

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, \dots, x_n)
$ \mathcal{X} $	the size (cardinality) of the set \mathcal{X}
$\text{Bern}(\lambda)$	Bernoulli (binary) random variable taking 1 with probability λ and 0 with probability $1 - \lambda$
$H_2(\lambda)$	entropy of $\text{Bern}(\lambda)$
$:=$	definition
Italic bold letters	Deterministic matrix \mathbf{M} / vector \mathbf{v}
Non-italic capital (bold) letters	Random variables X / Random vectors \mathbf{X}
$P(\cdot)$	Probability measure
$\mathbb{E}\{X\}$	Mean of the random variable X
$I(X; Y)$	Mutual information between X and Y
$\delta[\cdot]$	Kronecker delta function
$\mathbf{1}\{\cdot\}$	indicator function
$\log(x)$	Base-2 logarithm of x
\mathbf{I}	Identity matrix
$\ \mathbf{v}\ $	Euclidean (\mathcal{L}_2) norm of \mathbf{v}

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 (Ω, \mathcal{H}, 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 $\Omega \rightarrow E$ 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 $\mu(A) := P(X(\omega) \in A)$, which is a probability measure on (E, \mathcal{E}) . For discrete random variables, we call $P_X(x) := P(X(\omega) = x)$, $x \in E$, 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 $\frac{d\nu}{d\mu}$, such that

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

Note that the Radon-Nikodym derivative is positive and measurable, i.e., in \mathcal{E}_+ . If $\nu \ll \mu \ll \lambda$, we have $\frac{d\nu}{d\mu} = \frac{d\nu}{d\lambda} \frac{d\mu}{d\lambda}$. If $P \ll Q$ 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 $\mathbb{E}_P(f(X)) = \mathbb{E}_Q(f(X) \frac{dP}{dQ}(X))$.

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 $p_X(x) := \frac{d\mu}{d\lambda}(x)$ 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 $\pi f = \int_E \mu(dx) \int_F K(x, dy) f(x, y)$ defines the unique probability measure satisfying $\pi(A \times B) = \int_A \mu(dx) K(x, B)$ 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 probability measure μ on E and a transition probability kernel K from (E, \mathcal{E}) to (F, \mathcal{F}) such that $\int_{E \times F} \pi(dx \times dy) f(x, y) = \int_E \mu(dx) \int_F K(x, dy) f(x, y)$, 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 $P_{Y|X}$ to denote the transition probability kernel such that $P_{Y|X=x}(dy) = K(x, dy)$. We use $P_X P_{Y|X}$ to denote the measure π on the product space such that $\int_{E \times F} \pi(dx \times dy) f(x, y) = \int_E P_X(dx) \int_F P_{Y|X=x}(dy) f(x, y)$. Both notations $P_{Y|X}(dy|x)$ and $P_{Y|X=x}(dy)$ are equivalent and can be used interchangeably. In particular, in the discrete case, $P_{Y|X}(y|x) = P_{Y|X=x}(y)$.

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

$$(P_{Y|X} \circ P_X)(A) = \int_E P_X(dx) P_{Y|X=x}(A).$$

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 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

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

which is quite easy to visualize. A real function is called **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 $\mathbb{E}f(X) \geq f(\mathbb{E}X)$. For concave functions, we have $\mathbb{E}f(X) \leq f(\mathbb{E}X)$.

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

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

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 $P_{X_1 \dots X_n}$ (or, in short, P_{X^n}).

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

We also define the **conditional entropy** as

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

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

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

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

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

Similarly, the conditional version also holds.

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

1.2 Cross-entropy

Let us consider two probability measures P and Q defined on the same measurable space. Define the **cross-entropy** of P with respect to Q as¹

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

And the **conditional cross-entropy**

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

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

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

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

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

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

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

And the conditional version also holds.

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

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 $X \sim P_X$ with pdf p as

$$h(X) \equiv h(P_X) := \mathbb{E}_{P_X} \left[\log \frac{1}{p(X)} \right] = \int dx p(x) \log \frac{1}{p(x)}.$$

Similarly, we define the conditional differential entropy

$$h(X | Y) \equiv h(P_{X|Y} | P_Y) := \mathbb{E}_{Y \sim P_Y} [h(P_{X|Y=Y})] = \mathbb{E}_{P_Y P_{X|Y}} \left[\log \frac{1}{p(X|Y)} \right].$$

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

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

The differential cross-entropy $h(P \| Q)$ and the conditional version $h(P_{X|Y} \| Q_{X|Y} | P_Y)$ 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 $D(P \| Q)$, is defined as

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

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

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

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

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

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 arbitrary probability transition kernels $P_{X|Y}$ and $Q_{X|Y}$,

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

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

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

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

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

and the conditional version

$$D(P_{X^n|Y} \| Q_{X^n|Y} \mid P_Y) = \sum_{i=1}^n D(P_{X_i|X^{i-1}Y} \| Q_{X_i|X^{i-1}Y} \mid P_{X^{i-1}Y})$$

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

$$H(P_{X|Y} \| Q_{X|Y} \mid P_Y) = H(P_{X|Y} \mid P_Y) + D(P_{X|Y} \| Q_{X|Y} \mid P_Y).$$

And in particular,

$$D(P\|Q) = H(P\|Q) - H(P).$$

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 $P_{X|Y}$ and $P_{Y|X}$ be the transition probability kernels such that $P_{XY} = P_X P_{Y|X} = P_Y P_{X|Y}$.²

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

$$I(X; Y) := D(P_{XY} \| P_X P_Y) = D(P_{X|Y} \| P_X \mid P_Y) = D(P_{Y|X} \| P_Y \mid P_X),$$

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 $P_{Y|X}$ and $P_{X|Y}$ to the marginals P_Y and P_X . Sometimes mutual information is also denoted by $I(P_X, P_{Y|X})$, $I(P_Y, P_{X|Y})$, and $I(P_{XY})$

$$I(X; Y) \equiv I(P_{XY}) \equiv I(P_X, P_{Y|X}) \equiv I(P_Y, P_{X|Y}).$$

²We 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

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

It follows that

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (1.1)$$

$$= H(X) - H(X|Y) \quad (1.2)$$

$$= H(Y) - H(Y|X) \quad (1.3)$$

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

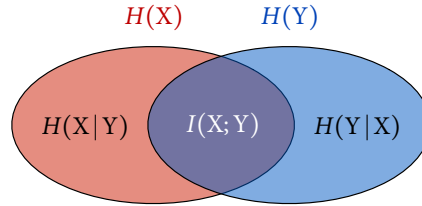


Figure 1.1: The Venn diagram.

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

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

It follows that

$$I(X; Y) = h(X) + h(Y) - h(X, Y) \quad (1.4)$$

$$= h(X) - h(X|Y) \quad (1.5)$$

$$= h(Y) - h(Y|X), \quad (1.6)$$

similar to the discrete case.

If X is discrete and Y continuous, then

$$I(X; Y) = D(P_{XY} \| P_X P_Y) \quad (1.7)$$

$$= D(P_{Y|X} \| P_Y | P_X) \quad (1.8)$$

$$= h(Y) - h(Y|X). \quad (1.9)$$

We also have

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

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 (X, Y, Z) . Then, we can define the **conditional mutual information** between X and Y given Z .

$$I(X; Y | Z) \equiv I(P_{XY|Z} | P_Z) := D(P_{XY|Z} \| P_{X|Z} P_{Y|Z} | P_Z) = D(P_{X|YZ} \| P_{X|Z} | P_{YZ}) = D(P_{Y|XZ} \| P_{Y|Z} | P_{XZ}),$$

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

The chain rule of mutual information is

$$I(X; Y^n) = \sum_{i=1}^n I(X; Y_i | Y^{i-1})$$

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

$$\begin{aligned} I(X; Y^n) &= D(P_{Y^n|X} \| P_{Y^n} | P_X) \\ &= \sum_{i=1}^n D(P_{Y_i|XY^{i-1}} \| P_{Y_i|Y^{i-1}} | P_{XY^{i-1}}) \\ &= \sum_{i=1}^n I(X; Y_i | Y^{i-1}) \end{aligned}$$

The conditional version follows in the same way.

$$I(X; Y^n | Z) = \sum_{i=1}^n I(X; Y_i | Y^{i-1}Z)$$

1.6 Some properties of information measures

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

$$\begin{aligned} H(P_{X^n}) &= \sum_{i=1}^n H(P_{X_i|X^{i-1}} | P_{X^{i-1}}), \quad h(P_{X^n}) = \sum_{i=1}^n h(P_{X_i|X^{i-1}} | P_{X^{i-1}}) \\ H(P_{X^n} \| Q_{X^n}) &= \sum_{i=1}^n H(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}} | P_{X^{i-1}}), \quad h(P_{X^n} \| Q_{X^n}) = \sum_{i=1}^n h(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}} | P_{X^{i-1}}) \\ D(P_{X^n} \| Q_{X^n}) &= \sum_{i=1}^n D(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}} | P_{X^{i-1}}) \\ I(P_{X, Y^n}) &= \sum_{i=1}^n I(P_{X, Y_i|Y^{i-1}} | P_{Y^{i-1}}) \end{aligned}$$

- Positivity

$$\begin{aligned} H(P) &\geq 0, \quad h(P) \not\leq 0 \\ H(P \| Q) &\geq 0, \quad h(P \| Q) \not\leq 0 \\ D(P \| Q) &\geq 0 \implies H(P) \leq H(P \| Q), \quad h(P) \leq h(P \| Q) \\ I(P_X, P_{Y|X}) &\geq 0 \end{aligned}$$

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

The positivity of mutual information is from that of divergence. It is 0 if and only if $P_{XY} = P_X P_Y$, i.e., X and Y are independent. □

- Conditioning

- Conditioning reduces (differential) entropy

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

- Conditioning increases divergence

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

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., $H(X) - H(X|Y) = I(X;Y) \geq 0$. Similarly for differential entropy. For divergence, we have $D(P_{X|Y} \| Q_{X|Y} | P_Y) = D(P_{X|Y}P_Y \| Q_{X|Y}P_Y) = D(\tilde{P}_X \| \tilde{Q}_X) + D(\tilde{P}_{Y|X} \| \tilde{Q}_{Y|X} | \tilde{P}_X) \geq D(\tilde{P}_X \| \tilde{Q}_X)$, where we use the decompositions $P_{X|Y}P_Y = \tilde{P}_X\tilde{P}_{Y|X}$ and $Q_{X|Y}P_Y = \tilde{Q}_X\tilde{Q}_{Y|X}$. \square

- 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 $P_X = (1 - \lambda)P_0 + \lambda P_1$. $\lambda H(P_1) + (1 - \lambda)H(P_0) = H(X|S) \leq H(X) = H(P_X)$, 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, $\lambda D(P_1 \| Q_1) + (1 - \lambda)D(P_0 \| Q_0) = (P_{X|S} \| Q_{X|S} | P_S) \geq D(P_X \| Q_X)$, 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

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

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

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

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

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

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

$$\begin{aligned} I(P_X, \lambda P_{Y|X}^1 + (1 - \lambda)P_{Y|X}^0) &= D(\lambda P_X P_{Y|X}^1 + (1 - \lambda)P_X P_{Y|X}^0 \| \lambda P_X P_Y^1 + (1 - \lambda)P_X P_Y^0) \\ &\leq \lambda D(P_X P_{Y|X}^1 \| P_X P_Y^1) + (1 - \lambda)D(P_X P_{Y|X}^0 \| P_X P_Y^0) \\ &= \lambda I(P_X, P_{Y|X}^1) + (1 - \lambda)I(P_X, P_{Y|X}^0) \end{aligned}$$

where P_Y^0 and P_Y^1 are the marginals of $P_X P_{Y|X}^0$ and $P_X P_{Y|X}^1$, respectively.

The second way is to introduce the same S as before, let $P_{Y|X,S=0} = P_{Y|X}^0$ and $P_{Y|X,S=1} = P_{Y|X}^1$, 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 $P_{Y|X} = \lambda P_{Y|X}^0 + (1 - \lambda)P_{Y|X}^1$. Therefore, we have $I(P_X, \lambda P_{Y|X}^1 + (1 - \lambda)P_{Y|X}^0) = I(X; Y)$. We also have $\lambda I(P_X, P_{Y|X}^1) + (1 - \lambda)I(P_X, P_{Y|X}^0) = I(X; Y | S)$. Therefore, it is enough to prove that $I(X; Y | S) \geq I(X; Y)$. To that end, apply the independence so that $I(X; S) = 0$, and thus

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

□

- Data processing inequality (DPI)

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

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 $X \rightarrow Y \rightarrow Z$, we have

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

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

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

Obviously, the conditional version also holds similarly.

For the mutual information, since $X \rightarrow Y \rightarrow Z$, 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,

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

□

1.7 Maximum entropy

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

1.7.1 Finite alphabet

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

$$\begin{aligned} H(P_X) &\leq H(P_X\|Q) \\ &= \mathbb{E}_{P_X} [\log M] \\ &= \log M, \end{aligned}$$

where the equality holds when $P_X = Q$. Therefore, we show that uniform distribution maximizes entropy among all distributions with bounded alphabet size.

1.7.2 Continuous alphabet, finite second moment

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

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

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

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 $\mathbb{E}[c(X)] \leq P$ 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 $\lambda \geq 0$ is such that $\mathbb{E}_Q[c(X)] = P$.

Then, we have

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

where the equalities holds when $P_X = Q$.

Exercises³

- 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 $H(X)$ and $H(Y)$ if
 - $Y = 2^X$?
 - $Y = \cos(X)$?
- Conditional mutual information vs. unconditional mutual information [CT 2.6]. Give examples of joint random variables X , Y , and Z such that
 - $I(X; Y | Z) < I(X; Y)$.
 - $I(X; Y | Z) > I(X; Y)$.

- Data processing [CT 2.15]. Let $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots \rightarrow X_n$ form a Markov chain in this order; that is, let

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

Reduce $I(X_1; X_2, \dots, X_n)$ to its simplest form.

- 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 $P(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \dots$ has $H(X) = +\infty$.
- Inequalities [CT 2.29]. Let X , Y , and Z be joint random variables. Prove the following inequalities and find conditions for equality.
 - $H(X, Y | Z) \geq H(X | Z)$.
 - $I(X, Y; Z) \geq I(X; Z)$.
 - $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.
 - $I(X; Z | Y) \geq I(Z; Y | X) - I(Z; Y) + I(X; Z)$.

- Convexity/Concavity of mutual information.

- Let $(S, X, Y) \sim P_{SXY} = P_S P_{X|S} P_{Y|X}$, i.e., $S \rightarrow X \rightarrow Y$ forms a Markov chain. Show that

$$I(X; Y) \geq I(X; Y | S).$$

Use the above inequality to show that mutual information is concave in P_X for a fixed $P_{Y|X}$.

- Let $(S, X, Y) \sim P_{SXY} = P_S P_X P_{Y|X, S}$. Show that

$$I(X; Y) \leq I(X; Y | S).$$

Use the above inequality to show that mutual information is convex in $P_{Y|X}$ for a fixed P_X .

- Maximum entropy. [CT 2.30] Find the probability mass function $P(x)$ that maximizes the entropy $H(X)$ of a nonnegative integer-valued random variable X subject to the constraint

$$E(X) = \sum_{n=0}^{\infty} nP(n) = A$$

for a fixed value $A > 0$. Evaluate this maximum $H(X)$.

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

³The 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 = $y \in \{0,1\}$):

$$P_{XY} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} \end{bmatrix}, \quad Q_{XY} = \begin{bmatrix} \frac{1}{8} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{8} \end{bmatrix}.$$

- Compute the marginals P_X, P_Y and Q_X, Q_Y .
- Compute the conditional kernels $P_{X|Y}$ and $Q_{X|Y}$.
- Compute the entropies (in bits): $H(P_X), H(P_Y), H(P_{XY}), H(P_{X|Y} | P_Y)$; and the corresponding quantities under Q .
- Compute the divergences: $D(P_{XY} \| Q_{XY}), D(P_X \| Q_X), D(P_Y \| Q_Y)$, and the conditional divergence $D(P_{X|Y} \| Q_{X|Y} | P_Y)$. Verify the chain rule

$$D(P_{XY} \| Q_{XY}) = D(P_Y \| Q_Y) + D(P_{X|Y} \| Q_{X|Y} | P_Y).$$

- Compute the cross-entropies $H(P_X \| Q_X), H(P_Y \| Q_Y), H(Q_X \| P_X), H(Q_Y \| P_Y)$ and the conditional cross-entropy $H(P_{X|Y} \| Q_{X|Y} | P_Y)$. Verify the identities

$$H(P \| Q) = H(P) + D(P \| Q), \quad H(P_{X|Y} \| Q_{X|Y} | P_Y) = H(P_{X|Y} | P_Y) + D(P_{X|Y} \| Q_{X|Y} | P_Y).$$

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

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

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

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

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

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

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

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

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

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

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

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

Check your result for $N = 3$.

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

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

for any $P_{Y^n|X^n}$, with equality if and only if $P_{X^n|Y^n} = \prod_i P_{X_i|Y_i}$ (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 $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$. Show that in this case we have the Markov chain $Y_i \rightarrow X_i \rightarrow (\{X_j, j \neq i\}, \{Y_j, j \neq i\})$. Show that

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

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

Quiz (unique correct answer)

- For a discrete random variable X taking on n possible values, which of the following statements is **TRUE** regarding its Shannon entropy $H(X)$?
 - $H(X)$ is maximized when X is deterministic.
 - $H(X)$ is minimized when X follows a uniform distribution.
 - $H(X)$ is always non-negative and less than or equal to $\log_2 n$.
 - $H(X)$ can be negative if X takes negative values.
 - $H(X)$ measures the variance of X .
- Which of the following expressions correctly represents the mutual information $I(X; Y)$ between two discrete random variables X and Y ?
 - $I(X; Y) = H(X, Y) - H(X) - H(Y)$
 - $I(X; Y) = H(X) + H(Y) - H(X, Y)$
 - $I(X; Y) = H(X|Y) - H(X)$
 - $I(X; Y) = H(X, Y) + H(X|Y)$
 - $I(X; Y) = H(X|Y) + H(Y|X)$
- Suppose X and Y are independent discrete random variables. Which of the following is **TRUE**?
 - $H(X|Y) = H(X)$
 - $H(X|Y) = 0$
 - $H(X, Y) = H(X)$
 - $I(X; Y) = H(X)$
 - $I(X; Y) = H(Y)$
- Which of the following is **TRUE** about the Kullback-Leibler divergence $D(P\|Q)$ between two discrete probability distributions P and Q ?
 - $D(P\|Q)$ is symmetric in P and Q .
 - $D(P\|Q) \geq 0$, and equals zero if and only if $P = Q$ almost everywhere.
 - $D(P\|Q)$ is always finite.
 - $D(P\|Q)$ measures the variance between P and Q .
 - $D(P\|Q)$ is the mutual information between $X \sim P$ and $Y \sim Q$.
- Which of the following information measures can be negative?
 - Shannon entropy $H(X)$
 - Mutual information $I(X; Y)$
 - Conditional entropy $H(X|Y)$
 - Differential entropy $h(X)$ of a continuous random variable X
 - Kullback-Leibler divergence $D(P\|Q)$
- For a continuous random variable X with probability density function $f(x)$, which of the following statements about the differential entropy $h(X)$ is **TRUE**?
 - $h(X)$ is always non-negative.
 - $h(X)$ is invariant under scaling of X .
 - $h(X)$ increases when X is scaled by a factor $a > 1$.

- D) $h(X)$ cannot be less than zero.
- E) $h(X)$ measures the variance of X .
7. The entropy $H(X)$ of a Bernoulli random variable X with parameter p (i.e., $P(X = 1) = p$) is given by:
- A) $H(X) = -p \log p$
- B) $H(X) = -p \log p - (1 - p) \log(1 - p)$
- C) $H(X) = p \log(1 - p) + (1 - p) \log p$
- D) $H(X) = -\log p$
- E) $H(X) = p$
8. Which of the following inequalities relates the conditional entropy $H(X|Y)$ and the entropy $H(X)$ of two discrete random variables X and Y ?
- A) $H(X|Y) \geq H(X)$
- B) $H(X|Y) \leq H(X)$
- C) $H(X|Y) = H(X)$
- D) $H(X|Y) = H(X) + H(Y)$
- E) $H(X|Y) = H(X, Y) - H(Y)$
9. The chain rule for entropy states that for discrete random variables X and Y :
- A) $H(X, Y) = H(X|Y) + H(Y)$
- B) $H(X, Y) = H(X) + H(Y)$
- C) $H(X, Y) = H(Y|X) - H(X)$
- D) $H(X, Y) = H(X|Y) - H(Y)$
- E) $H(X, Y) = H(X) - H(Y|X)$
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 $I(X; Y|Z)$ can be expressed in terms of entropies as:
- A) $I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z)$
- B) $I(X; Y|Z) = H(X, Y, Z) - H(Z)$
- C) $I(X; Y|Z) = H(X|Y, Z) - H(X|Z)$
- D) $I(X; Y|Z) = H(X, Z) + H(Y, Z) - H(Z)$
- E) $I(X; Y|Z) = H(X, Y) - H(Z)$
12. For a discrete random variable X and function $f(X)$, which of the following is **TRUE** regarding the entropy $H(f(X))$?
- A) $H(f(X)) \geq H(X)$
- B) $H(f(X)) \leq H(X)$
- C) $H(f(X)) = H(X)$

- 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 $X \rightarrow Y \rightarrow Z$, which of the following is **TRUE**?
- A) $I(X; Z) \geq I(X; Y)$
 B) $I(X; Z) \leq I(X; Y)$
 C) $I(X; Y) = I(Y; Z)$
 D) $I(X; Z) = I(X; Y) + I(Y; Z)$
 E) $I(X; Z) \geq I(Y; Z)$
14. Which distribution maximizes the entropy for a discrete random variable X with a fixed mean over the support $\{1, 2, \dots, n\}$?
- A) Uniform distribution over $\{1, 2, \dots, n\}$
 B) Geometric distribution
 C) Binomial distribution
 D) Discrete exponential distribution
 E) Poisson distribution
15. Define the cross-entropy $H(P, Q) := \mathbb{E}_P \log \frac{1}{Q(X)}$. Which of the following is a property of the cross-entropy $H(P, Q)$ between two probability distributions P and Q ?
- A) $H(P, Q) = H(Q, P)$
 B) $H(P, Q) \geq H(P)$
 C) $H(P, Q) \leq H(P)$
 D) $H(P, Q) = H(P) + D(P\|Q)$
 E) $H(P, Q) = D(P\|Q)$
16. For two independent continuous random variables X and Y , the differential entropy of their sum $Z = X + Y$ satisfies:
- A) $h(Z) = h(X) + h(Y)$
 B) $h(Z) = h(X) - h(Y)$
 C) $h(Z) = h(X) + h(Y) + \log 2\pi e$
 D) $h(Z) \leq h(X) + h(Y)$
 E) $h(Z) \geq h(X) + h(Y)$
17. Which of the following statements about mutual information $I(X; Y)$ is **TRUE**?
- A) Mutual information $I(X; Y)$ is always less than or equal to zero.
 B) Mutual information $I(X; Y)$ is zero if and only if X and Y are independent.
 C) Mutual information $I(X; Y)$ is the same as conditional entropy $H(X|Y)$.
 D) Mutual information $I(X; Y)$ is maximized when X and Y are independent.
 E) Mutual information $I(X; Y)$ is always greater than the joint entropy $H(X, Y)$.
18. The conditional entropy $H(Y|X)$ can be expressed in terms of joint entropy $H(X, Y)$ and marginal entropy $H(X)$ as:
- A) $H(Y|X) = H(X, Y) - H(X)$
 B) $H(Y|X) = H(X) - H(X, Y)$

- C) $H(Y|X) = H(Y) - H(X)$
 - D) $H(Y|X) = H(X, Y) + H(X)$
 - E) $H(Y|X) = H(Y) + H(X, Y)$
19. For a continuous random variable X uniformly distributed over the interval $[a, b]$, the differential entropy $h(X)$ is:
- A) $h(X) = \log(b - a)$
 - B) $h(X) = \log(b + a)$
 - C) $h(X) = \frac{1}{2} \log(b - a)$
 - D) $h(X) = -\log(b - a)$
 - E) $h(X) = \log\left(\frac{b}{a}\right)$
20. The **Chain Rule** for mutual information states that for random variables X, Y, Z :
- A) $I(X; Y, Z) = I(X; Y) + I(X; Z)$
 - B) $I(X; Y, Z) = I(X; Y|Z) + I(X; Z)$
 - C) $I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$
 - D) $I(X; Y, Z) = I(X; Y) - I(X; Z|Y)$
 - E) $I(X; Y, Z) = I(X; Y|Z) - I(X; Z|Y)$

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.,

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

Thus, $\hat{P}_{x^n} := [\hat{P}_{x^n}(a) : a \in \mathcal{X}]$ satisfies all the properties of a pmf. Similarly, we can define the **joint type** \hat{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.,

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

We can verify that

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

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

We say that \hat{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 $(n_1, \dots, n_M) \in \mathbb{Z}_+^M$ with $n_1 + \dots + n_M = n$ and $\hat{P} = [\frac{n_1}{n}, \dots, \frac{n_M}{n}]$. The **set of types** in \mathcal{X}^n is denoted by $\mathcal{P}_n^{\mathcal{X}}$ (or simply \mathcal{P}_n). Specifically, we define

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

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

$$\mathcal{P}^{\mathcal{X}} := \{[p_1, \dots, p_M] \in \mathbb{R}_+^M : p_1 + \dots + p_M = 1\}$$

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

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

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

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

since for each of the $M = |\mathcal{X}|$ components in a type $\hat{P} \in \mathcal{P}_n$ 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

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

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

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

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

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

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

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

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

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

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

where P is a pmf⁴ defined in \mathcal{X} . In words, $P^n(x^n)$ is the probability of the sequence x^n under the assumption that x_1, \dots, x_n are i.i.d. $\sim P$. Such a notation makes the distribution of the random variables involved explicit.

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

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

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

Or, we can concisely write

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

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 $Q = \hat{P}$, we have from (2.2)

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

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

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

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

$$\hat{P}^n(x^n) \not\geq \hat{P}^n(\tilde{x}^n).$$

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 $Q^n(\mathcal{T}(\hat{P}))$ be the probability of the set of all the sequences in the type class \hat{P} under the measure Q^n , i.e.,

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

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

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

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

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

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{P}^n , the type class $\mathcal{T}(\hat{P})$ has a higher probability than any other type classes.

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

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

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

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

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

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

For brevity, we can write^a

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

^awhere \doteq means equality in exponent in the large n regime, i.e., $f(n) \doteq g(n)$ means

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

Proof. The upper bound is straightforward from (2.3)

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

The lower bound can be obtained using (2.6):

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

□

The main message is that the size of a type class of \hat{P} is roughly $2^{nH(\hat{P})}$, 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 $\mathcal{T}(\hat{P})$ under the distribution Q^n .

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

implying

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

In particular, we have

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

Intuitively, the probability of generating a sequence with a mismatched type $Q \neq \hat{P}$ decreases exponentially with n as $2^{-nD(\hat{P} \| Q)}$, while the probability of generating a sequence 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

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

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

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

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

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

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

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

Typical sequences have the following properties.

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

$$2^{-n(1+\varepsilon)H(P)} \leq P^n(x^n) \leq 2^{-n(1-\varepsilon)H(P)}, \quad (2.8)$$

and

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

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

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

More specifically, we have

$$P^n(X^n \notin \mathcal{T}_\varepsilon^{(n)}(P)) \leq \delta_T(\varepsilon, n, P) := 2M^* e^{-2n\varepsilon^2 c_P} \quad (2.10)$$

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

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

$$2^{n(1+\varepsilon)H(P)} \geq |\mathcal{T}_\varepsilon^{(n)}(P)| \geq 2^{n(1-\varepsilon)H(P) - \delta'_T(\varepsilon, n, P)},$$

where

$$\delta'_T(\varepsilon, n, P) := -\log(1 - \delta_T(\varepsilon, n, P))^+ \quad (2.11)$$

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

Proof. 1. The bounds (2.8) are straightforward from the typical average lemma where we let $g(x) = \log P(x)$. And we apply the lemma on $\log P^n(x^n) = \sum_i g(x_i)$. Note that $\mathbb{E}|\log P(X)| = -\mathbb{E} \log P(X) = H(P)$. The divergence bound is from the fact that $\hat{P} \ll P$ and that $\log \frac{\hat{P}(a)}{P(a)} \leq \log(1 + \varepsilon)$. Taking the expectation, we obtain the upper bound. Finally, for the entropy bounds, we write

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

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

2. We check the probability

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

$$\leq \sum_{a \in \mathcal{X}} P^n\left(\left\{ \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i = a\} - P(a) \right| > \varepsilon P(a) \right\}\right) \quad (2.13)$$

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

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

where in (2.14), we use the fact that the event has probability 0 when $P(a) = 0$; the last inequality is from Hoeffding's inequality.

Hoeffding's inequality Let X_1, \dots, X_n be independent random variables such that $X_i \in [a_i, b_i]$ almost surely. Define $S_n := \frac{1}{n} \sum_{i=1}^n X_i$ and $\Delta_n := \frac{1}{n} \sum_{i=1}^n (a_i - b_i)^2$. We have

$$\mathbb{P}\{|S_n - \mathbb{E}(S_n)| \geq t\} \leq 2e^{-\frac{2nt^2}{\Delta_n}}. \quad (2.16)$$

Specifically, we have $\mathbf{1}\{X_1 = a\}, \dots, \mathbf{1}\{X_n = a\}$ i.i.d., bounded in $[0, 1]$, and with expectation $P(a)$, so the average concentrates around the mean exponentially fast. Replacing each term in the sum by the dominant one, i.e., when $P(a)$ is the smallest, we get the upper bound (2.10). (2.9) follows directly.

3. The upper bound is straightforward from $1 \geq P^n(\mathcal{T}_\varepsilon^{(n)}(P)) \geq 2^{-n(1+\varepsilon)H(P)} |\mathcal{T}_\varepsilon^{(n)}(P)|$. From (2.10), $P^n(X^n \in \mathcal{T}_\varepsilon^{(n)}(P)) \geq (1 - \delta_T(\varepsilon, n, P))^+$. Since we also have $P^n(\mathcal{T}_\varepsilon^{(n)}(P)) \leq |\mathcal{T}_\varepsilon^{(n)}(P)| 2^{-n(1-\varepsilon)H(P)}$, we prove that $|\mathcal{T}_\varepsilon^{(n)}(P)| 2^{-n(1-\varepsilon)H(P)} \geq (1 - 2|\mathcal{X}|e^{-2n\varepsilon^2 c_P})^+$ 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 $\mathcal{T}_\varepsilon^{(n)}(P)$ with approximately $2^{nH(P)}$ sequences, out of the set of all $M^n = 2^{n \log M}$ sequences. Moreover, every sequence (generated by the source P^n) inside $\mathcal{T}_\varepsilon^{(n)}(P)$ has approximately the same probability $2^{-nH(P)}$. 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} .

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

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

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

By definition, we have

$$(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(P_{XY}) \iff x^n \in \mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n) \iff y^n \in \mathcal{T}_\varepsilon^{(n)}(P_{XY} | x^n)$$

We have the following properties.⁵

1. If $(x^n, y^n) \in \mathcal{T}_\varepsilon^{(n)}(P_{XY})$, then
 - $x^n \in \mathcal{T}_\varepsilon^{(n)}(P_X)$, $y^n \in \mathcal{T}_\varepsilon^{(n)}(P_Y)$
 - $P_X^n(x^n) \approx 2^{-n(1 \pm \varepsilon)H(X)}$, $P_Y^n(y^n) \approx 2^{-n(1 \pm \varepsilon)H(Y)}$, $P_{XY}^n(x^n, y^n) \approx 2^{-n(1 \pm \varepsilon)H(X, Y)}$
 - $P_{Y|X}^n(y^n | x^n) \approx 2^{-n(1 \pm \varepsilon)H(Y|X)}$, $P_{X|Y}^n(x^n | y^n) \approx 2^{-n(1 \pm \varepsilon)H(X|Y)}$

2. $|\mathcal{T}_\varepsilon^{(n)}(P_{XY})| \leq 2^{n(1+\varepsilon)H(P_{XY})}$, and $|\mathcal{T}_\varepsilon^{(n)}(P_{XY})| \geq 2^{n(1-\varepsilon)H(P_{XY}) - \delta(\varepsilon, n, P_{XY})}$

3. For every $y^n \in \mathcal{Y}^n$,

$$|\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)| \leq 2^{n(1+\varepsilon)H(X|Y)}.$$

4. **Conditional typicality lemma.** Let $P_{X^n|Y^n} = P_{X|Y}^n$. For every $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(P_Y)$,

$$\lim_{n \rightarrow \infty} P_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) = 1, \quad \text{if } \lim_{n \rightarrow \infty} n\varepsilon^2 = \infty.$$

Specifically, we have

$$P_{X^n|Y^n=y^n}(X^n \notin \mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) \leq 2|\mathcal{X}| |\mathcal{Y}| e^{-n\varepsilon^2 c_P/2},$$

where $c_P := \min_{a,b: P_{XY}(a,b) > 0} P_{XY}(a,b)^2$.

5. Let $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(P_Y)$. Then,

$$|\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)| \geq 2^{n(1-\varepsilon)H(X|Y) - \delta'_T(\varepsilon/2, n, P_{XY})}. \quad (2.17)$$

where δ'_T is defined in (2.11).

6. **Joint typicality lemma.**

- For every $y^n \in \mathcal{Y}^n$,

$$P_X^n(\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) \leq 2^{-n(I(X;Y) - 2\varepsilon H(X))}. \quad (2.18)$$

- Let $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}$. Then,

$$P_X^n(\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) \geq 2^{-n(I(X;Y) + 2\varepsilon H(X)) - \delta'_T(\varepsilon/2, n, P_{XY})}. \quad (2.19)$$

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

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 $g(x, y) = \log P_{X|Y}(x|y)$, using the fact that $\log(P_{X^n|Y^n}(x^n | y^n)) = \sum_{i=1}^n \log P_{X|Y}(x_i | y_i)$.

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

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

$$1 \geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)} P_{X|Y}^n(x^n | y^n) \quad (2.20)$$

$$\geq \sum_{x^n \in \mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)} 2^{-n(1+\varepsilon)H(X|Y)} \quad (2.21)$$

$$= |\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)| 2^{-n(1+\varepsilon)H(X|Y)}. \quad (2.22)$$

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

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

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

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

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

$$|\hat{P}_{X^n,y^n}(a, b) - P_{XY}(a, b)| \quad (2.25)$$

$$\leq |\hat{P}_{X^n,y^n}(a, b) - \hat{P}_{y^n}(b) P_{X|Y}(a|b)| + |\hat{P}_{y^n}(b) P_{X|Y}(a|b) - P_{XY}(a, b)| \quad (2.26)$$

$$\leq |\hat{P}_{X^n,y^n}(a, b) - \hat{P}_{y^n}(b) P_{X|Y}(a|b)| + \frac{\varepsilon}{2} P_{XY}(a, b), \quad (2.27)$$

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

$$P_{X^n|Y^n=y^n}(|\hat{P}_{X^n,y^n}(a, b) - P_{XY}(a, b)| > \varepsilon P_{XY}(a, b)) \quad (2.28)$$

$$\leq P_{X^n|Y^n=y^n}(|\hat{P}_{X^n,y^n}(a, b) - \hat{P}_{y^n}(b) P_{X|Y}(a|b)| > (\varepsilon - \varepsilon/2) P_{XY}(a, b)) \quad (2.29)$$

$$\leq 2e^{-n\varepsilon^2 P_{XY}(a, b)^2/2} \quad (2.30)$$

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

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

where $c_P := \min_{a,b: P_{XY}(a,b)>0} P_{XY}(a, b)^2$. Note that this bound goes to 0 when $n\varepsilon^2 \rightarrow \infty$. Hence, as long as $n\varepsilon^2 \rightarrow \infty$, we have the condition (2.23).

5. On the one hand, from the conditional typicality, we have $P_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) \geq (1 - 2|\mathcal{X}||\mathcal{Y}|e^{-n\varepsilon^2 c_P/2})^+$. On the other hand, we have $P_{X^n|Y^n=y^n}(\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)) \leq |\mathcal{T}_\varepsilon^{(n)}(P_{XY} | y^n)| 2^{-n(1-\varepsilon)H(X|Y)}$. 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 $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)| \leq 2^{n(1+\epsilon)H(X|Y)}$ and the upper bound on the probability of $\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_X)$ (implied by the fact that $\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)$) under \mathbf{P}_X^n , i.e., $\mathbf{P}_X^n(\mathbf{x}^n) \leq 2^{-n(1-\epsilon)H(X)}$. We have

$$\mathbf{P}_X^n(\mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)) = \sum_{\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)} \mathbf{P}_X^n(\mathbf{x}^n) \quad (2.32)$$

$$\leq \sum_{\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)} 2^{-n(1-\epsilon)H(X)} \quad (2.33)$$

$$\leq 2^{n(1+\epsilon)H(X|Y)} 2^{-n(1-\epsilon)H(X)} \quad (2.34)$$

$$= 2^{-n(I(X;Y) - \epsilon H(X) - \epsilon H(X|Y))} \quad (2.35)$$

$$\leq 2^{-n(I(X;Y) - 2\epsilon H(X))}. \quad (2.36)$$

To show the lower bound, apply the lower bound (2.17) on size of the set $\mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)$ and the lower bound on the probability of $\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_X)$ (implied by the fact that $\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)$) under \mathbf{P}_X^n . We have

$$\mathbf{P}_X^n(\mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)) = \sum_{\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)} \mathbf{P}_X^n(\mathbf{x}^n) \quad (2.37)$$

$$\geq \sum_{\mathbf{x}^n \in \mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY}|\mathbf{y}^n)} 2^{-n(1+\epsilon)H(X)} \quad (2.38)$$

$$\geq 2^{n(1-\epsilon)H(X|Y) - \delta(\epsilon/2, n, \mathbf{P}_{XY})} 2^{-n(1+\epsilon)H(X)} \quad (2.39)$$

$$\geq 2^{-n(I(X;Y) + 2\epsilon H(X) - \delta(\epsilon/2, n, \mathbf{P}_{XY}))}. \quad (2.40)$$

□

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

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

if the following conditions are satisfied:^a

$$\lim_{n \rightarrow \infty} n\epsilon^2 = \infty \quad (2.41)$$

$$\lim_{n \rightarrow \infty} \mathbb{P} \{ \mathbf{Y}^n \in \mathcal{T}_{\epsilon/2}(\mathbf{P}_Y) \} = 1 \quad (2.42)$$

$$\lim_{n \rightarrow \infty} nR_n - nI(X;Y) - 2n\epsilon H(X) = \infty. \quad (2.43)$$

^aWe don't require $\mathbf{Y}^n \sim \mathbf{P}_Y^n$. Instead, we only need that \mathbf{Y}^n is typical with respect to \mathbf{P}_Y with high probability.

Proof.

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

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

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 $y^n \in \mathcal{T}_{\varepsilon/2}^{(n)}(\mathbb{P}_Y)$,

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

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

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

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

Therefore, we have

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

From conditions (2.41), (2.42), and (2.43), we have $\delta'_T(\varepsilon/2, n, \mathbb{P}_{XY}) \rightarrow 0$, $\mathbb{P}_{Y^n}(\mathcal{T}_{\varepsilon/2}^{(n)}(\mathbb{P}_Y)) \rightarrow 1$, and $n(R_n - I(X;Y) - 2\varepsilon H(X)) \rightarrow \infty$. It follows that the above probability lower bound goes to 1. \square

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

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

if

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

Proof. It is enough to apply the union bound and the probability upper bound (2.18). Indeed, from the union bound, we have

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

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

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

where the last inequality is from the joint typicality lemma (2.18). The upper bound goes to 0 if

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

\square

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 $x^n, \tilde{x}^n \in \mathcal{T}(\hat{P})$, 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 $\mathcal{T}(\hat{P})$ is

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

which is also known as the multinomial coefficient

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

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

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

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

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

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

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

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{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

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

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

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

- Show that $\left| H(P) - E_Q \log \frac{1}{P(X)} \right| \leq \varepsilon H(P)$
- Show that $\left| E_Q \log \frac{1}{P(X)} - H(Q) \right| \leq \log \frac{1}{1-\varepsilon}$
- Show that $|H(Q) - H(P)| \leq \delta(\varepsilon)$ for some $\delta(\varepsilon) \rightarrow 0$ when $\varepsilon \rightarrow 0$. *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{P}_{x^n} of x^n , the second part indicates the exact sequence within the type class $\mathcal{T}(\hat{P}_{x^n})$.

- 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 divided by n , is $H(P)$.
7. Consider a binary source with $P = [0.2, 0.8]$. Let $n = 1000$ and $\varepsilon = 0.2$
- Provide an upper bound that the probability that $X^n \sim P^n$ is not inside typical set $\mathcal{T}_\varepsilon^{(n)}(P)$.
 - Provide an upper bound on the size of the typical set $\mathcal{T}_\varepsilon^{(n)}(P)$.

Quiz (unique correct answer)

1. The **type** \hat{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{P}_{x^n} is approximately:
 - A) $P^n(x^n) \approx 2^{-nH(\hat{P}_{x^n})}$
 - B) $P^n(x^n) \approx 2^{-n[H(\hat{P}_{x^n}) + D(\hat{P}_{x^n} \| P)]}$
 - C) $P^n(x^n) \approx 2^{-nD(\hat{P}_{x^n} \| P)}$
 - D) $P^n(x^n) \approx 2^{-nH(P)}$
 - E) $P^n(x^n) \approx 2^{-nD(P \| \hat{P}_{x^n})}$
6. The Asymptotic Equipartition Property (AEP) states that for a memoryless stationary source, the sequences of length n fall into two categories as $n \rightarrow \infty$:
 - 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 $\mathcal{T}_\epsilon^{(n)}(\mathbf{P})$ for a discrete memoryless source \mathbf{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 $\mathbf{P}^n(x^n) \geq 1 - \epsilon$.
 - C) The set of sequences x^n such that $\mathbf{P}^n(x^n) = 2^{-nH(\mathbf{X})}$.
 - 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 $\mathcal{T}_\epsilon^{(n)}(\mathbf{P})$ for a source \mathbf{P} satisfies:
- A) $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P})| \approx 2^{nH(\mathbf{P})}$
 - B) $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P})| \approx nH(\mathbf{P})$
 - C) $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P})| \approx n$
 - D) $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P})| \approx H(\mathbf{P})$
 - E) $|\mathcal{T}_\epsilon^{(n)}(\mathbf{P})| \approx 1$
9. The probability that a sequence drawn from a discrete memoryless stationary source lies in the typical set $\mathcal{T}_\epsilon^{(n)}(\mathbf{P})$ with a fixed $\epsilon > 0$ approaches what value as $n \rightarrow \infty$?
- A) 0
 - B) 1
 - C) ϵ
 - D) Depends on the source distribution
 - E) Cannot be determined
10. The jointly typical set $\mathcal{T}_\epsilon^{(n)}(\mathbf{P}_{XY})$ 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 \mathbf{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 \mathbf{Q} , i.e., x^n such that $\hat{\mathbf{P}}_{x^n} = \mathbf{Q}$, under the true distribution \mathbf{P} is:
- A) $\mathbf{P}^n(x^n) = 2^{-nH(\mathbf{P})}$
 - B) $\mathbf{P}^n(x^n) = 2^{-nD(\mathbf{Q}||\mathbf{P})}$
 - C) $\mathbf{P}^n(x^n) = 2^{-n[H(\mathbf{Q})+D(\mathbf{Q}||\mathbf{P})]}$
 - D) $\mathbf{P}^n(x^n) = 2^{-nH(\mathbf{Q})}$
 - E) $\mathbf{P}^n(x^n) = 2^{-nD(\mathbf{P}||\mathbf{Q})}$
12. The method of types provides an estimate for the probability of a type class $\mathcal{T}(\mathbf{Q})$ under distribution \mathbf{P} as:
- A) $\mathbf{P}^n(\mathcal{T}(\mathbf{Q})) \approx 2^{-nD(\mathbf{Q}||\mathbf{P})}$
 - B) $\mathbf{P}^n(\mathcal{T}(\mathbf{Q})) \approx 2^{-nH(\mathbf{Q})}$
 - C) $\mathbf{P}^n(\mathcal{T}(\mathbf{Q})) \approx 2^{-n[H(\mathbf{Q})+D(\mathbf{Q}||\mathbf{P})]}$
 - D) $\mathbf{P}^n(\mathcal{T}(\mathbf{Q})) \approx 2^{-nD(\mathbf{P}||\mathbf{Q})}$

- E) $P^n(\mathcal{T}(Q)) \approx 2^{-nH(P)}$
13. According to the method of types, the total number of possible types over an alphabet \mathcal{X} for sequences of length n is:
- A) Polynomial in n
 - B) Exponential in n
 - C) Logarithmic in n
 - D) Constant, independent of n
 - E) Depends on the actual distribution P
14. The divergence $D(\hat{P}_{x^n} \| P)$ between the type \hat{P}_{x^n} and the true distribution P is always:
- A) Non-negative and zero if and only if $\hat{P}_{x^n} = P$
 - B) Non-positive and zero if and only if $\hat{P}_{x^n} = P$
 - C) Non-negative and zero if and only if $\hat{P}_{x^n} \neq P$
 - D) Negative when $\hat{P}_{x^n} = P$
 - E) Equal to the entropy of \hat{P}_{x^n}
15. The method of types can be used to show that the probability of observing a sequence x^n whose type \hat{P}_{x^n} is significantly different from P is:
- A) High, due to randomness
 - B) Zero, as such sequences cannot occur
 - C) Exponentially small in n , decreasing with n
 - D) Independent of n
 - E) Equal to $D(\hat{P}_{x^n} \| P)$
16. The joint typicality lemma states that the probability that $(X^n, Y^n) \sim P_X^n P_Y^n$ is jointly typical with respect to P_{XY} is approximately:
- A) $2^{-n[I(X;Y)]}$
 - B) 1
 - C) 0
 - D) Equal to the product of their marginal probabilities
 - E) $2^{-nH(X,Y)}$
17. Which of the following statements is **TRUE** about the empirical distribution of a sequence?
- A) The empirical distribution is always identical to the true source distribution.
 - B) The empirical distribution converges to the true distribution as the sequence length increases.
 - C) The empirical distribution is defined only for sequences of infinite length.
 - D) The empirical distribution is the expected value of the random variable.
 - E) The empirical distribution is irrelevant in calculating the probability of sequences.
18. For a discrete memoryless stationary source, the probability that the type of a sequence deviates from its true distribution decreases exponentially with sequence length due to:
- A) The Weak Law of Large Numbers
 - B) The Central Limit Theorem
 - C) The Strong Law of Large Numbers
 - D) The Chebyshev Inequality

E) Hoeffding's Inequality

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

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

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

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