

# Information Theory

Jayadev Naram

## Contents

<a href="#">1 Entropy</a>	<a href="#">1</a>
<a href="#">2 Relative Entropy and Variational Distance</a>	<a href="#">6</a>
<a href="#">3 Typicality</a>	<a href="#">6</a>
<a href="#">4 Source Coding</a>	<a href="#">7</a>
<a href="#">5 Joint Typical</a>	<a href="#">7</a>
<a href="#">6 Channel Coding</a>	<a href="#">7</a>

## 1 Entropy

**Definition 1.1.** The *uncertainty or entropy* of a discrete random variable  $U$  that takes values in the set  $\mathcal{U}$  (also called alphabet  $\mathcal{U}$ ) is defined as

$$H(U) = - \sum_{u \in \mathcal{U}} P_U(u) \log_b P_U(u),$$

where  $P_U(\cdot)$  denotes the probability mass function of the random variable  $U$ .

**Remark 1.2.** It should be noted that when  $P_U(u) = 0$ , the corresponding term does not contribute to entropy because  $\lim_{t \downarrow 0} t \log_b t = 0$ . In view of this result, one can equivalently define entropy on the support of  $P_U$  which is defined as

$$\text{supp}(P_U) = \{u : P_U(u) > 0\} \subseteq \mathcal{U}.$$

**Remark 1.3.** Entropy does not depend on different possible values that  $U$  can take on, but only on the probabilities of these values.

**Definition 1.4.** If  $U$  is binary with two possible values  $u_1$  and  $u_2$ , such that  $\mathbb{P}[U = u_1] = p$  and  $\mathbb{P}[U = u_2] = 1 - p$ , then

$$H(U) = H_b(p) = -p \log_2 p - (1 - p) \log_2 (1 - p), p \in [0, 1],$$

where  $H_b(\cdot)$  is called the **binary entropy function**.

**Definition 1.5.** Let  $f$  be a function from a convex set  $C$  to  $\mathbb{R}$ . Then  $f$  is said to be convex on  $C$  if for every  $x, y \in C$  and  $0 \leq \lambda \leq 1$ ,

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y).$$

A function is said to be strictly convex if equality holds only if  $\lambda = 0$  or  $\lambda = 1$ . A function  $f$  is concave if  $-f$  is convex.

**Lemma 1.6.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a twice continuously differentiable function on an open set  $E$ . Then  $f$  is convex on  $E$  iff its second derivative  $f''$  is nonnegative throughout  $E$ . If  $f''$  is positive on  $E$ , then  $f$  is strictly convex.

**Remark 1.7.** Notice that  $-\log x$  and  $x \log x$  are strictly convex on  $(0, \infty)$ .

**Lemma 1.8** (Jensen's Inequality). Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function. Then

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i),$$

where  $\lambda_i \geq 0 \forall i$ ,  $\sum_{i=1}^n \lambda_i = 1$  and equality holds iff  $x_1 = \dots = x_n$  or  $f$  is linear.

**Remark 1.9.** If  $f$  is strictly convex which rules out the linearity, the equality of Jensen inequality holds iff  $x_1 = \dots = x_n$ .

**Remark 1.10.** Suppose  $X$  is a discrete random variable over an alphabet  $\mathcal{X} = \{x_1, \dots, x_n\}$  and  $f$  is a strictly convex function on  $\mathbb{R}$ . Then by setting  $\lambda_i = P_X(x_i)$ , we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)],$$

where equality holds iff  $x_1 = \dots = x_n$ , i.e.,  $X$  is a constant.

**Theorem 1.11.** If  $U$  has  $r$  possible values, then

$$0 \leq H(U) \leq \log r,$$

where

$$\begin{aligned} H(U) = 0 &\iff \exists u \in \mathcal{U}, P_U(u) = 1, \\ H(U) = \log r &\iff \forall u \in \mathcal{U}, P_U(u) = \frac{1}{r}. \end{aligned}$$

*Proof.* Since  $0 \leq P_U(u) \leq 1$ , we have

$$-P_U(u) \log_2 P_U(u) \begin{cases} = 0 & \text{if } P_U(u) = 1, \\ > 0 & \text{if } 0 < P_U(u) < 1. \end{cases}$$

Hence,  $H(U) \geq 0$ . Equality can only be achieved if  $-P_U(u) \log_2 P_U(u) = 0$  for all  $u \in \text{supp}(P_U)$ , i.e.,  $P_U(u) = 1$  for all  $u \in \text{supp}(P_U)$ .

To derive the upper bound we use a trick that is quite common in information theory: We take the difference and try to show that it must be nonpositive:

$$\begin{aligned} H(U) - \log r &= -\sum_{u \in \mathcal{U}} P_U(u) \log P_U(u) - \log r \\ &= \sum_{u \in \mathcal{U}} P_U(u) \log \frac{1}{P_U(u)r} \\ &\leq \log \left( \sum_{u \in \mathcal{U}} P_U(u) \frac{1}{P_U(u)r} \right) = 0, \end{aligned}$$

where we have used the strict concavity of  $\log x$  and Jensen inequality. Equality holds iff  $\frac{1}{P_U(u)r} = 1$  for all  $u \in \mathcal{U}$ , i.e.,  $P_U(u) = \frac{1}{r}$  for all  $u \in \mathcal{U}$ .  $\square$

**Definition 1.12.** The **conditional entropy** of the random variable  $X$  given the event  $Y = y$  is defined as

$$H(X|Y = y) = - \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log P_{X|Y}(x|y) = -\mathbb{E} \left[ \log P_{X|Y}(X|Y) \middle| Y = y \right],$$

where the conditional probability distribution is given by

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}.$$

**Corollary 1.13.** If  $X$  has  $r$  possible values, then

$$0 \leq H(X|Y = y) \leq \log r,$$

where

$$\begin{aligned} H(X|Y = y) = 0 &\iff \exists x \in \mathcal{X}, P_{X|Y}(x|y) = 1, \\ H(X|Y = y) = \log r &\iff \forall x \in \mathcal{X}, P_{X|Y}(x|y) = \frac{1}{r}. \end{aligned}$$

**Definition 1.14.** The **conditional entropy** of the random variable  $X$  given the random variable  $Y$  is defined as

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y) \\ &= \mathbb{E}_Y [H(X|Y = y)] \\ &= - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log P_{X|Y}(x|y) \\ &= -\mathbb{E} \left[ \log P_{X|Y}(X|Y) \right]. \end{aligned}$$

**Corollary 1.15.** If  $X$  has  $r$  possible values, then

$$0 \leq H(X|Y) \leq \log r,$$

where

$$\begin{aligned} H(X|Y) = 0 &\iff \exists x \in \mathcal{X}, \forall y \in \mathcal{Y}, P_{X|Y}(x|y) = 1, \\ H(X|Y) = \log r &\iff \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, P_{X|Y}(x|y) = \frac{1}{r}. \end{aligned}$$

**Remark 1.16.** Generally,  $H(X|Y) \neq H(Y|X)$ .

**Theorem 1.17** (Conditioning Reduced Uncertainty). For any two discrete random variables  $X$  and  $Y$ ,

$$H(X|Y) \leq H(X),$$

where equality holds iff  $X$  and  $Y$  are independent, i.e.,  $X \perp Y$ .

*Proof.* Consider the following:

$$\begin{aligned}
H(X|Y) - H(X) &= - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log P_{X|Y}(x|y) + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log P_X(x) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_X(x)}{P_{X|Y}(x|y)} \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_X(x)P_Y(y)}{P_{X,Y}(x, y)} \\
&\leq \log \left( \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \frac{P_X(x)P_Y(y)}{P_{X,Y}(x, y)} \right) \\
&= \log \left( \left( \sum_{x \in \mathcal{X}} P_X(x) \right) \left( \sum_{y \in \mathcal{Y}} P_Y(y) \right) \right) = 0,
\end{aligned}$$

where we have used the strict concavity of  $\log x$  and Jensen inequality. Equality holds iff  $\frac{P_X(x)P_Y(y)}{P_{X,Y}(x, y)} = 1$  for all  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ , i.e.,  $X \perp Y$ .  $\square$

**Remark 1.18.** The conditioning reduces entropy-rule only applies to random variables, but not to events. In particular,

$$H(X|Y = y) \leq H(X).$$

To understand why this is the case, consider the following example. Suppose  $X, Y$  are random variables such that  $P_X(x_1) = 0.4, P_X(x_2) = 0.6, P_{X|Y}(x_1|y_1) = 1$  and  $P_{X|Y}(x_i|y_2) = 1/2, i = 1, 2$ . Then we see that

$$\begin{aligned}
H(X) &= H_b(0.4) \approx 0.97 \text{ bits}, \\
H(X|Y = y_1) &= H_b(1) = 0 \text{ bits}, \\
H(X|Y = y_2) &= H_b(0.5) = 1 \text{ bit}.
\end{aligned}$$

However from Theorem 1.17 we know that on average the knowledge of  $Y$  will reduce the uncertainty about  $X$ :  $H(X|Y) \leq H(X)$ .

**Theorem 1.19** (Chain Rule). *Let  $X_1, \dots, X_n$  be  $n$  discrete random variables. Then*

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1, \dots, X_{n-1}) = \sum_{k=1}^n H(X_k|X^{(k-1)}),$$

where  $X^{(k-1)} = X_{1:k-1}$ .

*Proof.* This follows directly from the chain rule for probability mass functions:

$$P_{X^{(n)}} = \prod_{k=1}^n P_{X_k|X^{(k-1)}}.$$

$\square$

**Definition 1.20.** The **mutual information** between the random variables  $X$  and  $Y$  is

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

**Remark 1.21.** Notice that mutual information is symmetric in its arguments:

$$\begin{aligned}
H(X, Y) &= H(X) + H(Y|X) = H(Y) + H(X|Y) \\
\Rightarrow H(X) - H(X|Y) &= H(Y) - H(Y|X) \\
\Rightarrow I(X; Y) &= I(Y; X).
\end{aligned}$$

**Remark 1.22.** When  $X \perp\!\!\!\perp Y$ , we have  $I(X; Y) = 0$ . Additionally,  $I(X; X) = H(X)$ .

**Remark 1.23.** From the chain rule it follows that

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

and thus we obtain

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

**Remark 1.24.** The mutual information can be expressed as follows.

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= \mathbb{E}[-\log P_X(X)] - \mathbb{E}[P_{X|Y}(X|Y)] \\ &= \mathbb{E}\left[\log \frac{P_{X|Y}(X|Y)}{P_X(X)}\right] \\ &= \mathbb{E}\left[\log \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)}\right] \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)}. \end{aligned}$$

**Theorem 1.25.** Let  $X$  and  $Y$  be two random variables. Then

$$0 \leq I(X; Y) \leq \min\{H(X), H(Y)\}.$$

where equality holds on the left-hand side iff  $P_{X,Y} = P_X P_Y$ , i.e., iff  $X \perp\!\!\!\perp Y$ , and equality holds on the right-hand side iff  $X$  determines  $Y$  or vice versa.

*Proof.* It follows directly from the definition of mutual information and nonnegativity of conditional entropy.  $\square$

**Theorem 1.26** (Chain Rule). Let  $X, Y_1, \dots, Y_n$  be  $n + 1$  discrete random variables. Then

$$I(X; Y_1, \dots, Y_n) = I(X; Y_1) + I(X; Y_2|Y_1) + \dots + I(X; Y_n|Y_1, \dots, Y_{n-1}) = \sum_{k=1}^n I(X; Y_k|Y^{(k-1)}).$$

*Proof.* From the chain rule of entropy we have

$$\begin{aligned} I(X; Y^{(n)}) &= H(Y^{(n)}) - H(Y^{(n)}|X) \\ &= \sum_{k=1}^n H(Y_k|Y^{(k-1)}) - H(Y_k|Y^{(k-1)}, X) \\ &= \sum_{k=1}^n I(X; Y_k|Y^{(k-1)}). \end{aligned}$$

$\square$

**Remark 1.27.** Sometimes it is convenient to use another notation for entropy and mutual information, which is explicit in the probability mass functions these quantities depend on. Sometimes, we shall write  $H(X)$  as  $H(P_X)$  and  $I(X; Y)$  as  $I(P_X, P_{Y|X})$ .

**Theorem 1.28** (Uniqueness of the Definition of Entropy).

## 2 Relative Entropy and Variational Distance

**Definition 2.1.** Relative Entropy.

**Remark 2.2.** Unboundedness of Relative Entropy.

**Theorem 2.3** (Gibbs' Inequality). *Nonnegativity of Relative Entropy.*

**Remark 2.4.** Relative Entropy is not a Norm.

**Remark 2.5.** Mutual Information and Relative Entropy.

**Definition 2.6.** Variational Distance.

**Remark 2.7.** Trivial bounds on variational distance.

**Remark 2.8.** Variational Distance is a Norm.

**Theorem 2.9.** *Upper bound of relative entropy in terms of variational distance and entropy.*

**Theorem 2.10** (Pinsker Inequality).

**Definition 2.11.** Total Variation Distance.

## 3 Typicality

**Definition 3.1.** We say that a sequence of random variables  $\{X_n\}$  **converges in probability** to a random variable  $X$  if for all  $\varepsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0.$$

**Lemma 3.2** (Markov Inequality). *Let  $X$  be a nonnegative random variable of finite mean  $\mathbb{E}[X] < \infty$ . Then for all  $a > 0$ , we have*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

**Lemma 3.3** (Chebyshev Inequality). *Let  $X$  be a random variable with finite mean  $\mu$  and finite variance  $\sigma^2$ . Then for all  $\varepsilon > 0$ , we have*

$$\mathbb{P}[|X - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}.$$

**Lemma 3.4** (Weak Law of Large Numbers). *Let  $\{Z_n\}$  be a sequence of independent and identically distributed (i.i.d.) random variables with mean  $\mu$  and variance  $\sigma^2$ . Let*

$$S_n = \frac{1}{n} \sum_{k=1}^n Z_k$$

*be the sample mean. Then  $\{S_n\}$  converges in probability to  $\mu$ . In particular,*

$$\mathbb{P}[|S_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

**Definition 3.5** (Type). Let  $x^{(n)}$  be a sequence of  $n$  elements drawn from a finite-cardinality alphabet  $\mathcal{X}$ . The **empirical probability mass function** of  $x^{(n)}$ , also referred to as its **type**, is defined for  $x \in \mathcal{X}$  as

$$\pi(x|x^{(n)}) = \frac{|\{i \in [n] : x_i = x\}|}{n},$$

where  $[n] = \{1, \dots, n\}$ .

**Theorem 3.6.** Let  $\{X_n\}$  be an i.i.d. sequence of random variables with  $X_i \sim P_X(x_i)$ . Then  $\forall x \in \mathcal{X}$  and for all  $\varepsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}[|\pi(x|X^{(n)}) - P_X(x)| > \varepsilon] = 0,$$

or in other words,  $\{\pi(x|X^{(n)})\}$  converges in probability to  $P_X(x)$  for all  $x \in \mathcal{X}$ .

**Definition 3.7** (Typical Set). The **set of  $\varepsilon$ -typical  $n$ -sequences** for a random variable  $X \sim P_X$  and  $\varepsilon \in (0, 1)$  (simply typical set) is defined as

$$\mathcal{T}_\varepsilon^{(n)}(X) = \{x^{(n)} : |\pi(x|x^{(n)}) - P_X(x)| \leq \varepsilon P_X(x), \forall x \in \mathcal{X}\}.$$

**Remark 3.8.** For an element  $x \in \mathcal{X}$  which has  $P_X(x)$  cannot be a part of typical sequence. Suppose such an  $x$  belonged to a sequence  $x^{(n)}$ , then  $\pi(x|x^{(n)}) > 0$ . Consequently, we have  $|\pi(x|x^{(n)}) - P_X(x)| = \pi(x|x^{(n)}) > 0 = \varepsilon P_X(x)$  for all  $\varepsilon > 0$ , which shows that  $x^{(n)}$  is not a typical sequence.

**Lemma 3.9** (Typical Average Lemma). Consider a typical sequence  $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$ . Then for any **nonnegative** function  $g(\cdot)$  on  $\mathcal{X}$ , we have

$$(1 - \varepsilon)\mathbb{E}[g(X)] \leq \frac{1}{n} \sum_{k=1}^n g(x_k) \leq (1 + \varepsilon)\mathbb{E}[g(X)].$$

## 4 Source Coding

## 5 Joint Typical

**Definition 5.1** (Joint Type). Let  $(x^{(n)}, y^{(n)})$  be a sequence of a pair of  $n$  length sequences from a finite-cardinality alphabet  $(\mathcal{X}, \mathcal{Y})$ . The **joint empirical probability mass function** of  $(x^{(n)}, y^{(n)})$ , also referred to as its **joint type**, is defined for  $x \in \mathcal{X}$  as

$$\pi(x, y|x^{(n)}, y^{(n)}) = \frac{|\{i \in [n] : (x_i, y_i) = (x, y)\}|}{n}.$$

**Remark 5.2.** The  $X$ -marginal of  $X, Y$ -joint empirical probability mass function is the  $X$ -empirical probability mass function.

**Definition 5.3** (Jointly Typical Set). The **set of  $\varepsilon$ -jointly typical  $n$ -sequences** for a random variable  $(X, Y) \sim (P_X, P_Y)$  and  $\varepsilon \in (0, 1)$  (simply jointly typical set) is defined as

$$\mathcal{T}_\varepsilon^{(n)}(X, Y) = \{(x^{(n)}, y^{(n)}) : |\pi(x, y|x^{(n)}, y^{(n)}) - P_{X,Y}(x, y)| \leq \varepsilon P_{X,Y}(x, y), \forall x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

**Remark 5.4.** If  $(x^{(n)}, y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X, Y)$ , then  $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$  and  $y^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(Y)$ .

## 6 Channel Coding

## References