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 $\mathbf{X}$
$\mathrm{P}(\cdot)$	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 $P_{Y|X}P_X$. It can be regarded as the mixture of different distributions $P_{Y|X=x}$ 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) $P_{Y|X}$ and $P_{X|Y}$, such that

$$P_{XY} = P_X P_{Y|X} = P_Y P_{X|Y}.$$

In this case, the notation $P_X, P_Y, P_{Y|X}, P_{X|Y}$ 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 $P_{Y|X}$ but not a joint law. Together with two different input distributions P_X and Q_X, we can form two distinct joint distributions, $P_{XY} := P_X P_{Y|X}, Q_{XY} := Q_X P_{Y|X}$. 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 $X \in \mathcal{X}$ and f is convex in $\mathcal{X}$, then $\mathbb{E}f(X) \ge f(\mathbb{E}X)$. For concave functions, we have $\mathbb{E}f(X) \le f(\mathbb{E}X)$.

1.1 Entropy

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

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

Sometimes, we use the notation $H(P_X)$ 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 $P_{X_1 \cdots X_n}$ (or, in short, P_{X^n}).

$$H(X_1, \ldots, X_n) := \mathbb{E}_{P_{X^n}} \log \frac{1}{p(X_1,\ldots,X_n)} = \sum_{x_1\in\mathcal{X}_1} \cdots \sum_{x_n\in\mathcal{X}_n} p(x_1,\ldots,x_n) \log \frac{1}{p(x_1,\ldots,x_n)}.$$

We also define the **conditional entropy** as

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

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

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

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

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

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

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

Similarly, the conditional version also holds.

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

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]

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

And the **conditional cross-entropy**

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

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

$$\boxed{H(\mathrm{P}_{X|Y} \| \mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) = \mathbb{E}_{\mathrm{P}_Y \mathrm{P}_{X|Y}}\left[\log \frac{1}{q(X|Y)}\right]}$$

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

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

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

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

And the conditional version also holds.

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

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

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

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.

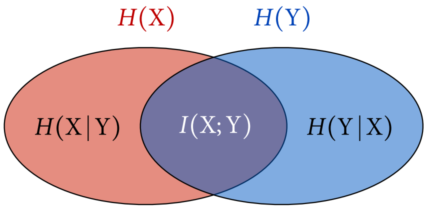


Figure 1.1: The Venn diagram.

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 $\mathrm{P} = \mathrm{Q}$.

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 $S \sim P_S := \text{Bern}(\lambda)$, 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 $I(P_X, \lambda P^1_{Y|X} + (1-\lambda)P^0_{Y|X}) \le \lambda I(P_X, P^1_{Y|X}) + (1-\lambda)I(P_X, P^0_{Y|X})$ for any kernels $P^0_{Y|X}$ and $P^1_{Y|X}$ and $\lambda \in [0,1]$. We can prove it in two different ways. First, we can apply the convexity of divergence. Indeed,

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

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

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

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

□

- Data processing inequality (DPI)

$$\begin{aligned} D(P_X \| Q_X) &\ge D(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X) \\ I(P_X, P_{Y|X}) &\ge I(P_X, P_{Z|Y} \circ P_{Y|X}) \end{aligned}$$

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

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

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

Proof. Let $\tilde{P}_Y := P_{Y|X} \circ P_X$ and $\tilde{Q}_Y := Q_{Y|X} \circ P_X$ be the marginals of Y from $P_{Y|X}P_X$ and $P_{Y|X}Q_X$, respectively. We have

$$\begin{aligned} D(P_X \| Q_X) &= D(P_X \| Q_X) + D(P_{Y|X} \| P_{Y|X} \mid P_X) \\ &= D(P_X P_{Y|X} \| Q_X P_{Y|X}) \\ &= D(\tilde{P}_Y \tilde{P}_{X|Y} \| \tilde{Q}_Y \tilde{Q}_{X|Y}) \\ &= D(\tilde{P}_Y \| \tilde{Q}_Y) + D(\tilde{P}_{X|Y} \| \tilde{Q}_{X|Y} \mid \tilde{P}_Y) \\ &\ge D(\tilde{P}_Y \| \tilde{Q}_Y) \end{aligned}$$

Obviously, the conditional version also holds similarly.

For the mutual information, since $X \to Y \to Z$, there exist $P_{Z|Y}$ such that $P_{Z|X} = P_{Z|Y} \circ P_{Y|X}$ and $P_Z = P_{Z|Y} \circ P_Y$. Thus,

$$\begin{aligned} I(X;Y) &= D(P_{Y|X} \| P_Y \mid P_X) \\ &\ge D(P_{Z|Y} \circ P_{Y|X} \| P_{Z|Y} \circ P_Y \mid P_X) \\ &= D(P_{Z|X} \| P_Z \mid P_X) \\ &= I(X;Z). \end{aligned}$$

□

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

   $$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2\,|\,x_1)\cdots p(x_n\,|\,x_{n-1}).$$

   Reduce $I(\mathrm{X}_1;\mathrm{X}_2,\ldots,\mathrm{X}_n)$ 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 $n = 2, 3, \ldots$ 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 $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$ be three binary Bernoulli(1/2) random variables that are pairwise independent; that is, $I(\mathrm{X};\mathrm{Y}) = I(\mathrm{X};\mathrm{Z}) = I(\mathrm{Y};\mathrm{Z}) = 0$.
   - Under this constraint, what is the minimum value for constraint, $H(\mathrm{X},\mathrm{Y},\mathrm{Z})$?
   - 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 $\mathrm{X} \in \{1,2,\ldots\}$, if $E \log \mathrm{X} < \infty$, then $H(\mathrm{X}) < \infty$.

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 $\{\mathrm{X}_i\}$. 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 $I(\mathrm{N};\mathrm{X}^\mathrm{N})$.
   - Find $H(\mathrm{X}^\mathrm{N} \,|\, \mathrm{N})$.
   - Find $H(\mathrm{X}^\mathrm{N})$.

   Let's now consider a different stopping time. For this part, again assume that $\mathrm{X} \sim$ Bernoulli(1/2) but stop at time $\mathrm{N} = 6$, with probability 1/3 and stop at time $\mathrm{N} = 12$ with probability 2/3. Let this stopping time be independent of the sequence $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{12}$.
   - Find $I(\mathrm{N};\mathrm{X}^\mathrm{N})$. Find $H(\mathrm{X} \,|\, \mathrm{N})$. Find $H(\mathrm{X}^\mathrm{N})$.

14. Function of variables from a Markov chain. [CK 3.7] Is it true that if $\mathrm{X}_1 \to \mathrm{X}_2 \to \cdots \to \mathrm{X}_n$, and f is an arbitrary function on the common range of the X_i's, then $f(\mathrm{X}_1) \to f(\mathrm{X}_2) \to \cdots \to f(\mathrm{X}_n)$? 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

$$I(\mathrm{X}_1;\mathrm{X}_2), I(\mathrm{X}_2;\mathrm{X}_3 \,|\, \mathrm{X}_1), \ldots, I(\mathrm{X}_{N-1};\mathrm{X}_N \,|\, \mathrm{X}_1, \ldots, \mathrm{X}_{N-2}).$$

   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 $\{1, 2, \ldots, n\}$?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.,

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

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

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

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.

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

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

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

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

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

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 $\mathrm{X} \sim \mathrm{P}$

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

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 $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})$,

$$2^{-n(1+\varepsilon)H(\mathrm{P})} \le \mathrm{P}^n(x^n) \le 2^{-n(1-\varepsilon)H(\mathrm{P})}, \tag{2.8}$$

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

$$\lim_{n\to\infty} \mathrm{P}^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) = 1, \quad \text{if } \lim_{n\to\infty} n\varepsilon^2 = \infty. \tag{2.9}$$

More specifically, we have

$$\mathrm{P}^n(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \le \delta_T(\varepsilon, n, \mathrm{P}) := 2M^* e^{-2n\varepsilon^2 c_{\mathrm{P}}} \tag{2.10}$$

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

$$\delta_T'(\varepsilon, n, \mathrm{P}) := -\log(1 - \delta_T(\varepsilon, n, \mathrm{P}))^+ \tag{2.11}$$

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

Proof. 1. The bounds (2.8) 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

$$\mathrm{P}^n(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) = \mathrm{P}^n\left(\bigcup_{a\in\mathcal{X}} \left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathrm{X}_i = a\} - \mathrm{P}(a)\right| > \varepsilon \mathrm{P}(a)\right\}\right) \tag{2.12}$$

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

$$= \sum_{a\in\mathcal{X},\mathrm{P}(a)>0} \mathrm{P}^n\left(\left\{\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathrm{X}_i = a\} - \mathrm{P}(a)\right| > \varepsilon \mathrm{P}(a)\right\}\right) \tag{2.14}$$

$$\le \sum_{a\in\mathcal{X},\mathrm{P}(a)>0} 2e^{-2n\varepsilon^2 \mathrm{P}(a)^2}, \tag{2.15}$$

where in (2.14), 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

$$\mathbb{P}\left\{|\mathrm{S_n} - \mathbb{E}(\mathrm{S}_n)| \ge t\right\} \le 2e^{-\frac{2nt^2}{\Delta_n}}. \tag{2.16}$$

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 (2.10). (2.9) 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 (2.10), $\mathrm{P}^n(\mathrm{X}^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P})) \ge (1-\delta_T(\varepsilon, n, \mathrm{P}))^+$. 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,
   $$|\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \,|\, y^n)| \ge 2^{n(1-\varepsilon)H(\mathrm{X}|\mathrm{Y}) - \delta'_T(\varepsilon/2, n, \mathrm{P}_{XY})}. \tag{2.17}$$

   where δ'_T is defined in (2.11).
6. **Joint typicality lemma**.
   - For every $y^n \in \mathcal{Y}^n$,
     $$\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) \le 2^{-n(I(\mathrm{X};\mathrm{Y}) - 2\varepsilon H(\mathrm{X}))}. \tag{2.18}$$
   - Let $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}$. Then,
     $$\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) \ge 2^{-n(I(\mathrm{X};\mathrm{Y}) + 2\varepsilon H(\mathrm{X})) - \delta'_T(\varepsilon/2, n, \mathrm{P}_{XY})}. \tag{2.19}$$

[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

$$\begin{aligned} 1 &\geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)} \mathrm{P}^n_{X|Y}(x^n|y^n) && (2.20) \\ &\geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)} 2^{-n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})} && (2.21) \\ &= |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)| 2^{-n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}. && (2.22) \end{aligned}$$

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

$$\mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \mathrm{P}_{XY}(a, b)| > \varepsilon \mathrm{P}_{XY}(a, b)) \to 0 \tag{2.23}$$

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

$$\pi_{\mathrm{X}^n,y^n}(a, b) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{(\mathrm{X}_i, y_i) = (a, b)\} \tag{2.24}$$

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

$$\begin{aligned} &|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \mathrm{P}_{XY}(a, b)| && (2.25) \\ &\leq |\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b)| + |\hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b) - \mathrm{P}_{XY}(a, b)| && (2.26) \\ &\leq |\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a \mid b)| + \frac{\varepsilon}{2}\mathrm{P}_{XY}(a, b), && (2.27) \end{aligned}$$

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

$$\begin{aligned} &\mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \mathrm{P}_{XY}(a, b)| > \varepsilon \mathrm{P}_{XY}(a, b)) && (2.28) \\ &\leq \mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \hat{\mathrm{P}}_{y^n}(b)\mathrm{P}_{X|Y}(a|b)| > (\varepsilon - \varepsilon/2)\mathrm{P}_{XY}(a, b)) && (2.29) \\ &\leq 2e^{-n\varepsilon^2 \mathrm{P}_{XY}(a,b)^2/2} && (2.30) \end{aligned}$$

according to Hoeffding's inequality (2.16). From the union bound, we have

$$\begin{aligned} \mathrm{P}_{X^n|Y^n=y^n}(\mathrm{X}^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)) &\leq \sum_{a,b:\mathrm{P}_{XY}(a,b)>0} \mathrm{P}_{X^n|Y^n=y^n}(|\hat{\mathrm{P}}_{\mathrm{X}^n,y^n}(a, b) - \mathrm{P}_{XY}(a, b)| > \varepsilon \mathrm{P}_{XY}(a, b)) \\ &\leq 2|\mathcal{X}|\,|\mathcal{Y}| e^{-2n\varepsilon^2 c_P/2}, && (2.31) \end{aligned}$$

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 (2.23).

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 $\mathrm{P}_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)) \leq |\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY} \mid y^n)| 2^{-n(1-\varepsilon)H(\mathrm{X}\mid\mathrm{Y})}$. Combining both inequalities, we prove the lower bound (2.17).

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

$$\begin{aligned}
\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) &= \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} \mathrm{P}_X^n(x^n) &(2.32)\\
&\le \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} 2^{-n(1-\varepsilon)H(\mathrm{X})} &(2.33)\\
&\le 2^{n(1+\varepsilon)H(\mathrm{X}|\mathrm{Y})}2^{-n(1-\varepsilon)H(\mathrm{X})} &(2.34)\\
&= 2^{-n(I(\mathrm{X};\mathrm{Y})-\varepsilon H(\mathrm{X})-\varepsilon H(\mathrm{X}|\mathrm{Y}))} &(2.35)\\
&\le 2^{-n(I(\mathrm{X};\mathrm{Y})-2\varepsilon H(\mathrm{X}))}. &(2.36)
\end{aligned}$$

To show the lower bound, apply the lower bound (2.17) 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

$$\begin{aligned}
\mathrm{P}_X^n(\mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)) &= \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} \mathrm{P}_X^n(x^n) &(2.37)\\
&\ge \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathrm{P}_{XY}|y^n)} 2^{-n(1+\varepsilon)H(\mathrm{X})} &(2.38)\\
&\ge 2^{n(1-\varepsilon)H(\mathrm{X}|\mathrm{Y})-\delta_T(\varepsilon/2,n,\mathrm{P}_{XY})}2^{-n(1+\varepsilon)H(\mathrm{X})} &(2.39)\\
&\ge 2^{-n(I(\mathrm{X};\mathrm{Y})+2\varepsilon H(\mathrm{X}))-\delta_T(\varepsilon/2,n,\mathrm{P}_{XY})}. &(2.40)
\end{aligned}$$

□

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,

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

if the following conditions are satisfied:[a]

$$\lim_{n\to\infty} n\varepsilon^2 = \infty \qquad (2.41)$$

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathrm{Y}^n \in \mathcal{T}_{\varepsilon/2}(\mathrm{P}_Y)\right\} = 1 \qquad (2.42)$$

$$\lim_{n\to\infty} nR_n - nI(\mathrm{X};\mathrm{Y}) - 2n\varepsilon H(\mathrm{X}) = \infty. \qquad (2.43)$$

[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.

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

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

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

For any given y^n, the events in the unions in the second probability are independent. From the probability lower bound (2.19), we have the following upper bound for $y^n\in\mathcal{T}^{(n)}_{\varepsilon/2}(\mathrm{P}_Y)$,

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

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

$$\mathcal{P}\left\{\bigcap_{m=1,\dots,2^{nR_n}}\left\{\mathrm{X}^n(m)\notin\mathcal{T}^{(n)}_{\varepsilon}(\mathrm{P}_{XY}|y^n)\right\}\right\}\leq\exp\left(-2^{nR}2^{-n(I(\mathrm{X};\mathrm{Y})+2\varepsilon H(X))-\delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})}\right)\tag{2.44}$$

$$=\exp\left(-2^{n(R_n-I(\mathrm{X};\mathrm{Y})-2\varepsilon H(\mathrm{X}))-\delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})}\right).\tag{2.45}$$

Therefore, we have

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

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

From conditions (2.41), (2.42), and (2.43), we have $\delta'_T(\varepsilon/2,n,\mathrm{P}_{XY})\to 0$, $\mathrm{P}_{Y^n}(\mathcal{T}^{(n)}_{\varepsilon/2}(\mathrm{P}_Y))\to 1$, 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 $\mathrm{X}^n(m)\sim\mathrm{P}^n_X$ for $m=1,\dots,2^{nR_n}$ (not necessarily independent for different m) be independent of an arbitrary random sequence Y^n. Then,

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

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 (2.18). Indeed, from the union bound, we have

$$\mathbb{P}\left\{\bigcup_{m=1,\dots,2^{nR_n}}\left\{(\mathrm{X}^n(m),\mathrm{Y}^n)\in\mathcal{T}^{(n)}_{\varepsilon}(\mathrm{P}_{XY})\right\}\right\}=\mathbb{P}\left\{\bigcup_{m=1,\dots,2^{nR_n}}\left\{\mathrm{X}^n(m)\in\mathcal{T}^{(n)}_{\varepsilon}(\mathrm{P}_{XY}\,|\,\mathrm{Y}^n)\right\}\right\}\tag{2.46}$$

$$\leq\sum_{m=1}^{2^{nR_n}}\mathbb{P}\left\{\mathrm{X}^n(m)\in\mathcal{T}^{(n)}_{\varepsilon}(\mathrm{P}_{XY}\,|\,\mathrm{Y}^n)\right\}\tag{2.47}$$

$$\leq 2^{nR_n-nI(\mathrm{X};\mathrm{Y})+2n\varepsilon H(\mathrm{X})},\tag{2.48}$$

where the last inequality is from the joint typicality lemma (2.18). 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) $\mathrm{P}^n(\mathcal{T}(\mathrm{Q})) \approx 2^{-nH(\mathrm{P})}$

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

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

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

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

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

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

15. The method of types can be used to show that the probability of observing a sequence x^n whose type $\hat{\mathrm{P}}_{x^n}$ 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 $D(\hat{\mathrm{P}}_{x^n} \| \mathrm{P})$

16. The joint typicality lemma states that the probability that $(\mathrm{X}^n, \mathrm{Y}^n) \sim \mathrm{P}_X^n \mathrm{P}_Y^n$ 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) $2^{-nH(\mathrm{X},\mathrm{Y})}$

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 $\mathcal{C} \subseteq \{0,1\}^* := \bigcup_{n\ge 0}\{0,1\}^n$. 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.

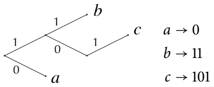
3.2.1 Prefix codes

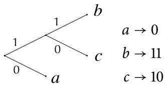
An important class of uniquely decodable codes is the **prefix code** (a.k.a. prefix-free code). For $m \leq n$, a string $y := (y_1 \cdots y_m)$ is called the **prefix** of another string $y' := (y'_1 \cdots y'_n)$ if $y_i = y'_i$ for $i = 1, \dots, m$, that is, the string y' starts with y, denoted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

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





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 $\mathcal{I}(y_i) \cap \mathcal{I}(y_j) = \emptyset$, $\forall\ i \neq j$.

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

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, $f_{\textbf{Shannon},\mathrm{Q}}$

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 $\{l_1, \ldots, l_M\}$ 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,

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

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 $i = 1, \ldots, M$. Since the Kraft-McMillan inequality ensures that $\sum_i q_i \leq 1$, we can apply the Shannon coding procedure with $\mathrm{Q} = [q_1, \ldots, q_M]$. The resulting code then assigns to the i-th symbol a codeword of length $\lceil \log \frac{1}{q_i} \rceil = l_i$, $i = 1, \ldots, M$, 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

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

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

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

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 $l_i = \lceil \log \frac{1}{p_i} \rceil \le \lceil \log \frac{1}{p_i} \rceil$, $i = 1, \ldots, M$. 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*:

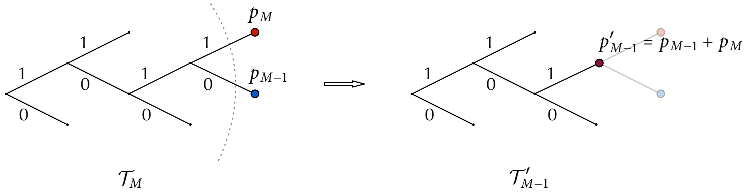
1. A symbol with higher probability should be assigned with a shorter codeword. That is, if $p_i \ge p_j$, then $l_i \le l_j$. Otherwise, swapping the codewords results in a smaller expected length.
2. The corresponding tree must be full, i.e., each node either is a leaf node or has two children. Otherwise,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

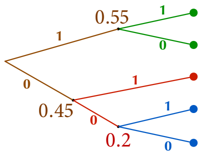
Let $\mathcal{T}^*$ be the tree corresponding to the optimal code, with ordered codeword length $\tilde{l}_1^* \le \tilde{l}_2^* \le \cdots \le \tilde{l}_M^*$. The corresponding probability is $\tilde{p}_1 \ge \tilde{p}_2 \ge \cdots \ge \tilde{p}_M$. At the beginning, nothing is known about $\mathcal{T}^*$ (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

$$\begin{aligned} \bar{L}(\mathcal{T}^*, (\tilde{p}_1, \ldots, \tilde{p}_M)) &:= \sum_{i=1}^{M} \tilde{l}_i^* \tilde{p}_i && (3.9) \\ &= \sum_{i=1}^{M-2} \tilde{l}_i^* \tilde{p}_i + (\tilde{l}_{M-1} - 1)(\tilde{p}_{M-1} + \tilde{p}_M) + (\tilde{p}_{M-1} + \tilde{p}_M) && (3.10) \\ &= \bar{L}(\mathcal{T}', (\tilde{p}_1, \ldots, \tilde{p}_{M-2}, \tilde{p}_{M-1} + \tilde{p}_M)) + (\tilde{p}_{M-1} + \tilde{p}_M) && (3.11) \end{aligned}$$

Since $\mathcal{T}^*$ is an optimal tree with respect to $(\tilde{p}_1, \ldots, \tilde{p}_M)$, $\mathcal{T}'$ must also be an optimal tree with respect to $(\tilde{p}_1, \ldots, \tilde{p}_{M-2}, \tilde{p}_{M-1} + \tilde{p}_M)$. 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., $X^n \sim P^n$ 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
> $$\lim_{n\to\infty} \frac{1}{n} \bar{L}^*(P^n) = \lim_{n\to\infty} \frac{1}{n} \bar{L}_{\mathrm{UD}}(P^n) = H(P).$$

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

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

since $H(P^n) = nH(P)$.

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

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

where $\delta_T(\varepsilon, n, \mathrm{P})$ is defined in (2.10) 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

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

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

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

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.

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

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

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

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 $1_1, l_2, \ldots$ (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) $\bar{L} = \min_i l_i$