# Note on Information Theory

Summary of *Information Theory From Coding to Learning*
by Yury Polyanskiy and Yihong Wu

Anthony Hong

February 14, 2024

## Contents

**1 Some Notations in Probability** **2**

**2 Information Measures** **3**
2.1 Entropy . . . . . 3
2.2 Divergence . . . . . 6

**Abstract**

This note is a short summary of *Information Theory From Coding to Learning* by Yury Polyanskiy and Yihong Wu plus self-contained preliminaries.

# 1 Some Notations in Probability

We refer to [2] to have a basic setup in probability. First, we denote the collection of all probability measures on space $\mathcal{X}$ as $\triangle(\mathcal{X})$. For finite spaces we abbreviate $\triangle_k \equiv \triangle([k])$, a $(k-1)$-dimensional simplex.

Let $(\mathcal{X}, \mathcal{E})$ be a measurable space, $(\mathcal{Y}, \mathcal{F}, \nu)$ be a measure space, and $f : \mathcal{Y} \to \mathcal{X}$ be a measurable function. We call $f_{\#}\nu := \nu \circ f^{-1} : \mathcal{E} \to [0, \infty]$ the **image of** $\nu$ **under** $f$. It is easy to verify $\mu = f_{\#}\nu$ is indeed a measure and that for any $g \in \mathcal{E}_+$ (i.e. $g$ is non-negative $\mathcal{E}$-measurbale), one has $\int_{\mathcal{X}} g(x)\, \mu(dx) = \int_{\mathcal{Y}} (g \circ f)(y)\, \nu(dy)$.

Another measure built from construction is given by the following. Let $(\mathcal{X}, \mathcal{E}, \nu)$ be a measure space and $f \in \mathcal{E}_+$. Define $\mu : \mathcal{E} \to [0, \infty]; \mu(A) = \nu(f\mathbb{1}_A) = \int_A f(x)\, \nu(dx)$. One deduces from monotone convergence theorem that $(\mathcal{X}, \mathcal{E}, \mu)$ is now a new measure space. $\mu$ is called the **indefinite integral of** $f$ **with respect to** $\nu$. It is easy to check that for any $g \in \mathcal{E}_+$, one has $\int_{\mathcal{X}} g(x)\, \mu(dx) = \int_{\mathcal{X}} f(x) g(x)\, \nu(dx)$.

Recall that for measures $\mu$ and $\nu$ on a measurable space $(\mathcal{X}, \mathcal{E})$, $\mu$ is said to be **absolutely continuous with respect to** $\nu$ ($\mu \ll \nu$) if $\forall A \in \mathcal{E} : \nu(A) = 0 \Rightarrow \mu(A) = 0$. This is the case when for example $\mu$ is the indefinite integral of $f$ with respect to $\nu$. The Radon-Nikodym theorem gives the converse:

**Theorem 1.1** (Radon-Nikodym Theorem)**.** *Let $(\mathcal{X}, \mathcal{E}, \nu)$ be a $\sigma$-finite measure space and $\mu$ be a $\sigma$-finite measure on $\mathcal{E}$ with $\mu \ll \nu$. Then, there exists $f \in \mathcal{E}_+$ such that*

$$\int_{\mathcal{X}} g(x)\, \mu(dx) = \int_{\mathcal{X}} f(x) g(x)\, \nu(dx) \quad g \in \mathcal{E}_+$$

*or equivalently,*

$$\mu(A) = \int_A f(x)\, \nu(dx) \quad A \in \mathcal{E} \tag{1}$$

*The function $f$ is unique in the following sense: if $f'$ is another function with the above property then $f = f'$ $\nu - a.e.$. Besides, $f$ is denoted as $d\mu/d\nu$, called the* ***Radon-Nikodym derivative*** *of $\mu$ with respect to $\nu$.* □

We thus sometimes use $\mu(dy) = f(x)\nu(dx)$ to define measure $\mu$ for known $f \in \mathcal{E}_+$ and $\nu$.The theorem can be extended with the concept of signed measure.

Let $(\Omega, \mathcal{D}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{E})$ be a measurable space. A map $X : \Omega \to \mathcal{X}$ is called a **random variable** in $(\mathcal{X}, \mathcal{E})$ if it is measurable relative to $\mathcal{D}$ and $\mathcal{E}$, i.e., $X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$ is an event for every $A$ in $\mathcal{E}$. A **distribution** of $X$ is the probability measure $\mu = X_{\#}\mathbb{P}$ on $(\mathcal{X}, \mathcal{E})$ as the image of $\mathbb{P}$ under $X$, i.e., $A \in \mathcal{E} : \mu(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}\{X \in A\}$. Notice that the knowledge of how $\mu$ acts on a $\pi$-system (a collection of subsets of $\mathcal{X}$ that is closed under intersection) that generates $\mathcal{E}$ characterizes $\mu$.

**Example 1.2** (cdf & pdf)**.** When $\mathcal{X} = \dot{\mathbb{R}} = [-\infty, \infty]$ and $\mathcal{E} = \mathcal{B}_{\mathcal{X}}$, the Borel $\sigma$-algebra of $\mathcal{X}$, the intervals $[-\infty, x]$ with $x$ in $\mathbb{R}$ form a convenient $\pi$-system. It is thus enough to specify $\mu = X_{\#}\mathbb{P} = \mathbb{P} \circ X^{-1}$ by evulating function $c(x) = \mu[-\infty, x] = \mathbb{P}\{X \leq x\}$ for each $x \in \mathbb{R}$. And $c : \mathbb{R} \to [0, 1]$ is called the **cumulative distribution function (cdf)** of $X$. For the probability measure $\mu$ on $\dot{\mathbb{R}}$, we call the Radon-Nikodym derivative $d\mu/d\nu$ **probabilitiy density function (pdf)**, where $\nu =$ Leb. ◊

**Example 1.3** (pmf)**.** When the space $\mathcal{X}$ where $X$ takes values is now any countable set $I$, and the $\sigma$-algebra $\mathcal{E}$ is the power set $\mathcal{I} = \wp(I)$, we use the $\pi$-system $\{i \in I\}$ to specify the measure $\lambda = X_{\#}\mathbb{P}$:

$$\lambda_i := \lambda(i) = \mathbb{P}(X = i) = \mathbb{P}\{\omega \in \Omega : X(\omega) = i\}$$

Since $\lambda$ is a measure on $(I, \mathcal{I})$, we see $1 = \lambda(I) = \lambda\left(\bigcup_{i \in I} i\right) = \sum_{i \in I} \lambda_i$. Letting $\nu$ be the counting measure, then the Radon-Nikodym derivative $d\lambda/d\nu$ is called **probabilitiy mass function (pmf)**. Because of the fact that for each $i \in I$,

$$\lambda_i \overset{\text{eq.(1)}}{=\!=\!=} \int_{\{i\}} \frac{d\lambda}{d\nu}(x)\, \nu(dx) \overset{\nu \text{ counting mes}}{=\!=\!=\!=} \underbrace{|\{i\}|}_{=1} \frac{d\lambda}{d\nu}(i) = \frac{d\lambda}{d\nu}(i)$$

we see the distribution $\lambda$ is often referred as pmf itself. ◊

More things need to be added: convergence and dist. etc.

# 2 Information Measures

## 2.1 Entropy

R.A. Fisher was among the first to identify connection between information and "entropy" (or transformative content) in thermodynamics. The second law of thermodynamics states that hot things always cool unless you do something to stop them. Boltzmann and Gibbs gave a microscopic description: low temperatures are accompanied by molecular inactivity and order, and entropy enters saying how chaotic a system is.

The theory of **Information**, an abstract object that can be described as quantitatively representing changes of beliefs, formally started with Shannon's foundational work "A Mathematical Theory of Communication." We shall now examine some of his ideas.

**Definition 2.1** (Shannon Entropy)**.** Let $X$ be a discrete r.v. with pmf $P_X(x), x \in \mathcal{X}$. The **Shannon entropy** of $X$ is defined as the expectation of function $f(\cdot) = \log \frac{1}{P_X(\cdot)} : \mathcal{X} \to [0, \infty]$ of r.v. $X : \Omega \to \mathcal{X}$. Namely,

$$H(X) \equiv H(P_X) = \mathbb{E}[f(X)] = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{x \in \mathcal{X}} P_X(x) \frac{1}{\log P_X(x)}$$

where $0 \log \frac{1}{0} := 0$. The basis of logorithm determines the units of entropy: $\log_2 \leftrightarrow$ bits; $\log_{256} \leftrightarrow$ bytes. ♦

**Definition 2.2** (Joint Entropy)**.** The **joint entropy** of $n$ discrete r.v. $X^n := (X_1, \cdots, X_n)$ is

$$H(X^n) \equiv H(X_1, \cdots, X_n) = \mathbb{E}\left[\log \frac{1}{P_{X_1, \cdots, X_n}(X_1, \cdots, X_n)}\right]$$

Note that joint entropy is a special case of defn 2.1 applied to the r.v. $X^n$ taking values in product space. ♦

**Example 2.3** ($X \sim$ Unif$(\mathcal{X})$)**.** The entropy of $X$ is simply given by log-cardinality: $H(X) = \log |\mathcal{X}|$. ◊

**Example 2.4** ($X \sim$ Ber$(p)$)**.** Let $X \sim$ Ber$(p)$, with $P_X(1) = p$ and $P_X(0) = \bar{p} := 1 - p$. Then

$$H(X) = h(p) := p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}}$$

Here $h(\cdot)$ is called the **binary entropy function**, which is continuous, concave on $[0, 1]$, symmetric around $\frac{1}{2}$, and satisfies $h'(p) = \log \frac{\bar{p}}{p}$, with infinite slope at 0 and 1. The highest entropy is attainced at $p = \frac{1}{2}$ (i.e., $X \sim$ Unif$(\{0, 1\})$), while the lowest entropy is attained at $p = 0$ or 1 (i.e., deterministic). It is instructive to compare the plot of the binary entropy function with the variance $p(1 - p)$ (Figure 1). ◊

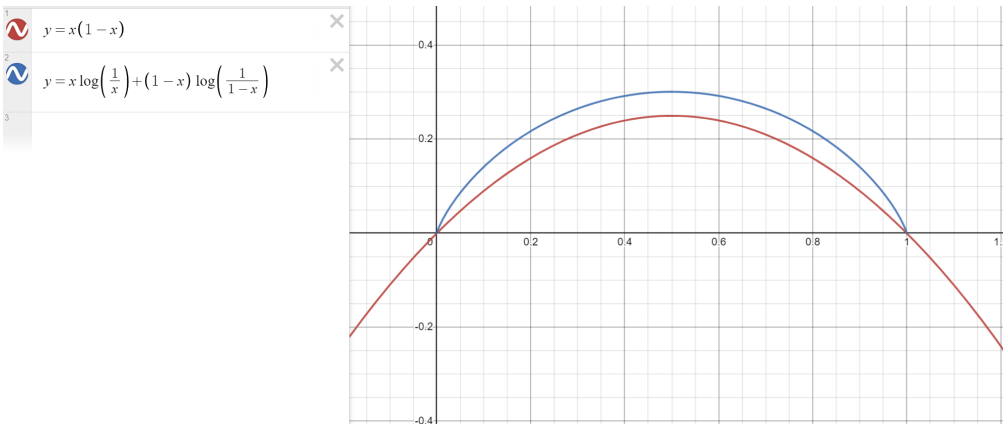


Figure 1: Binary entropy (blue) and variance (red)

**Definition 2.5** (Conditional Entropy)**.** Let $X$ be a discrete r.v. and $Y$ arbitrary. Denote by $P_{X|Y=y}(\cdot)$ or $P_{X|Y}(\cdot|y)$ the conditional distribution of $X$ given $Y = y$. The conditional entropy of $X$ given $Y$ is

$$H(X|Y) = \mathbb{E}_{y\sim P_Y}[H(P_{X|Y=y})] = \mathbb{E}_{y\sim P_Y}\left[\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log\frac{1}{P_{X|Y}(x|y)}\right] = \mathbb{E}\left[\log\frac{1}{P_{X|Y}(X|Y)}\right].$$

Note that $X|Y = y$ is a discrete r.v. ♦

Similary to entropy, conditional etropy measures the remaining randomness of a r.v. when another is revealed. As such, $Y \perp\!\!\!\perp X \Rightarrow H(X|Y) = H(X)$, because $P_{X|Y=y} = P_X \Rightarrow H(P_{X|Y=y}) = H(X)$ becomes a constant. However, when $Y$ depends on $X$, observing $Y$ does lower the entropy of $X$.

**Example 2.6** (Conditional entropy and noisy channel)**.** Let $Y$ be a noisy observation of $X \sim \text{Ber}(1/2)$ as follows. 1. $Y = X \oplus Z$, where $\oplus$ denotes binary addition (XOR) and $Z \sim \text{Ber}(\delta)$ independently of $X$.

**Review of Exclusive Or (XOR)**

Exclusive Or is a map $\oplus : \{0,1\}\times\{0,1\} \to \{0,1\}$ defined as

(1) $0 \oplus 0 = 0$
(2) $1 \oplus 0 = 1$
(3) $0 \oplus 1 = 1$
(4) $1 \oplus 1 = 0$

We observe that:
- for $y = x \oplus z$, $y$ agreees with $x$ if (1) or (3) happen; disagrees if (2) or (4) happen.
- $y = 0$ iff $x$ and $z$ agree; $y = 1$ iff $x$ and $z$ disagree.

By the first observation of above box, we see $Y$ agrees with $X$ with probability $\delta$ and disagrees with probability $\bar{\delta}$. The second observation shows that $P_{X|Y=0} = \text{Ber}(\delta)$ and $P_{X|Y=1} = \text{Ber}(\bar{\delta})$. Since $h(\delta) = h(\bar{\delta})$, $H(X|Y) = h(\delta)$. Note that when $\delta = 1/2$, $Y$ is independent of $X$ and $H(X|Y) = H(X) = 1$ `bits`; when $\delta = 0$ or 1, $X$ is completely determined by $Y$ and hence $H(X|Y) = 0$. ♢

We will need some facts in convexity to show several properties of entropy.

**Review of Convexity**

A subset $S$ of a vector space $V$ is called **convex** if $x, y \in S \Rightarrow tx + (1-t)y \in S$ for any $t \in [0,1]$. In particular, any vector space and its linear subspace are convex. Examples include $[0,1] \subset \mathbb{R}$, $S = \{\text{probabilitiy dist. on } \mathcal{X}\}$, $S = \{P_X : \mathbb{E}[X] = 0\}$.

Let $S \subseteq X$ be a convex set, then a function $f : X \to \mathbb{R}$ is **convex function** if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)\ \forall x, y \in S, t \in [0,1]$, and **strictly convex** if $f(tx+(1-t)y) < tf(x)+(1-t)f(y)\ \forall x \neq y \in S, t \in [0,1]$, and **(strictly) concave** if $-f$ is (strictly) convex. Examples include:
- $x \mapsto x\log x$ is strictly convex;
- $P \mapsto \int x\, dP$ is (nonstrictly) convex;
- variance is (nonstrictly) concave (use Jensen's inequality with $(\cdot)^2$'s convexity; consider zero-expectation dist.)

**Jensen Inequality**: For any $S$-valued r.v. $X$,
- $f$ is convex $\Rightarrow f(\mathbb{E}X) \leq \mathbb{E}f(X)$;
- $f$ is strictly convex $\Rightarrow f(\mathbb{E}X) < \mathbb{E}f(X)$, unless $X$ is a constant ($X = \mathbb{E}X$ a.s.)

**Proposition 2.7** (Properties of Entorpy)**.**

(a) *(Positivity)* $H(X) \geq 0$ *with equality iff* $X$ *is a constant (no randomness).*

(b) *(Uniform dist. maximizes entropy) For finite* $\mathcal{X}$, $H(X) \leq H(\textit{Unif}(\mathcal{X})) \overset{2.3}{=\!=} \log|\mathcal{X}|$ *with equality iff* $X \sim \textit{Unif}(\mathcal{X})$ *(i.e.* $\arg\max_{P_X \in \mathcal{P}(\mathcal{X})} H(P_X) = \textit{Unif}(\mathcal{X})$*)*

(c) *(Invariance under relabelling)* $H(X) = H(f(X))$ *for any bijective* $f$*.*

(d) *(Conditioning reduces entropy)* $H(X|Y) \leq H(X)$*, with equality iff* $X \perp\!\!\!\perp Y$*.*

(e) *(Simple chain rule)*
$$H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

(f) *(Entropy under deterministic transform)* $H(X) = H(X, f(X)) \geq H(f(X))$ *with equality iff* $f$ *is one-to-one on the support of* $P_X$*.*

(g) *(Full chain rule)*
$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \leq \sum_{i=1}^{n} H(X_i)$$

*Proof.* (a) Recall $f(\cdot)$ in defn 2.1 and thus observe that $\log\frac{1}{P_X(x)}$ is a nonnegative r.v. with a nonnegative expectation $H(X)$. $H(X) = 0$ iff $f(\cdot) = \log\frac{1}{P_X(\cdot)} = 0$ almost surely, namely, $P_X$ is a point mass.

(b) Apply Jensen's inequality to the strictly concave function $x \mapsto \log x$:
$$H(X) = \mathbb{E}\left[\log\frac{1}{P_X(X)}\right] \leq \log\mathbb{E}\left[\frac{1}{P_X(X)}\right] = \log\sum_{x\in\mathcal{X}} P_X(x)\frac{1}{P_X(x)} = \log|\mathcal{X}|$$

(c) Since $X$ is a discrete r.v. and $f$ is a bijective map, we see the pmf of $Y = f(X)$ is
$$P_Y(y) = \mathbb{P}[Y = y] = \mathbb{P}[f(X) = y] = \mathbb{P}\left[X = f^{-1}(y)\right] = P_X[f^{-1}(y)],$$
so
$$H(Y) = \sum_{y\in\mathcal{X}} P_Y(y)\log\frac{1}{P_Y(y)} = \sum_{y\in\mathcal{X}} P_X[\underbrace{f^{-1}(y)}_{=x}]\log\frac{1}{P_X[\underbrace{f^{-1}(y)}_{=x}]} = H(X).$$
Intuitively, the summation (expectation) goes through all locations on $\mathcal{X}$, the order of summation does not matter as long as each location appears exactly once, a property ensured by the bijectivity.

(d) Abbreviate $P_X(x)$ as $p(x)$, and similary for $p(y), p(x|y)$. By law of total probabilitiy,
$$p(x) = \mathbb{E}_Y[p(x|Y)] = \sum_{y\in\mathcal{Y}} p(y)p(x|y) \tag{2}$$
We apply Jensen's inequality to the strictly concave function $x \mapsto x\log\frac{1}{x}$,
$$\begin{aligned} H(X|Y) =& \mathbb{E}_{y\sim P_Y}\left[\sum_{x\in\mathcal{X}} p(x|y)\log\frac{1}{p(x|y)}\right] = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(y)p(x|y)\log\frac{1}{p(x|y)} = \sum_{x\in\mathcal{X}}\mathbb{E}_Y\left[p(x|Y)\log\frac{1}{p(x|Y)}\right] \\ \leq& \sum_{x\in\mathcal{X}}\mathbb{E}_Y[p(x|Y)]\log\frac{1}{\mathbb{E}_Y[p(x|Y)]} \overset{\text{eq.(2)}}{=\!=} \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} = H(X) \end{aligned}$$

(e) Telescoping $P_{X,Y}(X,Y) = P_{Y|X}(Y|X)P_X(X)$ and noting that both sides are positive $P_{X,Y}$-almost surely, we have

$$\mathbb{E}\left[\log\frac{1}{P_{X,Y}(X,Y)}\right] = \mathbb{E}\left[\log\frac{1}{P_X(X)P_{Y|X}(Y|X)}\right] = \underbrace{\mathbb{E}\left[\log\frac{1}{P_X(X)}\right]}_{H(X)} + \underbrace{\mathbb{E}\left[\log\frac{1}{P_{Y|X}(Y|X)}\right]}_{H(Y|X)}$$

(f) Use (c) and (e).

(g) Similary telescoping $P_{X_1X_2\cdots X_n} = P_{X_1}P_{X_2|X_1}\cdots P_{X_n|X^{n-1}}$

□

We end the section with Shannon's axiomatic characterization of entropy: denote a probabilitiy distribution on $m$ letters by $P = (p_1, \cdots, p_m)$ and consider a functional $H_m(p_1, \cdots, p_m)$. If $H_m$ obeys the following axioms:

(a) Permutation Invariance;

(b) Expansible: $H_m(p_1, \cdots, p_{m-1}, 0) = H_{m-1}(p_1, \cdots, p_{m-1})$;

(c) Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = \log 2$;

(d) Subadditivity: $H(X,Y) \leq H(X) + H(Y)$. Equivalently, $H_{mn}(r_{11}, \cdots, r_{mn}) \leq H_m(p_1, \cdots, p_m) + H_n(q_1, \cdots, q_n)$ whenever $\sum_{j=1}^n r_{ij} = p_i$ and $\sum_{i=1}^m r_{ij} = q_j$;

(e) Additivity: $H(X,Y) = H(X)+H(Y)$ if $X \perp\!\!\!\perp Y$. Equivalently, $H_{mn}(p_1q_1, \cdots, p_mq_n) = H_m(p_1, \cdots, p_m)+ H_n(q_1, \cdots, q_n)$;

(f) Continuity: $H_2(p, 1-p) \to 0$ as $p \to 0$;

then $H_m(p_1, \cdots, p_m) = \sum_{i=1}^m p_i \log\frac{1}{p_i}$ is the only possible choice.

## 2.2 Divergence

We begin with a review:

A measurable space $(\mathcal{X}, \mathcal{E})$ is said to be **standard Borel** if there exists a metric on space $\mathcal{X}$ that makes it a complete separable metric space in such a way that $\mathcal{E}$ is then the Borel $\sigma$-algebra of $\mathcal{X}$. A **Polish space** (i.e. separable $(\mathcal{X}, \mathcal{T})$ metrizable with $d$ s.t. $(\mathcal{X}, d)$ is complete and $\mathcal{T}_d = \mathcal{T}$) can thus be made into a standard Borel space by equipping $\mathcal{B}_\mathcal{T}$. Some of the nice properties of a standard Borel space are:

- All complete separable metric spaces, endowed with Borel $\sigma$-algebras are standard Borel. In particular, countable alphabets and $\mathbb{R}^n$ abd $\mathbb{R}^\infty$ (space of sequences) are standard Borel.
- If $\mathcal{X}_i, i = 1, \cdots$ are standard Borel, then so is $\prod_{i=1}^\infty \mathcal{X}_i$.
- Singletons $\{x\}$ are measurable sets.
- The diagonal $\{(x,x) : x \in \mathcal{X}\}$ is measurable in $\mathcal{X} \times \mathcal{X}$.

We now introduce the concept of KL divergence, or relative entropy.

**Definition 2.8** (Kullback-Leiber (KL) Divergence)**.** Let $P, Q$ be distributions on an alphabet (space where the r.v. takes values) $\mathcal{A}$, with $Q$ called the **reference measure**. The **KL divergence** between $P$ and $Q$ is

$$D(P\|Q) := \begin{cases} \mathbb{E}_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ}\right], & P \ll Q \\ +\infty, & \text{otherwise} \end{cases}$$

adopting again the convention that $0\log 0 = 0$. ♦

Two special cases are

- when $\mathcal{A}$ is a discrete (finite or countably infinite) alphabet:

$$D(P\|Q) = \begin{cases} \sum_{a\in\mathcal{A}:P(a),Q(a)>0} P(a)\log\frac{P(a)}{Q(a)}, & \text{supp}(P)\subset\text{supp}(Q) \\ +\infty, & \text{otherwise} \end{cases}$$

- when $\mathcal{A}=\mathbb{R}^k$, $P$ and $Q$ have pdfs $p$ and $q$ w.r.t. Leb (see example 1.2):

$$D(P\|Q) = \begin{cases} \int_{p>0,q>0} p(x)\log\frac{p(x)}{q(x)}\,dx, & \text{Leb}\{p>0,q=0\}=0 \\ +\infty, & \text{otherwise} \end{cases}$$

They are special by [7] Lemma 2.4. When $P\ll Q$ in particular,

$$\mathbb{E}_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ}\right] = \int_{\mathcal{A}}\frac{dP}{dQ}\log\frac{dP}{dQ}\,dQ \overset{\text{indefinite int}}{\underset{\text{R-N derivative}}{=\!=\!=}} \int_{\mathcal{A}}\log\frac{dP}{dQ}\,dP = \mathbb{E}_P\left[\log\frac{dP}{dQ}\right] \tag{3}$$

which coincides with the two cases with $Q$ being counting measure and Lebesgue measure respectively. Note that $D(P\|Q)$ is $+\infty$ when it is not the case $P\ll Q$. However, it can also be $+\infty$ even when $P\ll Q$. For example, $D(\text{Cauchy}\|\text{Gaussian})=\infty$. Our first observation is that $D(P\|Q)\neq D(Q\|P)$, so divergence is not a distance. We also see that generalizing entropy, $D$ is so called relative entropy.

**Theorem 2.9** ($H$ v.s. $D$)**.** *If distribution $P$ is supported on a finite alphabet $\mathcal{A}$, then*

$$H(P) = \log|\mathcal{A}| - D(P\|\mathit{Unif}(\mathcal{A}))$$

*Proof.* Using the first special case, we have

$$D(P\|\text{Unif}(\mathcal{A})) = \sum_{a\in\mathcal{A}:P(a),Q(a)>0} P(a)\log\frac{P(a)}{1/|\mathcal{A}|} = \sum_{a\in\mathcal{A}} P(a)\left(\log|\mathcal{A}| - \log\frac{1}{P(a)}\right) = \log|\mathcal{A}| - H(P)$$

□

**Example 2.10** (Binary divergence)**.** Consider $P=\text{Ber}(p)$ and $Q=\text{Ber}(q)$ on $\mathcal{A}=\{0,1\}$. Then

$$D(P\|Q) = d(p\|q) := p\log\frac{p}{q} + \bar{p}\log\frac{\bar{p}}{q}$$

Here is how $d(p\|q)$ depends on $p$ and $q$:

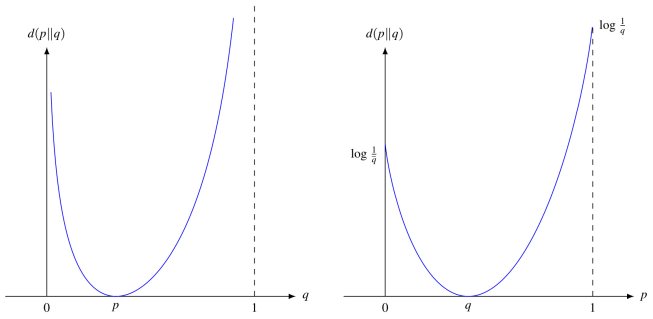


Figure 2: Binary divergence

It is easy to check the following quadratic lower bound, which is in fact a special case of Pinsker's inequality:

$$d(p\|q) \geq 2(p-q)^2\log e \quad \Diamond$$

**Example 2.11** (Real Gaussian)**.** For two Gaussians on $\mathcal{A} = \mathbb{R}$,

$$D(\mathcal{N}(m_1, \sigma_1^2) \| \mathcal{N}(m_0, \sigma_0^2)) = \frac{\log e}{2} \frac{(m_1 - m_0)^2}{\sigma_0^2} + \frac{1}{2} \left[ \log \frac{\sigma_0^2}{\sigma_1^2} + \left( \frac{\sigma_0^2}{\sigma_1^2} - 1 \right) \log e \right]$$

Here, the first and second term compares the means and the variances, respectively. Similary, in the vector case of $\mathcal{A} = \mathbb{R}^k$ and assuming $\det \boldsymbol{\Sigma_0} \neq 0$, we have

$$\begin{aligned} &D(\mathcal{N}(\boldsymbol{m_1}, \boldsymbol{\Sigma_1}) \| \mathcal{N}(\boldsymbol{m_0}, \boldsymbol{\Sigma_0})) \\ =&\frac{\log e}{2} (\boldsymbol{m_1} - \boldsymbol{m_0})^\top \boldsymbol{\Sigma_0}^{-1} (\boldsymbol{m_1} - \boldsymbol{m_0}) + \frac{1}{2} \left( \log \det \boldsymbol{\Sigma_0} - \log \det \boldsymbol{\Sigma_1} + \operatorname{tr}(\boldsymbol{\Sigma_0}^{-1} \boldsymbol{\Sigma_1} - I) \log e \right) \quad \lozenge \end{aligned}$$

We show a fundamental result.

**Theorem 2.12** (Information Inequality)**.**

$$D(P \| Q) \geq 0$$

*with equality iff $P = Q$.*

*Proof.* In view of defn 2.8, it suffices to consider $P \ll Q$. Let $\varphi(x) := x \log x$, which is strictly convex on $\mathbb{R}_+$. Applying Jensen's inequality:

$$D(P \| Q) = \mathbb{E}_Q \left[ \varphi \left( \frac{dP}{dQ} \right) \right] \geq \varphi \left( \mathbb{E}_Q \left[ \frac{dP}{dQ} \right] \right) = \varphi(1) = 0,$$

with equality iff $dP/dQ = 1\, Q - a.e.$, namely, $P = Q$. □

The definition of $D(P\|Q)$ extends verbatim to measures $P$ and $Q$ (not necessarily probabilitiy measures), in which case $D(P\|Q)$ can be nagative. A sufficient condition for $D(P\|Q) \geq 0$ is that $P$ is a probabilitiy measure and $Q$ is a sub-probabilitiy measure, i.e., $\int dQ \leq 1 = \int dP$. The notion of differential entropy is simply the divergence with respect to the Lebesgue measure:

**Definition 2.13** (Differential entropy)**.** The **differential entropy** of a random vector $X$ is

$$h(X) \equiv h(P_X) := -D(P_X \| \text{Leb})$$

In particular, if $X$ has pdf $p = dP_X/d\text{Leb}$ (i.e., $P_X \ll$ Leb), then $h(X) \overset{\text{eq.(3)}}{=\!=} -\mathbb{E}_{P_X}[\log \frac{dP_x}{d\text{Leb}}] = \mathbb{E}[\log \frac{1}{p(X)}]$; otherwise $h(X) = -\infty$. The **conditional differential entropy** is

$$h(X|Y) := \mathbb{E} \left[ \log \frac{1}{p_{X|Y}(X|Y)} \right]$$

where $p_{X|Y}$ is a conditional pdf. Compare this definition with conditional entropy. ♦

**Example 2.14** (Gaussian)**.** For $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned} h(X) =& \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \log \left( \sigma\sqrt{2\pi} \exp\left( \frac{(x-\mu)^2}{2\sigma^2} \right) \right) dx \\ \overset{\substack{\text{substitution} \\ t=(x-\mu)/\sqrt{2}\sigma}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=} & \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp(-t^2) \log(\sigma\sqrt{2\pi} \exp(t^2))\, dt \\ =& \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \log(\sigma\sqrt{2\pi}) e^{-t^2}\, dt + \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \log e^{t^2} e^{-t^2}\, dt \\ \overset{\substack{\text{Gaussian integral} \\ \text{int by parts}}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=} & \log(\sigma\sqrt{2\pi}) + \log(e^{1/2}) = \frac{1}{2} \log(2\pi e \sigma^2) \end{aligned}$$

More generally, for $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in $\mathbb{R}^d$,

$$h(\boldsymbol{X}) = \frac{1}{2}\log((2\pi e)^d \det \boldsymbol{\Sigma}) \quad \lozenge$$

**Warning:** Even for continuous r.v. $X$, $h(X)$ can be positive, negative, takes values of $\pm\infty$ or even undefined. For the last case, consider a piecewise-constant pdf taking value $e^{(-1)^n n}$ on the $n$-th interval of width $\Delta_n = \frac{c}{n^2} e^{-(-1)^n n}$. There are many differences between the Shannon entropy and the differential entropy. For example, from Proposition 2.7 we know that deterministic processing cannot increase the Shannon entropy, i.e., $H(f(X)) \leq H(X)$ for any discrete $X$, which is intuitively clear. However, this fails completely for differential entropy (e.g. consider scaling, see proposition 2.15). Furthermore, for sums of independent r.v., for integer-valued $X$ and $Y$, $H(X+Y)$ is finite whenever $H(X)$ and $H(Y)$ are, because $H(X+Y) \leq H(X,Y) = H(X) + H(Y)$. This again fails for differential entropy. In fact, there exists real-valued $X$ with finite $h(X)$ such that $h(X+Y) = \infty$ for any independent $Y$ such that $h(Y) > -\infty$; there also exist $X$ and $Y$ with finite differential entropy such that $h(X+Y)$ does not exist.

Nevertheless, differential entropy shares many functional properties with the usual Shannon entropy.

**Proposition 2.15** (Properties of differential entropy)**.** *Assuming that all differential entropies appearing below exist and are finite (in particular all r.v. have pdfs and conditional pdfs).*

*(a) (Uniform dist. maximizes differential entropy) If $\mathbb{P}[X^n \in S] = 1$ then $h(X^n) \leq \log \mathit{Leb}(S)$, with equality iff $X^n$ is uniform on $S$.*

*(b) (Scaling and shifting) $h(X^n + x) = h(X^n), h(\alpha X^n) = h(X^n) + k\log|\alpha|$ and for an invertible matrix $A$, $h(AX^n) = h(X^n) + \log|\det A|$.*

*(c) (Condtioning reduces differential entropy) $h(X|Y) \leq h(X)$ ($Y$ is arbitrary).*

*(d) (Chain rule) Let $X^n$ has a joint pdf. Then $h(X^n) = \sum_{k=1}^n h(X_k|X^{k-1})$.*

TO DO: Markov Kernel (along with standard Borel space, need to be put into the first section "notations in probabilitiy"); conditional divergence and data processing inequality; most importantly, Fisher information.

# References

[1] Benaïm, Michel, and Tobias Hurth. *Markov Chains on Metric Spaces: A Short Course*, Springer Nature Switzerland AG, 2022.

[2] Cinlar, Erhan. *Probability and Stochastics*, Springer Science+Business Media, LLC, 2011.

[3] Häaggström, Olle. *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press, 2002.

[4] Lee, John. M. *Introduction to Smooth Manifolds*, second edition, Springer Science+Business Media New York, 2013.

[5] Lee, John. M. *Introduction to Riemannian Manifolds*, second edition, Springer International Publishing AG, 2018.

[6] Norris, James. *Markov Chains*, Cambridge University Press, 1997.

[7] Wu, Yihong, and Yury Polyanskiy *Information Theory From Coding to Learning*, Cambridge University Press, 2022.