-information between X and Y to be

$$I_{\chi^2}(X;Y) = \chi^2\left(P_{XY} \| P_X P_Y\right).$$

More generally, if D_f is an f-divergence, the mutual f-information between X and Y is

$$I_f(X;Y) = D_f\left(P_{XY} \| P_X P_Y\right).$$

Theorem 6.34. *Under the above setting, we have the following upper bound for the estimation error:*

$$\mathbb{E}\left[D(P_{Y|X} \circ \widehat{P}_n \| P_Y)\right] \le \log\left(1 + \frac{1}{n} I_{\chi^2}(X;Y)\right);$$

Moreover, we have the following lower bound for the estimation error:

$$\lim_{n \to \infty} n\mathbb{E}\left[D(P_{Y|X} \circ \widehat{P}_n \| P_Y)\right] \ge \frac{1}{2} I_{\chi^2}(X;Y).$$

Proof. We first prove the upper bound. Note that

$$\begin{aligned}\mathbb{E}\left[\chi^2(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right] &= \mathbb{E}\left[\int\frac{(\widehat{p}_n(y)-p_Y(y))^2}{p_Y(y)}dy\right]\\ &= \mathbb{E}\left[\int\frac{1}{p_Y(y)}\left(\frac{1}{n}\sum_{j=1}^n p_{Y|X}(y|X_j)-p_Y(y)\right)^2 dy\right]\\ &= \int\frac{1}{p_Y(y)}\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^n p_{Y|X}(y|X_j)-p_Y(y)\right)^2\right]dy.\end{aligned}$$

For any $X_j\sim p_X$, we have

$$\mathbb{E}\left[p_{Y|X}(y|X_j)-p_Y(y)\right]=0,$$

and

$$\begin{aligned}\int\frac{\mathbb{E}\left[(p_{Y|X}(y|X_j)-p_Y(y))^2\right]}{p_Y(y)}\,dy &= \int\int\frac{(p_{Y|X}(y|x)-p_Y(y))^2}{p_Y(y)}p_X(x)\,dx\,dy\\ &= \int\int\frac{(p_{X,Y}(x,y)-p_X(x)p_Y(y))^2}{p_X(x)p_Y(y)}\,dx\,dy\\ &= I_{\chi^2}(X;Y).\end{aligned}$$

Hence

$$\mathbb{E}\left[\chi^2(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right]=\frac{1}{n}I_{\chi^2}(X;Y).$$

By (6.6) and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[D(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right] &\le \mathbb{E}\left[\log\left(1+\chi^2(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right)\right]\\ &\le \log\left(1+\mathbb{E}\left[\chi^2(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right]\right)=\log\left(1+\frac{1}{n}I_{\chi^2}(X;Y)\right).\end{aligned}$$

To prove the lower bound, we let $X^*\sim\mathrm{Unif}(X_1,\cdots,X_n)$, and let Y^* be the output of the channel $P_{Y|X}$ given the input X^*. Then $Y^*\overset{d}{=}Y$ marginally, and the joint distribution of $(X_{1:n},Y^*)$ is

$$p^*(x_1,\cdots,x_n,y)=p_X(x_1)\cdots p_X(x_n)\cdot\frac{1}{n}\sum_{j=1}^n\phi(y-x_j).$$

Then

$$\begin{aligned}I(X_{1:n};Y^*) &= \int\cdots\int\int p^*(x_1,\cdots,x_n,y)\log\frac{p^*(x_1,\cdots,x_n,y)}{p_X(x_1)\cdots p_X(x_n)p_Y(y)}\,dy\,dx_1\cdots dx_n\\ &= \int\cdots\int\int p_X(x_1)\cdots p_X(x_n)\cdot\frac{1}{n}\sum_{j=1}^n\phi(y-x_j)\log\frac{\sum_{j=1}^n\phi(y-x_j)}{np_Y(y)}\,dy\,dx_1\cdots dx_n\\ &= \mathbb{E}\left[\int\frac{1}{n}\sum_{j=1}^n\phi(y-X_j)\log\frac{\sum_{j=1}^n\phi(y-X_j)}{np_Y(y)}\,dy\right]\\ &= \mathbb{E}\left[D(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right].\end{aligned}\tag{6.22}$$

On the other hand, by the chain rule,

$$\begin{aligned} I(X_{1:n};Y^*) &= h(X_{1:n}) - h(X_{1:n}|Y^*) = \sum_{j=1}^n h(X_j) - \sum_{j=1}^n h(X_j|Y^*, X_{j-1}, \cdots, X_1) \\ &\geq \sum_{j=1}^n (h(X_j) - h(X_j|Y^*)) = \sum_{j=1}^n I(X_j;Y^*) = nI(X_1;Y^*). \end{aligned} \tag{6.23}$$

The joint distribution of X_1 and Y^* is

$$\begin{aligned} p^*(x_1,y) &= \int\cdots\int p_X(x_1)\cdots p_X(x_n)\cdot\frac{1}{n}\sum_{j=1}^n \phi(y-x_j)\,dx_2\cdots dx_n \\ &= \frac{1}{n}\int\cdots\int p_X(x_1)\cdots p_X(x_n)\phi(y-x_1)\,dx_2\cdots dx_n + \frac{1}{n}\sum_{j=2}^n\int\cdots\int p_X(x_1)\cdots p_X(x_n)\phi(y-x_j)\,dx_2\cdots dx_n \\ &= \frac{1}{n}p_X(x_1)\phi(y-x_1) + \frac{n-1}{n}p_X(x_1)p_Y(y). \end{aligned}$$

Hence by Theorem 6.12,

$$I(X_1;Y^*) = D\left(\frac{1}{n}P_{XY} + \frac{n-1}{n}P_XP_Y\middle\|P_XP_Y\right) = \frac{1}{2n^2}\chi^2(P_{XY}\|P_XP_Y) + o(n^{-2}). \tag{6.24}$$

Combining (6.22), (6.23) and (6.24),

$$\lim_{n\to\infty} n\mathbb{E}\left[D(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right] \geq \frac{1}{2}I_{\chi^2}(X;Y).$$

Thus we conclude the proof. □

Remark. We can summarize our result as follows:

- If $I_{\chi^2}(X;Y) < \infty$, we have $\mathbb{E}\left[D(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right] = O(n^{-1})$;
- If $I_{\chi^2}(X;Y) = \infty$, we have $\mathbb{E}\left[D(P_{Y|X}\circ\widehat{P}_n\|P_Y)\right] = \omega(n^{-1})$.

Discrete case. When X is a discrete random variable, we take $P_{Y|X}$ to be the identity δ_X to obtain the guarantee on the closeness between the empirical and the population distribution. This fact can be used to test whether the sample was truly generated by the distribution P_X.

Corollary 6.35. *Assume that P_X is supported on a discrete space $\mathcal{X}$. If $|\mathcal{X}| = \infty$, we have*

$$\lim_{n\to\infty} n\mathbb{E}\left[D(\widehat{P}_n\|P_X)\right] = \infty;$$

Otherwise,

$$\mathbb{E}\left[D(\widehat{P}_n\|P_X)\right] \leq \frac{|\mathcal{X}|-1}{n}.$$

Proof. We note that

$$I_{\chi^2}(X;Y) = \sum_{x\in\mathcal{X}}\frac{p_{X,Y}(x,x)^2}{p_X(x)p_Y(x)} - 1 = |\mathcal{X}| - 1.$$

The corollary then follows from Theorem 6.34. □