Information Theory

Jayadev Naram

Contents

1 Entropy

Definition 1.1. The ***uncertainity 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 $f : \mathbb{R} \to \mathbb{R}$ 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 $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then*

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

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

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

Remark 1.10. Suppose X is a discrete random variable over an alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$ 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 $x_1 = \cdots = x_n$, i.e., X is a constant.

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

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

where

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

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 $u \in \operatorname{supp}(P_U)$, i.e., $P_U(u) = 1$ for all $u \in \operatorname{supp}(P_U)$.

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:

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

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

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

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

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*

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

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

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

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

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

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 Uncertainity)**.** *For any two discrete random variables X and Y,*

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

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

Proof. Consider the following:

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

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., $X \perp\!\!\!\perp Y$. □

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

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

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 uncertainity about X: $H(X|Y) \le H(X)$.

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

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

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

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

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

□

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:

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

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

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

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.

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

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

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

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. □

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

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

Proof. From the chain rule of entropy we have

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

□

Theorem 1.27 (Data Processing Inequality)**.** *Let X, Y, Z be random variables that form a Markov chain, denoted by $X - Y - Z$, i.e., $X \perp\!\!\!\perp Z \,|\, Y$. Then*

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

Proof 1. We start by considering

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &\leq H(X) - H(X|Z,Y) \quad \text{(conditioning reduces entropy)}, \\ &= H(X) - H(X|Y) \qquad \text{(since } X \perp\!\!\!\perp Z \,|\, Y\text{)}, \\ &= I(X;Y). \end{aligned}$$

□

Proof 2. Another way of proving the inequality is to start by considering the following mutual information

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \\ &= I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} \\ \implies I(X;Z) &\leq I(X;Y). \end{aligned}$$

□

Remark 1.28. 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 1.29 (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. Conditional Divergence.

Remark 2.7. Conditional divergence as divergence of product of pmfs.

Theorem 2.8. *Chain rule.*

Theorem 2.9. *Conditioning increases divergence.*

Theorem 2.10. *Data Processing for Divergence.*

Theorem 2.11. *Data Processing for divergence implies Data Processing for mutual information.*

Interpretation.

Theorem 2.12. *Golden Formula.*

Definition 2.13. Variational Distance.

Remark 2.14. Trivial bounds on variational distance.

Remark 2.15. Variational Distance is a Norm.

Theorem 2.16. *Upper bound of relative entropy in terms of variational distance and entropy.*

Theorem 2.17 (Pinsker Inequality)**.**

Definition 2.18. 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

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

Lemma 3.2 (Markov Inequality)**.** *Let X be a nonnegtaive 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 μ and finite variance σ^2. Then for all $\varepsilon > 0$, we have*

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

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 μ an variance σ^2. Let*

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

be the sample mean. Then $\{S_n\}$ converges in probability to μ. 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 $[n] = \{1, \ldots, n\}$.

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*

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

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 ε-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 **nonnegtaive** 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 ε-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

[1] Giuseppe Durisi. Lecture notes in SSY210-Information Theory, Chalmers University of Technology, 19th May 2021.

[2] Stefan M Moser. Information theory: Lecture notes (version 6.12 from 222 March 2023, PDF), 6th edition, ETH Zürich, 2023.

[3] Thomas M. Cover and Joy A. Thomas. Elements of information theory, 2th edition. John Wiley & Sons, 2005.