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

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 $\mu(A) \coloneqq \mathrm{P}(X(\omega) \in A)$, which is a probability measure on $(E, \mathcal{E})$. For discrete random variables, we call $\mathrm{P}_X(x) \coloneqq \mathrm{P}(X(\omega) = x)$, $x \in E$, 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 $p_X(x) \coloneqq \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)$ is called the **probability density function (pdf)**. We call the random variable X a **continuous random variable**.

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

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

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

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

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

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

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

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

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

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

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

1.1 Entropy

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

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

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

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

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

We also define the **conditional entropy** as

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

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

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

For discrete $\mathrm{X}^n \sim \mathrm{P}_{X^n}$, 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

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

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.

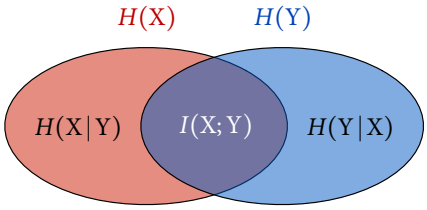

Figure 1.1: The Venn diagram.

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

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

- Positivity

$$\begin{aligned} H(\mathrm{P}) \ge 0, \quad h(\mathrm{P}) \not\ge 0 \\ H(\mathrm{P}\|\mathrm{Q}) \ge 0, \quad h(\mathrm{P}\|\mathrm{Q}) \not\ge 0 \\ D(\mathrm{P}\|\mathrm{Q}) \ge 0 \Longrightarrow H(\mathrm{P}) \le H(\mathrm{P}\|\mathrm{Q}),\ h(\mathrm{P}) \le h(\mathrm{P}\|\mathrm{Q}) \\ I(\mathrm{P}_X, \mathrm{P}_{Y|X}) \ge 0 \end{aligned}$$

  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 $D(\mathrm{P}\|\mathrm{Q}) \ge f(\mathbb{E}_\mathrm{Q}\frac{\mathrm{dP}}{\mathrm{dQ}}) = f(1) = 0$, 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

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

  - Conditioning increases divergence

$$D(\mathrm{P}_{X|Y} \| \mathrm{Q}_{X|Y} \mid \mathrm{P}_Y) \geq D(\tilde{\mathrm{P}}_X \| \tilde{\mathrm{Q}}_X)$$

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

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

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

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

Let $\mathrm{P}_{X|S=0} := \mathrm{P}_0$ and $\mathrm{P}_{X|S=1} := \mathrm{P}_1$. We have $\mathrm{P}_X = (1-\lambda)\mathrm{P}_0 + \lambda\mathrm{P}_1$. $\lambda H(\mathrm{P}_1) + (1-\lambda)H(\mathrm{P}_0) = H(\mathrm{X}|\mathrm{S}) \leq H(\mathrm{X}) = H(\mathrm{P}_X)$, proving the concavity of $\mathrm{P} \mapsto H(\mathrm{P})$ using conditioning reduces entropy.

For divergence, let $\mathrm{P}_{X|S=k} := \mathrm{P}_k$ and $\mathrm{Q}_{X|S=k} := \mathrm{Q}_k$ for $k = 0, 1$. Then, $\lambda D(\mathrm{P}_1 \| \mathrm{Q}_1) + (1-\lambda)D(\mathrm{P}_0 \| \mathrm{Q}_0) = (\mathrm{P}_{X|S} \| \mathrm{Q}_{X|S} \mid \mathrm{P}_S) \geq D(\mathrm{P}_X \| \mathrm{Q}_X)$, proving the convexity of divergence using conditioning increases divergence.

For the concavity of mutual information, for the given P_0, P_1, and $\mathrm{P}_{Y|X}$, let us set $\mathrm{P}_{X|S=0} := \mathrm{P}_0$ and $\mathrm{P}_{X|S=1} := \mathrm{P}_1$, and let $\mathrm{P}_{Y|XS} = \mathrm{P}_{Y|X}$, i.e., $\mathrm{P}_{Y|X,S=0} = \mathrm{P}_{Y|X,S=1} = \mathrm{P}_{Y|X}$. Hence, the joint distribution is $\mathrm{P}_{SXY} = \mathrm{P}_S\mathrm{P}_{X|S}\mathrm{P}_{Y|X}$. The conditional mutual information

$$\begin{aligned} I(\mathrm{X};\mathrm{Y} \mid \mathrm{S}) &= D(\mathrm{P}_{XY|S} \| \mathrm{P}_{X|S}\mathrm{P}_{Y|S} | \mathrm{P}_S) \\ &= \lambda D(\mathrm{P}_{Y|X,S=0} \| \mathrm{P}_{Y|S=0} \mid \mathrm{P}_{X|S=0}) + (1-\lambda) D(\mathrm{P}_{Y|X,S=1} \| \mathrm{P}_{Y|S=1} \mid \mathrm{P}_{X|S=1}) \\ &= \lambda D(\mathrm{P}_{Y|X} \| \mathrm{P}_{Y|S=0} \mid \mathrm{P}_0) + (1-\lambda) D(\mathrm{P}_{Y|X} \| \mathrm{P}_{Y|S=1} \mid \mathrm{P}_1) \\ &= \lambda I(\mathrm{P}_0, \mathrm{P}_{Y|X}) + (1-\lambda) I(\mathrm{P}_1, \mathrm{P}_{Y|X}). \end{aligned}$$

On the other hand,

$$I(\mathrm{X};\mathrm{Y}) = I(\mathrm{P}_X, \mathrm{P}_{Y|X}) = I(\lambda\mathrm{P}_0 + (1-\lambda)\mathrm{P}_1, \mathrm{P}_{Y|X})$$

To finish the proof, we write

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

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

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

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

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

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

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

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

□

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. 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.
10. 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)?
11. 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 $H(\mathrm{X}) < \infty$.
12. 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 $H(\mathrm{X}^{\mathrm{N}})$.
13. 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.
14. 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 $N = 3$.
15. 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}.
16. 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 P = 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) $H(f(\mathrm{X})) = 0$

E) $H(f(\mathrm{X})) = H(\mathrm{X}|f(\mathrm{X}))$

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

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

B) $I(\mathrm{X};\mathrm{Z}) \leq I(\mathrm{X};\mathrm{Y})$

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

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

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

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

A) $H(\mathrm{P},\mathrm{Q}) = H(\mathrm{Q},\mathrm{P})$

B) $H(\mathrm{P},\mathrm{Q}) \geq H(\mathrm{P})$

C) $H(\mathrm{P},\mathrm{Q}) \leq H(\mathrm{P})$

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

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

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

A) $h(\mathrm{Z}) = h(\mathrm{X}) + h(\mathrm{Y})$

B) $h(\mathrm{Z}) = h(\mathrm{X}) - h(\mathrm{Y})$

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

D) $h(\mathrm{Z}) \leq h(\mathrm{X}) + h(\mathrm{Y})$

E) $h(\mathrm{Z}) \geq h(\mathrm{X}) + h(\mathrm{Y})$

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

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

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

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

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

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

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

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

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

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