Elements of Information Theory

Sheng Yang
`sheng.yang@centralesupelec.fr`

> "*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.* "
>
> "A mathematical theory of communication", 1948, Claude Shannon (1916-2001)

References

- T. Cover and J. Thomas, "Elements of information theory"
- Y. Polyanskiy and Y. Wu, "Information theory"
- I. Csiszár and J. Körner, "Information theory: Coding theorems for discrete memoryless systems"
- R. Gallager, "Information theory and reliable communication"
- R. Yeung, "A first course in information theory"
- A. El Gamal and Y.-H. Kim, "Network information theory"

Notations and terms

Throughout this course, we use the following notations and terminologies.

$\mathbb{R}, \mathbb{C}, \mathbb{Z}, \mathbb{N}$	real, complex, integer, natural numbers
i	$\sqrt{-1}$
x^n	$(x_1, \ldots, x_n)$
$\lvert\mathcal{X}\rvert$	the size (cardinality) of the set $\mathcal{X}$
$\mathrm{Bern}(\lambda)$	Bernoulli (binary) random variable taking 1 with probability λ and 0 with probability $1-\lambda$
$H_2(\lambda)$	entropy of $\mathrm{Bern}(\lambda)$
:=	definition
Italic bold letters	Deterministic matrix $\boldsymbol{M}$ / vector $\boldsymbol{v}$
Non-italic capital (bold) letters	Random variables X / Random vectors **X**
$\mathrm{P}(\cdot)$	Probability measure
$\mathbb{E}\{\mathrm{X}\}$	Mean of the random variable X
$I(\mathrm{X};\mathrm{Y})$	Mutual information between X and Y
$\delta[\cdot]$	Kronecker delta function
$\mathbf{1}\{\cdot\}$	indicator function
$\log(x)$	Base-2 logarithm of x
$\boldsymbol{I}$	Identity matrix
$\lVert\boldsymbol{v}\rVert$	Euclidean ($\mathcal{L}_2$) norm of $\boldsymbol{v}$

Elements of Information Theory **2025-2026**

Lecture 1: Information Measures

Lecturer: S. Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

The goal is to introduce the basic measures of information on which we rely throughout the course.

Probability measure and preliminaries

In this course, we consider a probability space $(\Omega, \mathcal{H}, \mathrm{P})$ where Ω is the **sample space**, $\mathcal{H}$ is the σ-**algebra**, and P is the probability measure.

Let $(E, \mathcal{E})$ be a measurable space. A random variable X is a mapping $\Omega \to E$ that is measurable relative to $\mathcal{H}$ and $\mathcal{E}$. In particular, if E is countable, then we call X a **discrete random variable**. For any $A \in \mathcal{E}$, we can define $\mu(A) := \mathrm{P}(X(\omega) \in A)$, which is a probability measure on $(E, \mathcal{E})$. For discrete random variables, we call $\mathrm{P}_X(x) := \mathrm{P}(X(\omega) = x)$, $x \in E$, the **probability mass function (pmf)**.

Let μ and ν be two measures on a measurable space $(E, \mathcal{E})$, then ν is said to be **absolutely continuous** with respect to μ, denoted by $\nu \ll \mu$, if, for every set $A \in \mathcal{E}$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$. If $\nu \ll \mu$, then there exists a Radon-Nikodym derivative of ν with respect to μ, often denoted by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, such that

$$\int_A \nu(\mathrm{d}x) = \int_A \mu(\mathrm{d}x) \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x), \quad \forall A \in \mathcal{E}.$$

Note that the Radon-Nikodym derivative is positive and measurable, i.e., in $\mathcal{E}_+$. If $\nu \ll \mu \ll \lambda$, we have $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}$. If $\mathrm{P} \ll \mathrm{Q}$ are two probability measures defined on the same space $(E, \mathcal{E})$, and f is a P-measurable function, then we have the change of measure $\mathbb{E}_{\mathrm{P}} f(\mathrm{X}) = \mathbb{E}_{\mathrm{Q}}\big(f(\mathrm{X}) \frac{\mathrm{dP}}{\mathrm{dQ}}(\mathrm{X})\big)$.

Consider the case where E of the random variable X is the Euclidean space. If the probability measure μ is absolutely continuous with respect to the Lebesgue measure, then $p_X(x) := \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)$ is called the **probability density function (pdf)**. We call the random variable X a **continuous random variable**.

The mapping $(x, B) \mapsto K(x, B)$, $x \in E$ and $B \in \mathcal{F}$, is a **transition kernel** from $(E, \mathcal{E})$ to $(F, \mathcal{F})$. In particular, we consider **probability transition kernel** such that $K(x, \mathcal{F}) = 1$ for all $x \in E$. If μ is a probability measure in E, then $\pi f = \int_E \mu(\mathrm{d}x) \int_F K(x, \mathrm{d}y) f(x, y)$ defines the unique probability measure satisfying $\pi(A \times B) = \int_A \mu(\mathrm{d}x) K(x, B)$ for all $A \in \mathcal{E}, B \in \mathcal{F}$. Conversely, under some regularity conditions, for every probability measure on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$, there exist a proability measure μ on E and a transition probability kernel K from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ such that $\int_{E \times F} \pi(\mathrm{d}x \times \mathrm{d}y) f(x, y) = \int_E \mu(\mathrm{d}x) \int_F K(x, \mathrm{d}y) f(x, y)$, also known as “disintegration”. Throughout the course, we assume that such regularity conditions are met and ignore all measurability issues whenever possible. In most cases, we use $\mathrm{P}_{Y|X}$ to denote the transition probability kernel such that $\mathrm{P}_{Y|X=x}(\mathrm{d}y) = K(x, \mathrm{d}y)$. We use $\mathrm{P}_X \mathrm{P}_{Y|X}$ to denote the measure π on the product space such that $\int_{E \times F} \pi(\mathrm{d}x \times \mathrm{d}y) f(x, y) = \int_E \mathrm{P}_X(\mathrm{d}x) \int_F \mathrm{P}_{Y|X=x}(\mathrm{d}y) f(x, y)$. Both notations $\mathrm{P}_{Y|X}(\mathrm{d}y|x)$ and $\mathrm{P}_{Y|X=x}(\mathrm{d}y)$ are equivalent and can be used interchangeably. In particular, in the discrete case, $\mathrm{P}_{Y|X}(y|x) = \mathrm{P}_{Y|X=x}(y)$.

We use $\mathrm{P}_{Y|X} \circ \mathrm{P}_X$ to refer to the probability measure generated by the measure P_X and the transition kernel $\mathrm{P}_{Y|X}$:

$$(\mathrm{P}_{Y|X} \circ \mathrm{P}_X)(A) = \int_E \mathrm{P}_X(\mathrm{d}x) \mathrm{P}_{Y|X=x}(A).$$

It is the **marginalization** of the joint measure $\mathrm{P}_{Y|X}\mathrm{P}_X$. It can be regarded as the mixture of different distributions $\mathrm{P}_{Y|X=x}$ according to the measure P_X. In the discrete case, the transition probability kernel is a matrix and the pmf's are column vectors, and $\circ$ be be done as matrix multiplications. In most cases, we use P and Q to denote probability measures and p and q as the corresponding density function, i.e., pmf in the discrete case (w.r.t. the counting measure) and pdf in the continuous case (w.r.t. the Lebesgue measure). Finally, we remove the subscript of the pmf/pdf whenever ambiguity is not likely.

We will use the terms *distribution* and *probability measure* interchangeably. Unless the context makes it obvious, the underlying probability distribution will always be specified. Given a joint distribution P_{XY}, one can derive the marginals P_X and P_Y, as well as the conditional distributions (transition kernels) $\mathrm{P}_{Y|X}$ and $\mathrm{P}_{X|Y}$, such that

$$\mathrm{P}_{XY} = \mathrm{P}_X\mathrm{P}_{Y|X} = \mathrm{P}_Y\mathrm{P}_{X|Y}.$$

In this case, the notation $\mathrm{P}_X, \mathrm{P}_Y, \mathrm{P}_{Y|X}, \mathrm{P}_{X|Y}$ should always be understood as quantities induced by the given joint law P_{XY}. In more general situations, when several distributions or transition kernels are involved, we must be explicit about which objects are given and how others are constructed. For example, suppose we are given a conditional distribution $\mathrm{P}_{Y|X}$ but not a joint law. Together with two different input distributions P_X and Q_X, we can form two distinct joint distributions, $\mathrm{P}_{XY} := \mathrm{P}_X\mathrm{P}_{Y|X}, \mathrm{Q}_{XY} := \mathrm{Q}_X\mathrm{P}_{Y|X}$. In this case, P_{XY} and Q_{XY} are simply notations with explicit definition from the original distributions and transition kernels.

A real function f is called **convex** in a set $\mathcal{X}$ if for all $\lambda \in [0,1]$ and all $x_1, x_2 \in \mathcal{X}$, we have

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2),$$

which is quite easy to visualize. A real function is call **concave** if $-f$ is convex, or, equivalently, the above inequality changes direction. For example, $x \mapsto \log(x)$ is concave in $(0,\infty)$, $x \mapsto x\log(x)$ is convex in $(0,\infty)$.

One of the most important inequalities that we use in information theory is the so-called **Jensen's inequality**: Let $\mathrm{X} \in \mathcal{X}$ and f is convex in $\mathcal{X}$, then $\mathbb{E}f(\mathrm{X}) \ge f(\mathbb{E}\mathrm{X})$. For concave functions, we have $\mathbb{E}f(\mathrm{X}) \le f(\mathbb{E}\mathrm{X})$.

1.1 Entropy

Now, we introduce the first information measure. The **entropy** $H(\mathrm{X})$ of a discrete random variable $\mathrm{X} \sim \mathrm{P}_X$ is defined as

$$\boxed{H(\mathrm{X}) \equiv H(\mathrm{P}_X) := \mathbb{E}_{\mathrm{P}_X} \log \frac{1}{p(\mathrm{X})} = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)}.}$$

Sometimes, we use the notation $H(\mathrm{P}_X)$ to emphasize that entropy is a functional of the pmf. Intuitively, entropy measures the *uncertainty* (or amount of *information*) of a random variable.

Similarly, we define the **joint entropy** with the joint distribution $\mathrm{P}_{X_1\cdots X_n}$ (or, in short, P_{X^n}).

$$H(\mathrm{X}_1,\ldots,\mathrm{X}_n) := \mathbb{E}_{\mathrm{P}_{X^n}} \log \frac{1}{p(\mathrm{X}_1,\ldots,\mathrm{X}_n)} = \sum_{x_1\in\mathcal{X}_1} \cdots \sum_{x_n\in\mathcal{X}_n} p(x_1,\ldots,x_n) \log \frac{1}{p(x_1,\ldots,x_n)}.$$

We also define the **conditional entropy** as

$$\boxed{H(\mathrm{X}\,|\,\mathrm{Y}) = H(\mathrm{P}_{X|Y} \mid \mathrm{P}_Y) := \mathbb{E}_{y\sim\mathrm{P}_Y} H(\mathrm{P}_{X|Y=y}) = \mathbb{E}_{\mathrm{P}_Y\mathrm{P}_{X|Y}} \log \frac{1}{p(\mathrm{X}\,|\,\mathrm{Y})}.}$$

Here Y need not be discrete. If Y is also discrete, then

$$H(\mathrm{X}\,|\,\mathrm{Y}) := \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y) \log \frac{1}{p(x\,|\,y)}.$$

For discrete $X^n \sim P_{X^n}$, we have the following **chain rule**:

$$\boxed{H(X^n) = \sum_{i=1}^{n} H(X_i \mid X^{i-1})}$$

This can be shown with the chain rule of probability $p(x^n) = \prod_i p(x_i \mid x^{i-1})$:

$$\begin{aligned} H(X^n) &= \mathbb{E}_{P_{X^n}}\left[\log \frac{1}{p(X_1, \dots, X_n)}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{P_{X^n}}\left[\log \frac{1}{p(X_i \mid X^{i-1})}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{P_{X^i}}\left[\log \frac{1}{p(X_i \mid X^{i-1})}\right] \\ &= \sum_{i=1}^{n} H(X_i \mid X^{i-1}) \end{aligned}$$

Similarly, the conditional version also holds.

$$\boxed{H(X^n \mid Y) = \sum_{i=1}^{n} H(X_i \mid X^{i-1}, Y)}$$

1.2 Cross-entropy

Let us consider two probability measures P and Q defined on the same measurable space. Define the **cross-entropy** of P with respect to Q as[1]

$$H(P\|Q) := \begin{cases} \mathbb{E}_P \log \frac{1}{q(X)}, & P \ll Q \\ +\infty, & P \not\ll Q \end{cases}$$

And the **conditional cross-entropy**

$$H(P_{X|Y}\|Q_{X|Y} \mid P_Y) = \mathbb{E}_{y\sim P_Y}\left[H(P_{X|Y=y}\|Q_{X|Y=y})\right]$$

When $P_{X|Y=y} \ll Q_{X|Y=y}$, P_Y-a.s., we have

$$\boxed{H(P_{X|Y}\|Q_{X|Y} \mid P_Y) = \mathbb{E}_{P_Y P_{X|Y}}\left[\log \frac{1}{q(X|Y)}\right]}$$

As for entropy, the chain rule also holds for cross-entropy:

$$\boxed{H(P_{X^n}\|Q_{X^n}) = \sum_{i=1}^{n} H(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}} \mid P_{X^{i-1}})}$$

[1] Here, we use a non-standard notation $H(P\|Q)$ instead of $H(P, Q)$ to avoid confusion with the joint entropy $H(X, Y)$.

$$\begin{aligned}
H(\mathrm{P}_{X^n}\|\mathrm{Q}_{X^n}) &= \mathbb{E}_{\mathrm{P}_{X^n}}\left[\log\frac{1}{q(\mathrm{X}_1,\dots,\mathrm{X}_n)}\right] \\
&= \sum_{i=1}^n \mathbb{E}_{\mathrm{P}_{X^n}}\left[\log\frac{1}{q(\mathrm{X}_i \mid \mathrm{X}^{i-1})}\right] \\
&= \sum_{i=1}^n \mathbb{E}_{\mathrm{P}_{X^i}}\left[\log\frac{1}{q(\mathrm{X}_i \mid \mathrm{X}^{i-1})}\right] \\
&= \sum_{i=1}^n H(\mathrm{P}_{X_i|X^{i-1}}\|\mathrm{Q}_{X_i|X^{i-1}} \mid \mathrm{P}_{X^{i-1}})
\end{aligned}$$

And the conditional version also holds.

$$\boxed{H(\mathrm{P}_{X^n|Y}\|\mathrm{Q}_{X^n|Y}\,|\,\mathrm{P}_Y) = \sum_{i=1}^n H(\mathrm{P}_{X_i|X^{i-1}Y}\|\mathrm{Q}_{X_i|X^{i-1}Y} \mid \mathrm{P}_{X^{i-1}Y})}$$

1.3 Differential entropy, differential cross-entropy

There is an equivalent definition of the entropy for continuous random variables. We define the **differential entropy** of a continuous random variable $\mathrm{X} \sim \mathrm{P}_X$ with pdf p as

$$\boxed{h(\mathrm{X}) \equiv h(\mathrm{P}_X) := \mathbb{E}_{\mathrm{P}_X}\left[\log\frac{1}{p(\mathrm{X})}\right] = \int \mathrm{d}x\, p(x)\log\frac{1}{p(x)}.}$$

Similarly, we define the conditional differential entropy

$$h(\mathrm{X}\,|\,\mathrm{Y}) \equiv h(\mathrm{P}_{X|Y} \mid \mathrm{P}_Y) := \mathbb{E}_{y\sim\mathrm{P}_Y}\left[h(\mathrm{P}_{X|Y=y})\right] = \mathbb{E}_{\mathrm{P}_Y\mathrm{P}_{X|Y}}\left[\log\frac{1}{p(\mathrm{X}\,|\,\mathrm{Y})}\right].$$

Finally, if X^n has a density, we also define the **joint differential entropy**

$$h(\mathrm{X}^n) := \mathbb{E}\left[\log\frac{1}{p(\mathrm{X}^n)}\right].$$

The differential cross-entropy $h(\mathrm{P}\|\mathrm{Q})$ and the conditional version $h(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y)$ are defined as the cross-entropy but with the pdf's.

The chain rule holds as for entropy.

1.4 Divergence

Let us now introduce the divergence (aka. Kullback-Leibler divergence, KL divergence, or relative entropy). Consider two probability measures P and Q defined on the same measurable space. The **divergence** of P from Q, denoted by $D(\mathrm{P}\|\mathrm{Q})$, is defined as

$$\boxed{D(\mathrm{P}\|\mathrm{Q}) := \begin{cases} \mathbb{E}_{\mathrm{P}}\left[\log\frac{\mathrm{dP}}{\mathrm{dQ}}(\mathrm{X})\right] = \mathbb{E}_{\mathrm{Q}}\left[\frac{\mathrm{dP}}{\mathrm{dQ}}(\mathrm{X})\log\frac{\mathrm{dP}}{\mathrm{dQ}}(\mathrm{X})\right], & \mathrm{P} \ll \mathrm{Q} \\ +\infty, & \mathrm{P} \not\ll \mathrm{Q}. \end{cases}}$$

If P and Q are discrete distributions defined on the same set $\mathcal{X}$, then the divergence becomes

$$D(\mathrm{P}\|\mathrm{Q}) := \begin{cases} \mathbb{E}_{\mathrm{P}}\left[\log\frac{p(\mathrm{X})}{q(\mathrm{X})}\right] = \sum_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}, & \mathrm{P} \ll \mathrm{Q} \\ +\infty, & \mathrm{P} \not\ll \mathrm{Q} \end{cases}$$

where the Radon-Nikodym derivative is simply the ratio of two pmf's.

If P and Q are distributions of continuous random variables defined on the same set $\mathcal{X}$, then the divergence becomes

$$D(\mathrm{P}\|\mathrm{Q}) := \begin{cases} \mathbb{E}_{\mathrm{P}}\left[\log \frac{p(\mathrm{X})}{q(\mathrm{X})}\right] = \int_{\mathcal{X}} \mathrm{d}x p(x) \log \frac{p(x)}{q(x)}, & \mathrm{P} \ll \mathrm{Q} \\ +\infty, & \mathrm{P} \not\ll \mathrm{Q} \end{cases}$$

where the Radon-Nikodym derivative is the ratio of two pdf's.

As for cross-entropy, one can also define the **conditional divergence** for an arbitrary probability measure P_Y and arbitary probability transition kernels $\mathrm{P}_{X|Y}$ and $\mathrm{Q}_{X|Y}$,

$$D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) = \mathbb{E}_{y\sim\mathrm{P}_Y}\left[D(\mathrm{P}_{X|Y=y}\|\mathrm{Q}_{X|Y=y})\right]$$

When $\mathrm{P}_{X|Y=y} \ll \mathrm{Q}_{X|Y=y}$, P_Y-a.s., we have

$$\boxed{D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) = \mathbb{E}_{\mathrm{P}_Y\mathrm{P}_{X|Y}}\left[\log \frac{p(X \mid Y)}{q(X \mid Y)}\right]}$$

Exactly as for cross-entropy, the chain rule of divergence holds:

$$\boxed{D(\mathrm{P}_{X^n}\|\mathrm{Q}_{X^n}) = \sum_{i=1}^{n} D(\mathrm{P}_{X_i|X^{i-1}}\|\mathrm{Q}_{X_i|X^{i-1}} \mid \mathrm{P}_{X^{i-1}})}$$

and the conditional version

$$D(\mathrm{P}_{X^n|Y}\|\mathrm{Q}_{X^n|Y} \mid \mathrm{P}_Y) = \sum_{i=1}^{n} D(\mathrm{P}_{X_i|X^{i-1}Y}\|\mathrm{Q}_{X_i|X^{i-1}Y} \mid \mathrm{P}_{X^{i-1}Y})$$

For discrete distributions, the following relation between the entropy, cross-entropy, and divergence is straightforward

$$H(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) = H(\mathrm{P}_{X|Y} \mid \mathrm{P}_Y) + D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y).$$

And in particular,

$$\boxed{D(\mathrm{P}\|\mathrm{Q}) = H(\mathrm{P}\|\mathrm{Q}) - H(\mathrm{P}).}$$

The same holds for the continuous counterpart, by replacing entropy/cross-entropy by differential entropy and differential cross-entropy.

1.5 Mutual information

Let P_{XY} be the **joint distribution** of (X,Y) (e.g. probability measure of the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$. Similarly, let P_X and P_Y be the **marginal distributions** of X and Y, respectively. Further, let $\mathrm{P}_{X|Y}$ and $\mathrm{P}_{Y|X}$ be the transition probability kernels such that $\mathrm{P}_{XY} = \mathrm{P}_X\mathrm{P}_{Y|X} = \mathrm{P}_Y\mathrm{P}_{X|Y}$.[2]

Then, **mutual information** measures the dependence between X and Y:

$$\boxed{I(\mathrm{X};\mathrm{Y}) := D(\mathrm{P}_{XY}\|\mathrm{P}_X\mathrm{P}_Y) = D(\mathrm{P}_{X|Y}\|\mathrm{P}_X \mid \mathrm{P}_Y) = D(\mathrm{P}_{Y|X}\|\mathrm{P}_Y \mid \mathrm{P}_X),}$$

where the last two equalities can be proved with the chain rule of divergence. It compares the joint distribution to the one where X and Y are independent, or the conditional distributions $\mathrm{P}_{Y|X}$ and $\mathrm{P}_{X|Y}$ to the marginals P_Y and P_X. Sometimes mutual information is also denoted by $I(\mathrm{P}_X, \mathrm{P}_{Y|X})$, $I(\mathrm{P}_Y, \mathrm{P}_{X|Y})$, and $I(\mathrm{P}_{XY})$

$$\boxed{I(\mathrm{X};\mathrm{Y}) \equiv I(\mathrm{P}_{XY}) \equiv I(\mathrm{P}_X, \mathrm{P}_{Y|X}) \equiv I(\mathrm{P}_Y, \mathrm{P}_{X|Y}).}$$

[2] We assume that both kernels exists, which is guaranteed when both $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are *standard spaces*.

Indeed, mutual information is a functional of the joint distribution P_{XY}.

In particular, in the discrete case

$$I(\mathrm{X};\mathrm{Y}) = \mathbb{E}_{\mathrm{P}_{XY}} \log \frac{p_{XY}(\mathrm{X},\mathrm{Y})}{p_X(\mathrm{X})p_Y(\mathrm{Y})}.$$

It follows that

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}) &= H(\mathrm{X}) + H(\mathrm{Y}) - H(\mathrm{X},\mathrm{Y}) && (1.1)\\ &= H(\mathrm{X}) - H(\mathrm{X}\,|\,\mathrm{Y}) && (1.2)\\ &= H(\mathrm{Y}) - H(\mathrm{Y}\,|\,\mathrm{X}) && (1.3)\end{aligned}$$

The relationship between entropy and mutual information is best visualized with the Venn diagram below.

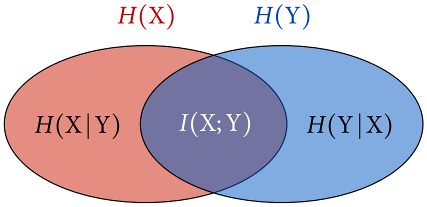

$H(\mathrm{X})$
$H(\mathrm{Y})$
$H(\mathrm{X}\,|\,\mathrm{Y})$
$I(\mathrm{X};\mathrm{Y})$
$H(\mathrm{Y}\,|\,\mathrm{X})$

Figure 1.1: The Venn diagram.

In the continuous case, the definition is the same with the pdf

$$I(\mathrm{X};\mathrm{Y}) = \mathbb{E}_{\mathrm{P}_{XY}} \log \frac{p_{XY}(\mathrm{X},\mathrm{Y})}{p_X(\mathrm{X})p_Y(\mathrm{Y})}.$$

It follows that

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}) &= h(\mathrm{X}) + h(\mathrm{Y}) - h(\mathrm{X},\mathrm{Y}) && (1.4)\\ &= h(\mathrm{X}) - h(\mathrm{X}\,|\,\mathrm{Y}) && (1.5)\\ &= h(\mathrm{Y}) - h(\mathrm{Y}\,|\,\mathrm{X}), && (1.6)\end{aligned}$$

similar to the discrete case.

If X is discrete and Y continuous, then

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}) &= D(\mathrm{P}_{XY}\|\mathrm{P}_X\mathrm{P}_Y) && (1.7)\\ &= D(\mathrm{P}_{Y|X}\|\mathrm{P}_Y\,|\,\mathrm{P}_X) && (1.8)\\ &= h(\mathrm{Y}) - h(\mathrm{Y}\,|\,\mathrm{X}). && (1.9)\end{aligned}$$

We also have

$$I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X}\,|\,\mathrm{Y}).$$

Note that in this case, although the conditional (differential) entropy exist, neither joint entropy nor joint differential entropy exists for (X,Y).

Let P_{XYZ} be some joint distribution of $(\mathrm{X},\mathrm{Y},\mathrm{Z})$. Then, we can define the **conditional mutual information** between X and Y given Z.

$$\boxed{I(\mathrm{X};\mathrm{Y}\,|\,\mathrm{Z}) \equiv I(\mathrm{P}_{XY|Z}\,|\,\mathrm{P}_Z) := D(\mathrm{P}_{XY|Z}\|\mathrm{P}_{X|Z}\mathrm{P}_{Y|Z}\,|\,\mathrm{P}_Z) = D(\mathrm{P}_{X|YZ}\|\mathrm{P}_{X|Z}\,|\,\mathrm{P}_{YZ}) = D(\mathrm{P}_{Y|XZ}\|\mathrm{P}_{Y|Z}\,|\,\mathrm{P}_{XZ}),}$$

where we replace the divergence in the definition of mutual information by the conditional divergence given Y.

The chain rule of mutual information is

$$\boxed{I(\mathrm{X};\mathrm{Y}^n) = \sum_{i=1}^{n} I(\mathrm{X};\mathrm{Y}_i \mid \mathrm{Y}^{i-1})}$$

Indeed, this can be proved from the (conditional) chain rule of divergence

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}^n) &= D(P_{Y^n|X}\|\mathrm{P}_{Y^n} \mid \mathrm{P}_X) \\ &= \sum_{i=1}^{n} D(P_{Y_i|XY^{i-1}}\|\mathrm{P}_{Y_i|Y^{i-1}} \mid \mathrm{P}_{XY^{i-1}}) \\ &= \sum_{i=1}^{n} I(\mathrm{X};\mathrm{Y}_i \mid \mathrm{Y}^{i-1}) \end{aligned}$$

The conditional version follows in the same way.

$$I(\mathrm{X};\mathrm{Y}^n \mid \mathrm{Z}) = \sum_{i=1}^{n} I(\mathrm{X};\mathrm{Y}_i \mid \mathrm{Y}^{i-1}\mathrm{Z})$$

1.6 Some properties of information measures

- General chain rule: Writing the chain rules in the same notational convention, we have

$$\begin{aligned} H(\mathrm{P}_{X^n}) &= \sum_{i=1}^{n} H(\mathrm{P}_{X_i|X^{i-1}} \,|\, \mathrm{P}_{X^{i-1}}), \quad h(\mathrm{P}_{X^n}) = \sum_{i=1}^{n} h(\mathrm{P}_{X_i|X^{i-1}} \,|\, \mathrm{P}_{X^{i-1}}) \\ H(\mathrm{P}_{X^n}\|\mathrm{Q}_{X^n}) &= \sum_{i=1}^{n} H(\mathrm{P}_{X_i|X^{i-1}}\|\mathrm{Q}_{X_i|X^{i-1}} \,|\, \mathrm{P}_{X^{i-1}}), \quad h(\mathrm{P}_{X^n}\|\mathrm{Q}_{X^n}) = \sum_{i=1}^{n} h(\mathrm{P}_{X_i|X^{i-1}}\|\mathrm{Q}_{X_i|X^{i-1}} \,|\, \mathrm{P}_{X^{i-1}}) \\ D(\mathrm{P}_{X^n}\|\mathrm{Q}_{X^n}) &= \sum_{i=1}^{n} D(\mathrm{P}_{X_i|X^{i-1}}\|\mathrm{Q}_{X_i|X^{i-1}} \,|\, \mathrm{P}_{X^{i-1}}) \\ I(\mathrm{P}_{X,Y^n}) &= \sum_{i=1}^{n} I(\mathrm{P}_{XY_i|Y^{i-1}}|\mathrm{P}_{Y^{i-1}}) \end{aligned}$$

- Positivity

$$\begin{aligned} H(\mathrm{P}) &\geq 0, \quad h(\mathrm{P}) \ngeq 0 \\ H(\mathrm{P}\|\mathrm{Q}) &\geq 0, \quad h(\mathrm{P}\|\mathrm{Q}) \ngeq 0 \\ D(\mathrm{P}\|\mathrm{Q}) &\geq 0 \Longrightarrow H(\mathrm{P}) \leq H(\mathrm{P}\|\mathrm{Q}),\ h(\mathrm{P}) \leq h(\mathrm{P}\|\mathrm{Q}) \\ I(\mathrm{P}_X, \mathrm{P}_{Y|X}) &\geq 0 \end{aligned}$$

Proof. For entropy, since probability is upper bounded by 1, entropy is nonnegative. Entropy is 0 if and only if the random variable is deterministic. The positivity does not hold for differential entropy. The same arguments apply for cross-entropy and differential cross-entropy.

For divergence, if $\mathrm{P} \not\ll \mathrm{Q}$, then $D(\mathrm{P}\|\mathrm{Q}) = +\infty > 0$. We assume therefore $\mathrm{P} \ll \mathrm{Q}$. Then, let us write $D(\mathrm{P}\|\mathrm{Q}) = \mathbb{E}_{\mathrm{Q}}\left(f(\frac{\mathrm{dP}}{\mathrm{dQ}})\right)$ where $f(x) := x \log x$. Finally, since $f(x)$ is strictly convex (check), we have $D(\mathrm{P}\|\mathrm{Q}) \geq f(\mathbb{E}_{\mathrm{Q}} \frac{\mathrm{dP}}{\mathrm{dQ}}) = f(1) = 0$, where we applied Jensen's inequality on f. The equality holds if and only if $\frac{\mathrm{dP}}{\mathrm{dQ}}$ is constant (Q-almost everywhere), impling that P = Q.

The positivity of mutual information is from that of divergence. It is 0 if and only if $\mathrm{P}_{XY} = \mathrm{P}_X\mathrm{P}_Y$, i.e., X and Y are independent.

□

- Conditioning

- Conditioning reduces (differential) entropy

$$H(X) \geq H(X \mid Y), \quad h(X) \geq h(X \mid Y)$$

- Conditioning increases divergence

$$D(P_{X|Y} \| Q_{X|Y} \mid P_Y) \geq D(\tilde{P}_X \| \tilde{Q}_X)$$

where $\tilde{P}_X$ and $\tilde{Q}_X$ are the marginals of $P_{X|Y}P_Y$ and $Q_{X|Y}P_Y$ respectively, i.e., $\tilde{P}_X = \mathbb{E}_{y \sim P_Y}[P_{X|Y=y}]$ and $\tilde{Q}_X = \mathbb{E}_{y \sim P_Y}[Q_{X|Y=y}]$.

Proof. Conditioning reduces entropy is from the positivity of mutual information, i.e., $H(X) - H(X \mid Y) = I(X;Y) \geq 0$. Similarly for differential entropy. For divergence, we have $D(P_{X|Y} \| Q_{X|Y} \mid P_Y) = D(P_{X|Y}P_Y \| Q_{X|Y}P_Y) = D(\tilde{P}_X \| \tilde{Q}_X) + D(\tilde{P}_{Y|X} \| \tilde{Q}_{Y|X} \mid \tilde{P}_X) \geq D(\tilde{P}_X \| \tilde{Q}_X)$, where we use the decompositions $P_{X|Y}P_Y = \tilde{P}_X\tilde{P}_{Y|X}$ and $Q_{X|Y}Q_Y = \tilde{Q}_X\tilde{Q}_{Y|X}$. □

- Convexity/Concavity
  - $P \mapsto H(P)$, $P \mapsto h(P)$ are both concave
  - $(P, Q) \mapsto D(P\|Q)$ is convex
  - $P_X \mapsto I(P_X, P_{Y|X})$ is concave
  - $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ is convex

Proof. Fix $\lambda \in [0,1]$. Let $S \sim P_S := \mathrm{Bern}(\lambda)$, i.e., $p_S(0) = \lambda$ and $p_S(1) = 1 - \lambda$.

Let $P_{X|S=0} := P_0$ and $P_{X|S=1} := P_1$. We have $P_X = (1-\lambda)P_0 + \lambda P_1$. $\lambda H(P_1) + (1-\lambda)H(P_0) = H(X \mid S) \leq H(X) = H(P_X)$, proving the concavity of $P \mapsto H(P)$ using conditioning reduces entropy.

For divergence, let $P_{X|S=k} := P_k$ and $Q_{X|S=k} := Q_k$ for $k = 0, 1$. Then, $\lambda D(P_1\|Q_1) + (1-\lambda)D(P_0\|Q_0) = (P_{X|S}\|Q_{X|S} \mid P_S) \geq D(P_X\|Q_X)$, proving the convexity of divergence using conditioning increases divergence.

For the concavity of mutual information, for the given P_0, P_1, and $P_{Y|X}$, let us set $P_{X|S=0} := P_0$ and $P_{X|S=1} := P_1$, and let $P_{Y|XS} = P_{Y|X}$, i.e., $P_{Y|X,S=0} = P_{Y|X,S=1} = P_{Y|X}$. Hence, the joint distribution is $P_{SXY} = P_S P_{X|S} P_{Y|X}$. The conditional mutual information

$$\begin{aligned}
I(X;Y \mid S) &= D(P_{XY|S}\|P_{X|S}P_{Y|S}|P_S) \\
&= \lambda D(P_{Y|X,S=0}\|P_{Y|S=0} \mid P_{X|S=0}) + (1-\lambda)D(P_{Y|X,S=1}\|P_{Y|S=1} \mid P_{X|S=1}) \\
&= \lambda D(P_{Y|X}\|P_{Y|S=0} \mid P_0) + (1-\lambda)D(P_{Y|X}\|P_{Y|S=1} \mid P_1) \\
&= \lambda I(P_0, P_{Y|X}) + (1-\lambda)I(P_1, P_{Y|X}).
\end{aligned}$$

On the other hand,

$$I(X;Y) = I(P_X, P_{Y|X}) = I(\lambda P_0 + (1-\lambda)P_1, P_{Y|X})$$

To finish the proof, we write

$$\begin{aligned}
I(X;Y) &= I(X;Y) + I(S;Y \mid X) \\
&= I(X, S; Y) \\
&= I(S;Y) + I(X;Y \mid S) \\
&\geq I(X;Y \mid S)
\end{aligned}$$

where the first equality holds since $I(S;Y \mid X) = 0$ due to the Markov chain $S \to X \to Y$. Indeed,

$$\begin{aligned}
I(S;Y \mid X) &= D(P_{Y|XS}\|P_{Y|X} \mid P_{XS}) \\
&= D(P_{Y|X}\|P_{Y|X} \mid P_{XS}) \\
&= 0
\end{aligned}$$

Finally, for the convexity of mutual information, we need to prove that given P_X $I(\mathrm{P}_X, \lambda \mathrm{P}^1_{Y|X} + (1-\lambda)\mathrm{P}^0_{Y|X}) \leq \lambda I(\mathrm{P}_X, \mathrm{P}^1_{Y|X}) + (1-\lambda) I(\mathrm{P}_X, \mathrm{P}^0_{Y|X})$ for any kernels $\mathrm{P}^0_{Y|X}$ and $\mathrm{P}^1_{Y|X}$ and $\lambda \in [0,1]$. We can prove it in two different ways. First, we can apply the convexity of divergence. Indeed,

$$\begin{aligned} I(\mathrm{P}_X, \lambda \mathrm{P}^1_{Y|X} + (1-\lambda)\mathrm{P}^0_{Y|X}) &= D(\lambda \mathrm{P}_X \mathrm{P}^1_{Y|X} + (1-\lambda)\mathrm{P}_X \mathrm{P}^0_{Y|X} \| \lambda \mathrm{P}_X \mathrm{P}^1_Y + (1-\lambda)\mathrm{P}_X \mathrm{P}^0_Y) \\ &\leq \lambda D(\mathrm{P}_X \mathrm{P}^1_{Y|X} \| \mathrm{P}_X \mathrm{P}^1_Y) + (1-\lambda) D(\mathrm{P}_X \mathrm{P}^0_{Y|X} \| \mathrm{P}_X \mathrm{P}^0_Y) \\ &= \lambda I(\mathrm{P}_X, \mathrm{P}^1_{Y|X}) + (1-\lambda) I(\mathrm{P}_X, \mathrm{P}^0_{Y|X}) \end{aligned}$$

where P^0_Y and P^1_Y are the marginals of $\mathrm{P}_X \mathrm{P}^0_{Y|X}$ and $\mathrm{P}_X \mathrm{P}^1_{Y|X}$, respectively.

The second way is to introduce the same S as before, let $\mathrm{P}_{Y|X,S=0} = \mathrm{P}^0_{Y|X}$ and $\mathrm{P}_{Y|X,S=1} = \mathrm{P}^1_{Y|X}$, so that $\mathrm{P}_{SXY} = \mathrm{P}_S \mathrm{P}_X \mathrm{P}_{Y|XS}$. Unlike the previous cases, here X and S are independent. It can be verified that $\mathrm{P}_{Y|X} = \lambda \mathrm{P}^0_{Y|X} + (1-\lambda)\mathrm{P}^1_{Y|X}$. Therefore, we have $I(\mathrm{P}_X, \lambda \mathrm{P}^1_{Y|X} + (1-\lambda)\mathrm{P}^0_{Y|X}) = I(\mathrm{X};\mathrm{Y})$. We also have $\lambda I(\mathrm{P}_X, \mathrm{P}^1_{Y|X}) + (1-\lambda) I(\mathrm{P}_X, \mathrm{P}^0_{Y|X}) = I(\mathrm{X};\mathrm{Y} \mid \mathrm{S})$. Therefore, it is enough to prove that $I(\mathrm{X};\mathrm{Y} \mid \mathrm{S}) \geq I(\mathrm{X};\mathrm{Y})$. To that end, apply the independence so that $I(\mathrm{X};\mathrm{S}) = 0$, and thus

$$\begin{aligned} I(\mathrm{X};\mathrm{Y} \mid \mathrm{S}) &= I(\mathrm{X};\mathrm{Y} \mid \mathrm{S}) + I(\mathrm{X};\mathrm{S}) \\ &= I(\mathrm{X};\mathrm{Y},\mathrm{S}) \\ &= I(\mathrm{X};\mathrm{Y}) + I(\mathrm{S};\mathrm{Y} \mid \mathrm{X}) \\ &\geq I(\mathrm{X};\mathrm{Y}). \end{aligned}$$

□

- Data processing inequality (DPI)

$$\begin{aligned} D(\mathrm{P}_X \| \mathrm{Q}_X) &\geq D(\mathrm{P}_{Y|X} \circ \mathrm{P}_X \| \mathrm{P}_{Y|X} \circ \mathrm{Q}_X) \\ I(\mathrm{P}_X, \mathrm{P}_{Y|X}) &\geq I(\mathrm{P}_X, \mathrm{P}_{Z|Y} \circ \mathrm{P}_{Y|X}) \end{aligned}$$

where $\mathrm{P}_{Y|X} \circ \mathrm{P}_X$ denotes the marginal on Y from joint distribution $\mathrm{P}_{Y|X}\mathrm{P}_X$; $\mathrm{P}_{Z|Y} \circ \mathrm{P}_{Y|X}$ is the conditional kernel defined as $\{\mathrm{P}_{Z|Y} \circ \mathrm{P}_{Y|X=x} : x \in \mathcal{X}\}$. In other words, if X → Y → Z, we have

$$I(\mathrm{X};\mathrm{Y}) \geq I(\mathrm{X};\mathrm{Z}).$$

The conditional version of DPI also holds in the same way.

Proof. Let $\tilde{P}_Y := \mathrm{P}_{Y|X} \circ \mathrm{P}_X$ and $\tilde{Q}_Y := \mathrm{Q}_{Y|X} \circ \mathrm{P}_X$ be the marginals of Y from $\mathrm{P}_{Y|X}\mathrm{P}_X$ and $\mathrm{P}_{Y|X}\mathrm{Q}_X$, respectively. We have

$$\begin{aligned} D(\mathrm{P}_X \| \mathrm{Q}_X) &= D(\mathrm{P}_X \| \mathrm{Q}_X) + D(\mathrm{P}_{Y|X} \| \mathrm{P}_{Y|X} \mid \mathrm{P}_X) \\ &= D(\mathrm{P}_X \mathrm{P}_{Y|X} \| \mathrm{Q}_X \mathrm{P}_{Y|X}) \\ &= D(\tilde{\mathrm{P}}_Y \tilde{\mathrm{P}}_{X|Y} \| \tilde{\mathrm{Q}}_Y \tilde{\mathrm{Q}}_{X|Y}) \\ &= D(\tilde{\mathrm{P}}_Y \| \tilde{\mathrm{Q}}_Y) + D(\tilde{\mathrm{P}}_{X|Y} \| \tilde{\mathrm{Q}}_{X|Y} \mid \tilde{\mathrm{P}}_Y) \\ &\geq D(\tilde{\mathrm{P}}_Y \| \tilde{\mathrm{Q}}_Y) \end{aligned}$$

Obviously, the conditional version also holds similarly.

For the mutual information, since X → Y → Z, there exist $\mathrm{P}_{Z|Y}$ such that $\mathrm{P}_{Z|X} = \mathrm{P}_{Z|Y} \circ \mathrm{P}_{Y|X}$ and $\mathrm{P}_Z = \mathrm{P}_{Z|Y} \circ \mathrm{P}_Y$. Thus,

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}) &= D(\mathrm{P}_{Y|X} \| \mathrm{P}_Y \mid \mathrm{P}_X) \\ &\geq D(\mathrm{P}_{Z|Y} \circ \mathrm{P}_{Y|X} \| \mathrm{P}_{Z|Y} \circ \mathrm{P}_Y \mid \mathrm{P}_X) \\ &= D(\mathrm{P}_{Z|X} \| \mathrm{P}_Z \mid \mathrm{P}_X) \\ &= I(\mathrm{X};\mathrm{Z}). \end{aligned}$$

□

1.7 Maximum entropy

In the following, we show how to apply the property $H(\mathrm{P}) \le H(\mathrm{P}\|\mathrm{Q})$ and $h(\mathrm{P}) \le h(\mathrm{P}\|\mathrm{Q})$ to find out maximum entropy in different cases.

1.7.1 Finite alphabet

Let $|\mathcal{X}| = M < \infty$. Fix $\mathrm{Q} = \mathrm{Unif}(\mathcal{X})$. Then, for any distribution P_X over $\mathcal{X}$,

$$\begin{aligned} H(\mathrm{P}_X) &\le H(\mathrm{P}_X\|\mathrm{Q}) \\ &= \mathbb{E}_{\mathrm{P}_X}[\log M] \\ &= \log M, \end{aligned}$$

where the equality holds when $\mathrm{P}_X = \mathrm{Q}$. Therefore, we show that uniform distribution maximizes entropy among all distributions with bounded aphabet size.

1.7.2 Continuous alphabet, finite second moment

Assume that P_X has a pdf and $\mathbb{E}\mathrm{X}^2 \le \sigma^2$. Fix $\mathrm{Q} \sim \mathcal{N}(0,\sigma^2)$, we have

$$\begin{aligned} h(\mathrm{P}_X) &\le h(\mathrm{P}_X\|\mathrm{Q}) \\ &= \mathbb{E}_{\mathrm{P}_X}\left[\log\sqrt{2\pi\sigma^2} + \frac{\mathrm{X}^2}{2\sigma^2}\log e\right] \\ &\le \frac{1}{2}\log(2\pi e\sigma^2), \end{aligned}$$

where the equalities hold when $\mathrm{P}_X = Q = \mathcal{N}(0,\sigma^2)$.

1.7.3 A general recipe

In general, if X has a density (e.g. pdf or pmf), then one can bound the (differential) entropy for a given expectation constraint $\mathbb{E}[c(\mathrm{X})] \le P$ for some positive function $x \mapsto c(x)$ such that the constraint can be satisfied with equality with some distribution.

Fix Q with

$$q(x) := \frac{e^{-\lambda c(x)}}{\int_{\mathcal{X}} e^{-\lambda c(x)}\mu(dx)} = \frac{1}{Z}e^{-\lambda c(x)},$$

where $\lambda \ge 0$ is such that $\mathbb{E}_{\mathrm{Q}}[c(\mathrm{X})] = P$.

Then, we have

$$\begin{aligned} h(\mathrm{P}_X) &\le h(\mathrm{P}_X\|\mathrm{Q}) \\ &= \mathbb{E}_{\mathrm{P}_X}[\log Z + \lambda\, c(\mathrm{X})\log e] \\ &\le \log Z + \lambda\, P\log e \end{aligned}$$

where the equalities holds when $\mathrm{P}_X = \mathrm{Q}$.

Exercises[3]

1. Entropy of functions [CT 2.2]. Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of $H(\mathrm{X})$ and $H(\mathrm{Y})$ if
   - $\mathrm{Y} = 2^{\mathrm{X}}$?
   - $\mathrm{Y} = \cos(\mathrm{X})$?
2. Conditional mutual information vs. unconditional mutual information [CT 2.6]. Give examples of joint random variables X, Y, and Z such that
   - $I(\mathrm{X};\mathrm{Y}\,|\,\mathrm{Z}) < I(\mathrm{X};\mathrm{Y})$.
   - $I(\mathrm{X};\mathrm{Y}\,|\,\mathrm{Z}) > I(\mathrm{X};\mathrm{Y})$.
3. Data processing [CT 2.15]. Let $\mathrm{X}_1 \to \mathrm{X}_2 \to \mathrm{X}_3 \to \cdots \to \mathrm{X}_n$ form a Markov chain in this order; that is, let
$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2\,|\,x_1)\cdots p(x_n\,|\,x_{n-1}).$$
   Reduce $I(\mathrm{X}_1;\mathrm{X}_2,\dots,\mathrm{X}_n)$ to its simplest form.
4. Infinite entropy. [CT 2.19] This problem shows that the entropy of a discrete random variable can be infinite. Let $A = \sum_{n=2}^{\infty}(n\log^2 n)^{-1}$. [It is easy to show that A is finite by bounding the infinite sum by the integral of $(x\log^2 x)^{-1}$.] Show that the integer-valued random variable X defined by $\mathrm{P}(\mathrm{X} = n) = (An\log^2 n)^{-1}$ for $n = 2, 3, \dots$ has $H(\mathrm{X}) = +\infty$.
5. Inequalities [CT 2.29]. Let X, Y, and Z be joint random variables. Prove the following inequalities and find conditions for equality.
   - $H(\mathrm{X},\mathrm{Y}\,|\,\mathrm{Z}) \geq H(\mathrm{X}\,|\,\mathrm{Z})$.
   - $I(\mathrm{X},\mathrm{Y};\mathrm{Z}) \geq I(\mathrm{X};\mathrm{Z})$.
   - $H(\mathrm{X},\mathrm{Y},\mathrm{Z}) - H(\mathrm{X},\mathrm{Y}) \leq H(\mathrm{X},\mathrm{Z}) - H(\mathrm{X})$.
   - $I(\mathrm{X};\mathrm{Z}\,|\,\mathrm{Y}) \geq I(\mathrm{Z};\mathrm{Y}\,|\,\mathrm{X}) - I(\mathrm{Z};\mathrm{Y}) + I(\mathrm{X};\mathrm{Z})$.
6. Convexity/Concavity of mutual information.
   - Let $(\mathrm{S},\mathrm{X},\mathrm{Y}) \sim P_{SXY} = P_S P_{X|S} P_{Y|X}$, i.e., $\mathrm{S} \to \mathrm{X} \to Y$ forms a Markov chain. Show that
   $$I(\mathrm{X};\mathrm{Y}) \geq I(\mathrm{X};\mathrm{Y}\,|\,\mathrm{S}).$$
     Use the above inequality to show that mutual information is concave in P_X for a fixed $P_{Y|X}$.
   - Let $(\mathrm{S},\mathrm{X},\mathrm{Y}) \sim P_{SXY} = P_S P_X P_{Y|X,S}$. Show that
   $$I(\mathrm{X};\mathrm{Y}) \leq I(\mathrm{X};\mathrm{Y}\,|\,\mathrm{S}).$$
     Use the above inequality to show that mutual information is convex in $P_{Y|X}$ for a fixed P_X.
7. Maximum entropy. [CT 2.30] Find the probability mass function $\mathrm{P}(x)$ that maximizes the entropy $H(\mathrm{X})$ of a nonnegative integer-valued random variable X subject to the constraint
$$E(\mathrm{X}) = \sum_{n=0}^{\infty} n\mathrm{P}(n) = A$$
   for a fixed value $A > 0$. Evaluate this maximum $H(\mathrm{X})$.
8. Relative entropy is not symmetric. [CT 2.35] Let the random variable X have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable:

Symbol	$\mathrm{P}(x)$	$\mathrm{Q}(x)$
a	$\frac{1}{2}$	$\frac{1}{3}$
b	$\frac{1}{4}$	$\frac{1}{3}$
c	$\frac{1}{4}$	$\frac{1}{3}$

[3] The citations “CT”, “CK” refer to Cover-Thomas, Csiszár-Körner, respectively.

9. Consider two joint distributions on $\{0,1\}^2$ represented as 2×2 tables (rows $= x \in \{0,1\}$, columns $= y \in \{0,1\}$):

$$\mathrm{P}_{XY} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} \end{bmatrix}, \qquad \mathrm{Q}_{XY} = \begin{bmatrix} \frac{1}{8} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{8} \end{bmatrix}.$$

   - Compute the marginals $\mathrm{P}_X, \mathrm{P}_Y$ and $\mathrm{Q}_X, \mathrm{Q}_Y$.
   - Compute the conditional kernels $\mathrm{P}_{X|Y}$ and $\mathrm{Q}_{X|Y}$.
   - Compute the entropies (in bits): $H(\mathrm{P}_X)$, $H(\mathrm{P}_Y)$, $H(\mathrm{P}_{XY})$, $H(\mathrm{P}_{X|Y} \mid \mathrm{P}_Y)$; and the corresponding quantities under Q.
   - Compute the divergences: $D(\mathrm{P}_{XY}\|\mathrm{Q}_{XY})$, $D(\mathrm{P}_X\|\mathrm{Q}_X)$, $D(\mathrm{P}_Y\|\mathrm{Q}_Y)$, and the conditional divergence $D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y)$. Verify the chain rule

$$D(\mathrm{P}_{XY}\|\mathrm{Q}_{XY}) = D(\mathrm{P}_Y\|\mathrm{Q}_Y) + D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y).$$

   - Compute the cross-entropies $H(\mathrm{P}_X\|\mathrm{Q}_X)$, $H(\mathrm{P}_Y\|\mathrm{Q}_Y)$, $H(\mathrm{Q}_X\|\mathrm{P}_X)$, $H(\mathrm{Q}_Y\|\mathrm{P}_Y)$ and the conditional cross-entropy $H(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y)$ Verify the identities

$$H(\mathrm{P}\|\mathrm{Q}) = H(\mathrm{P}) + D(\mathrm{P}\|\mathrm{Q}), \qquad H(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) = H(\mathrm{P}_{X|Y} \mid \mathrm{P}_Y) + D(\mathrm{P}_{X|Y}\|\mathrm{Q}_{X|Y} \mid \mathrm{P}_Y).$$

10. Entropy and pairwise independence. [CT 2.39] Let X, Y, Z be three binary Bernoulli(1/2) random variables that are pairwise independent; that is, $I(X;Y) = I(X;Z) = I(Y;Z) = 0$.

   - Under this constraint, what is the minimum value for constraint, $H(X,Y,Z)$?
   - Give an example achieving this minimum.

11. Mutual information of heads and tails [CT 2.43]

   - Consider a fair coin flip. What is the mutual information between the top and bottom sides of the coin?
   - A six-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)?

12. Finite entropy. [CT 2.45] Show that for a discrete random variable $X \in \{1,2,\ldots\}$, if $E \log X < \infty$, then $H(X) < \infty$.

13. Sequence length. [CT 2.48] How much information does the length of a sequence give about the content of a sequence? Suppose that we consider a Bernoulli(1/2) process $\{X_i\}$. Stop the process when the first 1 appears. Let N designate this stopping time. Thus, X^N is an element of the set of all finite-length binary sequences $\{0,1\}^* = \{0,1,00,01,10,11,000,\ldots\}$

   - Find $I(N;X^N)$.
   - Find $H(X^N \mid N)$.
   - Find $H(X^N)$.

   Let's now consider a different stopping time. For this part, again assume that $X \sim$ Bernoulli(1/2) but stop at time $N = 6$, with probability 1/3 and stop at time $N = 12$ with probability 2/3. Let this stopping time be independent of the sequence $X_1, X_2, \ldots, X_{12}$.

   - Find $I(N;X^N)$. Find $H(X \mid N)$. Find $H(X^N)$.

14. Function of variables from a Markov chain. [CK 3.7] Is it true that if $X_1 \to X_2 \to \cdots \to X_n$, and f is an arbitrary function on the common range of the X_i's, then $f(X_1) \to f(X_2) \to \cdots \to f(X_n)$? Give a counter example.

15. Mutual information. [Gallager 2.8] Consider an ensemble of sequences of N binary digits, $x_1, x_2, \ldots, x_N$. Each sequence containing an even number of 1's has probability 2^{-N+1} and each sequence with an odd number of 1's has probability zero. Find the mutual informations

$$I(X_1;X_2), I(X_2;X_3 \mid X_1), \ldots, I(X_{N-1};X_N \mid X_1, \ldots, X_{N-2}).$$

   Check your result for $N = 3$.

16. Memoryless source. Consider a sequence from the source contains independent symbols, i.e., $\mathrm{P}_{X^n} = \mathrm{P}_{X_1} \cdots \mathrm{P}_{X_n}$, also denoted by $\prod_{i=1}^n \mathrm{P}_{X_i}$. Show that

$$I(\mathrm{X}^n;\mathrm{Y}^n) \geq \sum_{i=1}^{n} I(\mathrm{X}_i;\mathrm{Y}_i),$$

for any $\mathrm{P}_{Y^n|X^n}$, with equality if and only if $\mathrm{P}_{X^n|Y^n} = \prod_i \mathrm{P}_{X_i|Y_i}$ (P_{Y^n}-almost surely). *Hint: Apply chain rule on* X^n*, then use the independence between* X_i *and* X^{i-1}.

17. Memoryless channels without feedback. We say that a channel is memoryless *without feedback* if $\mathrm{P}_{Y^n|X^n} = \prod_{i=1}^n \mathrm{P}_{Y_i|X_i}$. Show that in this case we have the Markov chain $\mathrm{Y}_i \to \mathrm{X}_i \to (\{\mathrm{X}_j, j \neq i\}, \{\mathrm{Y}_j, j \neq i\})$. Show that

$$I(\mathrm{X}^n;\mathrm{Y}^n) \leq \sum_{i=1}^{n} I(\mathrm{X}_i;\mathrm{Y}_i),$$

with equality if and only if $\mathrm{P}_{Y^n} = \prod_i \mathrm{P}_{Y_i}$. *Hint: Apply the chain rule and the Markov chain.*

Quiz (unique correct answer)

1. For a discrete random variable X taking on n possible values, which of the following statements is **TRUE** regarding its Shannon entropy $H(\mathrm{X})$?

   A) $H(\mathrm{X})$ is maximized when X is deterministic.

   B) $H(\mathrm{X})$ is minimized when X follows a uniform distribution.

   C) $H(\mathrm{X})$ is always non-negative and less than or equal to $\log_2 n$.

   D) $H(\mathrm{X})$ can be negative if X takes negative values.

   E) $H(\mathrm{X})$ measures the variance of X.

2. Which of the following expressions correctly represents the mutual information $I(\mathrm{X};\mathrm{Y})$ between two discrete random variables X and Y?

   A) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X},\mathrm{Y}) - H(\mathrm{X}) - H(\mathrm{Y})$

   B) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y}) - H(\mathrm{X},\mathrm{Y})$

   C) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y}) - H(\mathrm{X})$

   D) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X},\mathrm{Y}) + H(\mathrm{X}|\mathrm{Y})$

   E) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y}) + H(\mathrm{Y}|\mathrm{X})$

3. Suppose X and Y are independent discrete random variables. Which of the following is **TRUE**?

   A) $H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X})$

   B) $H(\mathrm{X}|\mathrm{Y}) = 0$

   C) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X})$

   D) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X})$

   E) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{Y})$

4. Which of the following is **TRUE** about the Kullback-Leibler divergence $D(\mathrm{P}\|\mathrm{Q})$ between two discrete probability distributions P and Q?

   A) $D(\mathrm{P}\|\mathrm{Q})$ is symmetric in P and Q.

   B) $D(\mathrm{P}\|\mathrm{Q}) \geq 0$, and equals zero if and only if $\mathrm{P} = \mathrm{Q}$ almost everywhere.

   C) $D(\mathrm{P}\|\mathrm{Q})$ is always finite.

   D) $D(\mathrm{P}\|\mathrm{Q})$ measures the variance between P and Q.

   E) $D(\mathrm{P}\|\mathrm{Q})$ is the mutual information between $\mathrm{X} \sim \mathrm{P}$ and $\mathrm{Y} \sim \mathrm{Q}$.

5. Which of the following information measures can be negative?

   A) Shannon entropy $H(\mathrm{X})$

   B) Mutual information $I(\mathrm{X};\mathrm{Y})$

   C) Conditional entropy $H(\mathrm{X}|\mathrm{Y})$

   D) Differential entropy $h(\mathrm{X})$ of a continuous random variable X

   E) Kullback-Leibler divergence $D(\mathrm{P}\|\mathrm{Q})$

6. For a continuous random variable X with probability density function $f(x)$, which of the following statements about the differential entropy $h(\mathrm{X})$ is **TRUE**?

   A) $h(\mathrm{X})$ is always non-negative.

   B) $h(\mathrm{X})$ is invariant under scaling of X.

   C) $h(\mathrm{X})$ increases when X is scaled by a factor $a > 1$.

D) $h(\mathrm{X})$ cannot be less than zero.

E) $h(\mathrm{X})$ measures the variance of X.

7. The entropy $H(\mathrm{X})$ of a Bernoulli random variable X with parameter p (i.e., $P(\mathrm{X} = 1) = p$) is given by:

A) $H(\mathrm{X}) = -p \log p$

B) $H(\mathrm{X}) = -p \log p - (1 - p) \log(1 - p)$

C) $H(\mathrm{X}) = p \log(1 - p) + (1 - p) \log p$

D) $H(\mathrm{X}) = -\log p$

E) $H(\mathrm{X}) = p$

8. Which of the following inequalities relates the conditional entropy $H(\mathrm{X}|\mathrm{Y})$ and the entropy $H(\mathrm{X})$ of two discrete random variables X and Y?

A) $H(\mathrm{X}|\mathrm{Y}) \geq H(\mathrm{X})$

B) $H(\mathrm{X}|\mathrm{Y}) \leq H(\mathrm{X})$

C) $H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X})$

D) $H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y})$

E) $H(\mathrm{X}|\mathrm{Y}) = H(\mathrm{X},\mathrm{Y}) - H(\mathrm{Y})$

9. The chain rule for entropy states that for discrete random variables X and Y:

A) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y}) + H(\mathrm{Y})$

B) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y})$

C) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{Y}|\mathrm{X}) - H(\mathrm{X})$

D) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y}) - H(\mathrm{Y})$

E) $H(\mathrm{X},\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{Y}|\mathrm{X})$

10. Which of the following distributions maximizes the entropy among all continuous distributions with a given variance?

A) Uniform distribution

B) Exponential distribution

C) Gaussian (Normal) distribution

D) Laplace distribution

E) Cauchy distribution

11. The conditional mutual information $I(\mathrm{X};\mathrm{Y}|\mathrm{Z})$ can be expressed in terms of entropies as:

A) $I(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = H(\mathrm{X}|\mathrm{Z}) + H(\mathrm{Y}|\mathrm{Z}) - H(\mathrm{X},\mathrm{Y}|\mathrm{Z})$

B) $I(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = H(\mathrm{X},\mathrm{Y},\mathrm{Z}) - H(\mathrm{Z})$

C) $I(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = H(\mathrm{X}|\mathrm{Y},\mathrm{Z}) - H(\mathrm{X}|\mathrm{Z})$

D) $I(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = H(\mathrm{X},\mathrm{Z}) + H(\mathrm{Y},\mathrm{Z}) - H(\mathrm{Z})$

E) $I(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = H(\mathrm{X},\mathrm{Y}) - H(\mathrm{Z})$

12. For a discrete random variable X and function $f(\mathrm{X})$, which of the following is **TRUE** regarding the entropy $H(f(\mathrm{X}))$?

A) $H(f(\mathrm{X})) \geq H(\mathrm{X})$

B) $H(f(\mathrm{X})) \leq H(\mathrm{X})$

C) $H(f(\mathrm{X})) = H(\mathrm{X})$

D) $H(f(X)) = 0$

E) $H(f(X)) = H(X|f(X))$

13. The Data Processing Inequality states that for random variables forming a Markov chain $X \to Y \to Z$, which of the following is **TRUE**?

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

B) $I(X;Z) \leq I(X;Y)$

C) $I(X;Y) = I(Y;Z)$

D) $I(X;Z) = I(X;Y) + I(Y;Z)$

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

14. Which distribution maximizes the entropy for a discrete random variable X with a fixed mean over the support $\{1, 2, \dots, n\}$?

A) Uniform distribution over $\{1, 2, \dots, n\}$

B) Geometric distribution

C) Binomial distribution

D) Discrete exponential distribution

E) Poisson distribution

15. Define the cross-entropy $H(P, Q) := \mathbb{E}_P \log \frac{1}{Q(X)}$. Which of the following is a property of the cross-entropy $H(P, Q)$ between two probability distributions P and Q?

A) $H(P, Q) = H(Q, P)$

B) $H(P, Q) \geq H(P)$

C) $H(P, Q) \leq H(P)$

D) $H(P, Q) = H(P) + D(P\|Q)$

E) $H(P, Q) = D(P\|Q)$

16. For two independent continuous random variables X and Y, the differential entropy of their sum $Z = X+Y$ satisfies:

A) $h(Z) = h(X) + h(Y)$

B) $h(Z) = h(X) - h(Y)$

C) $h(Z) = h(X) + h(Y) + \log 2\pi e$

D) $h(Z) \leq h(X) + h(Y)$

E) $h(Z) \geq h(X) + h(Y)$

17. Which of the following statements about mutual information $I(X;Y)$ is **TRUE**?

A) Mutual information $I(X;Y)$ is always less than or equal to zero.

B) Mutual information $I(X;Y)$ is zero if and only if X and Y are independent.

C) Mutual information $I(X;Y)$ is the same as conditional entropy $H(X|Y)$.

D) Mutual information $I(X;Y)$ is maximized when X and Y are independent.

E) Mutual information $I(X;Y)$ is always greater than the joint entropy $H(X,Y)$.

18. The conditional entropy $H(Y|X)$ can be expressed in terms of joint entropy $H(X,Y)$ and marginal entropy $H(X)$ as:

A) $H(Y|X) = H(X,Y) - H(X)$

B) $H(Y|X) = H(X) - H(X,Y)$

C) $H(Y|X) = H(Y) - H(X)$

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

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

19. For a continuous random variable X uniformly distributed over the interval $[a, b]$, the differential entropy $h(X)$ is:

A) $h(X) = \log(b - a)$

B) $h(X) = \log(b + a)$

C) $h(X) = \frac{1}{2}\log(b - a)$

D) $h(X) = -\log(b - a)$

E) $h(X) = \log\left(\frac{b}{a}\right)$

20. The **Chain Rule** for mutual information states that for random variables X, Y, Z:

A) $I(X; Y, Z) = I(X; Y) + I(X; Z)$

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

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

D) $I(X; Y, Z) = I(X; Y) - I(X; Z|Y)$

E) $I(X; Y, Z) = I(X; Y|Z) - I(X; Z|Y)$

Elements of Information Theory **2025-2026**

Lecture 2: Method of types

Lecturer: S. Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Types, type classes

Let $\mathcal{X}$ be a finite alphabet with $|\mathcal{X}| = M$, and consider a sequence $x^n \in \mathcal{X}^n$. The **type** of x^n is its **empirical pmf**, i.e.,

$$\hat{\mathrm{P}}_{x^n}(a) := \frac{1}{n}\left|\{i : x_i = a\}\right| = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i = a\},\ a \in \mathcal{X}.$$

Thus, $\hat{\mathrm{P}}_{x^n} := [\hat{\mathrm{P}}_{x^n}(a) : a \in \mathcal{X}]$ satisfies all the properties of a pmf. Similarly, we can define the **joint type** $\hat{\mathrm{P}}_{x^n,y^n}$ of (x^n, y^n) in $\mathcal{X}^n \times \mathcal{Y}^n$ by considering the couple (a, b) as a symbol, i.e.,

$$\hat{\mathrm{P}}_{x^n,y^n}(a,b) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i = a\}\mathbf{1}\{y_i = b\},\ a \in \mathcal{X}, b \in \mathcal{Y}.$$

We can verify that

$$\begin{aligned}\sum_{b\in\mathcal{Y}}\hat{\mathrm{P}}_{x^n,y^n}(a,b) &= \hat{\mathrm{P}}_{x^n}(a), \quad \forall a \in \mathcal{X}\\ \sum_{a\in\mathcal{X}}\hat{\mathrm{P}}_{x^n,y^n}(a,b) &= \hat{\mathrm{P}}_{y^n}(b), \quad \forall b \in \mathcal{Y}\end{aligned}$$

In words, $\hat{\mathrm{P}}_{x^n,y^n}$ is a joint pmf with marginal pmfs $\hat{\mathrm{P}}_{x^n}$ and $\hat{\mathrm{P}}_{y^n}$.

We say that $\hat{\mathrm{P}}$ is a type in $\mathcal{X}^n$ if it is a type of some sequence $x^n \in \mathcal{X}^n$, i.e., there exist $(n_1, \cdots, n_M) \in \mathbb{Z}_+^M$ with $n_1 + \cdots + n_M = n$ and $\hat{\mathrm{P}} = \left[\frac{n_1}{n}, \ldots, \frac{n_M}{n}\right]$. The **set of types** in $\mathcal{X}^n$ is denoted by $\mathcal{P}_n^{\mathcal{X}}$ (or simply $\mathcal{P}_n$). Specifically, we define

$$\mathcal{P}_n^{\mathcal{X}} := \left\{\left[\frac{n_1}{n}, \ldots, \frac{n_M}{n}\right] : \quad n_1 + \cdots + n_M = n, \quad n_i \in \mathbb{Z}_+, i = 1, \ldots, M\right\}$$

The set of types $\mathcal{P}_n^{\mathcal{X}}$ is a finite grid in the **probability simplex**

$$\mathcal{P}^{\mathcal{X}} := \left\{[p_1, \ldots, p_M] \in \mathbb{R}_+^M : \quad p_1 + \cdots + p_M = 1\right\}$$

All sequences with the same type form an equivalent class called **type class**. Specifically, the type class corresponding to a type $\hat{\mathrm{P}}$ is defined as

$$\mathcal{T}^{(n)}(\hat{\mathrm{P}}) := \{x^n \in \mathcal{X}^n : \hat{\mathrm{P}}_{x^n}(a) = \hat{\mathrm{P}}(a),\ \forall a \in \mathcal{X}\}.$$

2.2 Size and probability measure of type classes

In the following, we are interesting in finding out

- the number of type classes
- the size of each type class
- the probability of each type class, under a given probability measure

2.2.1 Number of type classes

We can show that the number of types is

$$\boxed{K_{n,M} := |\mathcal{P}_n^{\mathcal{X}}| = \binom{n+M-1}{M-1} \leq (n+1)^{M-1},}$$

since for each of the $M = |\mathcal{X}|$ components in a type $\hat{\mathrm{P}} \in \mathcal{P}_n$ we can at most have $n+1$ possible values, and only $M-1$ of the n_i's are free.

2.2.2 Size of each type class

If it is understood that each type class is defined for a given length n, we can remove the superscript for brevity. The size of each type class is

$$\boxed{|\mathcal{T}(\hat{\mathrm{P}})| = \binom{n}{n_1, \ldots, n_M} := \frac{n!}{\prod_{a \in \mathcal{X}} (n\hat{\mathrm{P}}(a))!}, \quad \hat{\mathrm{P}} \in \mathcal{P}_n}$$

Indeed, for any sequence $x^n \in \mathcal{T}(\hat{\mathrm{P}})$, the set of all $n!$ permutations can be partitioned according to the sequence after permutation. We can check that there are exactly $\prod_{a \in \mathcal{X}} (n\hat{\mathrm{P}}(a))!$ permutations that can transform x^n to $\tilde{x}^n \in \mathcal{T}(\hat{\mathrm{P}})$. Since there are exactly $|\mathcal{T}(\hat{\mathrm{P}})|$ different sequences $\tilde{x}^n$, we must have $n! = |\mathcal{T}(\hat{\mathrm{P}})| \prod_{a \in \mathcal{X}} (n\hat{\mathrm{P}}(a))!$

Let $\hat{\mathrm{P}}, \hat{\mathrm{Q}} \in \mathcal{P}_n$, then

$$\frac{|\mathcal{T}(\hat{\mathrm{P}})|}{|\mathcal{T}(\hat{\mathrm{Q}})|} \geq 2^{n(H(\hat{\mathrm{P}}) - H(\hat{\mathrm{Q}} \| \hat{\mathrm{P}}))}$$

In particular, if $\hat{\mathrm{P}}$ is uniform, then $|\mathcal{T}(\hat{\mathrm{P}})| \geq |\mathcal{T}(\hat{\mathrm{Q}})|$ for any $\hat{\mathrm{Q}}$.

Proof. From $|\mathcal{T}(\hat{\mathrm{P}})| = \frac{n!}{(n\hat{\mathrm{P}}(1))!\,(n\hat{\mathrm{P}}(2))!\cdots(n\hat{\mathrm{P}}(M))!}$ and $|\mathcal{T}(\hat{\mathrm{Q}})| = \frac{n!}{(n\hat{\mathrm{Q}}(1))!\,(n\hat{\mathrm{Q}}(2))!\cdots(n\hat{\mathrm{Q}}(M))!}$, we have

$$\begin{aligned}
\frac{|\mathcal{T}(\hat{\mathrm{P}})|}{|\mathcal{T}(\hat{\mathrm{Q}})|} &= \prod_{m=1}^{M} \frac{(n\hat{\mathrm{Q}}(m))!}{(n\hat{\mathrm{P}}(m))!} \\
&\geq \prod_{m=1}^{M} (n\hat{\mathrm{P}}(m))^{n(\hat{\mathrm{Q}}(m) - \hat{\mathrm{P}}(m))} \qquad \left(\frac{s!}{t!} \geq t^{s-t} \text{ for all } s, t \in \mathbb{Z}^+\right) \\
&= 2^{n(H(\hat{\mathrm{P}}) - H(\hat{\mathrm{Q}} \| \hat{\mathrm{P}}))}. \qquad (2.1)
\end{aligned}$$

If $\hat{\mathrm{P}}$ is uniform, then $H(\hat{\mathrm{P}}) - H(\hat{\mathrm{Q}} \| \hat{\mathrm{P}}) = \log M - \log M = 0$ and we have $\frac{|\mathcal{T}(\hat{\mathrm{P}})|}{|\mathcal{T}(\hat{\mathrm{Q}})|} \geq 1$. □

2.2.3 Probability of type classes

Let us use the notation P^n or Q^n to denote some product measure. It should be understood that

$$\mathrm{P}^n(x^n) = \prod_{i=1}^{n} \mathrm{P}(x_i), \quad x^n \in \mathcal{X}^n,$$

where P is a pmf[4] defined in $\mathcal{X}$. In words, $\mathrm{P}^n(x^n)$ is the probability of the sequence x^n under the assumption that $x_1, \ldots, x_n$ are i.i.d. ~ P. Such a notation makes the distribution of the random variables involved explicit.

[4] Here, we use the same notation for the probability measure and the pmf, which are the same for the discrete case.

If $x^n \in \mathcal{T}(\hat{\mathrm{P}})$, then its probability under the distribution Q^n, (i.e., the symbols are i.i.d. ~ Q), is exactly

$$\boxed{\mathrm{Q}^n(x^n) = 2^{-nH(\hat{\mathrm{P}}\|\mathrm{Q})}, \quad \forall\, x^n \in \mathcal{T}(\hat{\mathrm{P}}).} \tag{2.2}$$

Or, we can concisely write

$$\boxed{\mathrm{Q}^n(x^n) = 2^{-nH(\hat{\mathrm{P}}_{x^n}\|\mathrm{Q})}}$$

Now we see that under product pmf, sequences inside the same type class have the same probability, i.e., uniformly distributed inside each class. In particular, set $\mathrm{Q} = \hat{\mathrm{P}}$, we have from (2.2)

$$\boxed{\hat{\mathrm{P}}^n(x^n) = 2^{-nH(\hat{\mathrm{P}})}, \quad \forall\, x^n \in \mathcal{T}(\hat{\mathrm{P}}).} \tag{2.3}$$

Using $H(\mathrm{P}\|\mathrm{Q}) \geq H(\mathrm{P})$, we have the following.

$$\boxed{\hat{\mathrm{P}}^n(x^n) \geq \mathrm{Q}^n(x^n), \quad \forall\, x^n \in \mathcal{T}(\hat{\mathrm{P}})}$$

In words, a sequence of type $\hat{\mathrm{P}}$ has a larger pmf under the distribution $\hat{\mathrm{P}}^n$ than under any other distribution. However, under the same distribution $\hat{\mathrm{P}}^n$, a sequence of type $\hat{\mathrm{P}}$ does not necessarily has a pmf larger than a sequence outside of the type class. Namely, let $x^n \in \mathcal{T}(\hat{\mathrm{P}})$ and $\tilde{x}^n \in \mathcal{X}^n$ an arbitrary sequence, then

$$\hat{\mathrm{P}}^n(x^n) \ngeq \hat{\mathrm{P}}^n(\tilde{x}^n).$$

The message is that the most typical one is not necessarily the most probable one. Nevertheless, the conclusion is different if we consider the probability of an entire type class. For any Q, let $\mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}}))$ be the probability of the set of all the squences in the type class P under the measure Q^n, i.e.,

$$\mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}})) := \mathrm{Q}^n(\{x^n : x^n \in \mathcal{T}(\hat{\mathrm{P}})\}) = \sum_{x^n \in \mathcal{T}(\hat{\mathrm{P}})} \mathrm{Q}^n(x^n).$$

We also call it the Q^n-probability of the type class $\hat{\mathrm{P}}$. Since each sequence has the same probability, it is easily seen that

$$\mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}})) = |\mathcal{T}(\hat{\mathrm{P}})| 2^{-nH(\hat{\mathrm{P}}\|\mathrm{Q})} \tag{2.4}$$

$$\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) = |\mathcal{T}(\hat{\mathrm{P}})| 2^{-nH(\hat{\mathrm{P}})} \tag{2.5}$$

$$\boxed{\begin{aligned} \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) &\geq \mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}})), \quad \forall \hat{\mathrm{P}} \in \mathcal{P}_n, \mathrm{Q} \in \mathcal{P} \\ \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) &\geq \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{Q}})), \quad \forall \hat{\mathrm{P}}, \hat{\mathrm{Q}} \in \mathcal{P}_n \qquad (2.6) \end{aligned}}$$

In words, the probability of an entire type class is larger under the same distribution than under any other distribution. Furthermore, under the distribution $\hat{\mathrm{P}}^n$, the type class $\mathcal{T}(\hat{\mathrm{P}})$ has a higher probability than any other type classes.

Proof. The first inequality is straightforward from (2.4), (2.5), and $H(\mathrm{P}\|\mathrm{Q}) \geq H(\mathrm{P})$.

To prove the second one, taking the ratio, we have

$$\frac{\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}}))}{\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{Q}}))} = \frac{|\mathcal{T}(\hat{\mathrm{P}})|}{|\mathcal{T}(\hat{\mathrm{Q}})|} 2^{-n(H(\hat{\mathrm{P}}) - H(\hat{\mathrm{Q}}\|\hat{\mathrm{P}}))}. \tag{2.7}$$

From (2.1) and (2.7), we show that $\frac{\hat{\mathrm{Q}}^n(\mathcal{T}(\hat{\mathrm{Q}}))}{\hat{\mathrm{Q}}^n(\mathcal{T}(\hat{\mathrm{P}}))} \geq 1$. □

2.2.4 Size and probability of the type classes, revisited

Remarkably, the probability of the type classes can help us to obtain bounds on the size of each type class, although the latter does not depend on any probability distribution.

The size of each type class is bounded as

$$(n+1)^{-|\mathcal{X}|}2^{nH(\hat{\mathrm{P}})} \leq |\mathcal{T}(\hat{\mathrm{P}})| \leq 2^{nH(\hat{\mathrm{P}})}, \quad \hat{\mathrm{P}} \in \mathcal{P}_n.$$

For brevity, we can write[a]

$$|\mathcal{T}(\hat{\mathrm{P}})| \doteq 2^{nH(\hat{\mathrm{P}})}.$$

[a] where $\doteq$ means equality in exponent in the large n regime, i.e., $f(n) \doteq g(n)$ means

$$\lim_{n\to\infty} \frac{\log f(n)}{n} = \lim_{n\to\infty} \frac{\log g(n)}{n}.$$

Proof. The upper bound is straightforward from (2.3)

$$1 \geq \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) = |\mathcal{T}(\hat{\mathrm{P}})|2^{-nH(\hat{\mathrm{P}})}.$$

The lower bound can be obtained using (2.6):

$$\begin{aligned} 1 &= \sum_{\hat{\mathrm{Q}}\in\mathcal{P}_n} \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{Q}})) \\ &\leq \sum_{\hat{\mathrm{Q}}\in\mathcal{P}_n} \hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) \\ &= |\mathcal{P}_n|\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) \\ &\leq (n+1)^{|\mathcal{X}|}\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) \\ &\leq (n+1)^{|\mathcal{X}|}|\mathcal{T}(\hat{\mathrm{P}})|2^{-nH(\hat{\mathrm{P}})} \end{aligned}$$

□

The main message is that the size of a type class of $\hat{\mathrm{P}}$ is roughly $2^{nH(\hat{\mathrm{P}})}$, up to a polynomial factor in n. The size increases with the entropy given by the type. Intuitively, the more uniform the type, the larger the type class.

Then, as a simple consequence of the size bounds, combined with (2.4), we can have the following bounds on the probability measure of a given type class $\mathcal{T}(\hat{\mathrm{P}})$ under the distribution Q^n.

$$\boxed{(n+1)^{-|\mathcal{X}|}2^{-nD(\hat{\mathrm{P}}\|\mathrm{Q})} \leq \mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}})) \leq 2^{-nD(\hat{\mathrm{P}}\|\mathrm{Q})}}$$

implying

$$\mathrm{Q}^n(\mathcal{T}(\hat{\mathrm{P}})) \doteq 2^{-nD(\hat{\mathrm{P}}\|\mathrm{Q})}.$$

In particular, we have

$$\hat{\mathrm{P}}^n(\mathcal{T}(\hat{\mathrm{P}})) \geq (n+1)^{-|\mathcal{X}|}.$$

Intuitively, the probability of generating a sequence with a mismatched type $\mathrm{Q} \neq \hat{\mathrm{P}}$ decreases exponentially with n as $2^{-nD(\hat{\mathrm{P}}\|\mathrm{Q})}$, while the probability of generating a squence with a matched type decreases only polynomially with n not faster than $(n+1)^{-|\mathcal{X}|}$.

2.3 Strongly typical sequences

For each pmf P and $\varepsilon \in [0,1]$, we define the **strongly typical set**, or ε-typical n-sequences set, as

$$\begin{aligned}\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}) &:= \{x^n : \quad |\hat{\mathrm{P}}_{x^n}(a) - \mathrm{P}(a)| \le \varepsilon \mathrm{P}(a),\ \forall\, a \in \mathcal{X}\} \\ &= \{x^n : \quad (1-\varepsilon)\mathrm{P}(a) \le \hat{\mathrm{P}}_{x^n}(a) \le (1+\varepsilon)\mathrm{P}(a),\ \forall\, a \in \mathcal{X}\}\end{aligned}$$

Thus, the typical set contains sequences with types "close" to the given pmf P. In particular, when $\varepsilon = 0$, we have $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}) = \mathcal{T}(\mathrm{P})$ if P is a type and $\varnothing$ otherwise. Sometimes, we write $\mathcal{T}_\varepsilon^{(n)}(\mathrm{X})$ instead of $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ if $\mathrm{X} \sim \mathrm{P}$. Whenever confusion is unlikely, we can simply write $\mathcal{T}_\varepsilon^{(n)}$, or $\mathcal{T}_\varepsilon$.

The following **typical average lemma** shows the power of strong typicality.

For any function $g : \mathcal{X} \to \mathbb{R}$, and $\mathrm{X} \sim \mathrm{P}$

$$\mathbb{E}g(\mathrm{X}) - \delta(\varepsilon) \le \frac{1}{n}\sum_{i=1}^{n} g(x_i) \le \mathbb{E}g(\mathrm{X}) + \delta(\varepsilon), \quad \forall\, x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}), \tag{2.8}$$

where $\delta(\varepsilon) := \varepsilon\mathbb{E}|g(\mathrm{X})|$. Equivalently, we write

$$|\mathbb{E}_{\mathrm{P}} g(\mathrm{X}) - \mathbb{E}_{\hat{\mathrm{P}}_{x^n}} g(\mathrm{X})| \le \varepsilon \mathbb{E}_{\mathrm{P}}|g(\mathrm{X})|, \quad \forall\, x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}),$$

Typical sequences have the following properties.

1. **Sequence.** For every $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$,

$$2^{-n(1+\varepsilon)H(\mathrm{P})} \le \mathrm{P}^n(x^n) \le 2^{-n(1-\varepsilon)H(\mathrm{P})}, \tag{2.9}$$

   and

$$\begin{aligned} D(\hat{\mathrm{P}}_{x^n} \| \mathrm{P}) &\le \delta_D(\varepsilon) := \log(1+\varepsilon) \\ |H(\hat{\mathrm{P}}_{x^n}) - H(\mathrm{P})| &\le \delta_H(\varepsilon, \mathrm{P}) := \varepsilon H(\mathrm{P}) - \log(1-\varepsilon)\end{aligned}$$

2. **Probability of the set.** The probability of the typical set

$$\lim_{n\to\infty} \mathrm{P}^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) = 1, \quad \text{if } \lim_{n\to\infty} n\varepsilon^2 = \infty. \tag{2.10}$$

   More specifically, we have

$$\mathrm{P}^n(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \le \delta_T(\varepsilon, n, \mathrm{P}) := 2M^* e^{-2n\varepsilon^2 c_{\mathrm{P}}} \tag{2.11}$$

   where $c_P := \min_{a\in\mathcal{X}:\mathrm{P}(a)>0} \mathrm{P}(a)^2$ and $M^* := |\{a \in \mathcal{X} : \mathrm{P}(a) > 0\}|$.

3. **Size of the set.** The size of the typical set is bounded by

$$2^{n(1+\varepsilon)H(\mathrm{P})} \ge |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \ge 2^{n(1-\varepsilon)H(\mathrm{P}) - \delta_T'(\varepsilon,n,\mathrm{P})},$$

   where

$$\delta_T'(\varepsilon, n, \mathrm{P}) := -\log(1 - \delta_T(\varepsilon, n, \mathrm{P}))^+ \tag{2.12}$$

   that goes to 0 whenever $n\varepsilon^2 \to \infty$.

Proof. 1. The bounds (2.9) are straightforward from the typical average lemma where we let $g(x) = \log \mathrm{P}(x)$. And we apply the lemma on $\log \mathrm{P}^n(x^n) = \sum_i g(x_i)$. Note that $\mathbb{E}|\log \mathrm{P}(X)| = -\mathbb{E}\log \mathrm{P}(X) = H(\mathrm{P})$. The divergence bound is from the fact that $\hat{\mathrm{P}} \ll \mathrm{P}$ and that $\log \frac{\hat{\mathrm{P}}(a)}{\mathrm{P}(a)} \le \log(1+\varepsilon)$. Taking the expectation, we obtain the upper bound. Finally, for the entropy bounds, we write

$$\begin{aligned} H(\hat{\mathrm{P}}) &= \mathbb{E}_{\hat{\mathrm{P}}} \log \frac{1}{\hat{\mathrm{P}}(\mathrm{X})} \\ &\le \mathbb{E}_{\hat{\mathrm{P}}} \log \frac{1}{\mathrm{P}(\mathrm{X})} + \log \frac{1}{1-\varepsilon} \\ &\le (1+\varepsilon)\mathbb{E}_{\mathrm{P}} \log \frac{1}{\mathrm{P}(\mathrm{X})} + \log \frac{1}{1-\varepsilon} \\ H(\hat{\mathrm{P}}) &\ge \mathbb{E}_{\hat{\mathrm{P}}} \log \frac{1}{\mathrm{P}(\mathrm{X})} - \log \frac{1}{1+\varepsilon} \\ &\le (1-\varepsilon)\mathbb{E}_{\mathrm{P}} \log \frac{1}{\mathrm{P}(\mathrm{X})} - \log(1+\varepsilon) \end{aligned}$$

Therefore, $|H(\hat{\mathrm{P}}) - H(\mathrm{P})| \le \varepsilon H(\mathrm{P}) + \max\{\log(1+\varepsilon), -\log(1-\varepsilon)\} = \varepsilon H(\mathrm{P}) - \log(1-\varepsilon)$.

2. We check the probability

$$\mathrm{P}^n(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) = \mathrm{P}^n\left(\bigcup_{a\in\mathcal{X}} \left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathrm{X}_i = a\} - \mathrm{P}(a)\right| > \varepsilon \mathrm{P}(a)\right\}\right) \tag{2.13}$$

$$\le \sum_{a\in\mathcal{X}} \mathrm{P}^n\left(\left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathrm{X}_i = a\} - \mathrm{P}(a)\right| > \varepsilon \mathrm{P}(a)\right\}\right) \tag{2.14}$$

$$= \sum_{a\in\mathcal{X},\mathrm{P}(a)>0} \mathrm{P}^n\left(\left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathrm{X}_i = a\} - \mathrm{P}(a)\right| > \varepsilon \mathrm{P}(a)\right\}\right) \tag{2.15}$$

$$\le \sum_{a\in\mathcal{X},\mathrm{P}(a)>0} 2e^{-2n\varepsilon^2 \mathrm{P}(a)^2}, \tag{2.16}$$

where in (2.15), we use the fact that the event has probability 0 when $\mathrm{P}(a) = 0$; the last inequality is from Hoeffding's inequality.

Hoeffding's inequality Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be independent random variables such that $\mathrm{X}_i \in [a_i, b_i]$ almost surely. Define $\mathrm{S}_n := \frac{1}{n}\sum_{i=1}^n \mathrm{X}_i$ and $\Delta_n := \frac{1}{n}\sum_{i=1}^n (a_i - b_i)^2$. We have

$$\mathbb{P}\{|\mathrm{S_n} - \mathbb{E}(\mathrm{S}_n)| \ge t\} \le 2e^{-\frac{2nt^2}{\Delta_n}}. \tag{2.17}$$

Specifically, we have $\mathbf{1}\{\mathrm{X}_1 = a\}, \ldots, \mathbf{1}\{\mathrm{X}_n = a\}$ i.i.d., bounded in $[0,1]$, and with expectation $\mathrm{P}(a)$, so the average concentrates around the mean exponentially fast. Replacing each term in the sum by the dominant one, i.e., when $\mathrm{P}(a)$ is the smallest, we get the upper bound (2.11). (2.10) follows directly.

3. The upper bound is straightforward from $1 \ge \mathrm{P}^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \ge 2^{-n(1+\varepsilon)H(\mathrm{P}))}|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})|$.
From (2.11), $\mathrm{P}^n(\mathrm{X}^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \ge (1-\delta_T(\varepsilon, n, \mathrm{P}))^+$. Since we also have $\mathrm{P}^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \le |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})|\, 2^{-n(1-\varepsilon)H(\mathrm{P})}$, we prove that $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| 2^{-n(1-\varepsilon)H(\mathrm{P})} \ge (1 - 2|\mathcal{X}|e^{-2n\varepsilon^2 c_{\mathrm{P}}})^+$ which proves the lower bound on the size of the typical set.

□

From the above properties, we see that for a memoryless stationary source P^n, the probability mass is concentrated on a small subset $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ with approximately $2^{nH(\mathrm{P})}$ sequences, out of the set of all $M^n = 2^{n\log M}$ sequences. Moreover, every sequence (generated by the source P^n) inside $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ has approximately the same probability $2^{-nH(\mathrm{P})}$. This is also known as the **asymptotic equipartition property (AEP)**.

2.4 Jointly typical sequences, conditional typical sequences

In the exact same way, we can define the **jointly typical sequences** with respect to the joint pmf P_{XY}.

$$\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) := \{(x^n, y^n): \quad |\hat{\mathrm{P}}_{x^n,y^n}(a,b) - \mathrm{P}_{XY}(a,b)| \le \varepsilon \mathrm{P}_{XY}(a,b),\ \forall\, a \in \mathcal{X}, b \in \mathcal{Y}\}$$

We also define the **conditionally typical sequences** with respect to the joint pmf P_{XY}.

$$\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n) := \{x^n: \quad (x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\}.$$

By definition, we have

$$\boxed{(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) \quad \Leftrightarrow \quad x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n) \quad \Leftrightarrow \quad y^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,x^n)}$$

We have the following properties. [5]

1. If $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})$, then
   - $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X),\ y^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_Y)$
   - $\mathrm{P}_X^n(x^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{X})}$, $\mathrm{P}_Y^n(y^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{Y})}$, $\mathrm{P}_{XY}^n(x^n, y^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{X},\mathrm{Y})}$
   - $\mathrm{P}_{Y|X}^n(y^n|x^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{Y}\,|\,\mathrm{X})}$, $\mathrm{P}_{X|Y}^n(x^n|y^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{X}\,|\,\mathrm{Y})}$
2. $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})| \le 2^{n(1+\varepsilon)H(\mathrm{P}_{XY})}$, and $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})| \ge 2^{n(1-\varepsilon)H(\mathrm{P}_{XY}) - \delta_T(\varepsilon,n,\mathrm{P}_{XY})}$
3. For every $y^n \in \mathcal{Y}^n$,
$$|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n)| \le 2^{n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}.$$
4. **Conditional typicality lemma**. Let $\mathrm{P}_{X^n|Y^n} = \mathrm{P}_{X|Y}^n$. For every $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)$,
$$\lim_{n\to\infty} \mathrm{P}_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n)) = 1, \quad \text{if } \lim_{n\to\infty} n\varepsilon^2 = \infty.$$
   Specifically, we have
$$\mathrm{P}_{X^n|Y^n=y^n}(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n)) \le 2|\mathcal{X}|\,|\mathcal{Y}|e^{-n\varepsilon^2 c_P/2},$$
   where $c_P := \min_{a,b:\mathrm{P}_{XY}(a,b)>0} \mathrm{P}_{XY}(a,b)^2$.
5. Let $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)$. Then,
$$|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}\,|\,y^n)| \ge 2^{n(1-\varepsilon)H(\mathrm{X}|\mathrm{Y}) - \delta_T'(\varepsilon/2,n,\mathrm{P}_{XY})}. \tag{2.18}$$
   where δ_T' is defined in (2.12).
6. **Joint typicality lemma**.
   - For every $y^n \in \mathcal{Y}^n$,
$$\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) \le 2^{-n(I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X}))}. \tag{2.19}$$
   - Let $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}$. Then,
$$\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) \ge 2^{-n(I(\mathrm{X};\mathrm{Y}) + 2\varepsilon H(\mathrm{X})) - \delta_T'(\varepsilon/2,n,\mathrm{P}_{XY})}. \tag{2.20}$$

[5] With a slight abuse of notation, by $\mathrm{P}_X^n(x^n) \approx 2^{-n(1\pm\varepsilon)H(\mathrm{X})}$ we mean $\mathrm{P}_X^n(x^n) \in [2^{-n(1+\varepsilon)H(\mathrm{X})}, 2^{-n(1-\varepsilon)H(\mathrm{X})}]$.

Proof. 1. The first two items are straightforward from the definition of jointly typical sequences. Marginalizing, we verify that the sequences x^n and y^n are individually typical with the same ε. To prove the third item, we need to apply the typical average lemma to P_{XY} and the function $g(x, y) = \log \mathrm{P}_{X|Y}(x|y)$, using the fact that $\log(\mathrm{P}_{X^n|Y^n}(x^n \mid y^n)) = \sum_{i=1}^{n} \log \mathrm{P}_{X|Y}(x_i \mid y_i)$.

2. Same as for typical sequences. (cf. previous section)

3. Since $\mathrm{P}^n_{X|Y}(x^n|y^n) \geq 2^{-n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}$ from the first property, we have

$$1 \geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)} \mathrm{P}^n_{X|Y}(x^n|y^n) \tag{2.21}$$

$$\geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)} 2^{-n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})} \tag{2.22}$$

$$= |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)| 2^{-n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}. \tag{2.23}$$

4. It is enough to show that for any $(a, b) \in \mathcal{X} \times \mathcal{Y}$

$$\mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \mathrm{P}_{XY}(a,b)| > \varepsilon \mathrm{P}_{XY}(a,b)) \to 0 \tag{2.24}$$

and apply the union bound to finish the proof since $|\mathcal{X} \times \mathcal{Y}|$ is finite. When $\mathrm{P}_{XY}(a, b) = 0$, the above condition is verified trivially. In the following, we focus on the case where $\mathrm{P}_{XY}(a,b) > 0$, which is equivalent to $\mathrm{P}_X(a) > 0$, $\mathrm{P}_Y(b) > 0$, and $\mathrm{P}_{X|Y}(a|b) > 0$. Note that

$$\pi_{\mathrm{X}^n,y^n}(a,b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{(\mathrm{X}_i, y_i) = (a,b)\} \tag{2.25}$$

is a sum of n independent random variables. Indeed, X_i's are independent, each generated according to the distribution $\mathrm{P}_{X_i|Y_i=y_i}$. This implies that $\mathbf{1}\{(\mathrm{X}_i, y_i) = (a, b)\}$ are independent for different i's. The expectation of average is $\mathbb{E}_{\mathrm{P}_{X^n|Y^n=y^n}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{(\mathrm{X}_i, y_i) = (a,b)\}\right) = \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a|b)$. Due to the triangle inequality, we have

$$|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \mathrm{P}_{XY}(a,b)| \tag{2.26}$$

$$\leq |\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b)| + |\hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b) - \mathrm{P}_{XY}(a,b)| \tag{2.27}$$

$$\leq |\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b)| + \frac{\varepsilon}{2}\mathrm{P}_{XY}(a,b), \tag{2.28}$$

where we use the fact that $(1-\varepsilon/2)\mathrm{P}_Y(b) \leq \hat{\mathrm{P}}_{y^n}(b) \leq (1+\varepsilon/2)\mathrm{P}_Y(b)$ to establish the last inequality. Now, we can bound the probability

$$\mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \mathrm{P}_{XY}(a,b)| > \varepsilon \mathrm{P}_{XY}(a,b)) \tag{2.29}$$

$$\leq \mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a|b)| > (\varepsilon - \varepsilon/2)\mathrm{P}_{XY}(a,b)) \tag{2.30}$$

$$\leq 2e^{-n\varepsilon^2 \mathrm{P}_{XY}(a,b)^2/2} \tag{2.31}$$

according to Hoeffding's inequality (2.17). From the union bound, we have

$$\mathrm{P}_{X^n|Y^n=y^n}(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)) \leq \sum_{a,b:\mathrm{P}_{XY}(a,b)>0} \mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a,b) - \mathrm{P}_{XY}(a,b)| > \varepsilon \mathrm{P}_{XY}(a,b))$$

$$\leq 2|\mathcal{X}|\,|\mathcal{Y}|e^{-2n\varepsilon^2 c_P/2}, \tag{2.32}$$

where $c_P := \min_{a,b:\mathrm{P}_{XY}(a,b)>0} P_{XY}(a,b)^2$. Note that this bound goes to 0 when $n\varepsilon^2 \to \infty$. Hence, as long as $n\varepsilon^2 \to \infty$, we have the condition (2.24).

5. On the one hand, from the conditional typicality, we have $\mathrm{P}_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)) \geq (1-2|\mathcal{X}|\,|\mathcal{Y}|e^{-n\varepsilon^2 c_P/2})^+$. On the other hand, we have $\mathrm{P}_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)) \leq |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)| 2^{-n(1-\varepsilon)H(\mathrm{X}\,|\,\mathrm{Y})}$. Combining both inequalities, we prove the lower bound (2.18).

6. To show the upper bound, apply the upper bound on the size of the set $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)| \le 2^{n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}$ and the upper bound on the probability of $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)$ (implied by the fact that $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)$) under P_X^n, i.e., $\mathrm{P}_X^n(x^n) \le 2^{-n(1-\varepsilon)H(\mathrm{X})}$. We have

$$\begin{aligned}
\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) &= \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} \mathrm{P}_X^n(x^n) \qquad (2.33)\\
&\le \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} 2^{-n(1-\varepsilon)H(\mathrm{X})} \qquad (2.34)\\
&\le 2^{n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}2^{-n(1-\varepsilon)H(\mathrm{X})} \qquad (2.35)\\
&= 2^{-n(I(\mathrm{X};\mathrm{Y})-\varepsilon H(\mathrm{X})-\varepsilon H(\mathrm{X}|\mathrm{Y}))} \qquad (2.36)\\
&\le 2^{-n(I(\mathrm{X};\mathrm{Y})-2\varepsilon H(\mathrm{X}))}. \qquad (2.37)
\end{aligned}$$

To show the lower bound, apply the lower bound (2.18) on size of the set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)$ and the lower bound on the probability of $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)$ (implied by the fact that $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)$) under P_X^n. We have

$$\begin{aligned}
\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) &= \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} \mathrm{P}_X^n(x^n) \qquad (2.38)\\
&\ge \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} 2^{-n(1+\varepsilon)H(\mathrm{X})} \qquad (2.39)\\
&\ge 2^{n(1-\varepsilon)H(\mathrm{X}|\mathrm{Y})-\delta_T(\varepsilon/2,n,\mathrm{P}_{XY})}2^{-n(1+\varepsilon)H(\mathrm{X})} \qquad (2.40)\\
&\ge 2^{-n(I(\mathrm{X};\mathrm{Y})+2\varepsilon H(\mathrm{X}))-\delta_T(\varepsilon/2,n,\mathrm{P}_{XY})}. \qquad (2.41)
\end{aligned}$$

□

2.5 Covering lemma, packing lemma

From the joint typicality lemma, we can derive the following covering lemma and packing lemma that are useful for proving source coding and channel coding theorems, respectively.

Covering lemma. Let $\mathrm{X}^n(m) \sim \mathrm{P}_X^n$ for $m = 1, \ldots, 2^{nR_n}$ be mutually independent and also independent of some Y^n. Then,

$$\lim_{n\to\infty} \mathbb{P}\left\{\bigcup_{m=1,\ldots,2^{nR_n}} \left\{(\mathrm{X}^n(m), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\} = 1, \tag{2.42}$$

if the following conditions are satisfied:[a]

$$\begin{aligned}
\lim_{n\to\infty} n\varepsilon^2 &= \infty \qquad (2.43)\\
\lim_{n\to\infty} \mathbb{P}\left\{\mathrm{Y}^n \in \mathcal{T}_{\varepsilon/2}(\mathrm{P}_Y)\right\} &= 1 \qquad (2.44)\\
\lim_{n\to\infty} nR_n - nI(\mathrm{X};\mathrm{Y}) - 2n\varepsilon H(\mathrm{X}) &= \infty. \qquad (2.45)
\end{aligned}$$

[a] We don't require $\mathrm{Y}^n \sim \mathrm{P}_Y^n$. Instead, we only need that Y^n is typical with respect to P_Y with high probability.

Proof.

$$\mathbb{P}\left\{\bigcup_{m=1,\ldots,2^{nR_n}} \left\{(\mathrm{X}^n(m), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\} = \sum_{y^n} \mathrm{P}_{Y^n}(y^n)\mathbb{P}\left\{\bigcup_{m=1,\ldots,2^{nR_n}} \left\{(\mathrm{X}^n(m), y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\}$$

$$\geq \sum_{y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)} \mathrm{P}_{Y^n}(y^n) \mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ (\mathrm{X}^n(m), y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) \right\} \right\}$$

$$= \sum_{y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)} \mathrm{P}_{Y^n}(y^n) \mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ \mathrm{X}^n(m) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n) \right\} \right\}$$

For any given y^n, the events in the unions in the second probability are independent. From the probability lower bound (2.20), we have the following upper bound for $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)$,

$$\mathcal{P}\left\{ \mathrm{X}^n(m) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n) \right\} \leq 1 - 2^{-n(I(\mathrm{X};\mathrm{Y})+2\varepsilon H(\mathrm{X})) - \delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})}.$$

Applying the upper bound $1 - x \leq e^{-x}$, and the independence between $\mathrm{X}^n(m)$, we have

$$\mathcal{P}\left\{ \bigcap_{m=1,\ldots,2^{nR_n}} \left\{ \mathrm{X}^n(m) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n) \right\} \right\} \leq \exp\left(-2^{nR} 2^{-n(I(\mathrm{X};\mathrm{Y})+2\varepsilon H(X)) - \delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})} \right) \tag{2.46}$$

$$= \exp\left(-2^{n(R_n - I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X})) - \delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})} \right). \tag{2.47}$$

Therefore, we have

$$\mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ (\mathrm{X}^n(m), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) \right\} \right\} \geq \sum_{y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)} \mathrm{P}_{Y^n}(y^n) \left(1 - \exp\left(-2^{n(R_n - I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X})) - \delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})} \right) \right)$$

$$= \mathrm{P}_{Y^n}(\mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)) \left(1 - \exp\left(-2^{n(R_n - I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X})) - \delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})} \right) \right).$$

From conditions (2.43), (2.44), and (2.45), we have $\delta'_T(\varepsilon/2, n, \mathrm{P}_{XY}) \to 0$, $\mathrm{P}_{Y^n}(\mathcal{T}_{\varepsilon/2}^{(n)}(\mathrm{P}_Y)) \to 1$, and $n(R_n - I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X})) \to \infty$. It follows that the above probability lower bound goes to 1. □

Packing lemma. Let $\mathrm{X}^n(m) \sim \mathrm{P}_X^n$ for $m = 1, \ldots, 2^{nR_n}$ (not necessarily independent for different m) be independent of an arbitrary random sequence Y^n. Then,

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ (\mathrm{X}^n(m), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) \right\} \right\} = 0, \tag{2.48}$$

if

$$\lim_{n\to\infty} nR_n - nI(\mathrm{X};\mathrm{Y}) + 2n\varepsilon H(\mathrm{X}) = -\infty.$$

Proof. It is enough to apply the union bound and the probability upper bound (2.19). Indeed, from the union bound, we have

$$\mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ (\mathrm{X}^n(m), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}) \right\} \right\} = \mathbb{P}\left\{ \bigcup_{m=1,\ldots,2^{nR_n}} \left\{ \mathrm{X}^n(m) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \,|\, \mathrm{Y}^n) \right\} \right\} \tag{2.49}$$

$$\leq \sum_{m=1}^{2^{nR_n}} \mathbb{P}\left\{ \mathrm{X}^n(m) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \,|\, \mathrm{Y}^n) \right\} \tag{2.50}$$

$$\leq 2^{nR_n - nI(\mathrm{X};\mathrm{Y}) + 2n\varepsilon H(\mathrm{X})}, \tag{2.51}$$

where the last inequality is from the joint typicality lemma (2.19). The upper bound goes to 0 if

$$\lim_{n\to\infty} nR_n - nI(\mathrm{X};\mathrm{Y}) + 2n\varepsilon H(\mathrm{X}) = -\infty.$$

□

Exercises

1. Number of types [CK 2.1]. Show that the exact number of types is

$$|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1}.$$

Hint: Consider n stars on a line and put $|\mathcal{X}| - 1$ bars on the line to partition the stars.

2. Size of a type class. Let $x^n, \tilde{x}^n \in \mathcal{T}(\hat{\mathrm{P}})$, i.e., have the same type P.

    - Show that there are exactly $\prod_{a\in\mathcal{X}}(n\hat{P}(a))!$ permutations f such that $f(x^n) = \tilde{x}^n$.
    - Show that the number of sequences in $\mathcal{T}(\hat{\mathrm{P}})$ is

    $$|\mathcal{T}(\hat{\mathrm{P}})| = \frac{n!}{\prod_{a\in\mathcal{X}}(n\hat{\mathrm{P}}(a))!},$$

    which is also known as the multinomial coefficient

    $$\binom{n}{n\hat{\mathrm{P}}(a_1), \ldots, n\hat{\mathrm{P}}(a_M)}$$

    where $\mathcal{X} := \{a_1, \ldots, a_M\}$.

3. Asymptotic size of a type class [CK 2.2]. Prove that the size of $\mathcal{T}^{(n)}(\hat{\mathrm{P}})$ is of order of magnitude $n^{-\frac{s(\hat{\mathrm{P}})-1}{2}} 2^{nH(\hat{\mathrm{P}})}$, where $s(P)$ is the number of elements $a \in \mathcal{X}$ with $\mathrm{P}(a) > 0$. More precisely, show that

$$\log|\mathcal{T}^{(n)}(\hat{\mathrm{P}})| = nH(\hat{\mathrm{P}}) - \frac{s(\hat{\mathrm{P}}) - 1}{2}\log(2\pi n) - \frac{1}{2}\sum_{a:\,\hat{\mathrm{P}}(a)>0}\log\hat{\mathrm{P}}(a) - \frac{\theta(k,\hat{\mathrm{P}})}{12\ln 2}s(\hat{\mathrm{P}})$$

where $0 \le \theta(k, \hat{\mathrm{P}}) \le 1$. *Hint: Use Robbins' sharpening of Stirling's formula:*

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12(n+1)}} \le n! \le \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n}},$$

noticing that $\hat{\mathrm{P}}(a) \ge \frac{1}{n}$ whenever $\hat{\mathrm{P}}(a) > 0$.

4. Consider the alphabet $\mathcal{X} = \{0, 1, 2\}$ and $n = 6$.

    - How many *type classes* (distinct empirical distributions) are there?
    - For a given type $\hat{\mathrm{P}}$ with symbol counts (n_0, n_1, n_2) satisfying $n_0 + n_1 + n_2 = 6$, how many sequences belong to that type class?
    - Let

    $$\mathrm{Q} = \left[\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}\right].$$

    What is the *most probable sequence* under Q^n?
    - What is the *most probable type class* under Q^n?

5. Let P and Q be two pmf's defined on $\mathcal{X}$ with

$$|\mathrm{Q}(a) - \mathrm{P}(a)| \le \varepsilon\mathrm{P}(a), \quad a \in \mathcal{X}.$$

    - Show that $\left|H(\mathrm{P}) - E_{\mathrm{Q}}\log\frac{1}{\mathrm{P(X)}}\right| \le \varepsilon H(\mathrm{P})$
    - Show that $\left|E_{\mathrm{Q}}\log\frac{1}{\mathrm{P(X)}} - H(\mathrm{Q})\right| \le \log\frac{1}{1-\varepsilon}$
    - Show that $|H(\mathrm{Q}) - H(\mathrm{P})| \le \delta(\varepsilon)$ for some $\delta(\varepsilon) \to 0$ when $\varepsilon \to 0$. *Hint: Apply the triangle inequality.*

6. Universal source coding. Consider a sequence $x^n \in \mathcal{X}^n$ where $\mathcal{X}$ is a finite alphabet. One can encode the sequence into two parts: the first part indicates the type $\hat{\mathrm{P}}_{x^n}$ of x^n, the second part indicates the exact sequence within the type class $\mathcal{T}(\hat{\mathrm{P}}_{x^n})$.

    - Argue that this encoding scheme does not depend on the distribution of the source sequence.

- What is the length of the encoded binary sequence. corresponding to x^n?
- What is the expected length of the encoded sequence if the source is i.i.d. P?
- Show that in this case the encoding rate, defined as the expected length devided by n, is $H(\mathrm{P})$.

7. Consider a binary source with $\mathrm{P} = [0.2,\ 0.8]$. Let $n = 1000$ and $\varepsilon = 0.2$

    - Provide an upper bound that the probability that $\mathrm{X}^n \sim \mathrm{P}^n$ is not inside typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$.
    - Provide an upper bound on the size of the typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$.

Quiz (unique correct answer)

1. The **type** $\hat{\mathrm{P}}_{x^n}$ of a sequence x^n of length n over a finite alphabet $\mathcal{X}$ is defined as:

   A) The cumulative distribution function of x^n.
   B) The frequency of occurrence of each symbol in $\mathcal{X}$ within x^n.
   C) The expected value of x^n over $\mathcal{X}$.
   D) The joint distribution of x^n and $\mathcal{X}$.
   E) The probability distribution that minimizes the KL divergence to x^n.

2. Which of the following statements is **TRUE** about the relationship between types and sequences?

   A) Two sequences of the same type have the same empirical distribution.
   B) Two sequences of the same type must be identical.
   C) Sequences of different types can have the same empirical distribution.
   D) The number of types decreases exponentially with sequence length.
   E) Types are only defined for continuous random variables.

3. For a finite alphabet $\mathcal{X}$ with $|\mathcal{X}| = M$, the number of possible types for sequences of length n is:

   A) n^M
   B) $(n+1)^M$
   C) $\binom{n+M-1}{M-1}$
   D) M^n
   E) $\frac{n!}{M!}$

4. The method of types is particularly useful because:

   A) It allows us to bound probabilities involving sequences by considering their types.
   B) It provides exact probabilities for any sequence.
   C) It eliminates the need for large deviations theory.
   D) It simplifies continuous distributions into discrete ones.
   E) It maximizes the entropy of the source.

5. Under a given i.i.d. distribution P, the probability of observing a sequence x^n of type $\hat{\mathrm{P}}_{x^n}$ is approximately:

   A) $\mathrm{P}^n(x^n) \approx 2^{-nH(\hat{\mathrm{P}}_{x^n})}$
   B) $\mathrm{P}^n(x^n) \approx 2^{-n[H(\hat{\mathrm{P}}_{x^n})+D(\hat{\mathrm{P}}_{x^n}\|\mathrm{P})]}$
   C) $\mathrm{P}^n(x^n) \approx 2^{-nD(\hat{\mathrm{P}}_{x^n}\|\mathrm{P})}$
   D) $\mathrm{P}^n(x^n) \approx 2^{-nH(\mathrm{P})}$
   E) $\mathrm{P}^n(x^n) \approx 2^{-nD(\mathrm{P}\|\hat{\mathrm{P}}_{x^n})}$

6. The Asymptotic Equipartition Property (AEP) states that for a memoryless stationary source, the sequences of length n fall into two categories as $n \to \infty$:

   A) Typical sequences with probability close to zero and atypical sequences with probability close to one.
   B) Typical sequences with high probability and atypical sequences with low probability.
   C) All sequences become equally probable.
   D) The entropy of the source approaches zero.

E) The sequences can no longer be compressed.

7. The typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ for a discrete memoryless source P with alphabet $\mathcal{X}$ is defined as:

A) The set of sequences x^n whose empirical distribution is close to the source distribution.

B) The set of sequences x^n such that $\mathrm{P}^n(x^n) \geq 1 - \varepsilon$.

C) The set of sequences x^n such that $\mathrm{P}^n(x^n) = 2^{-nH(X)}$.

D) The set of sequences x^n whose empirical distribution equals the source distribution.

E) The set of sequences x^n whose probability is less than ε.

8. According to the AEP, the size of the typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ for a source P satisfies:

A) $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \approx 2^{nH(\mathrm{P})}$

B) $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \approx nH(\mathrm{P})$

C) $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \approx n$

D) $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \approx H(\mathrm{P})$

E) $|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})| \approx 1$

9. The probability that a sequence drawn from a discrete memoryless stationary source lies in the typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$ with a fixed $\varepsilon > 0$ approaches what value as $n \to \infty$?

A) 0

B) 1

C) ε

D) Depends on the source distribution

E) Cannot be determined

10. The jointly typical set $\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})$ is defined as the set of pairs (x^n, y^n) such that:

A) (x^n, y^n) are both individually typical sequences.

B) The joint empirical distribution of (x^n, y^n) is close to the joint distribution P_{XY}.

C) x^n and y^n are identical sequences.

D) x^n and y^n are uncorrelated.

E) The mutual information between x^n and y^n is zero.

11. In the method of types, the probability of observing a sequence of a particular type Q, i.e., x^n such that $\hat{\mathrm{P}}_{x^n} = \mathrm{Q}$, under the true distribution P is:

A) $\mathrm{P}^n(x^n) = 2^{-nH(\mathrm{P})}$

B) $\mathrm{P}^n(x^n) = 2^{-nD(\mathrm{Q}\|\mathrm{P})}$

C) $\mathrm{P}^n(x^n) = 2^{-n[H(\mathrm{Q})+D(\mathrm{Q}\|\mathrm{P})]}$

D) $\mathrm{P}^n(x^n) = 2^{-nH(\mathrm{Q})}$

E) $\mathrm{P}^n(x^n) = 2^{-nD(\mathrm{P}\|\mathrm{Q})}$

12. The method of types provides an estimate for the probability of a type class $\mathcal{T}(\mathrm{Q})$ under distribution P as:

A) $\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \approx 2^{-nD(\mathrm{Q}\|\mathrm{P})}$

B) $\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \approx 2^{-nH(\mathrm{Q})}$

C) $\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \approx 2^{-n[H(\mathrm{Q})+D(\mathrm{Q}\|\mathrm{P})]}$

D) $\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \approx 2^{-nD(\mathrm{P}\|\mathrm{Q})}$

E) $P^n(\mathcal{T}(Q)) \approx 2^{-nH(P)}$

13. According to the method of types, the total number of possible types over an alphabet $\mathcal{X}$ for sequences of length n is:

A) Polynomial in n

B) Exponential in n

C) Logarithmic in n

D) Constant, independent of n

E) Depends on the actual distribution P

14. The divergence $D(\hat{P}_{x^n}\|P)$ between the type $\hat{P}_{x^n}$ and the true distribution P is always:

A) Non-negative and zero if and only if $\hat{P}_{x^n} = P$

B) Non-positive and zero if and only if $\hat{P}_{x^n} = P$

C) Non-negative and zero if and only if $\hat{P}_{x^n} \neq P$

D) Negative when $\hat{P}_{x^n} = P$

E) Equal to the entropy of $\hat{P}_{x^n}$

15. The method of types can be used to show that the probability of observing a sequence x^n whose type $\hat{P}_{x^n}$ is significantly different from P is:

A) High, due to randomness

B) Zero, as such sequences cannot occur

C) Exponentially small in n, decreasing with n

D) Independent of n

E) Equal to $D(\hat{P}_{x^n}\|P)$

16. The joint typicality lemma states that the probability that $(X^n, Y^n) \sim P_X^n P_Y^n$ is jointly typical with respect to P_{XY} is approximately:

A) $2^{-n[I(X;Y)]}$

B) 1

C) 0

D) Equal to the product of their marginal probabilities

E) $2^{-nH(X,Y)}$

17. Which of the following statements is **TRUE** about the empirical distribution of a sequence?

A) The empirical distribution is always identical to the true source distribution.

B) The empirical distribution converges to the true distribution as the sequence length increases.

C) The empirical distribution is defined only for sequences of infinite length.

D) The empirical distribution is the expected value of the random variable.

E) The empirical distribution is irrelevant in calculating the probability of sequences.

18. For a discrete memoryless stationary source, the probability that the type of a sequence deviates from its true distribution decreases exponentially with sequence length due to:

A) The Weak Law of Large Numbers

B) The Central Limit Theorem

C) The Strong Law of Large Numbers

D) The Chebyshev Inequality

E) Hoeffding's Inequality

19. In the method of types, when considering sequences x^n and y^n of length n, the number of joint types is:

    A) Exponential in n
    B) Polynomial in n
    C) Independent of n
    D) Logarithmic in n
    E) Double exponential in n

20. The number of sequences jointly typical with a given sequence $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)$ under a joint distribution P_{XY} is approximately:

    A) $2^{nH(\mathrm{X}|\mathrm{Y})}$
    B) $2^{nH(\mathrm{Y})}$
    C) $2^{nI(\mathrm{X};\mathrm{Y})}$
    D) $2^{nH(\mathrm{Y}|\mathrm{X})}$
    E) $2^{n[H(\mathrm{Y})-I(\mathrm{X};\mathrm{Y})]}$

Elements of Information Theory **2025-2026**

Lecture 3: Lossless data compression

Lecturer: S. Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In this lecture, we consider the lossless data compression problem. That is, to represent a sequence of symbols from a discrete alphabet with a sequence of binary digits, in such a way that the source sequence can be recovered perfectly from the encoded sequence. We shall establish the fundamental limit of lossless data compression in terms of the minimum number of bits per source symbol.

3.1 Lossless codes

Let $\mathcal{X}$ be the set of source symbols, e.g., $\mathcal{X} = \{1, 2, 3, \dots\}$ or $\{a, b, c, \dots\}$ that is countable. An **encoding function** or **encoder** is a mapping from $\mathcal{X}$ to $\mathcal{C} \subseteq \{0,1\}^* := \bigcup_{n\ge0}\{0,1\}^n$. The set $\mathcal{C}$ is called a **codebook** or simply **code**

$$\mathcal{C} := \{f(x) : \quad x \in \mathcal{X}\}.$$

One can think of $\{0,1\}^*$ as an infinite binary tree, where each node corresponds to a sequence identified by the path from the root node. Such a tree is sometimes referred to as a **codetree**.

Each element in the codebook is called a **codeword**. The **codeword length** of $f(x)$ is the number of bits in $f(x)$, denoted by $l(f(x))$. The **maximum codeword length** is $L_{\max} := \sup_{x\in\mathcal{X}} l(f(x))$, while the **expected codeword length**, if the distribution is given, is

$$\bar{L}(f, \mathrm{P}) := \mathbb{E}_{\mathrm{P}}\big(l(f(\mathrm{X}))\big).$$

A code/encoder is called **lossless** if $f : \mathcal{X} \to \mathcal{C}$ is a bijection. In this case, $|\mathcal{C}| = |\mathcal{X}|$ and the **decoder** or **decoding function** g is simply the inverse function of f, i.e., $g(f(x)) = x$ for all $x \in \mathcal{X}$.

For example, for an alphabet of seven symbols, we can propose the following encoding

$$\{\text{a, b, c, d, e, f, g}\} \longleftrightarrow \{\varnothing, 0, 1, 00, 01, 10, 11\}.$$

Two codes $\mathcal{C}$ and $\mathcal{C}'$ are said to be **equivalent**, denoted by $\mathcal{C} \sim \mathcal{C}'$, if they contain the same number of codewords, i.e., $|\mathcal{C}| = |\mathcal{C}'| = M$, and if the lengths of their codewords, ordered from shortest to longest, coincide: $l(y_i) = l(y_i')$, $i = 1, \dots, M$, where y_i and y_i' are the i-th shortest codeword in $\mathcal{C}$ and $\mathcal{C}'$, respectively.

3.2 Uniquely decodable codes

A practical issue emerges when we wish to encode a sequence of symbols individually with the same encoder described in the previous section. Going back to the example with $\mathcal{X} = \{a, b, c, d, e, f, g\}$, and apply the variable-length code $\mathcal{C} = \varnothing, 0, 1, 00, 01, 10, 11$ on the letters in 'fade', we obtain an encoded sequence 100010. We now realize that there is no way to recover the original sequence from 100010 since there is an infinite number of possibilities [6]. One workaround is to introduce a separator in the encoded sequence, which would make the output essentially

[6] The infinity of solution in this example comes from the empty codeword $\varnothing$. Even without this codeword, though, the possible input sequence is not unique.

ternary, not binary. A more direct solution is to integrate the "uniquely decodability" constraint right into the encoding function.

A code is called **uniquely decodable** if two distinct input sequences cannot produce the same output sequence. Specifically, for any encoder f of a uniquely decodable code, $x_1, \dots, x_m$ and $x'_1, \dots, x'_n$ are two distinct input sequences if and only if the concatenation of output bits $(f(x_1), \dots, f(x_m)) \neq (f(x'_1), \dots, f(x'_n))$. Such a code is sometimes referred to as **separable** code.

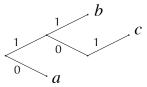
3.2.1 Prefix codes

An important class of uniquely decodable codes is the **prefix code** (a.k.a. prefix-free code). For $m \leq n$, a string $y := (y_1 \cdots y_m)$ is called the **prefix** of another string $y' := (y'_1 \cdots y'_n)$ if $y_i = y'_i$ for $i = 1, \dots, m$, that is, the string y' starts with y, denoted by $y \ll y'$.

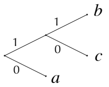
A code $\mathcal{C}$ is called a prefix code if no codeword is a prefix of another codeword. Such a code is sometimes referred to as **instantaneous** code.

Tree representation: It is convenient to associate the codewords with the **label** of nodes of a tree: the sequence of edge labels ($\{0, 1\}$) along the path from the root to the node. In particular, for a prefix code, there is a unique tree whose leaf nodes correspond to the set of codewords.

$$\text{Prefix code } \mathcal{C} \equiv \text{ Set of leaf nodes of a tree } \mathcal{T}$$



$a \to 0$
$b \to 11$
$c \to 101$



$a \to 0$
$b \to 11$
$c \to 10$

Interval representation: We can also associate the codewords with the intervals inside $[0, 1]$. For each $y = (y(1) \dots y(l(y))) \in \mathcal{C}$, we can define the real value

$$r(y) := y(1)2^{-1} + \cdots + y(l(y))2^{-l(y)} \equiv 0.y(1)y(2) \cdots y(l(y))$$

It follows that the interval $\mathcal{I}(y) := [r(y), r(y) + 2^{-l(y)}) \subseteq [0, 1]$ is the collection of all values whose binary representation has y as prefix. Therefore, $\mathcal{C} = \{y_1, y_2, \dots\}$ is a prefix code if and and only if $\mathcal{I}(y_i) \cap \mathcal{I}(y_j) = \varnothing, \ \forall \ i \neq j$.

$$\text{prefix code } \mathcal{C} \equiv \text{ Disjoint intervals } \{\mathcal{I}(y), y \in \mathcal{C}\}.$$

3.2.1.1 The Huffman Code $f_{\textbf{Huffman},\mathrm{Q}}$

Let $\mathrm{Q} := [q_1, \dots, q_M]$ be a given set of input parameters. A prefix code for a source alphabet of size M can be constructed by building a binary tree in $M - 1$ steps as follows:

- **Initialization:** Start with M leaf nodes $n_1, \dots, n_M$, each corresponding to one of the M source symbols. Assign node n_i the value q_i, and form the initial list

$$\mathcal{L} := \{(n_1, q_1), \dots, (n_M, q_M)\}.$$

- **Iterative construction:** For $k = M-1, M-2, \dots, 1$:
  - Identify two nodes a and b in $\mathcal{L}$ with the smallest q values, denoted q_a and q_b.
  - Create a new internal node $\tilde{n}_k$ as the parent of a and b, and assign it the combined value $\tilde{q}_k = q_a + q_b$.
  - Remove (a, q_a) and (b, q_b) from $\mathcal{L}$, and insert the new pair $(\tilde{n}_k, \tilde{q}_k)$.
- **Codeword assignment:** After $M-1$ iterations, the resulting binary tree $\mathcal{T}$ defines the code. Assign binary labels $0, 1$ to the edges. The codeword corresponding to each leaf node is given by the sequence of edge labels along the path from the root to that leaf.

This is known as the **Huffman algorithm**. Note that usually Q is a pmf, although it can be any real vector in theory.

3.2.1.2 The Shannon Code, $f_{\mathbf{Shannon},\mathrm{Q}}$

The Huffman code is constructed from a tree representation. In contrast, the Shannon code is based on an interval representation, which we describe below.

Assume that $\mathrm{Q} = [q_1, \dots, q_M]$ is a sequence of nonnegative numbers satisfying $q_1 + \cdots + q_M \le 1$, and that the elements are sorted in decreasing order: $q_1 \ge q_2 \ge \cdots \ge q_M$.

Define the cumulative sums

$$s_1 = 0, \qquad s_i = q_1 + \cdots + q_{i-1}, \quad i = 2, \dots, M,$$

so that $s_{i+1} - s_i = q_i$, $i = 1, \dots, M-1$.

For any positive real number q, define its l-bit quantization by

$$[q]_l = \lfloor q, 2^l \rfloor, 2^{-l}.$$

Then, the Shannon code associated with Q assigns to the i-th source symbol a binary codeword corresponding to the first

$$l_i := \left\lceil \log \frac{1}{q_i} \right\rceil$$

bits in the binary representation of s_i.

Indeed, the Shannon code can be shown to be a prefix code. From the interval representation, we associate each codeword y_i with an interval $\mathcal{I}(y_i)$ that starts at $r(y_i) = [s_i]_{l_i}$, and has length 2^{-l_i}. To prove the prefix property, it suffices to show that, for every $i = 1, \dots, M-1$, the starting point $r(y_{i+1})$ does not lie within the interval $\mathcal{I}(y_i)$. This ensures that the intervals $\mathcal{I}(y_i)$ are non-overlapping, and thus no codeword is a prefix of another. To that end, we check the difference

$$\begin{aligned} r(y_{i+1}) - r(y_i) &= [s_{i+1}]_{l_{i+1}} - [s_i]_{l_i} \\ &\ge [s_{i+1}]_{l_i} - [s_i]_{l_i} \\ &\ge [s_{i+1} - s_i]_{l_i} \\ &= [q_i]_{l_i} \\ &= \lfloor q_i 2^{l_i} \rfloor 2^{-l_i} \\ &\ge 2^{-l_i} \end{aligned}$$

3.2.2 Kraft-McMillan inequality

The main result of this section is the well-known **Kraft-McMillan** inequality (also referred to as the Kraft inequality or K-M inequality).

If a code $\mathcal{C}$ of size M is uniquely decodable, then the set of codeword lengths $\{l(y) : y \in \mathcal{C}\}$ satisfies the Kraft-McMillan inequality, namely,

$$\sum_{y \in \mathcal{C}} 2^{-l(y)} \leq 1.$$

Conversely, for any collection of positive integers $\{l_1, \dots, l_M\}$ satisfying the above inequality, there exists a uniquely decodable code — indeed, a prefix code — with these codeword lengths.

As a direct consequence, every uniquely decodable $\mathcal{C}$ has an equivalent prefix code $\mathcal{C}'$.

If f is an encoding function associated with $\mathcal{C}$, the K-M inequality can also be written as

$$\boxed{\sum_{x \in \mathcal{X}} 2^{-l(f(x))} \leq 1}$$

We call it the **K-M equality** when equality holds in the K-M inequality.

First, we shall show the converse ("only if" part) of the above result.

Necessity of the K-M inequality. For any integer k,

$$\begin{aligned}
\left(\sum_{y \in \mathcal{C}} 2^{-l(y)}\right)^k &= \left(\sum_{x \in \mathcal{X}} 2^{-l(f(x))}\right)^k && (3.1)\\
&= \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} 2^{-(l(f(x_1))+\cdots+l(f(x_k)))} && (3.2)\\
&= \sum_{x^k \in \mathcal{X}^k} 2^{-(l(f^k(x^k)))} \qquad \left(f^k(x^k) := (f(x_1), \dots, f(x_k)) \text{ is the output sequence}\right) && (3.3)\\
&= \sum_{\lambda=1}^{kL_{\max}} \sum_{x^k : l(f^k(x^k))=\lambda} 2^{-\lambda} \qquad \text{(rearrange according to codeword lengths)} && (3.4)\\
&= \sum_{\lambda=1}^{kL_{\max}} \left|\{x^k : l(f^k(x^k)) = \lambda\}\right| 2^{-\lambda} && (3.5)\\
&\leq \sum_{\lambda=1}^{kL_{\max}} 2^{\lambda} 2^{-\lambda} \qquad \left(\text{unique decodability: at most } 2^{\lambda} \text{ input sequences can have encoded length } \lambda\right) && (3.6)\\
&\leq kL_{\max}. && (3.7)
\end{aligned}$$

Since the above inequality holds for all k, we have

$$\log \sum_{y \in \mathcal{C}} 2^{-l(y)} \leq \inf_{k \geq 1} \frac{\log(kL_{\max})}{k} = 0,$$

proving the Kraft-McMillan inequality. □

Next, we show that a Shannon code can always be constructed with prescribed codeword lengths, provided that these lengths satisfy the Kraft-McMillan inequality. Specifically, let $q_i = 2^{-l_i}$ for $i = 1, \dots, M$. Since the Kraft-McMillan inequality ensures that $\sum_i q_i \leq 1$, we can apply the Shannon coding procedure with Q $= [q_1, \dots, q_M]$. The resulting code then assigns to the i-th symbol a codeword of length $\lceil \log \frac{1}{q_i} \rceil = l_i$, $i = 1, \dots, M$, as desired.

3.3 Minimum expected codeword length

Define

$$\bar{L}^*(\mathrm{P}) := \min_{f:\,\text{lossless}} \mathbb{E}_{\mathrm{P}}\left[l(f(\mathrm{X}))\right]$$
$$\bar{L}_{\mathrm{UD}}(\mathrm{P}) := \min_{f:\,\text{uniquely decodable}} \mathbb{E}_{\mathrm{P}}\left[l(f(\mathrm{X}))\right]$$

Then we have

$$\max_{N\le|\mathcal{X}|}\ \min_{\mathcal{A}\subseteq\mathcal{X}:|\mathcal{A}|=N}(1-\mathrm{P}(\mathcal{A}))\log(|\mathcal{A}|+1)\ \le\ \bar{L}^*(\mathrm{P})\ \le\ H(\mathrm{P})\ \le\ \bar{L}_{\mathrm{UD}}(\mathrm{P})\le H(\mathrm{P})+1 \qquad (3.8)$$

In addition, the Huffman code is an optimal uniquely decodable code in the following sense

$$\bar{L}_{\mathrm{UD}}(\mathrm{P}) = \mathbb{E}_{\mathrm{X}\sim\mathrm{P}}\left[l(f_{\mathrm{Huffman},\mathrm{P}}(\mathrm{X}))\right]$$

3.3.1 Lossless codes

To minimize the expected length, the optimal lossless code with respect to a given distribution P is straightforward: Assign the codewords $\{\varnothing, 0, 1, 00, 01, 10, 11, 000, \ldots\}$ to the symbols with decreasing probability. Without loss of generality, assume that $\mathcal{X} = \{1, \ldots, M\}$ with decreasing probability $p_1 \ge p_2 \ge \cdots \ge p_M$. Then, the i-th codeword has length $\lfloor \log i \rfloor$ bits. We have

$$\bar{L}^*(\mathrm{P}) = \sum_{i=1}^{M} p_i \lfloor \log i \rfloor$$

Note that $ip_i \le p_1 + \cdots + p_i \le 1$. We have $\lfloor \log i \rfloor \le \log i \le \log \frac{1}{p_i}$, from which we get

$$\bar{L}^*(\mathrm{P}) \le \sum_{i=1}^{M} p_i \log \frac{1}{p_i} = H(\mathrm{P}).$$

For the lower bound, we have

$$\begin{aligned}
\bar{L}^*(\mathrm{P}) &= \sum_{i=1}^{M} p_i \lfloor \log i \rfloor \\
&\ge \sum_{i=N+1}^{M} p_i \lfloor \log i \rfloor \\
&\ge \sum_{i=N+1}^{M} p_i \lfloor \log(N+1) \rfloor \\
&= \left(1 - \sum_{i=1}^{N} p_i\right) \lfloor \log(N+1) \rfloor \\
&= \min_{\mathcal{A}\subseteq\mathcal{X}:|\mathcal{A}|=N} (1-\mathrm{P}(\mathcal{A}))\ \lfloor \log(|\mathcal{A}|+1) \rfloor \\
&\ge \min_{\mathcal{A}\subseteq\mathcal{X}:|\mathcal{A}|=N} (1-\mathrm{P}(\mathcal{A}))\ \log \frac{|\mathcal{A}|+1}{2}
\end{aligned}$$

Since the above bound holds for any $N \le M$, we can maximize over N and get the desired bound.

3.3.2 Uniquely decodable codes

Since uniquely decodable codes are completely characterized by the Kraft-McMillan inequality, we can express the minimum achievable average codeword length as

$$\bar{L}_{\mathrm{UD}}(\mathrm{P}) = \min_{l_i \in \mathbb{Z}^+ : \sum_i 2^{-l_i} \le 1} \sum_i p_i 2^{-l_i}$$

Relaxing the integer constraint and substituting $q_i = 2^{-l_i}$, we obtain

$$\begin{aligned} \bar{L}_{\mathrm{UD}}(\mathrm{P}) &\ge \min_{\sum_i q_i \le 1} \sum_i p_i \log \frac{1}{q_i} \\ &\ge \min_{\sum_i q_i = 1} \sum_i p_i \log \frac{1}{q_i} \\ &= \min_{\sum_i q_i = 1} H(\mathrm{P} \| \mathrm{Q}) \\ &= H(\mathrm{P}) \end{aligned}$$

To establish an upper bound, consider the feasible choice $l_i = \lceil \log \frac{1}{p_i} \rceil \le \lceil \log \frac{1}{p_i} \rceil$, $i = 1, \dots, M$. Indeed, we can verify the K-M inequality with this choice:

$$\begin{aligned} \sum_{i=1}^{M} 2^{-l_i} &\le \sum_{i=1}^{M} 2^{-\log \frac{1}{p_i}} \\ &= \sum_{i=1}^{M} p_i = 1 \end{aligned}$$

Hence, we can upper bound the minimization by this feasible point:

$$\begin{aligned} \bar{L}_{\mathrm{UD}}(\mathrm{P}) &\le \sum_{i=1}^{M} p_i \lceil \log \frac{1}{p_i} \rceil \\ &\le \sum_{i=1}^{M} p_i (\log \frac{1}{p_i} + 1) \\ &= H(\mathrm{P}) + 1 \end{aligned}$$

3.3.3 Optimality of the Huffman code

In the following, we will show that Huffman code is an optimal uniquely decodable code in the sense of minimum expected codeword length. From the previous discussion, it is without loss of optimality to consider prefix codes.

It is not hard to check that the optimal prefix code must satisfy the following *necessary conditions*:

1. A symbol with higher probability should be assigned with a shorter codeword. That is, if $p_i \ge p_j$, then $l_i \le l_j$. Otherwise, swapping the codewords results in a smaller expected length.
2. The corresponding tree must be full, i.e., each node either is a leaf node or has two children. Otherwise, one can always shorten the codewords (by pruning) which can only reduce the expected length.

Let $\mathcal{T}^*$ be the tree corresponding to the optimal code, with ordered codeword length $\tilde{l}_1^* \le \tilde{l}_2^* \le \cdots \le \tilde{l}_M^*$. The corresponding probability is $\tilde{p}_1 \ge \tilde{p}_2 \ge \cdots \ge \tilde{p}_M$. At the beginning, nothing is known about $\mathcal{T}^*$ (and thus the l_i^*'s). But we do know that the codeword corresponding to the most improbable symbol, being a leaf node, must have a **sibling** since the tree is full according to condition 2. In addition, the sibling must be a leaf node. Indeed, if its sibling is an inner node, there must be other leaf nodes with a strictly larger depth than the one for the most improbable symbol, violating the condition 1. Therefore, the two longest codewords have the same codeword length, and are siblings, without loss of optimality. Therefore, we have $\tilde{l}_M^* = \tilde{l}_{M-1}^*$. It turns out that this information on the optimal property is enough to construct the optimal code.

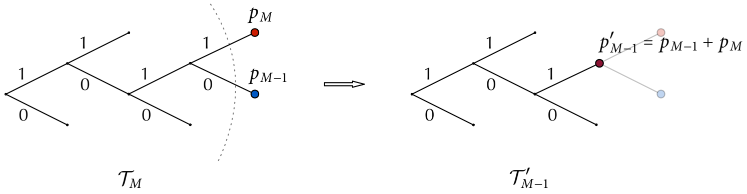
To see this, note that the expected depth of the tree, also corresponding to the expected codeword length of the code, is

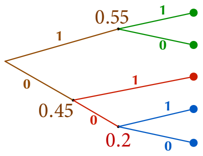
$$\bar{L}(\mathcal{T}^*, (\tilde{p}_1, \dots, \tilde{p}_M)) := \sum_{i=1}^{M} \tilde{l}_i^* \tilde{p}_i \tag{3.9}$$

$$= \sum_{i=1}^{M-2} \tilde{l}_i^* \tilde{p}_i + (\tilde{l}_{M-1} - 1)(\tilde{p}_{M-1} + \tilde{p}_M) + (\tilde{p}_{M-1} + \tilde{p}_M) \tag{3.10}$$

$$= \bar{L}(\mathcal{T}', (\tilde{p}_1, \dots, \tilde{p}_{M-2}, \tilde{p}_{M-1} + \tilde{p}_M)) + (\tilde{p}_{M-1} + \tilde{p}_M) \tag{3.11}$$

Since $\mathcal{T}^*$ is an optimal tree with respect to $(\tilde{p}_1, \dots, \tilde{p}_M)$, $\mathcal{T}'$ must also be an optimal tree with respect to $(\tilde{p}_1, \dots, \tilde{p}_{M-2}, \tilde{p}_{M-1} + \tilde{p}_M)$. This is precisely the idea of the **Huffman algorithm**: we start with the list of M nodes, let the two least probable nodes be siblings, replace both siblings by the parent node with combined probability, continue the procedure with the reduced list of $M - 1$ nodes, until there is only one node. The optimal tree structure is completely uncovered at the end of the procedure.





p_j	source symbol		codeword	length	$\lceil -\log p_j \rceil$
0.3	a	⟶	11	2	2
0.25	b	⟶	10	2	2
0.25	c	⟶	01	2	2
0.15	d	⟶	001	3	3
0.05	e	⟶	000	3	5

3.4 Shannon's lossless coding theorem

> Let X^n be a sequence of i.i.d. symbols from $\mathcal{X}$, i.e., X^n ~ P^n for some pmf P. Then, the minimum achievable average number of bits per symbol required to represent the sequence converges to the entropy of the source
>
> $$\lim_{n\to\infty} \frac{1}{n} \bar{L}^*(\mathrm{P}^n) = \lim_{n\to\infty} \frac{1}{n} \bar{L}_{\mathrm{UD}}(\mathrm{P}^n) = H(\mathrm{P}).$$

To prove the theorem, it suffices to apply (3.8) with P^n and show that both upper and lower bounds coincide asymptotically, after being normalized by n.

First, from the upper bound, we have

$$\lim_{n\to\infty} \frac{1}{n} \bar{L}_{\mathrm{UD}}(\mathrm{P}^n) \le \lim_{n\to\infty} \frac{1}{n} H(\mathrm{P}^n) + \frac{1}{n}$$
$$= H(\mathrm{P}),$$

since $H(\mathrm{P}^n) = nH(\mathrm{P})$.

Then, we derive the lower bound as follows. Note that for any $\mathcal{A} \subseteq \mathcal{X}^n$, we have

$$\begin{aligned} \mathrm{P}^n(\mathcal{A}) &= \mathrm{P}^n(\mathcal{A} \cap \mathcal{T}_\varepsilon^{(n)}(P)) + \mathrm{P}^n(\mathcal{A} \cap \overline{\mathcal{T}_\varepsilon^{(n)}(P)}) \\ &\le \mathrm{P}^n(\mathcal{A} \cap \mathcal{T}_\varepsilon^{(n)}(P)) + \mathrm{P}^n(\overline{\mathcal{T}_\varepsilon^{(n)}(P)}) \\ &\le |\mathcal{A}| 2^{-n(1-\varepsilon)H(\mathrm{P})} + \delta_T(\varepsilon, n, \mathrm{P}) \end{aligned}$$

where $\delta_T(\varepsilon, n, \mathrm{P})$ is defined in (2.11) and vanishes with n when $n\varepsilon^2 \to \infty$. Let $\varepsilon = n^{-\frac{1}{3}}$, so that

$$n\varepsilon^2 \to \infty, \quad n\varepsilon \to \infty, \quad \text{and } \varepsilon \to 0.$$

Let $|\mathcal{A}| = N = 2^{n(1-2\varepsilon)H(\mathrm{P})}$, and we have from the above bound

$$\mathrm{P}^n(\mathcal{A}) \le 2^{-n\varepsilon H(\mathrm{P})} + \delta_T(\varepsilon, n, \mathrm{P})$$

which vanishes as $n \to \infty$ with our choice of ε. Then, applying the lower bound in (3.8) with the upper bound on $\mathrm{P}^n(\mathcal{A})$ and $|\mathcal{A}| = 2^{n(1-2\varepsilon)H(\mathrm{P})}$, we have

$$\begin{aligned} \frac{1}{n}\bar{L}^*(\mathrm{P}^n) &\ge (1 - \mathrm{P}^n(\mathcal{A}))\frac{1}{n}\log\frac{2^{n(1-2\varepsilon)H(\mathrm{P})} + 1}{2} \\ &\ge (1 - \mathrm{P}^n(\mathcal{A}))(1 - 2\varepsilon)H(\mathrm{P}) - \frac{1}{n} \end{aligned}$$

which converges to $H(\mathrm{P})$.

3.4.1 Universal encoding using types

All the previous schemes use the knowledge on the source distribution P and are specifically designed for the distribution. In most cases, such knowledge is not available *a priori*. In the following, we present a *universal* scheme that works for any source distribution P.

For each type $\hat{\mathrm{P}} \in \mathcal{P}_n$, we assign an index $i(\hat{\mathrm{P}} \mid \mathcal{P}_n) \in \{1, \ldots, |\mathcal{P}_n|\}$. Then, for each sequence in the type class $\mathcal{T}(\hat{\mathrm{P}})$, we assign an index $i(x^n \mid \mathcal{T}(\hat{\mathrm{P}})) \in \{1, \ldots, |\mathcal{T}(\mathrm{Q})|\}$. Therefore, each sequence x^n is uniquely identified by the $(i(\hat{\mathrm{P}}_{x^n} \mid \mathcal{P}_n), i(x^n \mid \mathcal{T}(\hat{\mathrm{P}}_{x^n})))$. The universal encoding the following variable-length encoding:

$$f_{\mathrm{U}}(x^n) = (\mathrm{bin}(i(\hat{\mathrm{P}}_{x^n} \mid \mathcal{P}_n)), \mathrm{bin}(i(x^n \mid \mathcal{T}(\hat{\mathrm{P}}_{x^n}))))$$

where $\mathrm{bin}(i(\hat{\mathrm{P}}_{x^n} \mid \mathcal{P}_n))$ is the binary representation of $i(\hat{\mathrm{P}}_{x^n})$ in $\lceil \log |\mathcal{P}_n| \rceil$ bits, and $\mathrm{bin}(i(x^n \mid \mathcal{T}(\hat{\mathrm{P}}_{x^n})))$ is the binary representation of $i(x^n \mid \mathcal{T}(\hat{\mathrm{P}}_{x^n}))$ in $\lceil \log |\mathcal{T}(\hat{\mathrm{P}}_{x^n})| \rceil$ bits. Note that the encoding is uniquely decodable. The decoder first reads the first $\lceil \log |\mathcal{P}_n| \rceil$ bits and identify $\hat{\mathrm{P}}_{x^n}$. Then, the decoder goes on and read the following $\lceil \log |\mathcal{T}(\hat{\mathrm{P}}_{x^n})| \rceil$ bits and identify $i(x^n \mid \hat{\mathrm{P}}_{x^n})$. Looking up the table and the sequence x^n can be decoded.

Let us look at the expected length.

$$\begin{aligned} \mathbb{E}_{\mathrm{P}^n}(l(f_{\mathrm{U}}(\mathrm{X}^n))) &= \lceil \log |\mathcal{P}_n| \rceil + \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)} \mathrm{P}_X^n(x^n) \lceil \log |\mathcal{T}(\hat{\mathrm{P}}_{x^n})| \rceil + \sum_{x^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)} \mathrm{P}_X^n(x^n) \lceil \log |\mathcal{T}(\hat{\mathrm{P}}_{x^n})| \rceil \\ &\le \lceil \log |\mathcal{P}_n| \rceil + \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})} \mathrm{P}^n(x^n) \lceil \log |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)| \rceil + \sum_{x^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})} \mathrm{P}^n(x^n) \lceil \log M^n \rceil \\ &\le \lceil \log |\mathcal{P}_n| \rceil + \lceil \log |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)| \rceil + \sum_{x^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})} \mathrm{P}^n(x^n) \lceil \log M^n \rceil \\ &\le \lceil \log |\mathcal{P}_n| \rceil + 2 + n(1 + \varepsilon)H(\mathrm{P}) + \delta_T(\varepsilon, n, \mathrm{P}) \log M^n \\ &\le 3 + M \log(n + 1) + n(1 + \varepsilon)H(\mathrm{P}) + n\delta_T(\varepsilon, n, \mathrm{P}) \log M \end{aligned}$$

where we used the fact that $|\mathcal{P}_n| \le (n + 1)^M$. Let $\varepsilon = n^{-\frac{1}{3}}$, so that $n\varepsilon^2 \to \infty$ and $\varepsilon \to 0$, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{\mathrm{P}^n}(l(f_{\mathrm{U}}(\mathrm{X}^n))) &\le \lim_{n\to\infty} \frac{3}{n} + M\frac{\log(n + 1)}{n} + (1 + \varepsilon)H(\mathrm{P}) + \delta_T(\varepsilon, n, \mathrm{P}) \log M \\ &= H(P). \end{aligned}$$

Now we see that even without knowing the distribution of the source, one can achieve the entropy lower bound.

3.4.2 Practical codes

None of the above codes are practical when n is large. All of the codes presented above work with lookup tables. Such tables have size proportional to the number of all M^n sequences of codewords. For example, for $M = 2$ and $n = 100$, there are $2^{100} \approx 10^{30}$ codewords. In practice, it is desirable that encoding function "computes" the codeword on the fly. Arithmetic coding is such an encoding scheme that can encode a source sequence X^n successively using the conditional distributions $P_{X_1}, P_{X_2|X_1}, \ldots, P_{X_n|X^{n-1}}$ so that the complexity is linear in n. Moreover, the achievable rate is $\frac{1}{n}H(\mathrm{P}_{X^n}) + \frac{2}{n}$ provided the distribution P_{X^n} is known. For i.i.d. sources, i.e., $\mathrm{P}_{X^n} = \mathrm{P}_X^n$, the entropy lower bound is achieved with arithmetic coding when n is large.

When the source distribution is not known, there exist practical universal codes. One way is to apply a universal coding distribution that works well for almost all sources within a given class (e.g., i.i.d. sources, order-1 Markov sources). For instance, the Krichevsky-Trofimov distribution can be used for i.i.d. sources. Another way is to encode the sequence directly without an explicite coding distribution. A well-known example is the Lempel-Ziv coding, used in the original compression program in Unix.

Exercises[7]

1. Codes [CT 5.37]. Which of the following codes are

   - Uniquely decodable?
   - Instantaneous?

$$C_1 = \{00, 01, 0\} \tag{3.12}$$
$$C_2 = \{00, 01, 100, 101, 11\} \tag{3.13}$$
$$C_3 = \{0, 10, 110, 1110, \ldots\} \tag{3.14}$$
$$C_4 = \{0, 00, 000, 0000\} \tag{3.15}$$

2. Huffman coding [CT 5.4]. Consider the random variable

$$\mathrm{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

   - Find a binary Huffman code for X.
   - Find the expected code length for this encoding.
   - Find a ternary Huffman code for X.

3. Bad codes [CT 5.6]. Which of these codes cannot be Huffman codes for any probability assignment?

   - $\{0, 10, 11\}$
   - $\{00, 01, 10, 110\}$
   - $\{01, 10\}$

4. Shannon codes and Huffman codes [CT 5.12]. Consider a random variable X that takes on four values with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

   - Construct a Huffman code for this random variable.
   - Show that there exist two different sets of optimal lengths for the codewords; namely, show that codeword length assign- ments (1, 2, 3, 3) and (2, 2, 2, 2) are both optimal.
   - Conclude that there are optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\left\lceil \log \frac{1}{\mathrm{P}(x)} \right\rceil$

5. Data compression [CT 5.17]. Find an optimal set of binary codeword lengths $l_1, l_2, \ldots$ (minimizing $\sum_i p_i l_i$) for a prefix code for each of the following probability mass functions:

   - $\mathrm{P} = \left(\frac{10}{41}, \frac{9}{41}, \frac{8}{41}, \frac{7}{41}, \frac{7}{41}\right)$
   - $\mathrm{P} = \left(\frac{9}{10}, \frac{9}{10^2}, \frac{9}{10^3}, \frac{9}{10^4}, \cdots\right)$

6. Optimal codetree. Consider a codetree $\mathcal{T}_M$ with M leaf nodes. Suppose that $\mathcal{T}_M$ corresponds to an optimal prefix code with respect to (w.r.t.) the pmf $\{p_1, \ldots, p_M\}$. Let $\mathcal{S}_{M'} \subseteq \mathcal{T}_M$ be a subtree with M' leaf nodes, says, with indices in $\mathcal{I} \subseteq \{1, \ldots, M\}$ and $|\mathcal{I}| = M'$. Show that

   - $\mathcal{S}_{M'}$ corresponds to an optimal prefix code w.r.t. the (conditional) pmf $\{p_i/(\sum_{j\in\mathcal{I}} p_j) : i \in \mathcal{I}\}$
   - replacing $\mathcal{S}_{M'}$ by a leaf node, we obtain a tree $\mathcal{T}'_{M-M'+1}$ that is optimal with respect to the pmf $\{\{p_i : i \notin \mathcal{I}\}, \sum_{i\in\mathcal{I}} p_i\}$

7. Optimal codes for uniform distributions [CT 5.24]. Consider a random variable with M equiprobable outcomes. The entropy of this information source is obviously $\log M$ bits.

   - Describe the optimal prefix binary code for this source and compute the average codeword length L_M.

[7] The citations “CT”, “CK” refer to Cover-Thomas, Csiszár-Körner, respectively.

- For what values of M does the average codeword length L_M equal the entropy $H = \log M$?
- We know that $L < H + 1$ for any probability distribution. The redundancy of a variable-length code is defined to be $\rho = L - H$. For what value(s) of M, where $2^k \leq M \leq 2^{k+1}$, is the redundancy of the code maximized? What is the limiting value of this worst-case redundancy as $M \to \infty$?

8. Shannon code [CT 5.28]. Consider the following method for generating a code for a random variable X that takes on M values $\{1, 2, \ldots, M\}$ with probabilities $p_1, p_2, \ldots, p_M$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_M$. Define

$$F_i = \sum_{k=1}^{i-1} p_k,$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

- Show that the code constructed by this process is prefix-free and that the average length satisfies

$$H(\mathrm{X}) \leq L < H(\mathrm{X}) + 1.$$

- Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

9. Relative entropy is cost of miscoding [CT 5.30]. Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions $\mathrm{P}(x)$ and $\mathrm{Q}(x)$ on this random variable.

Symbol	$\mathrm{P}(x)$	$\mathrm{Q}(x)$	$C_1(x)$	$C_2(x)$
1	$\frac{1}{2}$	$\frac{1}{2}$	0	0
2	$\frac{1}{4}$	$\frac{1}{8}$	10	100
3	$\frac{1}{8}$	$\frac{1}{8}$	110	101
4	$\frac{1}{16}$	$\frac{1}{8}$	1110	110
5	$\frac{1}{16}$	$\frac{1}{8}$	1111	111

- Calculate $H(\mathrm{P}), H(\mathrm{Q}), D(\mathrm{P}\|\mathrm{Q})$, and $D(\mathrm{Q}\|\mathrm{P})$.
- The last two columns represent codes for the random variable. Verify that the average length of C_1 under P is equal to the entropy $H(\mathrm{P})$. Thus, C_1 is optimal for P. Verify that C_2 is optimal for Q.
- Now assume that we use code C_2 when the distribution is P. What is the average length of the codewords. By how much does it exceed the entropy $H(\mathrm{P})$?
- What is the loss if we use code C_1 when the distribution is Q?

Quiz (unique correct answer)

1. The Kraft inequality states that for any D-ary uniquely decodable code for a source with alphabet size M, e.g., $D = 2$ for binary codes:

   A) $\sum_{i=1}^{M} D^{-l_i} \leq 1$

   B) $\sum_{i=1}^{M} D^{l_i} \geq 1$

   C) $\sum_{i=1}^{M} D^{-l_i} = 1$

   D) $\sum_{i=1}^{M} l_i D^{-1} \leq 1$

   E) $\sum_{i=1}^{M} D^{-l_i} \geq 1$

2. For a given discrete memoryless stationary source P, the average codeword length $\bar{L}$ of any uniquely decodable code satisfies:

   A) $\bar{L} \geq H(\mathrm{P})$

   B) $\bar{L} \leq H(\mathrm{P})$

   C) $\bar{L} = H(\mathrm{P})$

   D) $\bar{L} \geq H(\mathrm{P}) - 1$

   E) $\bar{L} \leq H(\mathrm{P}) + 1$

3. Which of the following statements about Huffman coding is **TRUE**?

   A) Huffman coding always produces codes with equal-length codewords.

   B) Huffman coding is optimal for any arbitrary source.

   C) Huffman coding can produce non-prefix codes.

   D) Huffman coding minimizes the average codeword length among all prefix codes.

   E) Huffman coding is optimal only for sources with equiprobable symbols.

4. The redundancy of a code is defined as:

   A) The difference between the average codeword length and the entropy, $\bar{L} - H(X)$

   B) The ratio of the entropy to the average codeword length, $H(X)/\bar{L}$

   C) The sum of codeword lengths, $\sum_i l_i$

   D) The difference between the maximum and minimum codeword lengths

   E) The variance of the codeword lengths

5. Which of the following is **NOT** a property of prefix codes?

   A) No codeword is a prefix of any other codeword.

   B) They are instantaneous codes.

   C) They can be uniquely decoded without delimiters.

   D) They may require lookahead during decoding.

   E) They satisfy the Kraft inequality with equality if the code is complete.

6. According to Shannon's Source Coding Theorem, for a discrete memoryless source with entropy $H(\mathrm{P})$:

   A) It is impossible to compress the source below $H(\mathrm{P})$ bits per symbol.

   B) It is possible to compress the source to $H(\mathrm{P})$ bits per symbol with zero error.

   C) Any code with average length $\bar{L} < H(\mathrm{P})$ is uniquely decodable.

   D) The entropy $H(\mathrm{P})$ is always an integer value.

   E) The minimal average codeword length $\bar{L}$ can be made arbitrarily close to $H(\mathrm{P})$.

7. For a binary symmetric source with $\mathrm{P}(0) = \mathrm{P}(1) = 0.5$, the optimal code is:

   A) A code with codeword lengths of 1 for both symbols.

   B) A code with codeword lengths of 0 for both symbols.

   C) A code with average codeword length 0.5.

   D) Any code, since all codes are equally efficient.

   E) Not possible to code optimally due to equal probabilities.

8. The term ”uniquely decodable code” refers to:

   A) A code where codewords are of equal length.

   B) A code where each codeword can be uniquely mapped to a source symbol sequence.

   C) A code where no codeword is a prefix of any other codeword.

   D) A code that can be decoded without errors in the presence of noise.

   E) A code that satisfies the equality in the Kraft-McMillan inequality.

9. The Huffman coding algorithm constructs the code tree by:

   A) Starting from the most probable symbols and assigning them the longest codewords.

   B) Merging the two symbols with the highest probabilities at each step.

   C) Merging the two symbols with the lowest probabilities at each step.

   D) Assigning codewords randomly and checking for prefix property.

   E) Balancing the code tree to minimize the variance of codeword lengths.

10. Let us define the redundancy as $\mathbb{E}_{\mathrm{P}}[l(f(\mathrm{X}))] - H(\mathrm{P})$ where f is the encoding function. The redundancy of a Huffman code for a given source is:

    A) Always zero, since Huffman coding is optimal.

    B) Always positive, since the average codeword length is greater than the entropy.

    C) Always negative, since the average codeword length is less than the entropy.

    D) Dependent on the probabilities being powers of $\frac{1}{2}$.

    E) Zero only when the probabilities are powers of $\frac{1}{2}$.

11. The main reason that the average codeword length in Huffman coding may exceed the entropy $H(\mathrm{P})$ is because:

    A) Huffman coding does not account for source redundancy.

    B) The codeword lengths must be integer values, while entropy is generally an average of non-integers.

    C) Huffman coding is sub-optimal compared to arithmetic coding.

    D) It uses fixed-length codewords.

    E) It cannot handle sources with a large number of symbols.

12. In the context of Huffman coding, if all symbol probabilities are negative powers of 2 (i.e., $\mathrm{P}(x_i) = 2^{-k_i}$), then the average codeword length $\bar{L}$ is:

    A) Equal to the entropy $H(X)$

    B) Less than the entropy $H(X)$

    C) Greater than the entropy $H(X)$

    D) Dependent on the variance of the source

    E) Unrelated to the entropy $H(X)$

13. The **Entropy of an Extension** of a discrete memoryless stationary source P when grouping symbols into blocks of length n is:

    A) Equal to $nH(\mathrm{P})$
    B) Less than $nH(\mathrm{P})$
    C) Greater than $nH(\mathrm{P})$
    D) Equal to $H(\mathrm{P})$
    E) Independent of n

14. The **expected codeword length** $\bar{L}$ can be minimized by:

    A) Assigning longer codewords to more probable symbols
    B) Assigning shorter codewords to more probable symbols
    C) Making all codeword lengths equal
    D) Randomly assigning codewords
    E) Maximizing the codebook size

15. In the context of lossless compression, **blocking** (i.e., grouping symbols into blocks) can improve compression efficiency in general because:

    A) It reduces the entropy of the source
    B) It increases the redundancy in the source
    C) It allows the coder to exploit inter-symbol correlations
    D) It simplifies the coding algorithm
    E) It reduces the size of the codebook

16. The **main disadvantage** of increasing block size in block coding is:

    A) Decreased compression efficiency
    B) Increased computational complexity and memory requirements
    C) Loss of data due to larger codewords
    D) Inability to decode the source sequence
    E) Reduced error detection capabilities

17. For a given code with codeword lengths $\{l_1, l_2, \ldots, l_M\}$, the **average codeword length** $\bar{L}$ is calculated as:

    A) $\bar{L} = \sum_{i=1}^{M} l_i$
    B) $\bar{L} = \frac{1}{M} \sum_{i=1}^{M} l_i$
    C) $\bar{L} = \sum_{i=1}^{M} \mathrm{P}(x_i) l_i$
    D) $\bar{L} = \max_i l_i$
    E) $\bar{L} = \min_i l_i$

Elements of Information Theory **2025-2026**

Lecture 4: Lossy data compression

Lecturer: S. Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In this lecture, we consider the encoding problem of a source from an alphabet $\mathcal{X}$. For simplicity, we consider **memoryless stationary sources** (or, i.i.d. source), i.e.,

$$\boxed{\mathrm{P}_{X^n} = \prod_{i=1}^{n} \mathrm{P}_{X_i} = \mathrm{P}_X^n}$$

In addition, we focus on finite alphabets, while the conclusion can be applied to more general alphabets and distributions. As in the lossless data compression problem, we define the encoder/decoder. The **encoder** f_n maps the source sequence from $\mathcal{X}^n$ into a set of binary codewords, say, $\mathcal{C}_n$. The **decoder** $g_n : \mathcal{C}_n \to \hat{\mathcal{X}}_n \subseteq \mathcal{X}^n$ reconstructs the source from the binary codeword, where $\hat{\mathcal{X}}_n := \{g_n(y) : y \in \mathcal{C}_n\}$ is the **reconstruction codebook**. We denote $\mathrm{Y} = f_n(\mathrm{X}^n) \in \mathcal{C}_n$ and $\hat{\mathrm{X}}^n = g_n(\mathrm{Y}) = g_n(f_n(\mathrm{X}^n))$.

Without loss of optimality, we may assume that g_n is invertible. Indeed, suppose that g_n is not invertible, i.e., there exist $y \neq y' \in \mathcal{C}_n$ such that $g_n(y) = g_n(y')$. In this case, one can merge the corresponding codewords by keeping only one of them (for instance, the shorter one) and reassigning all source sequences that map to y or y' to this codeword. The reconstruction remains unchanged, while the average codeword length is reduced.

Let us define the **separable distortion function**

$$d_n(x^n, \hat{x}^n) := \sum_{i=1}^{n} d(x_i, \hat{x}_i), \quad \forall\, n \in \mathbb{N},$$

where $d : (\mathcal{X}, \mathcal{X}) \to \mathbb{R}^+$ is a given single-letter distortion.

A code is called a (n, R_n, D_n)-code[8] if

$$\frac{1}{n}\mathbb{E}\left[l(f_n(\mathrm{X}^n))\right] = R_n$$
$$\frac{1}{n}\mathbb{E}\left[d\left(\mathrm{X}^n, g_n(f_n(\mathrm{X}^n))\right)\right] = D_n$$

4.1 Motivating examples

4.1.1 Quantizing a continuous source

The problem of quantization is to represent a sequence of analog symbols (i.e., from a continuous alphabet), with a sequence of discrete symbols. In this section, we are interested in quadratic distortion function.

Let us first look at the scalar quantization problem. We partition the real field $\mathbb{R}$ into M **quantization regions**: $\mathcal{R}_1, \ldots, \mathcal{R}_M$, and assign a **representation point** c_j to each region $\mathcal{R}_j$.

[8] In most text books, it is assumed that a fixed-length code with length nR_n bits is used. In this case, $R_n = \frac{\log|\mathcal{C}_n|}{n}$.

The **quantization function** is the encoder $f : \mathbb{R} \to \{1, \dots, M\}$, such that

$$f(x) = j, \quad \text{if } x \in \mathcal{R}_j.$$

The decoder is $g(i) := c_i$ Here we consider the distortion function $d(x, y) := |x - y|^2$.

For a given source sequence $\{X_j\}$, the average distortion is

$$\frac{1}{n} \sum_{j=1}^{n} |X_j - g(f(X_j))|^2.$$

When the source symbols X_j's are i.i.d., the average quantization error converges, due to the WLLN, to the expectation when $n \to \infty$

$$\frac{1}{n} \sum_{j=1}^{n} |X_j - g(f(X_j))|^2 \xrightarrow{n\to\infty} \mathbb{E}\left[|X - g(f(X))|^2\right]$$

which is called **mean-squared distortion** or **mean-squared error (MSE)**.

The objective now is to find a quantizer that minimizes the MSE. We have the following necessary conditions for the optimal scalar quantizers with minimum MSE.

For a given set of representation points $c_1 < c_2 < \cdots < c_M$, The optimal quantization region must be the **Voronoi regions**

$$\mathcal{R}_j = \mathcal{V}(c_j) := \left\{ x \in \mathbb{R} : \quad |x - c_j| = \min_{l=1,\dots,M} |x - c_l| \right\}$$

For a given set of quantization region $\{\mathcal{R}_j\}$, the optimal representation point of $\mathcal{R}_j$ must be the conditional mean of X conditional on $X \in \mathcal{R}_j$

$$c_j = \mathbb{E}[X \,|\, X \in \mathcal{R}_j]$$

The proof is not difficult and is omitted here.

For scalar input, the Voronoi quantization region are intervals, i.e.,

$$\mathcal{R}_j = [b_{j-1}, b_j)$$

where $b_1 < \cdots < b_{M-1}$ are the boundaries between intervals. From the above necessary condition, the following iterative algorithm is straightforward.

An iterative algorithm: Lloyd algorithm

- Choose an arbitrary initial set of M points $c_1 < \cdots < c_M$
- Find the quantization region $\mathcal{R}_j = [b_{j-1}, b_j)$ with

$$b_j = (c_j + c_{j+1})/2, \quad j = 1, \dots, M - 1$$

- Update the representation points for $\{\mathcal{R}_j\}$ with

$$c_j = \frac{1}{\mathbb{P}\{U \in \mathcal{R}_j\}} \int_{b_{j-1}}^{b_j} x p_X(x) \mathrm{d}x, \quad j = 1, \dots, M$$

- Iterate until the improvement in MSE is negligible.

It is easy to verify that the Lloyd algorithm converges, since 1) the MSE is non-increasing with iterations and that 2) it is lower bounded by 0. But since the problem is not convex, there is no guarantee of global optimality and it may only converge to a local optimum.

The generalization to **vector quantization** is straightforward. We can take a block of n samples $\mathrm{X}^n \in \mathbb{R}^n$ and apply the encoding function: $f(x^n) : \mathbb{R}^n \to \{1, \ldots, M\}$. The only differences are now the quantization regions are not intervals, and the quadratic distortion becomes $d(x^n, y^n) := \|x^n - y^n\|^2 = \sum_i |x_i - y_i|^2$. The n-dimensional Voronoi regions is

$$\mathcal{R}_j = \mathcal{V}(c_j) := \{x^n \in \mathbb{R}^n : \|x^n - c_j\| = \min_l \|x^n - c_l\|\}$$

The quantizer maps $f(x^n) = j$, if $x^n \in \mathcal{R}_j$. And the Lloyd algorithm extends to the vector case trivially. As in the scalar case, the Lloyd algorithm is not optimal in general. More importantly, since the codebook size M increases with the block size n, exponentially in general, the Lloyd algorithm becomes infeasible for large n. Is it possible to find out the minimum MSE for a given codebook size? We shall answer this question in a more general setting later in this lecture.

4.1.2 Compressing a discrete source

Another example is the compression of a binary sequence with the Hamming distortion $d(x, y) := \mathbf{1}\{x \neq y\}$. The expected distortion is the bit error rate. For instance, compressing three bits from Bern($\frac{1}{2}$) to one bit, what is the bit error rate after decompression?

One way to encode is to keep the first bit and discard the rest. The decoder recovers the first bit without error and guesses the other two bits at random. The error rate is thus 1/3.

A better way is to use vector quantization. With one bit, we can have two codewords in the codebook. Let $c_1 = [000]$ and $c_2 = [111]$. The encoder is the minimum Hamming distance encoder, i.e., if there are more 0 than 1, then $f(x^3) = 1$, otherwise $f(x^3) = 2$. Note that the output can be represented with one bit. The chance without decoding error is 25% while the chance with one-bit errors is 75%. Thus, the bit error rate is $\frac{3}{4} \times \frac{1}{3} = \frac{1}{4}$. It is easy to verify (by exhaustion) in this case that we cannot do better.

In general, though, the answer is not clear. For instance, when $n = 30000$, what is the minimum bit error rate if it is to be compressed to 10000 bits? We shall have a clear answer at the end of this lecture.

4.2 Information rate-distortion function

Let us define the **information rate-distortion function** for a given source P_X.

$$\boxed{\phi(D) \equiv \phi(D, \mathrm{P}_X) := \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}_{\mathrm{P}_X \mathrm{P}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) \leq D} I(\mathrm{X}; \hat{\mathrm{X}}) = \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}_{\mathrm{P}_X \mathrm{P}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) \leq D} I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X})}$$

Note that the minimum should be replaced by infimum in the general case[9].

$D \mapsto \phi(D)$ is decreasing, convex, and continuous in $(0, \infty)$.

Proof. 1. From the definition, it is obvious that $\phi(D)$ is decreasing.

[9] For the discrete case, the constraint set $\{\mathrm{P}_{\hat{X}|X} : \mathbb{E} d(\mathrm{X}, \hat{\mathrm{X}}) \leq D\}$ is compact. Hence, the infimum is achievable and can be replaced with minimum.

2. The convexity is from the convexity of $\mathrm{P}_{\hat{X}|X} \mapsto I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X})$. Let $\mathrm{P}^{(\lambda)}_{\hat{X}|X} = \lambda \mathrm{P}'_{\hat{X}|X} + (1-\lambda)\mathrm{P}''_{\hat{X}|X}$ for some $\lambda \in [0,1]$, where $\mathrm{P}'_{\hat{X}|X}$ and $\mathrm{P}''_{\hat{X}|X}$ are two conditional probability kernels such that

$$\mathbb{E}_{\mathrm{P}_X \mathrm{P}'_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) \le D_1 \tag{4.1}$$

$$\mathbb{E}_{\mathrm{P}_X \mathrm{P}''_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) \le D_2 \tag{4.2}$$

Hence, $\mathbb{E}_{\mathrm{P}_X \mathrm{P}^{(\lambda)}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) = \lambda \mathbb{E}_{\mathrm{P}_X \mathrm{P}'_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) + (1-\lambda)\mathbb{E}_{\mathrm{P}_X \mathrm{P}''_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}) \le \lambda D_1 + (1-\lambda) D_2$, and

$$\begin{aligned} \phi(\lambda D_1 + (1-\lambda) D_2) &\le I(\mathrm{P}_X, \mathrm{P}^{(\lambda)}_{\hat{X}|X}) && \left(\text{definition of } \phi \text{ and } \mathrm{P}_{\hat{X}|X}\right) && (4.3) \\ &\le \lambda I(\mathrm{P}_X, \mathrm{P}'_{\hat{X}|X}) + (1-\lambda) I(\mathrm{P}_X, \mathrm{P}''_{\hat{X}|X}) && \left(\text{convexity of } \mathrm{P}_{\hat{X}|X} \mapsto I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X})\right). && (4.4) \end{aligned}$$

Since the above inequality holds for any $\mathrm{P}'_{\hat{X}|X}$ and $\mathrm{P}''_{\hat{X}|X}$ such that (4.1) and (4.2) are true, we can taking the minimum over $\mathrm{P}'_{\hat{X}|X}$ and $\mathrm{P}''_{\hat{X}|X}$, and have

$$\phi(\lambda D_1 + (1-\lambda) D_2) \le \lambda \phi(D_1) + (1-\lambda)\phi(D_2).$$

This proves the convexity of $\phi(D)$.

3. Any convex function defined in an open interval (a, b) is continuous inside it.

□

4.3 Operational rate-distortion function

A **rate-distortion pair** (R, D) is said to be achievable for a given source P^n_X if there exist a sequence of (n, R_n, D_n)-codes, $n = 1, \ldots,$ such that

$$\limsup_{n\to\infty} R_n \le R \tag{4.5}$$

$$\limsup_{n\to\infty} D_n \le D \tag{4.6}$$

In words, it means that when n is large enough, we can find a code with both rate smaller than $R+\varepsilon$ and distortion smaller than $D+\varepsilon$ for any $\varepsilon > 0$. The **rate-distortion region** $\mathcal{R} \equiv \mathcal{R}(\mathrm{P}_X)$ for a given source P^n_X is the closure of the set of achievable rate-distortion pairs.

The rate-distortion region $\mathcal{R}$ is convex.

Proof. Let $(R', D') \in \mathcal{R}$ and $(R'', D'') \in \mathcal{R}$. Consider the two sequence of codes that achieve (R', D') and (R'', D''), respectively. We argue that for any $\lambda \in [0,1]$, one can construct a sequence of codes that achieves $(\lambda R' + (1-\lambda)R'', \lambda D' + (1-\lambda)D'')$. For each n, the new encoder encodes the first $n_1 := \lceil \lambda n \rceil$ symbols using the first codebook[10] with $n_1 R'_{n_1}$ bits. The remaining $n_2 := n - n_1$ symbols are encoded with the second codebook with $n_2 R''_{n_2}$ bits. The number of encoded bits is $n_1 R'_{n_1} + n_2 R''_{n_2}$ and

$$\begin{aligned} \limsup_{n\to\infty} \frac{n_1 R'_{n_1} + n_2 R''_{n_2}}{n} &\le \limsup_{n\to\infty} \frac{n_1 R'_{n_1}}{n} + \limsup_{n\to\infty} \frac{n_2 R''_{n_2}}{n} && (4.7) \\ &\le \lambda R' + (1-\lambda) R'' && (4.8) \end{aligned}$$

The distortion is

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^{n_1} d(x_i, \hat{x}_i) + \sum_{i=n_1+1}^{n} d(x_i, \hat{x}_i) \right) &= \limsup_{n\to\infty} \frac{n_1 D'_{n_1}}{n} + \frac{n_2 D''_{n_2}}{n} && (4.9) \\ &\le \lambda D' + (1-\lambda) D''. && (4.10) \end{aligned}$$

□

[10] Precisely, the n_1-th codebook in the first sequence of codebooks.

The **(operational) rate-distortion function** $R(D)$ is the minimum of rate R such that (R, D) is achievable. Similarly, the **distortion-rate function** $D(R)$ is the minimum of distortion D such that (R, D) is achievable. The main result of the lecture is the following **rate-distortion theorem**.

Shannon's rate-distortion theorem For a memoryless stationary source $\mathrm{P}_{X^n} = \mathrm{P}_X^n$, the information and operational rate-distortion function coincide, i.e.,

$$R(D) = \phi(D).$$

In the following, we prove the theorem in two parts: the converse (lower bound, or necessary condition) and the achievability (upper bound, or sufficient condition).

4.3.1 Converse

The converse consists in proving that for any (n, R_n, D_n) code we must have $R_n \geq \phi(D_n) + o(1)$ where $o(1)$ vanishes with $n \to \infty$. With (4.5) and (4.6), this in turn implies that for any achievable rate-distortion pair (R, D), we must have $R \geq \phi(D)$.

For any (n, R_n, D_n) code and the corresponding decoder, $\frac{1}{n}\mathbb{E}[d_n(\mathrm{X}^n, \hat{\mathrm{X}}^n)] = D_n$. Then, we have

$$\begin{aligned} R_n &= \frac{1}{n}\mathbb{E}\left[l(f_n(\mathrm{X}^n))\right] \\ &= \frac{1}{n}\mathbb{E}\left[l(g_n^{-1}(\hat{\mathrm{X}}^n))\right] \\ &\geq \frac{1}{n}H(\hat{\mathrm{X}}^n) + o(1) \\ &\geq \frac{1}{n}I(\mathrm{X}^n;\hat{\mathrm{X}}^n) + o(1) && \text{(positivity of conditional entropy)} \\ &\geq \frac{1}{n}\min_{\mathrm{P}_{\hat{X}^n|X^n}:\mathbb{E}d_n(\mathrm{X}^n,\hat{\mathrm{X}}^n)\leq D_n} I(\mathrm{X}^n;\hat{\mathrm{X}}^n) + o(1) && \left(\text{relaxation to any } \mathrm{P}_{\hat{X}^n|X^n}\right) \\ &\geq \frac{1}{n}\min_{\mathrm{P}_{\hat{X}^n|X^n}:\mathbb{E}d_n(\mathrm{X}^n,\hat{\mathrm{X}}^n)\leq D_n} \sum_{i=1}^n I(\mathrm{X}_i;\hat{\mathrm{X}}_i) + o(1), && (4.11) \end{aligned}$$

where the last inequality is from the following result.

For a memoryless stationary source $\mathrm{P}_{X^n} = \mathrm{P}_X^n$, we have, for any $\mathrm{P}_{\hat{X}^n|X^n}$,

$$I(\mathrm{X}^n;\hat{\mathrm{X}}^n) \geq \sum_{i=1}^n I(\mathrm{X}_i;\hat{\mathrm{X}}_i).$$

Proof.

$$\begin{aligned} I(\mathrm{X}^n;\hat{\mathrm{X}}^n) &= \sum_{i=1}^n I(\mathrm{X}_i;\hat{\mathrm{X}}^n\,|\,\mathrm{X}^{i-1}) && \text{(chain rule)} && (4.12) \\ &= \sum_{i=1}^n I(\mathrm{X}_i;\hat{\mathrm{X}}^n\,|\,\mathrm{X}^{i-1}) + I(\mathrm{X}_i;\mathrm{X}^{i-1}) && \text{(independence between the } \mathrm{X}_i\text{'s)} && (4.13) \\ &= \sum_{i=1}^n I(\mathrm{X}_i;\hat{\mathrm{X}}^n, \mathrm{X}^{i-1}) && \text{(chain rule)} && (4.14) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} I(\mathrm{X}_i;\hat{\mathrm{X}}_i) + I(\mathrm{X}_i;\{\hat{\mathrm{X}}_j, j \neq i\}, \mathrm{X}^{i-1}\,|\,\hat{\mathrm{X}}_i) \qquad \text{(chain rule)} && (4.15)\\
&\geq \sum_{i=1}^{n} I(\mathrm{X}_i;\hat{\mathrm{X}}_i) \qquad \text{(positivity of mutual information)} && (4.16)
\end{aligned}
$$

□

Note that each $I(\mathrm{X}_i;\hat{\mathrm{X}}_i)$ only depends on the marginal distribution $\mathrm{P}_{X_i\hat{X}_i} = \mathrm{P}_{X_i}\mathrm{P}_{\hat{X}_i|X_i}$ with $\mathrm{P}_{X_i} = \mathrm{P}_X$ by assumption. Therefore, we can write $I(\mathrm{X}_i;\hat{\mathrm{X}}_i) = I(\mathrm{P}_{X_i}, \mathrm{P}_{\hat{X}_i|X_i}) = I(\mathrm{P}_X, \mathrm{P}_{\hat{X}_i|X_i})$. Here, $\mathrm{P}_{\hat{X}_i|X_i}$ is the conditional pmf, i.e., a collection of M pmf $\mathrm{P}_{\hat{X}_i|X_i=x}$ for $x \in \mathcal{X}$. We recall that this can be represented by a matrix where each column is a pmf. We have

$$
\begin{aligned}
\sum_{i=1}^{n} I(\mathrm{X}_i;\hat{\mathrm{X}}_i) &= \sum_{i=1}^{n} I(\mathrm{P}_{\mathrm{X}}, \mathrm{P}_{\hat{X}_i|X_i}) && (4.17)\\
&\geq nI\left(\mathrm{P}_{\mathrm{X}}, \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i}\right) \qquad \text{(convexity of mutual information)} && (4.18)\\
&= nI(\mathrm{P}_{\mathrm{X}}, \tilde{\mathrm{P}}_{\hat{X}|X}), && (4.19)
\end{aligned}
$$

where $\tilde{\mathrm{P}}_{\hat{X}|X} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i}$ means that $\tilde{\mathrm{P}}_{\hat{X}|X=x} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i=x}$, $\forall\, x \in \mathcal{X}$. Let us now look at the distortion.

> For a memoryless stationary source $\mathrm{P}_{X^n} = \mathrm{P}_X^n$, we have, for any $\mathrm{P}_{\hat{X}^n|X^n}$,
>
> $$\frac{1}{n}\mathbb{E}_{\mathrm{P}_{X^n}\mathrm{P}_{\hat{X}^n|X^n}} d_n(\mathrm{X}^n, \hat{\mathrm{X}}^n) = \mathbb{E}_{\mathrm{P}_X\tilde{\mathrm{P}}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}),$$
>
> where $\tilde{\mathrm{P}}_{\hat{X}|X} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i}$.

Proof.

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}_{\mathrm{P}_{X^n}\mathrm{P}_{\hat{X}^n|X^n}} d_n(\mathrm{X}^n, \hat{\mathrm{X}}^n) &= \mathbb{E}_{\mathrm{P}_{X^n}\mathrm{P}_{\hat{X}^n|X^n}} \frac{1}{n}\sum_{i=1}^{n} d(\mathrm{X}_i, \hat{\mathrm{X}}_i)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathrm{P}_{X_i}\mathrm{P}_{\hat{X}_i|X_i}} d(\mathrm{X}_i, \hat{\mathrm{X}}_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{X}}\sum_{\hat{x}\in\mathcal{X}} \mathrm{P}_{X_i}(x)\mathrm{P}_{\hat{X}_i|X_i}(\hat{x}|x)d(x,\hat{x})\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{X}}\sum_{\hat{x}\in\mathcal{X}} \mathrm{P}_{X}(x)\mathrm{P}_{\hat{X}_i|X_i}(\hat{x}|x)d(x,\hat{x})\\
&= \sum_{x\in\mathcal{X}} \mathrm{P}_X(x) \sum_{\hat{x}\in\mathcal{X}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i}(\hat{x}|x)\right) d(x,\hat{x})\\
&= \sum_{x\in\mathcal{X}} \mathrm{P}_X(x) \sum_{\hat{x}\in\mathcal{X}} \tilde{\mathrm{P}}_{\hat{X}|X}(\hat{x}|x)d(x,\hat{x})\\
&= \mathbb{E}_{\mathrm{P}_X\tilde{\mathrm{P}}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}).
\end{aligned}
$$

□

Now, we can continue the lower bound (4.11) and write

$$
\begin{aligned}
R_n &\geq \frac{1}{n} \min_{\mathrm{P}_{\hat{X}^n|X^n} : \frac{1}{n}\mathbb{E}d_n(\mathrm{X}^n,\hat{\mathrm{X}}^n)\leq D_n} \sum_{i=1}^{n} I(\mathrm{X}_i;\hat{\mathrm{X}}_i) + o(1) \\
&\geq \min_{\mathrm{P}_{\hat{X}^n|X^n} : \frac{1}{n}\mathbb{E}d_n(\mathrm{X}^n,\hat{\mathrm{X}}^n)\leq D_n} I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X}) + o(1) && \left(\text{convexity of mutual information, } \mathrm{P}_{\hat{X}|X} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\hat{X}_i|X_i}\right) \\
&= \min_{\mathrm{P}_{\hat{X}^n|X^n} : \mathbb{E}d(\mathrm{X},\hat{\mathrm{X}})\leq D_n} I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X}) + o(1) \\
&= \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}d(\mathrm{X},\hat{\mathrm{X}})\leq D_n} I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X}) + o(1) \\
&= \phi(D_n) + o(1).
\end{aligned}
$$

Finally, for any achievable rate-distorsion pair (R, D), there exist a sequence of (n, R_n, D_n) codes such that (4.5) and (4.6) hold. Therefore,

$$
\begin{aligned}
R &\geq \limsup_{n\to\infty} R_n \\
&\geq \limsup_{n\to\infty} \phi(D_n) \\
&\geq \liminf_{n\to\infty} \phi(D_n) \\
&= \phi\left(\limsup_{n\to\infty} D_n\right) && (4.20) \\
&\geq \phi(D), && (4.21)
\end{aligned}
$$

where in (4.20) we use the fact that $\phi(\cdot)$ is decreasing and continuous; in (4.21) we use the fact that $\phi(\cdot)$ is decreasing and $\limsup_{n\to\infty} D_n \leq D$ from (4.6).

4.3.2 Achievability

The achievability part consists in proving the existence of a sequence of (n, R_n, D_n) codes such that (4.5) and (4.6) are satisfied. The proof is based on Shannon's random coding argument. Instead of showing explicitly that a code can achieve the given performance, Shannon introduced a brilliant idea of showing that the average performance of a randomly generated code ensemble is good enough, implying that there must exist a particular code in the random ensemble that is good. Note that the analysis of the average performance of a random code is much easier than analyzing a particular code.

In the following, we show that for any P_X and $\mathrm{P}_{\hat{X}|X}$, $(R = I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X}), D = \mathbb{E}_{\mathrm{P}_X \mathrm{P}_{\hat{X}|X}} d(\mathrm{X}, \hat{\mathrm{X}}))$ is an achievable rate-distortion pair. Let us fix n and $\varepsilon = n^{-\frac{1}{3}}$. Then, let us define $\mathrm{P}_{\hat{X}} := \mathrm{P}_{\hat{X}|X} \circ \mathrm{P}_X$. And set

$$R_n = I(P_X, P_{\hat{X}|X}) + 2\varepsilon H(\mathrm{P}_{\hat{X}}) + \varepsilon.$$

Random codebook generation Let us generate the reconstruction codebook $\hat{\mathcal{X}}_n$ by independently and randomly generating 2^{nR_n} codewords

$$\hat{\mathcal{X}}_n := \left\{\hat{\mathrm{X}}^n(1), \ldots, \hat{\mathrm{X}}^n(2^{nR_n})\right\}$$

each one according to the same distribution $\mathrm{P}_{\hat{X}}^n$. That is, each codeword contains i.i.d. symbols from the distribution $\mathrm{P}_{\hat{X}}$. The codebook is revealed to the encoder and decoder.

Encoding function The encoding function f_n returns the nR_n-bit binary representation of the index of the first codeword that is jointly typical with X^n, i.e., the first i such that $(\mathrm{X}^n, \hat{\mathrm{X}}^n(i)) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{X\hat{X}})$. If no such codeword exists, it returns the binary representation of the index $i = 1$. It is therefore a fixed-length coding scheme with codeword length nR_n bits. We can verify that

$$\limsup_{n\to\infty} R_n = I(\mathrm{P}_X, \mathrm{P}_{\hat{X}|X}).$$

Decoding function From the binary sequence, the decoding function recovers the index i and returns the codeword $\hat{\mathrm{X}}^n(i)$.

Distortion analysis Let us define

$$\mathcal{A}_{n,\varepsilon} \equiv \mathcal{A}_{n,\varepsilon}(\mathrm{X}^n, \hat{\mathcal{X}}_n) := \bigcup_{m=1}^{2^{nR_n}} \left\{ (\mathrm{X}^n, \hat{\mathrm{X}}^n(m)) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{X\hat{X}}) \right\}.$$

We have $1 = \mathbf{1}\{\mathcal{A}_{n,\varepsilon}\} + \mathbf{1}\{\overline{\mathcal{A}}_{n,\varepsilon}\}$. First, let us bound

$$\begin{aligned} \frac{1}{n}\mathbb{E}_{X^n,\hat{\mathcal{X}}_n}\left\{ d_n(\mathrm{X}^n, g_n(f_n(\mathrm{X}^n)))\mathbf{1}\{\mathcal{A}_{n,\varepsilon}\} \right\} &\le \mathbb{E}_{X^n,\hat{\mathcal{X}}_n}\left(D(1+\varepsilon)\mathbf{1}\{\mathcal{A}_{n,\varepsilon}\} \right) \quad \text{typical average lemma on } d(x,\hat{x}) \\ &\le D(1+\varepsilon). \end{aligned}$$

Then,

$$\begin{aligned} \frac{1}{n}\mathbb{E}_{X^n,\hat{\mathcal{X}}_n}\left\{ d_n(\mathrm{X}^n, g_n(f_n(\mathrm{X}^n)))\mathbf{1}\{\overline{\mathcal{A}}_{n,\varepsilon}\} \right\} &\le D_{\max}\mathbb{E}\mathbf{1}\{\overline{\mathcal{A}}_{n,\varepsilon}\} \\ &= D_{\max}\mathbb{P}\{\overline{\mathcal{A}}_{n,\varepsilon}\}, \end{aligned}$$

where $D_{\max} := \max_{x,\hat{x}\in\mathcal{X}} d(x,\hat{x})$ that is finite since $\mathcal{X}$ is finite. From both bounds, we have

$$\frac{1}{n}\mathbb{E}_{X^n,\hat{\mathcal{X}}_n} d_n(\mathrm{X}^n, g_n(f_n(\mathrm{X}^n))) \le D(1+\varepsilon) + D_{\max}\mathbb{P}\{\overline{\mathcal{A}}_{n,\varepsilon}\} =: D_n. \tag{4.22}$$

Note that the expectation in the above distortion analysis is over the source X^n and the codebook $\hat{\mathcal{X}}_n$. Therefore, there must exist at least a particular code from the random ensemble such that (4.22) is satisfied. In other words, we showed the existence of a (n, R_n, D_n) code for every n.

Since we can check that

- $\lim_{n\to\infty} n\varepsilon_n^2 = \infty$;
- $\lim_{n\to\infty} \mathbb{P}\left\{ \mathrm{X}^n \in \mathcal{T}_{\varepsilon_n/2}^{(n)}(\mathrm{P}_X) \right\} = 1$, from (2.10) since $n(\varepsilon_n/2)^2 \to \infty$;
- $\lim_{n\to\infty} n(R_n - I(\mathrm{X};\hat{\mathrm{X}}) - 2\varepsilon_n H(\hat{\mathrm{X}})) = \lim_{n\to\infty} n\varepsilon_n = \infty$,

we can apply the covering lemma (2.42) to have

$$\lim_{n\to\infty} \mathbb{P}\{\overline{\mathcal{A}}_{n,\varepsilon_n}\} = 0.$$

Therfore, we have

$$\limsup_{n\to\infty} R_n = I(\mathrm{X};\hat{\mathrm{X}}) = R \tag{4.23}$$

$$\limsup_{n\to\infty} D_n = \limsup_{n\to\infty} D(1+\varepsilon_n) + D_{\max}\mathbb{P}\{\overline{\mathcal{A}}_{n,\varepsilon_n}\} = D, \tag{4.24}$$

which proves the achievability of the rate-distortion pair (R, D).

4.4 Evaluation of R(D)

We have characterized the exact rate-distortion function $R(D)$ as an optimization problem.

$$R(D) = \phi(D) = \min_{\mathrm{P}_{\hat{X}|X}:\mathbb{E}d(\mathrm{X},\hat{\mathrm{X}})\le D} I(\mathrm{X};\hat{\mathrm{X}}).$$

In the following, we look at two examples for which we can solve the optimization problem explicitly.

4.4.1 Bernoulli source with Hamming distortion

Let us consider $\mathrm{X} \sim \mathrm{Bern}(\lambda)$ with $\lambda \in [0, \frac{1}{2}]$, and the Hamming distortion function $d(x, \hat{x}) = \mathbf{1}(x \neq \hat{x}) = x \oplus \hat{x}$ where the sum is in the binary field. We first consider $D \leq \frac{1}{2}$.

We have the following lower bound

$$\begin{aligned} I(\mathrm{X};\hat{\mathrm{X}}) &= H(\mathrm{X}) - H(\mathrm{X}|\hat{\mathrm{X}}) \\ &= H_2(\lambda) - H(\mathrm{X} \oplus \hat{\mathrm{X}}|\hat{\mathrm{X}}) \qquad \text{(translation preserves entropy)} \\ &\geq H_2(\lambda) - H(\mathrm{X} \oplus \hat{\mathrm{X}}) \qquad \text{(conditioning reduces entropy)} \\ &= H_2(\lambda) - H_2(\mathbb{E}d(\mathrm{X}, \hat{\mathrm{X}})), \qquad (\mathrm{X} \oplus \hat{\mathrm{X}} \text{ is binary, } \mathbb{E}d(\mathrm{X}, \hat{\mathrm{X}}) = \mathbb{E}(\mathrm{X} \oplus \hat{\mathrm{X}})) \\ &\geq H_2(\lambda) - H_2(D), \qquad \left(H_2(\cdot) \text{ is increasing in } [0, \frac{1}{2}]\right), \end{aligned}$$

which implies that

$$R(D) \geq (H_2(\lambda) - H_2(D))^+, \quad D \leq \frac{1}{2}$$

since mutual information is non-negative.

Then, we shall show that the lower bound is achievable. When $D \geq \lambda$, we simply let $\hat{\mathrm{X}} = 0$. The distortion is $\mathbb{E}d(\mathrm{X}, \hat{\mathrm{X}}) = \mathrm{P}(\mathrm{X} = 1) = \lambda \leq D$ and $I(\mathrm{X}; \hat{\mathrm{X}}) = 0$, implying the achievability of the lower bound. It is thus without loss of generality to consider $D < \lambda$. To achieve the lower bound, we need to find $\mathrm{P}_{\hat{X}|X}$ such that $\mathrm{X} \oplus \hat{\mathrm{X}} \sim \mathrm{Bern}(D)$ and is independent of $\hat{\mathrm{X}}$. It follows that let the joint probability of $(\mathrm{X}, \hat{\mathrm{X}})$ can be represented by the following

	$\mathrm{X} = 0$	$\mathrm{X} = 1$
$\hat{\mathrm{X}} = 0$	$1 - \lambda - D + a$	a
$\hat{\mathrm{X}} = 1$	$D - a$	$\lambda - a$

From the assumption,

$$\begin{aligned} \mathrm{P}(\mathrm{X} = 0, \hat{\mathrm{X}} = 0) &= \mathrm{P}(\mathrm{X} \oplus \hat{\mathrm{X}} = 0, \hat{\mathrm{X}} = 0) &(4.25) \\ &= \mathrm{P}(\mathrm{X} \oplus \hat{\mathrm{X}} = 0)\mathrm{P}(\hat{\mathrm{X}} = 0), &(4.26) \end{aligned}$$

implying $1 - \lambda - D + a = (1 - D)(1 - \lambda - D + 2a)$, or explicitly,

$$a = \frac{1 - \lambda - D}{1 - 2D} D.$$

We just proved that $R(D) = (H_2(\lambda) - H_2(D))^+$ when $D \leq \frac{1}{2}$. Since $R(D)$ is decreasing, we must have $R(D) = 0$ for $D > \frac{1}{2}$. Hence,

$$R(D) = (H_2(\lambda) - H_2(D))^+, \quad D \geq 0.$$

Finally, if $\lambda > \frac{1}{2}$, we define $\mathrm{X}' = \mathrm{X} \oplus 1 \sim \mathrm{Bern}(\lambda')$ with $\lambda' = 1 - \lambda$. The rate-distortion function for X' is known from above $(H_2(\lambda') - H_2(D))^+ = (H_2(\lambda) - H_2(D))^+$. We argue that this is also the rate-distortion function of X, since

$$\begin{aligned} \phi(D) &= \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}d(\mathrm{X}, \hat{\mathrm{X}}) \leq D} I(\mathrm{X}; \hat{\mathrm{X}}) &(4.27) \\ &= \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}d(\mathrm{X} \oplus 1, \hat{\mathrm{X}} \oplus 1) \leq D} I(\mathrm{X} \oplus 1; \hat{\mathrm{X}} \oplus 1) &(4.28) \\ &= \min_{\mathrm{P}_{\hat{X}|X} : \mathbb{E}d(\mathrm{X}', \hat{\mathrm{X}}') \leq D} I(\mathrm{X}'; \hat{\mathrm{X}}') &(4.29) \\ &= \min_{\mathrm{P}_{\hat{X}'|X'} : \mathbb{E}d(\mathrm{X}', \hat{\mathrm{X}}') \leq D} I(\mathrm{X}'; \hat{\mathrm{X}}') &(4.30) \\ &= \phi(D, \mathrm{P}_{X'}) &(4.31) \end{aligned}$$

The rate-distortion function for $X \sim \text{Bern}(\lambda)$, $\lambda \in [0,1]$, with Hamming distortion function is

$$R(D) = (H_2(\lambda) - H_2(D))^+, \quad D \geq 0,$$

The optimal $P_{\hat{X}|X}$ is such that, $X \oplus \hat{X}$ and $\hat{X}$ are independent.

When $D = 0$ with Hamming distortion, we have $X = \hat{X}$ almost surely. We recover the result of lossless data compression:

$$R(0) = I(X;\hat{X}) = I(X;X) = H(X).$$

Now we are ready to answer the question raised at the beginning of the lecture: what is the minimum Hamming distortion when we compresse $n = 30000$ uniform bits into 10000 bits. We have $n = 30000$ and $R_n = 10000/30000 = 1/3$. An optimal (n, R_n, D_n) code at this length should be close to the rate distortion function, i.e., $R_n \approx R(D_n) = 1 - H_2(D_n)$. The distortion D_n is such that $H_2(D_n) = 2/3$, i.e., $D_n \approx 0.175$ which is smaller than the 0.25 achieved with $n = 3$.

4.4.2 Gaussian source with quadratic distortion

Let $X \sim \mathcal{N}(0, \sigma^2)$. We define $d(x, y) := (x - y)^2$, i.e., the quadratic distortion function. Then,

$$\begin{aligned}
I(X;Y) &= h(X) - h(X|Y) \\
&= \frac{1}{2}\log(2\pi e\sigma^2) - h(X - Y|Y) && \text{(translation preserves differential entropy)} \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - h(X - Y) && \text{(conditioning reduces differential entropy)} \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi e \mathsf{Var}(X - Y)) && \text{(Gaussian dist. maximizes diff. entropy under variance constraint)} \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi e D) && \left(\mathsf{Var}(U) \leq \mathbb{E}U^2\right) \\
&= \frac{1}{2}\log\frac{\sigma^2}{D},
\end{aligned}$$

implying that

$$R(D) \geq \frac{1}{2}\log^+\frac{\sigma^2}{D},$$

since mutual information is always positive.

If $D \geq \sigma^2$, we simply let $Y = 0$, then $I(X;Y) = 0$ and $\mathbb{E}d(X,Y) = \sigma^2 \leq D$, implying $R(D) = 0$. Let us now consider the non-trivial case $\sigma^2 \geq D$. Note that the above lower bound could actually be achieved if there exists $P_{Y|X}$ such that $X - Y \sim \mathcal{N}(0, D)$ and is independent of Y. Thus, (X, Y) are jointly Gaussian. Without loss of generality, we have $Y = \alpha X + Z$ where $Z \sim \mathcal{N}(0, \beta)$ is independent of X. We should have

$$\mathbb{E}((X - Y)Y) = \alpha\sigma^2 - (\alpha^2\sigma^2 + \beta) = 0 \tag{4.32}$$

$$(1 - \alpha)^2\sigma^2 + \beta = D \tag{4.33}$$

from which we have $\alpha = 1 - \frac{D}{\sigma^2}$ and $\beta = \alpha D$.

The rate-distortion function for $\mathrm{X} \sim \mathcal{N}(0, \sigma^2)$ with quadratic distortion function is

$$R(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}, \quad D > 0.$$

The optimal $\mathrm{P}_{Y|X}$ is such that, conditional on $\mathrm{X} = x$,

$$\mathrm{Y} \sim \mathcal{N}\left((1 - \frac{D}{\sigma^2})x, (1 - \frac{D}{\sigma^2})D\right).$$

Exercises[11]

1. One-bit quantization of a single Gaussian random variable [CT 10.1]. Let $\mathrm{X} \sim \mathcal{N}(0,\sigma^2)$ and let the distortion measure be squared error. Here we do not allow block descriptions. Show that the optimum reproduction points for 1-bit quantization are $\pm\sqrt{\frac{2}{\pi}}\sigma$ and that the expected distortion for 1-bit quantization is $\frac{\pi-2}{\pi}\sigma^2$. Compare this with the distortion-rate function $D = \sigma^2 2^{-2R}$ for $R = 1$.

2. Rate distortion for uniform source with Hamming distortion [CT 10.5]. Consider a source X uniformly distributed on the set $\{1,2,...,m\}$. Find the rate-distortion function for this source with the Hamming distortion.

3. Simplification [CT 10.13]. Suppose that $\hat{\mathcal{X}} = \mathcal{X} = \{1,2,3,4\}$, $\mathrm{P}(i) = \frac{1}{4}$, $i = 1,2,3,4$, and $\mathrm{X}_1, \mathrm{X}_2, \ldots$ are are i.i.d.~ P. The distortion function $d(x,\hat{x})$ is given by

	1	2	3	4
1	0	0	1	1
2	0	0	1	1
3	1	1	0	0
4	1	1	0	0

   - Find $R(0)$, the rate necessary to describe the process with zero distortion.
   - Find the rate-distortion function $R(D)$. There are some irrelevant distinctions in alphabets $\mathcal{X}$ and $\hat{\mathcal{X}}$, which allow the problem to be collapsed.
   - Suppose that we have a nonuniform distribution $\mathrm{P}(i) = p_i$, $i = 1,2,3,4$. What is $R(D)$?

4. Rate-distortion [CT 10.18]. Let $d(x,\hat{x})$ be a distortion function. We have a source $\mathrm{X} \sim \mathrm{P}$. Let $R(D)$ be the associated rate-distortion function.

   - Find $\tilde{R}(D)$ in terms of $R(D)$, where $\tilde{R}(D)$ is the rate-distortion function associated with the distortion $\tilde{d}(x,\hat{x}) = d(x,\hat{x}) + a$ for some constant $a > O$. (They are not equal.)
   - Now suppose that $d(x,\hat{x}) \geq 0$ for all $x,\hat{x}$ and define a new distortion function $d^*(x,\hat{x}) = bd(x,\hat{x})$, where b is some number ≥ 0. Find the associated rate distortion function $R^*(D)$ in terms of R(D).
   - Let $\mathrm{X} \sim \mathcal{N}(0,\sigma^2)$ and $d(x,\hat{x}) = 5(x-\hat{x})^2 + 3$. What is $R(D)$?

5. Rate-distortion [CT 10.20]. Consider the standard rate distortion problem, X_i i.i.d. ~ P, $\mathrm{X}^n \to i(\mathrm{X}^n) \to \hat{\mathrm{X}}^n$, the alphabet size $|i(\cdot)| = 2^{nR}$. Consider two distortion criteria $d_1(x,\hat{x})$ and $d_1(x,\hat{x})$. Suppose that $d_1(x,\hat{x}) \leq d_2(x,\hat{x})$ for all $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$. Let $R_1(D)$ and $R_2(D)$ be the corresponding rate-distortion functions.

   - Find the inequality relationship between $R_1(D)$ and $R_2(D)$.
   - Suppose that we must describe the source $\{\mathrm{X}_i\}$ at the minimum rate R achieving $d_1(\mathrm{X}^n, \hat{\mathrm{X}}_1^n) \leq D$ and $d_2(\mathrm{X}^n, \hat{\mathrm{X}}_2^n) \leq D$. Thus,
$$\mathrm{X}^n \to i(\mathrm{X}^n) \to \begin{cases} \hat{\mathrm{X}}_1^n(i(\mathrm{X}^n)) \\ \hat{\mathrm{X}}_2^n(i(\mathrm{X}^n)) \end{cases}$$
     and $|i(\cdot)| = 2^{nR}$.
     Find the minimum rate R.

6. Different alphabets [Gallager 9.1]. Let the source alphabet be $\mathcal{X} = \{0,1\}$ and the reconstruction alphabet be $\hat{\mathcal{X}} = \{0,1,2\}$. Define the distortion function
$$d(x,\hat{x}) = \begin{cases} 0, & x = \hat{x}, \\ 1, & \hat{x} = 2, \\ +\infty, & \text{otherwise} \end{cases}$$

   Find the rate-distortion function $R(D)$.

[11] The citations "CT", "CK" refer to Cover-Thomas, Csiszár-Körner, respectively.

Quiz (unique correct answer)

1. The **rate-distortion function** $R(D)$ for a source describes:

   A) The minimum achievable rate R for a given maximum distortion D.

   B) The maximum achievable rate R for a given minimum distortion D.

   C) The minimum achievable distortion D for a given rate R.

   D) The maximum achievable distortion D for a given rate R.

   E) The average mutual information between the source and its reconstruction.

2. In the context of rate-distortion theory, a **distortion measure** is:

   A) A function that quantifies the similarity between two probability distributions.

   B) A metric that measures the loss of information during transmission over a noisy channel.

   C) A function $d(x, \hat{x})$ that quantifies the cost of representing symbol x by $\hat{x}$.

   D) The difference between the entropy of the source and the mutual information.

   E) A parameter that adjusts the trade-off between rate and distortion.

3. The **rate-distortion theorem** states that for a memoryless stationary source P and a given distortion measure d, the minimal rate $R(D)$ is given by:

   A) $R(D) = H(\mathrm{P}) - D$

   B) $R(D) = \min_{\mathrm{P}(\hat{\mathrm{X}}|\mathrm{X}):\mathbb{E}[d(\mathrm{X},\hat{\mathrm{X}})]\leq D} I(\mathrm{X};\hat{\mathrm{X}})$

   C) $R(D) = \max_{P(\hat{\mathrm{X}}|\mathrm{X}):\mathbb{E}[d(\mathrm{X},\hat{\mathrm{X}})]\leq D} H(\hat{\mathrm{X}})$

   D) $R(D) = D \cdot H(\mathrm{X})$

   E) $R(D) = \mathbb{E}[d(\mathrm{X}, \hat{\mathrm{X}})]$

4. The function $R(D)$ is **convex** in D because:

   A) Mutual information is a convex function of the joint distribution.

   B) The set of achievable rate-distortion pairs is convex.

   C) Distortion measures are always convex functions.

   D) The entropy function is concave.

   E) The convexity of $R(D)$ depends on the specific source and distortion measure.

5. For a Gaussian memoryless source with variance σ^2 and mean squared error distortion measure, the rate-distortion function is:

   A) $R(D) = \frac{1}{2} \log\left(\frac{\sigma^2}{D}\right)$ bits per symbol

   B) $R(D) = \sigma^2 D$

   C) $R(D) = \log(\sigma^2 + D)$

   D) $R(D) = \frac{1}{2} \log(\sigma^2 D)$

   E) $R(D) = \frac{\sigma^2}{D}$

6. The operational meaning of the rate-distortion function is:

   A) It provides the exact number of bits needed to encode the source without any distortion.

   B) It represents the maximum distortion achievable at a given coding rate.

C) It characterizes the fundamental limit on how much a source can be compressed subject to a distortion constraint.

D) It is a theoretical concept without practical relevance.

E) It gives the error probability of the optimal code.

7. The **distortion-rate function** $D(R)$ is:

A) The inverse function of the rate-distortion function $R(D)$.

B) Always equal to $R(D)$ for any source.

C) The maximum distortion that can be achieved for a given rate R.

D) The difference between the source entropy and the rate.

E) Defined only for Gaussian sources.

8. In vector quantization, increasing the vector dimension (block length):

A) Decreases computational complexity.

B) Decreases the compression efficiency.

C) Does not affect the rate-distortion performance.

D) Can improve compression efficiency but increases computational complexity exponentially.

E) Makes the quantizer design easier.

9. For a discrete memoryless source with alphabet $\mathcal{X}$ and distortion measure $d(x, \hat{x})$, the rate-distortion function can be computed by solving:

A) A linear programming problem.

B) A convex optimization problem over $\mathrm{P}_{\hat{X}|X}$.

C) A set of differential equations.

D) An eigenvalue problem.

E) A combinatorial optimization problem.

10. The **rate-distortion function** $R(D)$ for a memoryless uniform binary source with Hamming distortion measure and distortion level $D \leq \frac{1}{2}$ is:

A) $R(D) = 1 - H_2(D)$, where $H_2(D)$ is the binary entropy function.

B) $R(D) = H_2(D)$

C) $R(D) = D$

D) $R(D) = \log_2 D$

E) $R(D) = 1 + H_2(D)$

11. In the context of lossy compression, a **codebook** refers to:

A) A set of codewords used to represent the source symbols exactly.

B) A dictionary of source sequences used in lossless compression.

C) A set of reproduction symbols $\hat{x}$ used to approximate the source symbols.

D) The set of all possible sequences generated by the source.

E) A lookup table for error correction codes.

12. The **trade-off** between rate and distortion in lossy compression implies that:

A) Increasing the rate always increases the distortion.

B) Decreasing the distortion requires increasing the coding rate.

C) There is no relationship between rate and distortion.

D) The optimal rate is achieved at maximum distortion.

E) The distortion is minimized when the rate is zero.

13. The **operational meaning** of the rate-distortion function $R(D)$ is:

A) It represents the maximum rate at which data can be transmitted with zero distortion.

B) It characterizes the minimum achievable rate for encoding a source within a specified average distortion.

C) It provides the exact distortion for any given coding rate.

D) It is the entropy of the source conditioned on the distortion.

E) It is always equal to the mutual information between the source and its reconstruction.

14. The **rate-distortion function** $R(D)$ is:

A) Always a concave function of D.

B) Always a convex function of D.

C) Always a linear function of D.

D) Neither convex nor concave in general.

E) Always constant, independent of D.

15. In rate-distortion theory, the **distortion measure** $d(x, \hat{x})$ is significant because:

A) It determines the codebook size.

B) It affects the source entropy.

C) It defines how errors between the source and reconstruction are quantified.

D) It has no impact on the rate-distortion function.

E) It is used to adjust the transmission power.

16. For a discrete memoryless stationary source, the rate-distortion function $R(D)$ becomes zero when the distortion D is:

A) Equal to zero.

B) Equal to the variance of the source.

C) Equal to its maximum allowable value $D_{\max}$.

D) Less than the entropy of the source.

E) Greater than the entropy of the source.

17. In source coding, **entropy** represents the fundamental limit for:

A) Lossy compression.

B) Lossless compression.

C) Both lossy and lossless compression.

D) Channel coding.

E) Error probabilities.

18. The rate-distortion function $R(D)$ equals the entropy $H(\mathrm{P})$ of the source when:

A) The distortion D is zero.

B) The distortion D is at its maximum allowable value.

C) The source is Gaussian.

- D) The source symbols are independent and identically distributed.
- E) The distortion measure is irrelevant.

19. The relationship between **mutual information** $I(\mathrm{X};\hat{\mathrm{X}})$ and the rate-distortion function $R(D)$ is that:

   - A) $R(D) = H(\mathrm{X}) - I(\mathrm{X};\hat{\mathrm{X}})$
   - B) $R(D) = I(\mathrm{X};\hat{\mathrm{X}})$ minimized over all conditional distributions satisfying $\mathbb{E}[d(\mathrm{X},\hat{\mathrm{X}})] \leq D$
   - C) $R(D) = I(\mathrm{X};\hat{\mathrm{X}})$ maximized over all conditional distributions
   - D) $R(D) = H(\hat{\mathrm{X}})$
   - E) $R(D) = D \cdot I(\mathrm{X};\hat{\mathrm{X}})$

20. In lossy compression, a **codebook** is used to:

   - A) Map source sequences to unique binary codewords.
   - B) Map source sequences to reproduction sequences with acceptable distortion.
   - C) Encrypt the source data.
   - D) Detect and correct errors during transmission.
   - E) Store the original source sequences.

Elements of Information Theory **2025-2026**

Lecture 5: Channel coding

Lecturer: S. Yang

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In this lecture, we consider the channel coding problem. In a nutshell, the channel coding problem is to encode a (discrete) message such that the decoder can recover it reliably from a *noisy* observation of the encoded signal.

5.1 Communication channels

One can describe the communication problem as follows. The **source** selects a **message** W uniformly from a **source alphabet** $\mathcal{W} := \{1, \ldots, 2^{nR_n}\}$, and encodes it into a **input sequence, or transmitted sequence** X^n with symbols from a given **input alphabet** $\mathcal{X}$; the **destination** observes accordingly a **output sequence, or received sequence** Y^n with symbols from a given **output alphabet** $\mathcal{Y}$, and compute an estimate $\hat{\mathrm{W}}$ of the source message. The sequence can be understood to take place in time, with n being the duration of the communication.

Specifically, for a given message w, the general encoding function generates a sequence of symbols such that each input symbol at time instant i can depend on the message w, the past transmitted symbols x^{i-1}, and the past observations y^{i-1} of the receiver when feedback is available

$$x_i = f_i(w, x^{i-1}, y^{i-1}), i = 1, \ldots, n.$$

If feedback is not available, then the encoded sequence only depends on the message w and the past transmitted symbols, not the observations of the receiver. $R_n = \frac{\log|\mathcal{W}|}{n}$ is called the **coding rate** or **communication rate**, i.e., the average number of bits **per channel use** of the encoding scheme. We call such a coding scheme a (n, R_n)-scheme.

The **decoding function** is $g_n : \mathcal{Y}^n \to \mathcal{W}$, and the **decoded message** is

$$\hat{\mathrm{W}} := g_n(\mathrm{Y}^n).$$

The **decoding error event** is $\mathrm{E} := \mathbf{1}\{\mathrm{W} \neq g_n(\mathrm{Y}^n)\}$. The **probability of error** is $P_{e,n} := \mathbb{P}\{\mathrm{W} \neq g_n(\mathrm{Y}^n)\}$. We can also call a (n, R_n)-coding scheme a $(n, R_n, P_{e,n})$-coding scheme if there exists a decoding function such that $\mathbb{P}\{\mathrm{W} \neq g_n(\mathrm{Y}^n)\} \leq P_{e,n}$.

More generally, the whole communication process can be described by the joint distribution of $(\mathrm{W}, \mathrm{X}^n, \mathrm{Y}^n, \hat{\mathrm{W}})$ as

$$\boxed{\mathrm{P}_{WX^nY^n\hat{W}} = \mathrm{P}_W \prod_{i=1}^{n} \mathrm{P}_{X_iY_i|X^{i-1}Y^{i-1}W}\mathrm{P}_{\hat{W}|Y^n} = \mathrm{P}_W \prod_{i=1}^{n} \mathrm{P}_{X_i|X^{i-1}Y^{i-1}W}\mathrm{P}_{Y_i|X^iY^{i-1}W}\mathrm{P}_{\hat{W}|Y^n}}$$

where P_W is a uniform distribution over $\mathcal{W}$; $\mathrm{P}_{X_i|X^{i-1}Y^{i-1}W}$ describes how the input symbol at time instant i depends on the available information at the encoder;[12]; $\mathrm{P}_{Y_i|X^iY^{i-1}W}$ describes how the nature selects the output at time instant i given the past inputs, outputs, and the message. In most cases, it is without loss of generality to consider that the current output only depends on the past inputs and outputs. Specifically, we assume that

$$\boxed{\mathrm{W} \to (\mathrm{X}^i, \mathrm{Y}^{i-1}) \to \mathrm{Y}_i, \quad \text{or, equivalently} \quad \mathrm{P}_{Y_i|X^iY^{i-1}W} = \mathrm{P}_{Y_i|X^iY^{i-1}}} \tag{5.1}$$

[12] In most practical cases, $f_i(w, x^{i-1}, y^{i-1})$ is a deterministic function and $\mathrm{P}_{X_i|X^{i-1}Y^{i-1}W}$ is just a set of Dirac functions.

For simplicity, we consider both $\mathcal{X}$ and $\mathcal{Y}$ discrete, unless otherwise is specified. The conclusions in this lecture also apply to general alphabets.

5.2 Memoryless stationary channels

In particular, we are interested in **memoryless stationary channels**, i.e.,

$$\boxed{\mathrm{P}_{Y_i|X^iY^{i-1}} = \mathrm{P}_{Y_i|X_i} = \mathrm{P}_{Y|X}, \quad i = 1, \dots, n} \tag{5.2}$$

for some $\mathrm{P}_{Y|X}$, where the first equality means **memoryless** and the second one means **stationary**. In words, the current output of a memoryless channel only depends on the current input, given the past inputs and outputs. Therefore, the channel is completely parameterized by the "single-letter" conditional distribution $\mathrm{P}_{Y|X}$, which can be understood as the instantaneous input-output relationship. From (5.1) and the first equality in (5.2), we have the following fundamental property of a memoryless channel.

For a memoryless channel, we have

$$(\mathrm{W}, \mathrm{X}^{i-1}, \mathrm{Y}^{i-1}) \to \mathrm{X}_i \to \mathrm{Y}_i$$

For a memoryless channel without feedback, we have

$$\mathrm{Y}^{i-1} \to (\mathrm{W}, \mathrm{X}^{i-1}) \to \mathrm{X}_i \to \mathrm{Y}_i$$

and

$$\mathrm{P}_{Y^n|X^n} = \prod_{i=1}^{n} \mathrm{P}_{Y_i|X_i}$$

Indeed, without feedback, we have $\mathrm{P}_{X_i|Y^{i-1}X^{i-1}} = \mathrm{P}_{X_i|X^{i-1}}$, so that

$$\begin{aligned}
\mathrm{P}_{X^nY^n} &= \prod_{i=1}^{n} \mathrm{P}_{X_iY_i|X^{i-1}Y^{i-1}} \\
&= \prod_{i=1}^{n} \mathrm{P}_{X_i|X^{i-1}Y^{i-1}} \mathrm{P}_{Y_i|X^iY^{i-1}} \\
&= \prod_{i=1}^{n} \mathrm{P}_{X_i|X^{i-1}} \mathrm{P}_{Y_i|X_i} \\
&= \mathrm{P}_{X^n} \prod_{i=1}^{n} \mathrm{P}_{Y_i|X_i}.
\end{aligned}$$

A rate R is **achievable** in a given channel if there exist a sequence of $(n, R_n, P_{e,n})$-schemes such that

$$\liminf_{n\to\infty} R_n \geq R, \tag{5.3}$$

$$\lim_{n\to\infty} P_{e,n} = 0. \tag{5.4}$$

That is, one can communication R bits per channel use in average with an arbitrarily small probability of error.

The **(operational) channel capacity** $C \equiv C(\mathrm{P}_{Y|X})$, of the channel $\mathrm{P}_{Y|X}$ is the supremum of all achievable rates. The **information channel capacity** $C_i \equiv C_i(\mathrm{P}_{Y|X})$, of the channel $\mathrm{P}_{Y|X}$ is defined as

$$C_i := \max_{\mathrm{P}_X} I(\mathrm{P}_X, \mathrm{P}_{Y|X}).$$

The main result of this lecture is Shannon's channel coding theorem.

Shannon's channel coding theorem For a memoryless stationary channel $\mathrm{P}_{Y|X}$, we have

$$C = C_i.$$

5.3 Converse

The converse consists in proving that any achievable rate must be smaller than C_i. The proof only relies on the following two Markov chains:

$$\mathrm{W} \to \mathrm{Y}^n \to \hat{\mathrm{W}} \tag{5.5}$$

$$(\mathrm{W}, \mathrm{X}^{i-1}, \mathrm{Y}^{i-1}) \to \mathrm{X}_i \to \mathrm{Y}_i \tag{5.6}$$

which are satisfied by any encoding and decoding schemes and memoryless channels.

Fano's inequality relates the probability of error to the (conditional) entropy:

$$H(\mathrm{W} \mid \hat{\mathrm{W}}) \le H_2(P_{e,n}) + P_{e,n} \log(|\mathcal{W}| - 1). \tag{5.7}$$

Proof of Fano's inequality (5.7). Let $M := |\mathcal{W}|$. First, we have

$$\begin{aligned} H(\mathrm{W} \mid \hat{\mathrm{W}}) &\le H(\mathrm{P}_{W|\hat{W}} \| \mathrm{Q}_{W|\hat{W}} \mid \mathrm{P}_{\hat{W}}) \\ &= \mathbb{E}_{\mathrm{P}}\left[\log \frac{1}{q(\mathrm{W} \mid \hat{\mathrm{W}})}\right] \\ &= \mathrm{P}(\mathrm{W} = \hat{\mathrm{W}}) \log \frac{1}{1 - P_{e,n}} + \mathrm{P}(\mathrm{W} \neq \hat{\mathrm{W}}) \log \frac{M-1}{P_{e,n}} \\ &= H_2(P_{e,n}) + P_{e,n} \log(M - 1) \end{aligned}$$

where we let $q(w \mid \hat{w}) := \frac{P_{e,n}}{M-1}\mathbf{1}(w \neq \hat{w}) + (1 - P_{e,n})\mathbf{1}(w = \hat{w})$.

An alternative (and more standard) proof is also shown here.

$$\begin{aligned} H(\mathrm{W} \mid \hat{\mathrm{W}}) &= H(\mathrm{W} \mid \hat{\mathrm{W}}) + H(\mathrm{E} \mid \hat{\mathrm{W}}, \mathrm{W}) && (\text{E is a function of } (\mathrm{W}, \hat{\mathrm{W}})) \\ &= H(\mathrm{E}, \mathrm{W} \mid \hat{\mathrm{W}}) && (\text{chain rule}) \\ &= H(\mathrm{E} \mid \hat{\mathrm{W}}) + H(\mathrm{W} \mid \hat{\mathrm{W}}, \mathrm{E}) && (\text{chain rule}) \\ &= H(\mathrm{E} \mid \hat{\mathrm{W}}) + P_{e,n} H(\mathrm{W} \mid \hat{\mathrm{W}}, \mathrm{E} = 1) + (1 - P_{e,n}) H(\mathrm{W} \mid \hat{\mathrm{W}}, \mathrm{E} = 0) \\ &= H(\mathrm{E} \mid \hat{\mathrm{W}}) + P_{e,n} H(\mathrm{W} \mid \hat{\mathrm{W}}, \mathrm{E} = 1) && (\mathrm{W} = \hat{\mathrm{W}} \text{ when } \mathrm{E} = 0) \\ &\le H_2(P_{e,n}) + P_{e,n} \log(|\mathcal{W}| - 1) && (\text{entropy upper bound with finite alphabet}) \end{aligned}$$

□

Intuitively, when $P_{e,n} \to 0$, $H(\mathrm{W}\,|\,\hat{\mathrm{W}}) \to 0$, i.e., W becomes deterministic given $\hat{\mathrm{W}}$.

Note that from (5.5) $I(\mathrm{W};\mathrm{Y}^n) \geq I(\mathrm{W};\hat{\mathrm{W}}) = H(\mathrm{W}) - H(\mathrm{W}\,|\,\hat{\mathrm{W}})$ due to data processing. We have

$$\boxed{H(\mathrm{W}) \leq I(\mathrm{W};\mathrm{Y}^n) + \underbrace{H_2(P_{e,n}) + P_{e,n}\log(|\mathcal{W}|-1)}_{\delta(P_{e,n})}.}$$

Note that the above inequalities hold for any encoding/decoding schemes. Therefore, they are useful to establish fundamental limits of the related setting.

Let us proceed with the converse proof.

$$\begin{align}
nR_n &= \log|\mathcal{W}| \tag{5.8}\\
&= H(\mathrm{W}) \qquad \text{(uniform distribution of the message)} \notag\\
&\leq I(\mathrm{W};\mathrm{Y}^n) + H_2(P_{e,n}) + P_{e,n}\log(|\mathcal{W}|-1) \qquad \text{(Fano's inequality)} \notag\\
&\leq I(\mathrm{W};\mathrm{Y}^n) + 1 + P_{e,n}nR_n \qquad \text{(from (5.8) and } H_2(\cdot) \leq 1\text{)} \notag
\end{align}$$

which implies that

$$R_n \leq \frac{I(\mathrm{W};\mathrm{Y}^n)}{n(1-P_{e,n})} + \frac{1}{n(1-P_{e,n})}$$

where the second term goes to 0 when n is large. Next, we bound the mutual information.

$$\begin{align}
I(\mathrm{W};\mathrm{Y}^n) &= \sum_{i=1}^{n} I(\mathrm{W};\mathrm{Y}_i\,|\,\mathrm{Y}^{i-1}) \qquad \text{(chain rule)} \notag\\
&\leq \sum_{i=1}^{n} \left(I(\mathrm{W};\mathrm{Y}_i\,|\,\mathrm{Y}^{i-1}) + I(\mathrm{Y}^{i-1};\mathrm{Y}_i)\right) \qquad \text{(positivity of mutual information)} \notag\\
&= \sum_{i=1}^{n} I(\mathrm{W},\mathrm{Y}^{i-1};\mathrm{Y}_i) \qquad \text{(chain rule)} \notag\\
&\leq \sum_{i=1}^{n} I(\mathrm{W},\mathrm{Y}^{i-1},\mathrm{X}^{i-1};\mathrm{Y}_i) \notag\\
&\leq \sum_{i=1}^{n} I(\mathrm{X}_i;\mathrm{Y}_i) \qquad \left(\text{data processing with Markovity } (\mathrm{W},\mathrm{X}^{i-1},\mathrm{Y}^{i-1}) \to \mathrm{X}_i \to \mathrm{Y}_i\text{, (5.6)}\right) \notag\\
&= \sum_{i=1}^{n} I(\mathrm{P}_{X_i},\mathrm{P}_{Y|X}) \qquad \left(\text{memoryless stationary channel } \mathrm{P}_{Y|X}\right) \notag\\
&\leq nI\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{P}_{X_i},\mathrm{P}_{Y|X}\right) \qquad \text{(concavity of mutual information on } \mathrm{P}_X\text{)} \tag{5.9}\\
&\leq nC_i \qquad \left(C_i := \max_{\mathrm{P}_X} I(\mathrm{P}_X,\mathrm{P}_{Y|X})\right) \notag
\end{align}$$

which implies that

$$R_n \leq \frac{1}{1-P_{e,n}}\left(C_i + \frac{1}{n}\right).$$

From the conditions (5.3) and (5.4), we have

$$R \leq \liminf_{n\to\infty} R_n \leq C_i.$$

5.4 Achievability

In the following, we show the existence of a sequence of $(n, R_n, P_{e,n})$-schemes such that (5.3) and (5.4) are satisfied. In fact, we shall show the achievability even without feedback, which in turn implies that feedback does not increase the capacity of the channel.

The proof is based on random coding. We would like to show that for any input distribution P_X, the rate $R = I(\mathrm{P}_X, \mathrm{P}_{Y|X})$ is achievable for the memoryless statationay channel $\mathrm{P}_{Y|X}$. Let $\mathrm{P}_{XY} := \mathrm{P}_{Y|X}\mathrm{P}_X$.

Random codebook generation Let us generate the codebook $\mathcal{C}_n$ by independently and randomly generating 2^{nR_n} codewords

$$\mathcal{C}_n := \left\{\mathrm{X}^n(1), \dots, \mathrm{X}^n(2^{nR_n})\right\}$$

each one according to the same distribution P_X^n. That is, each codeword contains i.i.d. symbols from the distribution P_X. The codebook is revealed to the encoder and decoder.

Encoding function For the message $w \in \mathcal{W}$, the encoding function returns the w-th codeword $\mathrm{X}^n(w)$.

Decoding function Let us consider the following suboptimal decoder. The decoder returns the first codeword that is jointly typical with the output sequence Y^n, i.e., the first $\hat{w}$ such that $(\mathrm{X}^n(\hat{w}), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})$. If no such codeword exists, the decoder returns $\hat{w} = 1$.

Decoding error analysis Note that

$$\mathbf{1}\left\{\hat{\mathrm{W}}(\mathrm{Y}^n) \neq \mathrm{W}\right\} \leq \mathbf{1}\left\{(\mathrm{X}^n(\mathrm{W}), \mathrm{Y}^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\} + \mathbf{1}\left\{\bigcup_{\hat{w} \neq \mathrm{W}} \left\{(\mathrm{X}^n(\hat{w}), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\}.$$

We analyze the probability of error averaging over all codebooks generated according to the procedure described at the beginning, i.e.,

$$\mathbb{E}_{\mathcal{C}_n, \mathrm{W}, \mathrm{Y}^n}\left[\mathbf{1}\left\{\hat{\mathrm{W}}(\mathrm{Y}^n) \neq \mathrm{W}\right\}\right].$$

By showing that this probability goes to 0, we prove that there exists at least a code $\mathcal{C}_n$ inside the ensemble that has vanishing probability of error. In the following, we can assume that the codewords are random and independent.

In order to show that the transmitted codeword has a high probability to be jointly typical with the received signal, we need to use the following observation.

Let the input sequence X^n be i.i.d.$\sim$ P_X, i.e., $\mathrm{X}^n \sim \mathrm{P}_X^n$, and Y^n the corresponding output of the memoryless stationary channel $\mathrm{P}_{Y|X}$. Then, the joint distribution of the input and output sequences is

$$\mathrm{P}_{X^nY^n} = \mathrm{P}_{XY}^n,$$

where $\mathrm{P}_{XY} := \mathrm{P}_{Y|X}\mathrm{P}_X$. This also implies that

$$\mathrm{P}_{Y^n} = \mathrm{P}_Y^n$$

where $\mathrm{P}_Y := \mathrm{P}_{Y|X} \circ \mathrm{P}_X$.

Indeed, we have $\mathrm{P}_{X^nY^n} = \prod_{i=1}^n \mathrm{P}_{X_iY_i|X^{i-1}Y^{i-1}} = \prod_{i=1}^n \mathrm{P}_{X_i|X^{i-1}Y^{i-1}}\mathrm{P}_{Y_i|X_iX^{i-1}Y^{i-1}} = \prod_{i=1}^n \mathrm{P}_{X_i}\mathrm{P}_{Y_i|X_i} = \mathrm{P}_{XY}^n$. Hence, we have

$$\mathcal{P}\left\{(\mathrm{X}^n(\mathrm{W}), \mathrm{Y}^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\} \to 0,$$

if $n\varepsilon^2 \to \infty$ when $n \to \infty$. Then, since for any $\hat{w} \neq \mathrm{W}$, we have $(\mathrm{X}^n(\hat{w}), \mathrm{Y}^n) \sim \mathrm{P}_X^n\mathrm{P}_Y^n$, i.e., $\mathrm{X}^n(\hat{w})$ and Y^n are independent due to the random codebook generation, we apply the packing lemma (2.48), and have

$$\mathbb{P}\left\{\bigcup_{\hat{w} \neq \mathrm{W}} \left\{(\mathrm{X}^n(\hat{w}), \mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\} \to 0,$$

if

$$\lim_{n\to\infty} \left(nR_n - nI(\mathrm{X};\mathrm{Y}) + 2n\varepsilon H(\mathrm{X})\right) = -\infty. \tag{5.10}$$

Now, let us fix $\varepsilon = \varepsilon_n := n^{-\frac{1}{3}}$, and $R_n = I(\mathrm{X};\mathrm{Y}) - 3\varepsilon_n H(\mathrm{X})$, which verifies (5.10). Hence, we have

$$\mathbb{E}_{\mathcal{C}_n,\mathrm{W},\mathrm{Y}^n}\mathbf{1}\left\{\hat{\mathrm{W}}(\mathrm{Y}^n) \neq \mathrm{W}\right\} \leq \mathcal{P}\left\{(\mathrm{X}^n(\mathrm{W}),\mathrm{Y}^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\} + \mathbb{P}\left\{\bigcup_{\hat{w}\neq \mathrm{W}}\left\{(\mathrm{X}^n(\hat{w}),\mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\}$$
$$\to 0$$

This implies that there exist a sequence of $n, R_n, P_{e,n}$-schemes such that $P_{e,n} \to 0$, and the rate satisfies

$$\liminf_{n\to\infty} R_n = \liminf_{n\to\infty} I(\mathrm{X};\mathrm{Y}) - 3n\varepsilon_n H(\mathrm{X}) = I(\mathrm{X};\mathrm{Y}),$$

which completes the achievability proof of the rate $R = I(\mathrm{X};\mathrm{Y})$.

5.5 Channel capacity with input constraints

Let us impose that the channel input must satisfy some **input constraint** $\frac{1}{n}s_n(x^n) \leq S_n$. In particular, we consider the **separable input constraint function** such that

$$s_n(x^n) := \sum_{i=1}^{n} s(x_i)$$

for some function $s : \mathcal{X} \to \mathbb{R}^+$. For instance, the Hamming weight $s(x) = \mathbf{1}\{x \neq 0\}$ and the quadratic function $s(x) = |x|^2$ whenever it is well defined. In communication, such constraints become natural in different context. A $(n, R_n, P_{e,n}, S_n)$ code $\mathcal{C}_n$ is defined as the set of 2^{nR_n} n-length codewords such that $\frac{1}{n}s_n(x^n) \leq S_n$ for all $x^n \in \mathcal{C}_n$. A rate R with input constraint S is achievable if there exists a sequence of $(n, R_n, P_{e,n}, S_n)$ code with probability of error $P_{e,n}$ such that

$$\liminf_{n\to\infty} R_n \geq R$$
$$\limsup_{n\to\infty} S_n \leq S$$
$$\lim_{n\to\infty} P_{e,n} = 0.$$

Let $C(\mathrm{P}_{Y|X}, S)$ be the maximum rate R with input constraint S in the above sense. We call $C(\mathrm{P}_{Y|X}, S)$ the **(operational) channel capacity** for input constraint S. Let us define the **information channel capacity** with constraint S as

$$C_i(\mathrm{P}_{Y|X}, S) := \max_{\mathrm{P}_X:\mathbb{E}(s(\mathrm{X}))\leq S} I(\mathrm{P}_X, \mathrm{P}_{Y|X}).$$

The function $S \mapsto C_i(\mathrm{P}_{Y|X}, S)$ is increasing, concave, and continuous.

The proof is left as an exercise. Fortunately, Shannon's channel coding theorem extends straightforwardly to the case with input constraints.

For a memoryless stationary channel $\mathrm{P}_{Y|X}$, the channel capacity with input constraint S is

$$C(\mathrm{P}_{Y|X}, S) = C_i(\mathrm{P}_{Y|X}, S).$$

5.5.1 Converse

The converse is almost the same as in the case without input constraint, until the step (5.9), after which we need to impose the input constraint $\frac{1}{n}\sum_{i=1}^{n} s(\mathrm{X}_i) \le S_n$ almost surely.

$$\begin{aligned} I(\mathrm{W};\mathrm{Y}^n) &\le nI\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{P}_{X_i},\mathrm{P}_{Y|X}\right) && \text{(concavity of mutual information on } \mathrm{P}_X\text{)} \\ &\le C_i(\mathrm{P}_{Y|X},S_n) && \left(C_i(\mathrm{P}_{Y|X},S_n) := \max_{\mathrm{P}_X:\mathbb{E}(s(X))\le S_n} I(\mathrm{P}_X,\mathrm{P}_{Y|X}) \text{ is the information capacity}\right) \end{aligned}$$

Note that to obtain the second inequality, we use the fact that $\frac{1}{n}\sum_{i=1}^{n} s(\mathrm{X}_i) \le S_n$ almost surely implies that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(s(\mathrm{X}_i)) \le S_n$, or equivalently,

$$\mathbb{E}_{\frac{1}{n}\sum_i \mathrm{P}_{X_i}} s(\mathrm{X}) = \sum_{x\in\mathcal{X}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{P}_{X_i}(x)\right)s(x) \le S_n.$$

We have

$$R_n \le \frac{1}{1-P_{e,n}}\left(C_i(\mathrm{P}_{Y|X},S_n)+\frac{1}{n}\right)$$

From the conditions (5.3) and (5.4), and the continuity of $S \mapsto C_i(\mathrm{P}_{Y|X},S)$, we have

$$R \le \liminf_{n\to\infty} R_n \le C_i(\mathrm{P}_{Y|X},S).$$

5.5.2 Achievability

The achievability is also similar to the one without input constraint. Let us first generate the same codebook $\mathcal{C}_n$ randomly according to P_X. Take any deterministic sequence $\tilde{x}^n$ such that $\frac{1}{n}s_n(\tilde{x}^n) \le S_n$. Define a set

$$\mathcal{B}_n(S_n) := \{x^n : \quad \frac{1}{n}s_n(x^n) \le S_n\},$$

and a new codebook $\tilde{\mathcal{C}}_n$ with codewords $\tilde{\mathrm{X}}^n(w)$, $w = 1,\ldots,2^{nR_n}$, such that

$$\tilde{\mathrm{X}}^n(w) = \begin{cases} \mathrm{X}^n(w), & \text{if } \mathrm{X}^n(w) \in \mathcal{B}_n(S_n) \\ \tilde{x}^n, & \text{otherwise} \end{cases}$$

The encoder uses the codebook $\tilde{\mathcal{C}}_n$ to encode. But the decoder uses the codebook $\mathcal{C}_n$ to decode as in the case without input constraint.

To analyze the probability of error, we write

$$\mathbf{1}\{\hat{\mathrm{W}}(\mathrm{Y}^n) \ne \mathrm{W}\} \le \mathbf{1}\{\mathrm{X}^n(\mathrm{W}) \notin \mathcal{B}_n(S_n)\} \tag{5.11}$$

$$+\mathbf{1}\{\mathrm{X}^n(\mathrm{W}) \in \mathcal{B}_n(S_n)\}\mathbf{1}\left\{(\mathrm{X}^n(\mathrm{W}),\mathrm{Y}^n) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\} + \mathbf{1}\left\{\bigcup_{\hat{w}\ne\mathrm{W}}\left\{(\mathrm{X}^n(\hat{w}),\mathrm{Y}^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY})\right\}\right\} \tag{5.12}$$

Since $(\mathrm{X}^n(\mathrm{W}),\mathrm{Y}^n) \sim \mathrm{P}_X^n\mathrm{P}_{Y|X}^n = \mathrm{P}_{XY}^n$, when $\mathrm{X}^n(W) \in \mathcal{B}_n(S_n)$, the last two events have the same bound as the case without input constraint. We just need to control the probability of the first event. Let

$$S_n = (1-\varepsilon)\mathbb{E}\left(s(\mathrm{X})\right).$$

Then, we have

$$\mathcal{P}\left\{\mathrm{X}^n(\mathrm{W}) \notin \mathcal{B}_n(S_n)\right\} \le \mathcal{P}\left\{\mathrm{X}^n(\mathrm{W}) \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_X)\right\},$$

since any typical sequence should satisfy the input constraint due to the typical average lemma (2.8). This probability vanishes with n if $n\varepsilon^2 \to \infty$. Now, let $\varepsilon = \varepsilon_n := n^{-\frac{1}{3}}$. We have $S_n \to \mathbb{E}\left(s(\mathrm{X})\right)$, $R_n \to I(\mathrm{X};\mathrm{Y})$, and $P_{e,n} \to 0$. This shows the existence of a sequence of $(n,R_n,P_{e,n},S_n)$ codes that achieves rate $I(\mathrm{X};\mathrm{Y})$ with input constraint $\mathbb{E}\left(s(\mathrm{X})\right)$.

5.6 Examples of memoryless stationary channels

5.6.1 Capacity of the binary symmetric channel (BSC)

Let $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and let $\oplus$ be the binary sum. Then, the **binary symmetric channel (BSC)** with parameter $p \in [0,1]$ can be represented by

$$\mathrm{Y} = \mathrm{X} \oplus \mathrm{Z},$$

where $\mathrm{Z} \sim \mathrm{Bern}(p)$ is independent of X. One can also write the conditional distribution explicitly as

$$\mathrm{P}_{Y|X=x}(y) = (1-p)\mathbf{1}\{y = x\} + p\mathbf{1}\{y \neq x\}, \quad x, y \in \{0,1\}$$

The capacity of the BSC(p) is

$$C(p) = 1 - H_2(p).$$

Proof. The capacity is $C(p) := \max_{P_X} I(\mathrm{X};\mathrm{Y})$ where P_X is a Bernoulli distribution, say, Bern(λ). The objective function is the mutual information that can be written as

$$\begin{aligned} I(\mathrm{X};\mathrm{Y}) &= H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X}) \\ &= H(\mathrm{Y}) - H(\mathrm{Y} \oplus \mathrm{X}|\mathrm{X}) \\ &= H(\mathrm{Y}) - H(\mathrm{Z}|\mathrm{X}) \\ &= H(\mathrm{Y}) - H(\mathrm{Z}) \\ &= H_2(p * \lambda) - H_2(p) \end{aligned}$$

where $\mathrm{Y} \sim \mathrm{Bern}(p * \lambda)$ with $p * \lambda := p(1-\lambda) + (1-p) * \lambda$. The maximizer of the mutual information is the λ that maximizes $H_2(p * \lambda)$. We see that $\lambda^* = \frac{1}{2}$ such that $p * \lambda = \frac{1}{2}$ and $H_2(p * \lambda) = 1$. □

If we impose the input constraint that the average Hamming weight cannot be larger than S, then we can verify the following.

The capacity of the BSC(p) with input Hamming constraint S is

$$C(p, S) = H_2\left(\min\left\{S * \tilde{p}, \frac{1}{2}\right\}\right) - H_2(p),$$

where $\tilde{p} := \min\{p, 1-p\}$.

The proof is left as an exercise.

5.6.2 Capacity of the binary erasure channel (BEC)

Let $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,?\}$ and let $\oplus$ be the binary sum. Then, the **binary erasure channel (BEC)** with parameter $\delta \in [0,1]$ can be represented by the conditional distribution explicitly as

$$\mathrm{P}_{Y|X=x}(y) = (1-\delta)\mathbf{1}\{y = x\} + \delta\mathbf{1}\{y =?\}, \quad x \in \{0,1\}$$

The capacity of the BSC(p) is

$$C(p) = 1 - \delta.$$

Proof. The capacity is $C(p) := \max_{P_X} I(X;Y)$ where P_X is a Bernoulli distribution, say, Bern(λ). The objective function is the mutual information that can be written as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \delta H(X|Y = ?) \qquad \text{(When Y = 0 or 1, X is deterministic)} \\ &= (1-\delta)H(X) \qquad (P_{X|Y=?} = P_X) \end{aligned}$$

The maximizer of the mutual information is the λ that maximizes $H_2(\lambda)$. We see that $\lambda^* = \frac{1}{2}$. □

5.6.3 Capacity of the Gaussian channel

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Then, the **Gaussian channel** with parameter $\sigma^2 \in \mathbb{R}^+$ can be represented by

$$Y = X + Z,$$

where $Z \sim \mathcal{N}(0, \sigma^2)$ is independent of X. One can also write the conditional distribution explicitly with the density function

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}}, \quad x, y \in \mathbb{R}.$$

Consider the input power constraint $\frac{1}{n}\sum_{i=1}^n x_i^2 \leq S$ for all input sequence x^n.

The capacity of the Gaussian channel with parameter σ^2 and input constraint S is

$$C(\sigma^2, S) = \frac{1}{2}\log\left(1 + \frac{S}{\sigma^2}\right).$$

Proof. The capacity is $C(\sigma^2, S) := \max_{P_X : EX^2 \leq S} I(X;Y)$ where the objective function is the mutual information that can be written as

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) && (5.13) \\ &= h(Y) - h(Y - X|X) && (5.14) \\ &= h(Y) - h(Z|X) && (5.15) \\ &= h(Y) - h(Z) && (5.16) \\ &\leq \frac{1}{2}\log(2\pi e \mathsf{Var}(Y)) - h(Z) && (5.17) \\ &= \frac{1}{2}\log\frac{\mathsf{Var}(Y))}{\mathsf{Var}(Z)} && (5.18) \\ &\leq \log\left(1 + \frac{S}{\sigma^2}\right) \qquad \left(\mathsf{Var}(Y) \leq \mathsf{Var}(X) + \mathsf{Var}(Z) \leq \mathbb{E}X^2 + \sigma^2 \leq S + \sigma^2\right) && (5.19) \end{aligned}$$

Note that the first inequality can be achieved if Y is Gaussian, which would imply that X is Gaussian; the second inequality can be achieved when $\mathrm{Var}(Y) = S + \sigma^2$. Indeed, if we choose $X \sim \mathcal{N}(0, S)$, which satisfies the input constraint with equality. □

Exercises[13]

1. Additive noise channel [CT 7.2]. Find the channel capacity of the following discrete memoryless channel:

$$\mathrm{Y} = \mathrm{X} + \mathrm{Z}$$

where $\mathrm{P}(\mathrm{Z} = 0) = \mathrm{P}(\mathrm{Z} = a) = 1/2$ and $\mathrm{X} \in \{0, 1\}$. Assume that Z is independent of X. Observe that the channel capacity depends on the value of a.

2. Channel capacity [CT 7.4]. Consider the discrete memoryless channel $\mathrm{Y} = X + Z(\mathrm{mod}\ 11)$, where Z is uniformly distributed in $\{1, 2, 3\}$ and $\mathrm{X} \in \{0, 1, 2, \ldots, 10\}$. Assume that Z is independent of X.

   - Find the capacity.
   - What is the optimal input distribution P_X^*?

3. Cascade of binary symmetric channels [CT 7.7]. Show that a cascade of n identical independent binary symmetric channels,

$$\mathrm{X}_0 \to \boxed{\mathrm{BSC}} \to \mathrm{X}_1 \to \boxed{\mathrm{BSC}} \to \mathrm{X}_2 \to \cdots \to \mathrm{X}_{n-1} \to \boxed{\mathrm{BSC}} \to \mathrm{X}_n$$

each with raw error probability p, is equivalent to a single BSC with error probability $\frac{1}{2}(1 - (1 - 2p)^n)$ and hence that $\lim_{n\to\infty} I(\mathrm{X}_0; \mathrm{X}_n) = 0$ if $p \neq 0, 1$. No encoding or decoding takes place at the intermediate terminals $\mathrm{X}_1, \ldots, \mathrm{X}_{n-1}$. Thus, the capacity of the cascade tends to zero.

4. Z-channel [CT 7.8]. The Z-channel has binary input and output alphabets and transition probabilities are $\mathrm{P}(\mathrm{Y} = 0 \mid \mathrm{X} = 0) = 1$ and $\mathrm{P}(\mathrm{Y} = 0 \mid \mathrm{X} = 1) = \mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{X} = 1) = \frac{1}{2}$. Find the capacity of the Z-channel and the optimal input distribution P_X^*.

5. Time-varying channels [CT 7.11]. Consider a time-varying discrete memoryless channel. Let $\mathrm{Y}_1, \mathrm{Y}_2, \cdots, \mathrm{Y}_n$ be conditionally independent given $\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n$, i.e., $\mathrm{P}_{Y^n|X^n} = \prod_{i=1}^n \mathrm{P}_{Y_i|X_i}$ where each $\mathrm{P}_{Y_i|X_i}$ is a BSC with parameter p_i. Find $\max_{\mathrm{P}_{X^n}} I(\mathrm{X}^n; \mathrm{Y}^n)$.

6. Unused symbols [CT 7.12]. Show that the capacity of the channel with prob- ability transition matrix

$$\mathrm{P}_{Y|X} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

is achieved by a distribution that places zero probability on one of input symbols. What is the capacity of this channel? Give an intuitive reason why that letter is not used.

7. Encoder and decoder as part of the channel [CT 7.16] Consider a binary symmetric channel with crossover probability 0.1. A possible coding scheme for this channel with two codewords of length 3 is to encode message a_1 as 000 and a_2 as 111. With this coding scheme, we can consider the combination of encoder, channel, and decoder as forming a new BSC, with two inputs a_1 and a_2 and two outputs a_1 and a_2.

   - What is the optimal decoder that minimizes the probability of error?
   - Calculate the crossover probability of the new channel.
   - What is the capacity of the new channel in bits per transmission of the original channel?
   - What is the capacity of the original BSC with crossover probability 0.1?
   - Prove a general result that for any channel, considering the encoder, channel, and decoder together as a new channel from messages to estimated messages will not increase the capacity in bits per transmission of the original channel.

8. Channel with two independent looks at Y [CT 7.20]. Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X according to $\mathrm{P}_{Y|X}$.

   - Show that $I(\mathrm{X}; \mathrm{Y}_1, \mathrm{Y}_2) = 2I(\mathrm{X}; \mathrm{Y}_1) - I(\mathrm{Y}_1, \mathrm{Y}_2)$.

[13] The citations “CT”, “CK” refer to Cover-Thomas, Csiszár-Körner, respectively.

- Conclude that the capacity of the channel $P_{Y_1 Y_2|X}$ is less than twice the capacity of $P_{Y|X}$

9. Noise alphabets [CT 7.24]. Consider the channel $Y = X + Z$ where $X \in \mathcal{X} = \{0, 1, 2, 3\}$, Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

   - What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.
   - What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

10. Binary multiplier channel [CT 7.23].

   - Consider the channel $Y = XZ$, where X and Z are independent binary random variables that take on values 0 and 1. Z is $\text{Bern}(a)$, i.e., $P(Z = 1) = a$. Find the capacity of this channel and the maximizing distribution on X.
   - Now suppose that the receiver can observe Z as well as Y. What is the capacity?

Quiz (unique correct answer)

1. The **Channel Capacity** of a discrete memoryless stationary channel is defined as:

   A) The maximum mutual information over all possible input distributions.

   B) The minimum mutual information over all possible input distributions.

   C) The mutual information for the uniform input distribution.

   D) The maximum entropy of the channel output.

   E) The minimum conditional entropy of the output given the input.

2. **Shannon's Channel Coding Theorem** states that for a discrete memoryless stationary channel with capacity C:

   A) Reliable communication is possible at any rate $R > C$.

   B) Reliable communication is impossible at any rate $R < C$.

   C) Reliable communication is possible at any rate $R < C$.

   D) The error probability cannot be made arbitrarily small for any rate R.

   E) The capacity C is always equal to the entropy $H(\mathrm{X})$.

3. The **mutual information** $I(\mathrm{X};\mathrm{Y})$ between input X and output Y of a discrete channel is:

   A) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}) + H(\mathrm{Y})$

   B) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}|\mathrm{Y})$

   C) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X})$

   D) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X},\mathrm{Y})$

   E) $I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{Y})$

4. The **capacity** C of a **Binary Symmetric Channel (BSC)** with crossover probability p is:

   A) $C = 1 - H_2(p)$, where $H_2(p)$ is the binary entropy function.

   B) $C = H_2(p)$

   C) $C = \log_2(p)$

   D) $C = 1 + H_2(p)$

   E) $C = 2H_2(p)$

5. In the **random coding** argument for proving the achievability of capacity, codewords are selected:

   A) Deterministically based on the channel characteristics.

   B) Randomly according to a uniform distribution over the codebook.

   C) Randomly according to the optimal input distribution that maximizes mutual information.

   D) By choosing codewords that are orthogonal.

   E) By selecting the most probable sequences.

6. The **jointly typical decoding** method involves decoding to the codeword that is:

   A) Closest to the received sequence in Hamming distance.

   B) Jointly typical with the received sequence.

   C) The most frequently occurring codeword.

   D) Matched to the received sequence via maximum likelihood.

   E) The first codeword in the codebook.

7. The **Binary Erasure Channel (BEC)** with erasure probability δ has capacity:

   A) $C = 1 - \delta$
   B) $C = \delta$
   C) $C = H_2(\delta)$
   D) $C = \log_2(1 - \delta)$
   E) $C = 1 + \delta$

8. The **channel coding theorem** is applicable to:

   A) Only discrete memoryless channels.
   B) Only continuous channels.
   C) Both discrete memoryless channels and continuous channels.
   D) Only channels without noise.
   E) Channels with feedback only.

9. The **capacity** of a **memoryless stationary channel** is determined solely by:

   A) The transition probabilities $\mathrm{P}_{Y|X}$.
   B) The input distribution P_X.
   C) The output distribution P_Y.
   D) The joint distribution P_{XY}.
   E) The conditional distribution $\mathrm{P}_{X|Y}$.

10. In the random coding proof of the channel coding theorem, the **union bound** is used to:

    A) Calculate the exact error probability.
    B) Upper bound the probability of decoding error.
    C) Lower bound the mutual information.
    D) Estimate the capacity of the channel.
    E) Optimize the codebook.

11. The **Channel Coding Theorem** implies that for rates $R < C$:

    A) The error probability can be made arbitrarily small by increasing the block length.
    B) The error probability remains constant regardless of block length.
    C) The capacity can be increased by decreasing the rate.
    D) Reliable communication is impossible.
    E) The entropy of the source decreases.

12. In a **Gaussian Channel**, the optimal input distribution to achieve capacity is:

    A) Uniform distribution over the input alphabet.
    B) Gaussian distribution with zero mean and variance equal to the power constraint.
    C) Exponential distribution.
    D) Any arbitrary distribution, since capacity is independent of input distribution.
    E) A discrete distribution with equiprobable mass points.

13. The **Additive White Gaussian Noise (AWGN)** channel is characterized by:

    A) Noise that is dependent on the input signal.

B) Noise that is Gaussian and uncorrelated with the input.

C) Noise that is uniformly distributed.

D) No noise at all.

E) Noise that has memory (dependent on previous inputs).

14. For the **Binary Symmetric Channel (BSC)** with crossover probability $p = 0.5$, the capacity C is:

A) $C = 1$

B) $C = 0.5$

C) $C = 0$

D) $C = H_2(0.5)$

E) $C = -1$

15. The **channel capacity with feedback** of a memoryless channel compared to its capacity without feedback is:

A) Always greater.

B) Always less.

C) Sometimes greater, sometimes equal.

D) Always equal.

E) Undefined, since feedback cannot be applied to memoryless channels.

16. In a **Binary Erasure Channel (BEC)**, when the erasure probability $\delta = 0$, the capacity is:

A) $C = 1$

B) $C = 0$

C) $C = \delta$

D) $C = 1 - \delta$

E) Undefined

17. The **Channel Coding Theorem** relies on which of the following properties for its proof?

A) Deterministic codebook construction.

B) Random selection of codebooks and averaging over ensembles.

C) The law of large numbers for small block lengths.

D) The ability to correct all errors.

E) The uniqueness of codebooks.

Elements of Information Theory **2025-2026**

Lecture 6: Information Theory and Statistics

Lecturer:

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Large deviation theory and Sanov's theorem

We have seen that the probability to have a typical sequence is high. That is, there is a high probability that the generated sequence has a behavior close to the average one with respect to the distribution. The probability that a sequence is inside a type class Q is

$$\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \doteq 2^{-nD(\mathrm{Q}\|\mathrm{P})}$$

that is, exponentially small when n is large. In this section, we are interested in non-typical sequences and characterize the probability of such sequences when n is large. Let $\mathcal{E}$ be some given set of pmf's. We would like to approximate the probability $\sum_{\mathrm{Q}\in\mathcal{E}\cap\mathcal{P}_n}\mathrm{P}^n(\mathcal{T}(\mathrm{Q}))$, that is, the probability of a sequence has a type that satisfies the constraint in $\mathcal{E}$. With some abuse of notation, we write

$$\mathrm{P}^n(\mathcal{E}) := \mathrm{P}^n(\hat{\mathrm{P}}_{\mathrm{X}^n}\in\mathcal{E}) = \sum_{\mathrm{Q}\in\mathcal{E}\cap\mathcal{P}_n}\mathrm{P}^n(\mathcal{T}(\mathrm{Q})).$$

Sanov's theorem. Let $\mathcal{E}$ be a set of pmf, then

$$\mathrm{P}^n(\mathcal{E}) \le K_{n,|\mathcal{X}|}2^{-nD(\mathrm{Q}^*\|\mathrm{P})} \tag{6.1}$$

where $K_{n,|\mathcal{X}|} := |\mathcal{P}_n^{\mathcal{X}}|$ is the number of types, and

$$\mathrm{Q}^* := \arg\inf_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P}).$$

If $\mathcal{E}$ is closed, then

$$\mathrm{P}^n(\mathcal{E}) \doteq 2^{-n\min_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P})}, \tag{6.2}$$

i.e.,

$$\lim_{n\to\infty} -\frac{\log \mathrm{P}^n(\mathcal{E})}{n} = \min_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P}).$$

Proof. The upper bound (6.1) can be obtained as follows.

$$\begin{aligned}
\mathrm{P}^n(\mathcal{E}) &= \sum_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n}\mathrm{P}^n(\mathcal{T}(\hat{\mathrm{P}}))\\
&\le \sum_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} 2^{-nD(\hat{\mathrm{P}}\|\mathrm{P})}\\
&\le \sum_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} 2^{-n\inf_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P})}\\
&\le K_{n,|\mathcal{X}|}2^{-n\inf_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P})}
\end{aligned}$$

If $\mathcal{E}$ is closed, then the infimum is attainable and can be replaced by the minimum. From the above, we have

$$\liminf_{n\to\infty} -\frac{\log \mathrm{P}^n(\mathcal{E})}{n} \geq \min_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P}) = D(\mathrm{Q}^\star\|P). \tag{6.3}$$

where $\mathrm{Q}^\star$ be such that $D(\mathrm{Q}^\star\|\mathrm{P}) = \min_{\mathrm{Q}\in\mathcal{E}} D(\mathrm{Q}\|\mathrm{P})$. To prove (6.2), we need a lower bound on the probability. Indeed,

$$\begin{aligned}\mathrm{P}^n(\mathcal{E}) &= \sum_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} \mathrm{P}^n(\mathcal{T}(\hat{\mathrm{P}})) \\ &\geq \max_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} \mathrm{P}^n(\mathcal{T}(\hat{\mathrm{P}})) \\ &\geq K_{n,|\mathcal{X}|}^{-1} \max_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} 2^{-nD(\hat{\mathrm{P}}\|\mathrm{P})} \\ &= K_{n,|\mathcal{X}|}^{-1} 2^{-n\min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P})} \end{aligned} \tag{6.4}$$

Since $\mathcal{P}_n$ is dense in the simplex $\mathcal{P}$, there exist a sequence $\hat{\mathrm{P}}_n \in \mathcal{E}\cap\mathcal{P}_n$ with $\hat{\mathrm{P}}_n \to \mathrm{Q}^\star$. If $D(\mathrm{Q}^\star\|\mathrm{P}) = +\infty$, we have $\min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) \geq \min_{\hat{\mathrm{P}}\in\mathcal{E}} D(\hat{\mathrm{P}}\|\mathrm{P}) = +\infty$. Otherwise, $\mathrm{Q} \mapsto D(\mathrm{Q}\|\mathrm{P})$ is continuous at Q^*, since the divergence is the sum of a finite number of continuous functions. This implies that for any $\varepsilon > 0$, when n is large enough, $\hat{\mathrm{P}}_n$ is close enough to $\mathrm{Q}^\star$ such that $|D(\hat{\mathrm{P}}_n\|\mathrm{P}) - D(\mathrm{Q}^\star\|\mathrm{P})| \leq \varepsilon$. Therefore,

$$\begin{aligned}\min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) &\leq D(\hat{\mathrm{P}}_n\|\mathrm{P}) \\ &\leq D(\mathrm{Q}^\star\|\mathrm{P}) + \varepsilon \qquad \text{when } n \text{ is large enough.}\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} \min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) \leq D(\mathrm{Q}^*\|P) + \varepsilon$$

for any $\varepsilon > 0$. Taking the infimum over ε and we have $\limsup_{n\to\infty} \min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) \leq D(\mathrm{Q}^\star\|P)$. This implies that

$$\lim_{n\to\infty} \min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) = D(\mathrm{Q}^*\|P).$$

From the lower bound (6.4), we have

$$\limsup_{n\to\infty} -\frac{\log \mathrm{P}^n(\mathcal{E}\cap\mathcal{P}_n)}{n} \leq \limsup_{n\to\infty} \min_{\hat{\mathrm{P}}\in\mathcal{E}\cap\mathcal{P}_n} D(\hat{\mathrm{P}}\|\mathrm{P}) = D(\mathrm{Q}^*\|P). \tag{6.5}$$

From (6.3) and (6.5), we have

$$\lim_{n\to\infty} -\frac{\log \mathrm{P}^n(\mathcal{E}\cap\mathcal{P}_n)}{n} = D(\mathrm{Q}^*\|P).$$

□

In particular, if the set $\mathcal{E}$ is in the following form

$$\mathcal{E} := \bigcap_{j=1,\ldots,k} \{\mathrm{Q} : \mathbb{E}_{X\sim\mathrm{Q}}[g_j(\mathrm{X})] \geq T_j\}$$

Then, the minimizing distibution is

$$\mathrm{Q}^*(x) := \frac{\mathrm{P}(x)2^{\sum_j \lambda_j g_j(x)}}{\sum_{a\in\mathcal{X}} \mathrm{P}(a)2^{\sum_j \lambda_j g_j(a)}} \tag{6.6}$$

where $\lambda_j \geq 0$, $j = 1, \ldots, k$, are such that the constraints are satisfied.[a]

[a] λ_j is the Lagrangian multiplier corresponding to the j's constraint. According to the KKT conditions, $\lambda_j = 0$ if the constraint is satisfied with strict inequality.

Example

What is the probability that the average outcome of n dice tosses is larger than 4?

From (6.6), we obtain, by setting $k = 1$ and $g(a) = a$ for $a = 1, \ldots, 6$,

$$\mathrm{Q}^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda a}}$$

where λ is such that $\mathbb{E}_{\mathrm{Q}^*}(\mathrm{X}) = 4$. We can find numerically that $\lambda \approx 0.2519$ and $D(\mathrm{Q}^* \| \mathrm{P}) \approx 0.0624$ bits. Hence, the probability that the outcome of 10000 dice tosses is larger than 4 is roughly $2^{-10000 \times 0.0624} \approx 2^{-624}$, up to a factor of $K_{n,M} = \binom{10005}{5} \approx 2^{10005\, H_2(5/10005)} \approx 2^{60}$.

Lemma 6.1 (Three precision levels for $\log_2 \binom{n}{k}$)**.** *Let $1 \le k \le n-1$ and $p := k/n \in (0,1)$. Then:*

$$\log_2 \binom{n}{k} \approx \underbrace{n\, H_2(p)}_{\textit{entropy term}} - \underbrace{\tfrac{1}{2} \log_2\big(2\pi n p(1-p)\big)}_{\textit{Stirling}} + \underbrace{\frac{1}{12 \ln 2}\left(\frac{1}{n} - \frac{1}{k} - \frac{1}{n-k}\right)}_{\textit{refinement}}.$$

Here $H_2(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy. The accuracy of each term is $\Theta(\log n)$, $O(1/n)$, and $O(1/n^3)$, respectively.

6.2 Binary hypothesis testing

One of the important problems in statistics is hypothesis testing. We consider the simplest setting with only two hypothesis. $\mathrm{H} = 0$ and $\mathrm{H} = 1$ where H need not be random. Let $\mathrm{X}_1, \ldots, \mathrm{X}_n \sim \mathrm{Q}$ be independent.

- $\mathrm{H} = 0$: $\mathrm{Q} = \mathrm{P}_0$
- $\mathrm{H} = 1$: $\mathrm{Q} = \mathrm{P}_1$

Let Z be the guess of the hypothesis defined by $\mathrm{P}_{Z|X^n}$, i.e., a randomized guess. Note that this includes the deterministic guess as a special case. For simplicity, we consider the deterministic guess with $\mathrm{Z} = g(\mathrm{X}^n)$. We can define two types of error:

- $\alpha := \mathrm{P}(\mathrm{Z} = 1 \,|\, \mathrm{H} = 0)$ which is sometimes called **false alarm** or **type I error**
- $\beta := \mathrm{P}(\mathrm{Z} = 0 \,|\, \mathrm{H} = 1)$ which is sometimes called **missed detection** or **type II error**

Some applications are highly asymmetric in the sensibility to both types of error. For instance, in most cases, missed detection is more dangerous than false alarm and should be given more attention.

For deterministic test, one can always define test regions, say, $\mathcal{D}_0$ and $\mathcal{D}_1$ such that $\mathcal{D}_0 \bigcup \mathcal{D}_1 = \mathcal{X}^n$, and

$$g(x^n) = i, \quad \text{if } x^n \in \mathcal{D}_i$$

Therefore,

$$\alpha = \mathrm{P}_0^n(\mathcal{D}_1), \quad \beta = \mathrm{P}_1^n(\mathcal{D}_0).$$

It is without loss of generality to let $\mathcal{D}_0 = \mathcal{D}$ and $\mathcal{D}_1 = \bar{\mathcal{D}}_1$.

It is clear that for each region $\mathcal{D} \subseteq \mathcal{X}^n$, we have a pair (α, β), or, explicitly, $(\alpha(\mathcal{D}), \beta(\mathcal{D}))$. Each region can be regarded as a test strategy. Let

$$\mathcal{R}_{\det} := \{(\alpha(\mathcal{D}), \beta(\mathcal{D})) : \quad \mathcal{D} \subseteq \mathcal{X}^n\}$$

be the set of achievable pairs (α, β) with deterministic tests. The convex hull of $\mathcal{R}_{\text{det}}$ is the set of achievable pairs with randomized tests.

In the following, we focus on the case with $\mathrm{P}_0 \ll \mathrm{P}_1$ and $\mathrm{P}_1 \ll \mathrm{P}_0$. Indeed, if $\mathrm{P}_0 \not\ll \mathrm{P}_1$, then for some $x \in \mathcal{X}$, $\mathrm{P}_1(x) = 0$ and $\mathrm{P}_0(x) \neq 0$. Clearly, all sequences containing at least one x should be in $\mathcal{D}$ since the probability of such sequences under P_1 would be 0. It remains to partition the remaining sequences, with alphabet $\mathcal{X} \setminus \{x\}$. Repeating the process leads to a problem with a smaller alphabet in which $\mathrm{P}_0 \ll \mathrm{P}_1$ and $\mathrm{P}_1 \ll \mathrm{P}_0$.

6.2.1 Likelihood ratio test (LRT), Neyman-Pearson lemma

Instead of looking at all the $2^{|\mathcal{X}|^n}$ strategies, we consider a subset of strategies call **likelihood ratio test** (LRT). Each strategy is parameterized by a nonnegative value τ, with the corresponding test region:

$$\mathcal{D}_\tau := \left\{ x^n : \frac{\mathrm{P}_0^n(x^n)}{\mathrm{P}_1^n(x^n)} > \tau \right\}$$

It follows that the number of different *LRT*'s is upper bounded by $|\mathcal{X}|^n + 1$, much lower than the number of all strategies $2^{|\mathcal{X}|^n}$. It turns out that it is without loss of optimality to restrict ourselves to LRT's.

Neyman-Pearson Lemma. Let (α^*, β^*) be achievable by the LRT with some τ, i.e., $\alpha^* := \mathrm{P}_0^n(\bar{\mathcal{D}}_\tau)$ and $\beta^* := \mathrm{P}_1^n(\mathcal{D}_\tau)$ for some $\tau \geq 0$. Let $\mathcal{D}' \subseteq \mathcal{X}^n$ be any region with $\alpha := \mathrm{P}_0^n(\bar{\mathcal{D}}')$ and $\beta := \mathrm{P}_1^n(\mathcal{D}')$. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Proof. When $\alpha^* = 0$, we have $\mathrm{P}_0^n(\mathcal{D}_\tau) = 1$, implying that $\mathsf{Supp}(\mathrm{P}_0^n) \subseteq \mathcal{D}_\tau$. But we also know that $\mathcal{D}_\tau \subseteq \mathsf{Supp}(\mathrm{P}_0^n)$ for any $\tau \geq 0$. Therefore, in the case with $\alpha^* = 0$, we must have $\mathcal{D}_\tau = \mathsf{Supp}(\mathrm{P}_0^n)$. For any $\alpha \leq \alpha^*$ achieved with strategy $\mathcal{D}'$, we have $\alpha = 0$, implying that $\mathsf{Supp}(\mathrm{P}_0^n) \subseteq \mathcal{D}'$. Thus, we have $\mathcal{D}_\tau \subseteq \mathcal{D}'$. In turn, we have $\beta = \mathrm{P}_1^n(\mathcal{D}') \geq \mathrm{P}_1^n(\mathcal{D}_\tau) = \beta^*$.

Next, we consider $\alpha^* > 0$ with corresponding strategy $\mathcal{D}_\tau$. Then, we must have $\tau > 0$, for otherwise $\alpha^* = 0$. Notice that

$$\begin{aligned}
\beta - \beta^* &= \mathrm{P}_1^n(\mathcal{D}') - \mathrm{P}_1^n(\mathcal{D}_\tau) \\
&= \mathrm{P}_1^n(\mathcal{D}' \cap \bar{\mathcal{D}}_\tau) - \mathrm{P}_1^n(\mathcal{D}_\tau \cap \bar{\mathcal{D}}') \\
&\geq \tau^{-1}\mathrm{P}_0^n(\mathcal{D}' \cap \bar{\mathcal{D}}_\tau) - \tau^{-1}\mathrm{P}_0^n(\mathcal{D}_\tau \cap \bar{\mathcal{D}}') \\
&= \tau^{-1}\mathrm{P}_0^n(\bar{\mathcal{D}}_\tau) - \tau^{-1}\mathrm{P}_0^n(\bar{\mathcal{D}}') \\
&= \tau^{-1}(\alpha^* - \alpha).
\end{aligned}$$

From $\tau > 0$, it is clear that $\alpha^* \geq \alpha$ implies $\beta \geq \beta^*$. □

6.2.2 Asymptotic regimes

We are interested in the error behavior in the asymptotic regimes when n is large. We consider the Neyman-Pearson formulation and look at two asymptotic regimes:

- Stein's regime: α_n remains bounded by some $\varepsilon \in (0, 1)$
- Chernoff's regime: α_n goes to 0 exponentially, i.e., $\alpha_n \doteq 2^{-n\eta_\alpha}$ for some $\eta_\alpha > 0$.

We are then interested in the asymptotic behavior of β_n.

6.2.2.1 Stein's regime

Stein's regime. For any $\varepsilon \in (0,1)$, we have

$$\lim_{n\to\infty} -\frac{\log \min_{\mathcal{D}:\alpha_n(\mathcal{D})\le\varepsilon} \beta_n(\mathcal{D})}{n} = D(\mathrm{P}_0\|\mathrm{P}_1)$$

Proof. Let us define a ε-typical set:

$$\mathcal{A}_\varepsilon := \left\{x^n : \left|\frac{1}{n}\log\frac{p_0^n(x^n)}{p_1^n(x^n)} - D(\mathrm{P}_0\|\mathrm{P}_1)\right| \le \varepsilon\right\}$$

Note that for $\mathrm{X}^n \sim \mathrm{P}_0^n$, $\frac{1}{n}\log\frac{p_0^n(\mathrm{X}^n)}{p_1^n(\mathrm{X}^n)} \to D(\mathrm{P}_0\|\mathrm{P}_1)$ and the probability $\mathrm{P}_0^n(\bar{\mathcal{A}}_\varepsilon) \to 0$. Therefore, when n is large enough, we have $\mathrm{P}_0^n(\mathcal{A}_\varepsilon) \ge 1-\varepsilon$. Let $\mathcal{D} = \mathcal{A}_\varepsilon$, then

$$\begin{aligned}
\beta_n(\mathcal{D}) &:= \mathrm{P}_1^n(\mathcal{D})\\
&= \mathrm{P}_1^n(\mathcal{A}_\varepsilon)\\
&= \sum_{x^n\in\mathcal{A}_\varepsilon} \mathrm{P}_1^n(x^n)\\
&\le \sum_{x^n\in\mathcal{A}_\varepsilon} \mathrm{P}_0^n(x^n)2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)+n\varepsilon}\\
&\le 2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)+n\varepsilon}.
\end{aligned}$$

We have

$$\liminf_{n\to\infty} -\frac{\log \min_{\mathcal{D}:\alpha_n(\mathcal{D})\le\varepsilon} \beta_n(\mathcal{D})}{n} \ge \liminf_{n\to\infty} -\frac{\log \beta_n(\mathcal{A}_\varepsilon)}{n} \ge D(\mathrm{P}_0\|\mathrm{P}_1) - \varepsilon. \tag{6.7}$$

To find an upper bound, let $\mathcal{D}$ be such that $\alpha_n(\mathcal{D}) \le \varepsilon$, i.e., $\mathrm{P}_0^n(\mathcal{D}) \ge 1-\varepsilon$. Then, we have

$$\begin{aligned}
\beta_n(\mathcal{D}) &:= \mathrm{P}_1^n(\mathcal{D})\\
&\ge \mathrm{P}_1^n(\mathcal{D}\cap\mathcal{A}_\varepsilon)\\
&= \sum_{x^n\in\mathcal{D}\cap\mathcal{A}_\varepsilon} \mathrm{P}_1^n(x^n)\\
&\ge \sum_{x^n\in\mathcal{D}\cap\mathcal{A}_\varepsilon} \mathrm{P}_0^n(x^n)2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)-n\varepsilon}\\
&= \mathrm{P}_0^n(\mathcal{D}\cap\mathcal{A}_\varepsilon)2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)-n\varepsilon}\\
&\ge (\mathrm{P}_0^n(\mathcal{D}) - \mathrm{P}_0^n(\bar{\mathcal{A}}_\varepsilon))2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)-n\varepsilon}\\
&\ge (1-2\varepsilon)2^{-nD(\mathrm{P}_0\|\mathrm{P}_1)-n\varepsilon}
\end{aligned}$$

for n sufficiently large. Thus, we have

$$\begin{aligned}
\limsup_{n\to\infty} -\frac{\log \min_{\mathcal{D}:\alpha_n(\mathcal{D})\le\varepsilon} \beta_n(\mathcal{D})}{n} &\le D(\mathrm{P}_0\|\mathrm{P}_1) + \varepsilon + \limsup_{n\to\infty} -\frac{\log(1-2\varepsilon)}{n}\\
&= D(\mathrm{P}_0\|\mathrm{P}_1) + \varepsilon
\end{aligned} \tag{6.8}$$

Since (6.7) and (6.8) are both valid for any $\varepsilon > 0$, we conclude that

$$\lim_{n\to\infty} -\log\frac{\min_{\mathcal{D}:\alpha_n(\mathcal{D})\le\varepsilon} \beta_n(\mathcal{D})}{n} = D(\mathrm{P}_0\|\mathrm{P}_1).$$

□

For example, if we want the false alarm probability $\alpha \le 10^{-2}$ and missed detection probability $\beta \le 10^{-30}$. Then, how many samples do we need? Applying the above result, we know that essentially it does not depend on α, but only on β when n is large:

$$n \approx -\frac{\log 10^{-30}}{D(\mathrm{P}_0\|\mathrm{P}_1)} \approx \frac{100}{D(\mathrm{P}_0\|\mathrm{P}_1)}.$$

When $\mathrm{P}_0 = \mathrm{Bern}(0.45)$ and $\mathrm{P}_1 = \mathrm{Bern}(0.5)$, we have $D(\mathrm{P}_0\|\mathrm{P}_1) \approx 0.0072$ and $n \approx 14000$. We see that the number of samples needed is inversely proportional to the divergence between the two hypotheses.

In a similar way, one can show that the following.

For any $\varepsilon \in (0,1)$, we have

$$\lim_{n\to\infty} -\frac{\log \min_{\mathcal{D}:\beta_n(\mathcal{D})\le\varepsilon} \alpha_n(\mathcal{D})}{n} = D(\mathrm{P}_1\|\mathrm{P}_0).$$

6.2.2.2 Chernoff's regime

In some cases, we want both α_n and β_n decay exponentially, say $\alpha_n \doteq 2^{-n\eta_\alpha}$ and $\beta_n \doteq 2^{-n\eta_\beta}$, where η_α and η_β are the corresponding error exponents. This is known as Chernoff's regime. From the previous discussion, any $\eta_\alpha > D(\mathrm{P}_1\|\mathrm{P}_0)$ would imply $\beta_n \to 1$ and $\eta_\beta > D(\mathrm{P}_0\|\mathrm{P}_1)$ would imply $\alpha_n \to 1$. Therefore, to avoid triviality, we consider the case

$$\eta_\alpha \in (0, D(\mathrm{P}_1\|\mathrm{P}_0)], \quad \eta_\beta \in (0, D(\mathrm{P}_0\|\mathrm{P}_1)].$$

We say that a pair of error exponents $(\eta_\alpha, \eta_\beta)$ is **achievable** if there exist a sequence of strategies so that

$$\liminf_{n\to\infty} -\frac{\log \alpha_n}{n} \ge \eta_\alpha, \qquad \liminf_{n\to\infty} -\frac{\log \beta_n}{n} \ge \eta_\beta$$

LRT exponents. For each $\lambda \in (0,1)$, there exists a LRT such that

$$\alpha_n \doteq 2^{-nD(\mathrm{P}_\lambda\|\mathrm{P}_0)}, \quad \beta_n \doteq 2^{-nD(\mathrm{P}_\lambda\|\mathrm{P}_1)}.$$

where

$$p_\lambda(x) := \frac{p_0^{1-\lambda}(x)p_1^\lambda(x)}{\sum_{a\in\mathcal{X}} p_0^{1-\lambda}(a)p_1^\lambda(a)}$$

Moreover, the corresponding test threshold is

$$\tau = 2^{n(D(\mathrm{P}_\lambda\|\mathrm{P}_1)-D(\mathrm{P}_\lambda\|\mathrm{P}_0))} \doteq \frac{\alpha_n}{\beta_n}. \tag{6.9}$$

Therefore, the error exponent pair

$$\big(\eta_\alpha^*(\lambda) := D(\mathrm{P}_\lambda\|\mathrm{P}_0), \eta_\beta^*(\lambda) := D(\mathrm{P}_\lambda\|\mathrm{P}_1)\big)$$

is achievable for every $\lambda \in (0,1)$.

Proof. First, notice that each $\mathcal{D}_\tau$ only depends on the type of the sequence. The LRT is therefore a test of types.

Indeed, we have

$$\begin{aligned}\frac{\mathrm{P}_0^n(x^n)}{\mathrm{P}_1^n(x^n)} &= 2^{\sum_i \log \frac{\mathrm{P}_0(x_i)}{\mathrm{P}_1(x_i)}} \\ &= 2^{n\sum_{a\in\mathcal{X}} \hat{\mathrm{P}}_{x^n}(a)\log\frac{\mathrm{P}_0(a)}{\mathrm{P}_1(a)}} \\ &= 2^{nD(\hat{\mathrm{P}}_{x^n}\|\mathrm{P}_1)-nD(\hat{\mathrm{P}}_{x^n}\|\mathrm{P}_0)}\end{aligned}$$

where $\hat{\mathrm{P}}_{x^n} := (\hat{\mathrm{P}}_{x^n}(a) : a \in \mathcal{X})$ is the type of x^n. Therefore, we can rewrite the region $\mathcal{D}_\tau$ as a set of type classes

$$\mathcal{D}_\tau = \left\{\mathcal{T}(\mathrm{P}) : D(\mathrm{P}\|\mathrm{P}_1) - D(\mathrm{P}\|\mathrm{P}_0) > \frac{1}{n}\log\tau\right\}$$

Reparameterizing with $T = \frac{1}{n}\log\tau$, we define

$$\begin{aligned}\mathcal{E}_T &:= \{\mathrm{Q} : D(\mathrm{Q}\|\mathrm{P}_1) - D(\mathrm{Q}\|\mathrm{P}_0) > T\} \\ &= \left\{\mathrm{Q} : \mathbb{E}_{X\sim\mathrm{Q}}\left[\log\frac{p_0(X)}{p_1(X)}\right] > T\right\}\end{aligned}$$

Therefore, for every T, we have

$$\alpha_n = \mathrm{P}_0^n(\bar{\mathcal{E}}_T) \doteq 2^{-nD(\mathrm{P}_0^*\|\mathrm{P}_0)}$$

where we have applied Sanov's theorem, and

$$\mathrm{P}_0^* := \arg\min_{\mathrm{Q}\in\bar{\mathcal{E}}_T} D(\mathrm{Q}\|\mathrm{P}_0).$$

Similarly, for every T, we have

$$\beta_n = \mathrm{P}_1^n(\mathcal{E}_T) \doteq 2^{-nD(\mathrm{P}_1^*\|\mathrm{P}_1)}$$

with

$$\mathrm{P}_1^* := \arg\min_{\mathrm{Q}\in\mathcal{E}_T} D(\mathrm{Q}\|\mathrm{P}_1).$$

In the Chernoff regime, we assume that α_n decays exponentially with n. This implies that $\mathrm{P}_0 \in \mathcal{E}_T$, for otherwise $\min_{\mathrm{Q}\in\bar{\mathcal{E}}_T} D(\mathrm{Q}\|\mathrm{P}_0) = D(\mathrm{P}_0\|\mathrm{P}_0) = 0$. Therefore, we have $D(\mathrm{P}_0\|\mathrm{P}_1) - D(\mathrm{P}_0\|\mathrm{P}_0) > T$, i.e.,

$$T < D(\mathrm{P}_0\|\mathrm{P}_1).$$

Similarly, β_n also decays exponentially with n, implying that

$$T \geq -D(\mathrm{P}_1\|\mathrm{P}_0).$$

To find out P_0^*, we apply (6.6) with $g(x) = \log\frac{p_1(x)}{p_0(x)}$ and $\alpha = -T$, and obtain

$$p_0^*(x) \propto 2^{\lambda g(x)} p_0(x) \propto p_\lambda(x)$$

where λ is such that

$$D(\mathrm{P}_\lambda\|\mathrm{P}_1) - D(\mathrm{P}_\lambda\|\mathrm{P}_0) = T. \tag{6.10}$$

Since the left-hand side of (6.10) is a continuous (and even monotone) function of λ and equals $D(\mathrm{P}_0\|\mathrm{P}_1)$ when $\lambda = 0$ and $-D(\mathrm{P}_1\|\mathrm{P}_0)$ when $\lambda = 1$. There exists $\lambda \in [0,1]$ such that (6.10) is satisfied.

To find out P_1^*, we apply (6.6) with $g(x) = \log\frac{p_0(x)}{p_1(x)}$ and $\alpha = T$, and obtain

$$p_1^*(x) \propto 2^{\lambda' g(x)} p_1(x) \propto p_{1-\lambda'}(x)$$

where λ is such that

$$D(\mathrm{P}_{1-\lambda'}\|\mathrm{P}_1) - D(\mathrm{P}_{1-\lambda'}\|\mathrm{P}_0) = T. \tag{6.11}$$

From (6.10) and (6.11), we have $\lambda = 1 - \lambda'$, and

$$\mathrm{P}_0^* = \mathrm{P}_1^* = \mathrm{P}_\lambda.$$

The threshold (6.9) is straightforward from (6.10). □

Intuitively, P_λ is on the geodesic path (with respect to the divergence) between P_0 and P_1 in the simplex. Note that $\lambda \mapsto (\eta^*_\alpha(\lambda), \eta^*_\beta(\lambda))$ is continuous in $(0,1)$. Therefore, the set $\mathcal{L} := \{(\eta^*_\alpha(\lambda), \eta^*_\beta(\lambda)) : \lambda \in (0,1)\}$ is a curve from $(0, D(P_0 \| P_1))$ to $(D(P_1 \| P_0), 0)$.

Upper boundary of error exponents. $\mathcal{L}$ is the upper boundary of all achievable error exponent pairs.

Proof. Assume that $(\eta_\alpha, \eta_\beta)$ is achievable and lies above the boundary define by $\mathcal{L}$, then, there must exist a $\lambda \in (0,1)$ such that $\eta_\alpha = \eta^*_\alpha(\lambda) + \varepsilon$ and $\eta_\beta = \eta^*_\beta(\lambda) + \varepsilon'$ with $\varepsilon, \varepsilon' > 0$.

When n is large enough, there exist LRT schemes such that $\alpha^*_n \geq 2^{-n(\eta^*_\alpha(\lambda)+\varepsilon/3)}$ and $\beta^*_n \geq 2^{-n(\eta^*_\beta(\lambda)+\varepsilon'/3)}$, for otherwise we don't achieve the exponent pair $(\eta^*_\alpha(\lambda), \eta^*_\beta(\lambda)$.

Similarly, when n is large enough, since $(\eta_\alpha, \eta_\beta)$ is achievable, there exist a scheme such that $\alpha_n \leq 2^{-n(\eta_\alpha - \varepsilon/3)} \leq 2^{-n(\eta^*_\alpha(\lambda)+2\varepsilon/3)} < 2^{-n(\eta^*_\alpha(\lambda)+\varepsilon/3)} \leq \alpha^*_n$, and $\beta_n < \beta^*$ for the same reason. This is a contradiction to the Neyman-Pearson lemma. Therefore, $\mathcal{L}$ is indeed the upper boundary of all achievable exponent pairs.

□

Exercises[14]

1. Large deviations [CT 11.18]. Let $X_1, X_2, \dots$ be i.i.d. random variables drawn according to the geometric distribution
$$P(X = k) = p^{k-1}(1 - p), \quad k = 1, 2, \dots$$
Find good estimates (to first order in the exponent) of:
   - $P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha)$
   - $P(X_1 = k \,|\, \frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha)$
   - Evaluate the previous two questions for $p = \frac{1}{2}$ and $\alpha = 4$.

2. Sanov [CT 11.14]. Let X_i be i.i.d.$\sim \mathcal{N}(0, \sigma^2)$.
   - Find the exponent in the behavior of $P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha^2)$. This can be done from first principles (since the normal dis- tribution is nice) or by using Sanov's theorem.
   - What do the data look like if $\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha^2$? That is what is the P^* that minimizes $D(P\|Q)$?

3. Counting states [CT 11.15]. Suppose that an atom is equally likely to be in each of six states, $X \in \{s_1, s_2, s_3, \dots, s_6\}$. One observes n atoms $X_1, X_2, \dots, X_n$ independently drawn according to this uniform distribution. It is observed that the frequency of occurrence of state s_1 is twice the frequency of occurrence of state s_2.
   - To first order in the exponent, what is the probability of observing this event?
   - Assuming n large, find the conditional distribution of the state of the first atom X_1, given this observation.

4. Hypothesis testing [CT 11.16]. Let $\{X_i\}$ be i.i.d.$\sim P(x)$, $x \in \{1, 2, \dots\}$. Consider two hypotheses, $H_0 : P = P_0$ vs. $H_1 : P = P_1$, where $P_0(x) = 2^{-x}$ and $P_1(x) = p^{x-1}(1 - p)$.
   - Find $D(P_0\|P_1)$.
   - Let $P(H_0) = \frac{1}{2}$. Find the minimal probability of error test for H_0 vs. H_1 given data $X_1, X_2, \dots, X_n \sim P$

5. Hypothesis testing [CT 11.12]. Let $X_1, X_2, \dots, X_n$ be i.i.d.$\sim$ P. Consider the hypothesis test $H_0 : P = P_0$ vs. $H_1 : P = P_1$. Let
$$P_0(x) = \begin{cases} \frac{1}{2}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{4}, & x = 1 \end{cases}$$
and
$$P_1(x) = \begin{cases} \frac{1}{4}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{2}, & x = 1 \end{cases}$$
Find the error exponent for $P(\text{Decide } H_1 \,|\, H_0 \text{ is true})$ in the best hypothesis test of H_0 vs. H_1 subject to $P(\text{Decide } H_0 \,|\, H_1 \text{ is true}) \leq \frac{1}{2}$.

[14] The citations "CT", "CK" refer to Cover-Thomas, Csiszár-Körner, respectively.

Elements of Information Theory **2025-2026**

Lecture 7: Information Theory and Machine Learning

Lecturer:

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In this lecture, we apply information measures to characterize performance bounds on machine learning algorithms in a theoretical setup.

Consider a training set $\mathrm{Z}^n \in \mathcal{Z}^n$ with i.i.d. samples from the same distribution P_Z, i.e., $\mathrm{Z}^n \sim \mathrm{P}_Z^n$. The **learning algorithm** $\mathrm{P}_{W|Z^n}$ "learns" the **hypothesis** $\mathrm{W} \in \mathcal{W}$ from the data Z^n. In the deterministic setting, W is a function of Z^n. For instance, W can be the vector of coefficients of a neural network, or a linear classifier, etc. The **loss function**

$$l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$$

measures the performance of a given hypothesis $w \in \mathcal{W}$ on a data point $z \in \mathcal{Z}$. For instance, let $z := (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where x is an image and y is a label, w be the parameter of a classification algorithm $f_w : \mathcal{X} \to \mathcal{Y}$. Then, the classification error can be $l(w, z) := \mathbf{1}\{f_w(x) \neq y\}$. For regression problems, we can define $l(w, z) := |f_w(x) - y|^2$.

We define the **population loss** as the expectation of the loss function for a given hypothesis

$$L_{\mathrm{P}_Z}(w) := \mathbb{E}_{\mathrm{P}_Z}(l(w, \mathrm{Z})).$$

The **training loss** is the defined with the training data

$$L_{Z^n}(w) := \frac{1}{n} \sum_{i=1}^{n} l(w, \mathrm{Z}_i) = L_{\hat{\mathrm{P}}_{Z^n}}(w).$$

The **generalization error** with respect to the distribution P_Z is defined as

$$\mathsf{gen}(w, \mathrm{Z}^n) := L_{\mathrm{P}_Z}(w) - L_{Z^n}(w).$$

The average population loss and average training loss are defined as

$$L := \mathbb{E}_{\mathrm{P}_{WZ^n}}\left[L_{\mathrm{P}_Z}(\mathrm{W})\right], \quad \hat{L} := \mathbb{E}_{\mathrm{P}_{WZ^n}}\left[L_{Z^n}(\mathrm{W})\right],$$

and the average generalization error

$$\overline{\mathsf{gen}} := \mathbb{E}_{\mathrm{P}_{WZ^n}}[\mathsf{gen}(w, \mathrm{Z}^n)] = L - \hat{L}.$$

In most applications, the ultimate goal is to minimize the population loss. In this lecture, we present an information-theretic analysis such a problem via bounds on the generalization error.

7.1 Donsker-Varadhan variational representation of divergence

Donsker-Varadhan. For any $\mathrm{P} \ll \mathrm{Q}$ and f such that $\mathbb{E}_{\mathrm{Q}}\left[e^{f(\mathrm{X})}\right] < \infty$, we have

$$D(\mathrm{P}\|\mathrm{Q}) \geq \mathbb{E}_{\mathrm{P}}\left[f(\mathrm{X})\right] - \log\left(\mathbb{E}_{\mathrm{Q}}\left[e^{f(\mathrm{X})}\right]\right) \tag{7.1}$$

with equality if $f(x) = \frac{d\mathrm{P}}{d\mathrm{Q}}(x)$, Q-a.s. As a result, we have

$$D(\mathrm{P}\|\mathrm{Q}) = \sup_{f:\mathbb{E}_{\mathrm{Q}}\left[e^{f(\mathrm{X})}\right]<\infty} \left\{\mathbb{E}_{\mathrm{P}}\left[f(\mathrm{X})\right] - \log\left(\mathbb{E}_{\mathrm{Q}}\left[e^{f(\mathrm{X})}\right]\right)\right\}$$

Proof. Let Q^f be the tilted distribution, i.e., $q^f(x) := q(x)e^{f(x)}/\mathbb{E}_{\mathrm{Q}}[e^{f(\mathrm{X})}]$. Then, the inequality (7.1) is straightforward from $D(\mathrm{P}\|\mathrm{Q}^f) \geq 0$. Equality condition can be directly checked as well. □

Since $(\mathrm{P},\mathrm{Q}) \mapsto \mathbb{E}_{\mathrm{P}}\left[f(\mathrm{X})\right] - \log\left(\mathbb{E}_{\mathrm{Q}}\left[e^{f(\mathrm{X})}\right]\right)$ is continuous under the weak topology of measures, the supremum of continuous functions are lower semicontinuous[15]

Lower semicontinuity of divergence. The divergence $(\mathrm{P},\mathrm{Q}) \mapsto D(\mathrm{P}\|\mathrm{Q})$ is lower semicontinuous (l.s.c.), i.e., if $(\mathrm{P}_n,\mathrm{Q}_n) \to (\mathrm{P},\mathrm{Q})$ weakly, then

$$\liminf_{n\to\infty} D(\mathrm{P}_n\|\mathrm{Q}_n) \geq D(\mathrm{P}\|\mathrm{Q}).$$

Relaxed binary divergence. Define the binary divergence $d(p\|q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$, and the relaxed binary divergence

$$d_\gamma(p\|q) := p\gamma - \log(1-q+qe^\gamma),$$

for any $\gamma \in \mathbb{R}$. We have

$$d(p\|q) = \sup_\gamma d_\gamma(p\|q)$$

The function $(p,q) \mapsto d_\gamma(p\|q)$ is convex.

Proof. Apply DV (7.1) with $f(0) = 0$ and $f(1) = \gamma$. □

7.2 Some concentration inequalities

Hoeffding's lemma. For any bounded $\mathrm{X} \in [a,b]$, we have

$$\mathbb{E}\left[e^{\lambda(\mathrm{X}-\mathbb{E}[\mathrm{X}])}\right] \leq e^{\frac{\lambda^2(a-b)^2}{8}}.$$

Proof. Let $\mathrm{Y} \in [a,b]$ with $\mathbb{E}[\mathrm{Y}] = 0$ and $b \geq 0 \geq a$. Note that by convexity $e^{\lambda\mathrm{Y}} \leq \frac{b-\mathrm{Y}}{b-a}e^{\lambda a} + \frac{\mathrm{Y}-a}{b-a}e^{\lambda b}$, implying

$$\mathbb{E}[e^{\lambda\mathrm{Y}}] \leq \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b} = \gamma e^{s(\gamma-1)} + (1-\gamma)e^{s\gamma}$$

[15] A function is lower semicontinuous if $\liminf_{n\to\infty} f(x_n) \geq f(x)$ for any $\{x_n\}$ converging to x.

where $\gamma := \frac{b}{b-a}$ and $s := \lambda(b-a)$. We would like to show that $f(s) := \log(\gamma e^{s(\gamma-1)} + (1-\gamma)e^{s\gamma}) \le \frac{s^2}{8}$. Note that

$$\begin{aligned} f(0) &= 0 \\ f'(0) &= 0 \\ f''(s) &= \frac{\gamma(1-\gamma)e^{s(\gamma-1)}e^{s\gamma}}{\left(\gamma e^{s(\gamma-1)} + (1-\gamma)e^{s\gamma}\right)^2} \le \frac{1}{4} \end{aligned}$$

where the last inequality is the AM-GM inequality. From Taylor-Lagrange, we have $f(s) = f(0) + f'(0)s + \frac{f''(\tilde{s})}{2}s^2 \le \frac{s^2}{8}$, which completes the proof. □

X is called σ-**sub-Gaussian** if

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall \lambda \in \mathbb{R}. \tag{7.2}$$

For any σ-**sub-Gaussian** X, we have

$$\mathbb{E}\left[e^{\frac{\lambda^2(X-\mathbb{E}[X])^2}{2\sigma^2}}\right] \le \frac{1}{\sqrt{1-\lambda^2}}.$$

Proof. Let $Y := (X - \mathbb{E}[X])/\sigma$, i.e., 1-sub-Gaussian. Introducing $Z \sim \mathcal{N}(0,1)$,

$$\begin{aligned} \mathbb{E}_X\left[e^{\frac{\lambda^2(X-\mathbb{E}[X])^2}{2\sigma^2}}\right] &= \mathbb{E}_Y\left[e^{\frac{\lambda^2 Y^2}{2}}\right] \\ &= \mathbb{E}_Y\left[\mathbb{E}_Z\left[e^{\lambda YZ}\right]\right] \\ &\le \mathbb{E}_Z\left[e^{\frac{\lambda^2 Z^2}{2}}\right] \\ &= \frac{1}{\sqrt{1-\lambda^2}}. \end{aligned}$$

□

Now, we see that any bounded X is also $\frac{b-a}{2}$-sub-Gaussian.

Let $X \in [0,1]$. Then, for any γ,

$$\mathbb{E}\left[e^{d_\gamma(X\|\mathbb{E}[X])}\right] \le 1.$$

Let $X_i \in [0,1]$, $i = 1, \dots, n$, be independent. Then, by convexity and independence, we have

$$\mathbb{E}\left[e^{nd_\gamma(\frac{1}{n}\sum_{i=1}^n X_i \| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i])}\right] \le 1. \tag{7.3}$$

When $X_i \in [0,1]$, $i = 1, \dots, n$, are i.i.d. with mean μ, we have

$$\mathbb{E}\left[e^{nd(\frac{1}{n}\sum_{i=1}^n X_i \| \mu)}\right] \le 2\sqrt{n}. \tag{7.4}$$

Proof. From the convexity of $x \mapsto e^{yx}$,

$$\begin{aligned}\mathbb{E}\left[e^{d_y(\mathrm{X}\|\mathbb{E}[\mathrm{X}])}\right] &= \frac{\mathbb{E}\left[e^{y\mathrm{X}}\right]}{1-\mathbb{E}[\mathrm{X}]+\mathbb{E}[\mathrm{X}]e^y} \\ &\le \frac{\mathbb{E}\left[1-\mathrm{X}+\mathrm{X}e^y\right]}{1-\mathbb{E}[\mathrm{X}]+\mathbb{E}[\mathrm{X}]e^y} \\ &= 1.\end{aligned}$$

Then, from the convexity of $(p,q) \mapsto d_y(p\|q)$, we have

$$\begin{aligned}\mathbb{E}\left[e^{nd_y(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathrm{X}_i])}\right] &\le \mathbb{E}\left[e^{\sum_{i=1}^n d_y(\mathrm{X}_i\|\mathbb{E}[\mathrm{X}_i])}\right] \\ &= \prod_{i=1}^n \mathbb{E}\left[e^{d_y(\mathrm{X}_i\|\mathbb{E}[\mathrm{X}_i])}\right] \\ &\le 1.\end{aligned}$$

To show (7.4), we first notice that $\mathbb{E}\left[e^{nd(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\mu)}\right]$ is maximized when X_i's are Bernoulli. This is due to Hoeffding's reduction. Indeed, since $\mathrm{X}_i \mapsto e^{nd(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\mu)}$ is convex, the expectation with respect to X_i is maximized when X_i is Bernoulli with the same mean. Applying the same argument and we can replace all X_i's by Bernoulli's and obtain an upper bound. Note that $e^{nd(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\mu)} = \left(\frac{k}{n\mu}\right)^k\left(\frac{n-k}{n(1-\mu)}\right)^{n-k}$ where $k := \frac{1}{n}\sum_{i=1}^n \mathrm{X}_i$ follows the Binomial distribution when X_i's are Bernoulli's. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[e^{nd(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\mu)}\right] &\le \sum_{k=0}^n \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} \\ &= 2 + \sum_{k=1}^{n-1} \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}\end{aligned}$$

where the bound becomes 2 and 2.5 for $n=1$ and $n=2$, respectively. When $n \ge 3$, we further bound

$$\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} \le \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}, \quad k = 1,\ldots,n-1,$$

using Stirling. Note that one can bound the sum by a Riemann integral as

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\sum_{k=1}^{n-1}\sqrt{\frac{n}{k(n-k)}} &= \sqrt{\frac{n}{2\pi}}\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{(k/n)(1-k/n)}} \\ &\le \sqrt{\frac{n}{2\pi}}\int_0^1 \frac{1}{\sqrt{t(1-t)}}\,\mathrm{d}t \\ &= \sqrt{\frac{n\pi}{2}}\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[e^{nd(\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i\|\mu)}\right] \le 2 + \sqrt{\frac{n\pi}{2}}$$

that is smaller than $2\sqrt{n}$ for $n \ge 3$. □

7.3 Generalization bounds (in expectation)

The main tool that we use in the following is the DV representation (7.1). For any $f(\mathrm{W},\mathrm{Z}^n)$, we have

$$\mathbb{E}_{\mathrm{P}_{WZ^n}}\left[f(\mathrm{W},\mathrm{Z}^n)\right] \le \log \mathbb{E}_{\mathrm{Q}_W\mathrm{P}_{Z^n}}\left[e^{f(\mathrm{W},\mathrm{Z}^n)}\right] + D(\mathrm{P}_{WZ^n}\|\mathrm{Q}_W\mathrm{P}_{Z^n})$$

In particular, if we let $\mathrm{Q}_W = \mathrm{P}_W$, we have

$$\mathbb{E}_{\mathrm{P}_{WZ^n}}\left[f(\mathrm{W},\mathrm{Z}^n)\right] \le \log \mathbb{E}_{\mathrm{P}_W\mathrm{P}_{Z^n}}\left[e^{f(\mathrm{W},\mathrm{Z}^n)}\right] + I(\mathrm{W};\mathrm{Z}^n) \tag{7.5}$$

7.3.1 Sub-Gaussian loss function

Fix $\lambda > 0$ and let $f(\mathrm{W}, \mathrm{Z}^n) = \lambda\,\mathsf{gen}(\mathrm{W}, \mathrm{Z}^n)$. For any given $\mathrm{W} = w$, we have

$$\mathsf{gen}(w, \mathrm{Z}^n) = \frac{1}{n}\sum_{i=1}^{n}\big(l(w, Z_i) - \mathbb{E}_{\mathrm{P}_{Z_i}}[l(w, Z_i)]\big)$$

Assume that $l(w, Z_i)$, $i = 1, \ldots, n$, is σ-sub-Gaussian. Then, $\mathsf{gen}(w, \mathrm{Z}^n)$ is $\frac{\sigma}{\sqrt{n}}$-sub-Gaussian. From the definition (7.2), we have

$$\mathbb{E}_{\mathrm{P}_{Z^n}}\left[e^{\lambda\,\mathsf{gen}(w, \mathrm{Z}^n)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2n}}, \quad \forall\, w$$

Plugging this bound back to (7.5), we have

$$\mathbb{E}_{\mathrm{P}_{WZ^n}}\left[\mathsf{gen}(\mathrm{W}, \mathrm{Z}^n)\right] \leq \frac{\lambda\sigma^2}{2n} + \frac{1}{\lambda} I(\mathrm{W}; \mathrm{Z}^n)$$

Taking the infimum over all $\lambda > 0$, we have

$$\mathbb{E}_{\mathrm{P}_{WZ^n}}\left[\mathsf{gen}(\mathrm{W}, \mathrm{Z}^n)\right] \leq \sqrt{\frac{2\sigma^2 I(\mathrm{W}; \mathrm{Z}^n)}{n}}.$$

Similarly, we can fix $\lambda < 0$ and get the same negative lower bound. We have

$$\hat{L} - \sqrt{\frac{2\sigma^2 I(\mathrm{W}; \mathrm{Z}^n)}{n}} \leq L \leq \hat{L} + \sqrt{\frac{2\sigma^2 I(\mathrm{W}; \mathrm{Z}^n)}{n}}.$$

7.3.2 Bounded loss function

In particular, if the loss function is bounded in $[0, 1]$, it is also $\frac{1}{2}$-sub-Gaussian, and we have

$$L \leq \hat{L} + \sqrt{\frac{I(\mathrm{W}; \mathrm{Z}^n)}{2n}}.$$

It is also possible to derive a different bound via the binary divergence function. Specifically, let

$$f(w, \mathrm{Z}^n) = n d_\gamma(L_{Z^n}(w) \| L_{\mathrm{P}_Z}(w)) = n d_\gamma\left(\frac{1}{n}\sum_{i=1}^{n} l(w, \mathrm{Z}_i) \| E_{\mathrm{P}_{Z_i}}[l(w, \mathrm{Z}_i)]\right)$$

We have the following bound on $d_\gamma(\hat{L}\|L)$.

$$\begin{aligned} n d_\gamma(\hat{L}\|L) &= \mathbb{E}_{\mathrm{P}_{WZ^n}}\left[n d_\gamma(L_{Z^n}(w) \| L_{\mathrm{P}_Z}(w))\right] \\ &\leq \log \mathbb{E}_{\mathrm{P}_W \mathrm{P}_{Z^n}}\left[e^{n d_\gamma(L_{Z^n}(w) \| L_{\mathrm{P}_Z}(w))}\right] + I(\mathrm{W}; \mathrm{Z}^n) \\ &\leq I(\mathrm{W}; \mathrm{Z}^n) \end{aligned}$$

where the last inequality is from (7.3). Since the above inequality holds for any γ, we have

$$d(\hat{L}\|L) \leq \frac{1}{n} I(\mathrm{W}; \mathrm{Z}^n)$$