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 function*** 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:

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

where we have used the strict concavity of $\log x$ and Jensen inequality. Equality holds iff $\dfrac{1}{P_U(u) r} = 1$ 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 (DPI)***)**.** *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. Let P and Q be two probability mass functions over the same finite (or countably infinite) alphabet $\mathcal{X}$. The ***relative entropy*** or ***Kullback-Leibler divergence*** between P and Q is defined as

$$\mathscr{D}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}_P\left[\log \frac{P(x)}{Q(x)}\right].$$

Remark 2.2. Note that $\mathscr{D}(P\|Q) = \infty$ if there exists an $x \in \operatorname{supp}(P)$ such that $Q(x) = 0$. So, strictly speaking, we should defined relative entropy as follows:

$$\mathscr{D}(P\|Q) = \begin{cases} \sum_{x\in\operatorname{supp}(P)} P(x) \log \dfrac{P(x)}{Q(x)} & \text{if } \operatorname{supp}(P) \subseteq \operatorname{supp}(Q), \\ \infty & \text{otherwise.} \end{cases}$$

Theorem 2.3 (***Gibbs' Inequality***)**.**

$$\mathscr{D}(P\|Q) \geq 0,$$

where equality holds iff $P(x) = Q(x)\ \forall\ x \in \mathcal{X}$.

Proof. In the case when $\operatorname{supp}(P) \not\subseteq \operatorname{supp}(Q)$, we have $\mathscr{D}(P\|Q) = \infty > 0$ trivially. So, we assume that $\operatorname{supp}(P) \subseteq \operatorname{supp}(Q)$. Then,

$$\begin{aligned} -\mathscr{D}(P\|Q) &= \sum_{x\in\operatorname{supp}(P)} P(x) \log \frac{Q(x)}{P(x)} \\ &\leq \log\left(\sum_{x\in\operatorname{supp}(P)} P(x) \frac{Q(x)}{P(x)}\right) \leq 0. \end{aligned}$$

Equality holds in both the inequalities iff $\frac{Q(x)}{P(x)} = 1$ and $\text{supp}(P) = \text{supp}(Q)$, i.e., $P(x) = Q(x)$ for all $x \in \mathcal{X}$. □

Remark 2.4. Relative Entropy is not a norm as it is not symmetric and does not satisfy triangle inequality.

Remark 2.5. From Remark 1.24, it can be see that mutual information is the relative entropy between the joint $P_{X,Y}$ and the product of its marginals:

$$I(X;Y) = \mathscr{D}(P_{X,Y} \| P_X P_Y).$$

Definition 2.6. The ***conditional divergence*** between two discrete probability distributions $P_{Y|X}$ and $Q_{Y|X}$ is defined as

$$\mathscr{D}(P_{Y|X} \| Q_{Y|X} | P_X) = \sum_{x \in \mathcal{X}} P_X(x) \mathscr{D}(P_{Y|X=x} \| Q_{Y|X=x}).$$

Remark 2.7. The conditional divergence can be represented in terms of divergence as follows:

$$\begin{aligned}
\mathscr{D}(P_{Y|X} \| Q_{Y|X} | P_X) &= \sum_x P_X(x) \mathscr{D}(P_{Y|X=x} \| Q_{Y|X=x}) \\
&= \sum_x P_X(x) \sum_y P_{Y|X=x}(y) \log \frac{P_{Y|X=x}(y)}{Q_{Y|X=x}(y)} \\
&= \sum_{x,y} P_X(x) P_{Y|X=x}(y) \log \frac{P_{Y|X=x}(y) P_X(x)}{Q_{Y|X=x}(y) P_X(x)} \\
&= \mathscr{D}(P_{Y|X} P_X \| Q_{Y|X} P_X).
\end{aligned}$$

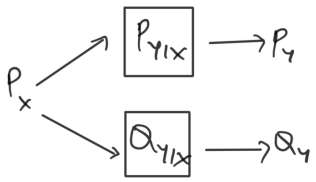
Theorem 2.8 (***Chain Rule***). $\mathscr{D}(P_{X,Y} \| Q_{X,Y}) = \mathscr{D}(P_{Y|X} \| Q_{Y|X} | P_X) + \mathscr{D}(P_X \| Q_X)$.

Proof.

$$\begin{aligned}
\mathscr{D}(P_{X,Y} \| Q_{X,Y}) &= \sum_{X,Y} P_{Y|X}(y) P_X(x) \log \frac{P_{Y|X=x}(y) P_X(x)}{Q_{Y|X}(y) Q_X(x)} \frac{P_X(x)}{P_X(x)} \\
&= \sum_{X,Y} P_{Y|X}(y) P_X(x) \log \frac{P_{Y|X=x}(y)}{Q_{Y|X=x}(y)} + \sum_{X,Y} P_{Y|X=x}(y) P_X(x) \log \frac{P_X(x)}{Q_X(x)} \\
&= \mathscr{D}(P_{Y|X} \| Q_{Y|X} | P_X) + \mathscr{D}(P_X \| Q_X).
\end{aligned}$$

□

Theorem 2.9 (***Conditioning Increases Divergence***). *Given $P_{Y|X}, Q_{Y|X}$ and P_X, let $P_Y = P_{Y|X} P_X$ and $Q_Y = Q_{Y|X} P_X$, as represented by the diagram.*



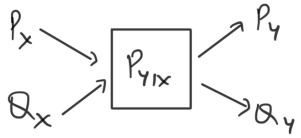
Then $\mathscr{D}(P_Y \| Q_Y) \leq \mathscr{D}(P_{Y|X} \| Q_{Y|X} | P_X)$, with equality iff $\mathscr{D}(P_{X|Y} \| Q_{X|Y} | P_Y) = 0$.

Proof.

$$\begin{aligned}\mathscr{D}\left(P_{X,Y}\|Q_{X,Y}\right) &= \mathscr{D}\left(P_{Y|X}\|Q_{Y|X}|P_X\right) + \underbrace{\mathscr{D}\left(P_X\|P_X\right)}_{=0} \\ &= \underbrace{\mathscr{D}\left(P_{X|Y}\|Q_{X|Y}|P_Y\right)}_{\geq 0} + \mathscr{D}\left(P_Y\|Q_Y\right) \\ \implies \mathscr{D}\left(P_Y\|Q_Y\right) &\leq \mathscr{D}\left(P_{Y|X}\|Q_{Y|X}|P_X\right).\end{aligned}$$

□

Theorem 2.10 (***DPI for Divergence***)**.** *Given $P_{Y|X}, P_X$ and Q_X, let $P_Y = P_{Y|X}P_X$ and $Q_Y = P_{Y|X}Q_X$, as represented by the diagram.*



Then $\mathscr{D}\left(P_Y\|Q_Y\right) \leq \mathscr{D}\left(P_X\|Q_X\right)$, with equality iff $\mathscr{D}\left(P_{X|Y}\|Q_{X|Y}|P_Y\right) = 0$.

Proof.

$$\begin{aligned}\mathscr{D}\left(P_{X,Y}\|Q_{X,Y}\right) &= \underbrace{\mathscr{D}\left(P_{Y|X}\|Q_{Y|X}|P_X\right)}_{=0} + \mathscr{D}\left(P_X\|Q_X\right) \\ &= \underbrace{\mathscr{D}\left(P_{X|Y}\|Q_{X|Y}|P_Y\right)}_{\geq 0} + \mathscr{D}\left(P_Y\|Q_Y\right) \\ \implies \mathscr{D}\left(P_Y\|Q_Y\right) &\leq \mathscr{D}\left(P_X\|Q_X\right).\end{aligned}$$

□

Theorem 2.11. *DPI for Divergence implies DPI for Mutual Information.*

Proof. Now suppose we consider a Markov chain $X - Y - Z$ over discrete random variables. Then notice that

$$P_{Z|X} = P_{Z|X,Y}P_{Y|X} = P_{Z|Y}P_{Y|X}, \quad P_Z = P_{Z|Y}P_Y.$$

We apply DPI for divergence on $P_{Z|Y}, P_{Y|X=x}$ and P_Y to get

$$\begin{aligned}\mathscr{D}\left(P_{Y|X=x}\|P_Y\right) \geq \mathscr{D}\left(P_{Z|X=x}\|P_Z\right) &\implies \mathbb{E}_X\left[\mathscr{D}\left(P_{Y|X=x}\|P_Y\right)\right] \geq \mathbb{E}_X\left[\mathscr{D}\left(P_{Z|X=x}\|P_Z\right)\right] \\ &\implies \mathscr{D}\left(P_{Y|X}\|P_Y|P_X\right) \geq \mathscr{D}\left(P_{Z|X}\|P_Z|P_X\right). \qquad (1)\end{aligned}$$

Now consider the mutual information

$$\begin{aligned}I(X;Z) &= \mathscr{D}\left(P_{X,Z}\|P_XP_Z\right) \\ &= \mathscr{D}\left(P_{Z|X}\|P_Z|P_X\right) \\ &\leq \mathscr{D}\left(P_{Y|X}\|P_Y|P_X\right) \qquad \text{(using (1))} \\ &= \mathscr{D}\left(P_{X,Y}\|P_XP_Y\right) \\ &= I(X;Y).\end{aligned}$$

□

Theorem 2.12 (***Golden Formula***)**.**

$$I(X;Y) = \min_{Q_Y} \mathscr{D}\left(P_{Y|X} \| Q_Y | P_X\right).$$

Proof.

$$\begin{aligned}
\mathscr{D}\left(P_{Y|X} \| Q_Y | P_X\right) &= \mathscr{D}\left(P_{Y|X} P_X \| Q_Y P_X\right) \\
&= \sum_{x,y} P_{X,Y} \log \frac{P_{X,Y}}{Q_Y P_X} \frac{P_Y}{P_Y} \\
&= \sum_{x,y} P_{X,Y} \log \frac{P_{X,Y}}{P_X P_Y} + \sum_{x,y} P_{X,Y} \log \frac{P_Y}{Q_Y} \\
&= \mathscr{D}\left(P_{X,Y} \| P_X P_Y\right) + \mathscr{D}\left(P_Y \| Q_Y\right) \\
&= I(X;Y) + \mathscr{D}\left(P_Y \| Q_Y\right).
\end{aligned}$$

We get the result by noting that $D(P_Y||Q_Y) \geq 0$ where equality holds iff $Q_Y = P_Y$. □

Definition 2.13. Let P and Q be two probability mass functions over the same finite (or countably infinite) alphabet $\mathcal{X}$. The ***variational distance*** between P and Q is defined as

$$V(P,Q) = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.$$

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***)**.**

3 Typicality

Definition 3.1. We say that a sequence of random variables $\{X_n\}$ ***converges in probability*** to a random variable X if

$$\forall\, \varepsilon > 0, \delta > 0, \exists\, N \in \mathbb{Z}_+ \quad \text{such that} \quad n \geq N \implies \mathbb{P}\left[|X_n - X| > \varepsilon\right] < \delta,$$

or using the definition of limit,

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

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*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}, \quad a > 0.$$

Proof. Fix some $a > 0$ and define

$$Y_a = \begin{cases} 0 & \text{if } X < a, \\ a & \text{if } X \geq a. \end{cases}$$

Since X is nonnegative by assumption, it follows that $Y_a \leq X$ or equivalently $\mathbb{E}[Y_a] \leq \mathbb{E}[X]$. On other hand, we have

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

□

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

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \varepsilon] \leq \frac{\mathbb{V}ar[X]}{\varepsilon^2}, \quad \varepsilon > 0.$$

Proof. This follows directly from applying the Markov inequality to $(X - \mu)^2$ with $a = \varepsilon^2$. □

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 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,

$$\mathbb{P}[|S_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}, \quad \varepsilon > 0.$$

Proof. Observe that

$$\mathbb{E}[S_n] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[Z_k\right] = \mu,$$

$$\mathbb{V}\mathrm{ar}\left(\frac{Z_k}{n}\right) = \mathbb{E}\left[\left(\frac{Z_k}{n} - \mathbb{E}\left[\frac{Z_k}{n}\right]\right)^2\right] = \mathbb{E}\left[\left(\frac{Z_k}{n} - \frac{\mu}{n}\right)^2\right] = \frac{\mathbb{E}\left[(Z_k - \mu)^2\right]}{n^2} = \frac{\sigma^2}{n^2},$$

$$\mathbb{V}\mathrm{ar}(S_n) = \sum_{k=1}^{n} \mathbb{V}\mathrm{ar}\left(\frac{Z_k}{n}\right) = \frac{\sigma^2}{n},$$

where we used the property that variance of sum of i.i.d. random variables is sum of variance of each random variable.

By applying Chebyshev inequality on S_n for some $\varepsilon > 0$, we get the required bound on tail probability. The convergence of S_n is a direct consequence of setting $N = 1$ in the tail bound for all $\varepsilon > 0$. □

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 $X_i \sim P_X$. Then $\forall\, x \in \mathcal{X}$ and for all $\varepsilon > 0$, we have*

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

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

Proof. We can rewrite empirical pmf as

$$\pi(x|X^{(n)}) = \sum_{k=1}^{n} \frac{\mathbb{I}_{[X_k = x]}}{n},$$

where $\mathbb{I}$ is the indicator function. Notice that

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{[X_k=x]}\right] &= P_X(x),\\ \mathbb{V}\text{ar}\left[\mathbb{I}_{[X_k=x]}\right] &= \mathbb{E}\left[\left(\mathbb{I}_{[X_k=x]}\right)^2\right] - \left(\mathbb{E}\left[\mathbb{I}_{[X_k=x]}\right]\right)^2\\ &= P_X(x) - P_X(x)^2\\ &= P_X(x)(1-P_X(x)).\end{aligned}$$

Applying Weak Law of Large Numbers on $\{\pi(x|X^{(n)})\}$, we get the required bound. □

Definition 3.7 (***Typical Set***)**.** The set of ε-typical n-sequences (simply typical set) for a random variable $X \sim P_X$, $n \in \mathbb{Z}_+$ and $\varepsilon \in (0,1)$ 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 $P_X(x) = 0$ cannot be a part of typical sequence. To see why, suppose on contrary 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 $g : \mathcal{X} \to \mathbb{R}$, we have*

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

Proof. Since $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}$, we have

$$(1-\varepsilon)P_X(x) \leq \pi(x|x^{(n)}) \leq (1+\varepsilon)P_X(x).$$

Summing the above inequality over $\mathcal{X}$ and multiplying by $g(x)$, we get

$$\begin{aligned}&\sum_{x\in\mathcal{X}}(1-\varepsilon)P_X(x)g(x) \leq \sum_{x\in\mathcal{X}}\pi(x|x^{(n)})g(x) \leq \sum_{x\in\mathcal{X}}(1+\varepsilon)P_X(x)g(x)\\ \implies\ &(1-\varepsilon)\mathbb{E}[g(X)] \leq \sum_{x\in\mathcal{X}}\pi(x|x^{(n)})g(x) \leq (1+\varepsilon)\mathbb{E}[g(X)],\end{aligned}$$

where we used the nonnegativity of g to retain the inequality on mutliplication. We obtain the result by noting that

$$\sum_{x\in\mathcal{X}}\pi(x|x^{(n)})g(x) = \frac{1}{n}\sum_{k=1}^{n} g(x_k),$$

where x_k is the k-th element of the sequence $x^{(n)}$. □

Theorem 3.10 (***Properties of Typical Sequence***)**.** *Let $X^{(n)}$ have i.i.d. entries with $X_i \sim P_X$ and suppose $\varepsilon > 0$.*

(a) All typical sequences are essentially equiprobable: If $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$, then

$$2^{-n(1+\varepsilon)H(X)} \leq \mathbb{P}[x^{(n)}] \leq 2^{-n(1-\varepsilon)H(X)},$$

where we define $\mathbb{P}[x^{(n)}] = \prod_{i=1}^{n} P_X(x_i)$.

(b) Almost all probability mass is in the typical set:

$$\mathbb{P}\left[X^{(n)} \notin \mathcal{T}_\varepsilon^{(n)}\right] \leq \left(\frac{1}{n}\right).$$

(c) Bounds on cardinality of typical set:

$$\left(1-\frac{1}{n}\right)2^{n(1-\varepsilon)H(X)} \leq \left|\mathcal{T}_\varepsilon^{(n)}\right| \leq 2^{n(1+\varepsilon)H(X)}.$$

4 Lossless Source Coding

Problem (Refer Fig. 1)**.** The source sequence $X^{(n)}$ is encoded into an index M at rate R bits per source symbol, and the receiver decodes the index to find the estimate $\hat{X}^{(n)}$ of the source sequence. The lossless source coding problem is to find the lowest compression rate in bits per source symbol such that the probability of decoding error decays asymptotically to zero with the code block length n.

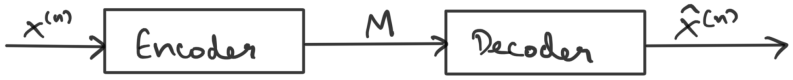


Figure 1: Point-to-point compression system.

Definition 4.1. A ***discrete memoryless source (DMS)***, denoted by $(\mathcal{X}, P_X)$ and informally referred to as X, consists of

- a finite alphabet $\mathcal{X}$ and
- a probability mass function P_X over $\mathcal{X}$.

The source is *stationary* and *memoryless* in the sense that it is time-invariant and generates i.i.d. random symbols $\{X_i\}$ with $X_i \sim P_X$.

Remark 4.2. The prefix "discrete memoryless" refers to "finite-alphabet, time-invariant and memoryless".

Definition 4.3. A $(2^{nR}, n)$ ***lossless source code*** of rate R bits per source symbol consists of

- an encoder $m : \mathcal{X}^n \to [2^{nR}]$ that assigns an index $m(x^{(n)})$, a codeword of length nR bits, to each source n-sequence $x^{(n)}$, and
- a decoder $\hat{x}^{(n)} : [2^{nR}] \to \mathcal{X}^n \cup \{e\}$ that assigns an estimate $\hat{x}^{(n)}(m) \in \mathcal{X}^n$ or an error message e to each index $m \in [2^{nR}]$.

Definition 4.4. The ***probability of decoding error*** for a $(2^{nR}, n)$ lossless source code is defined as $P_e^{(n)} = \mathbb{P}[\hat{X}^{(n)} \neq X^n]$. A rate R is said to be ***achievable*** if there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$. The ***optimal rate*** R^* for lossless source coding is the infimum of all achievable rates.

Remark 4.5. The lossless source coding scheme is required to be only asymptotically error-free (lossless).

Theorem 4.6 (***Fano's Inequality***)**.** *Let X, Y be two random variables on $\mathcal{X}$ such that $\hat{X}$ is an estimator of X and $X - Y - \hat{X}$. Let the error probability be defined as $P_e = \mathbb{P}[X \neq \hat{X}]$. Then*

$$H(X|Y) \leq H(X|\hat{X}) \leq H_b(P_e) + P_e log(|\mathcal{X}| - 1) \leq 1 + P_e log|\mathcal{X}|.$$

Proof. Define the indicator random variable

$$Z = \begin{cases} 1 & \text{if } \hat{X} \neq X, \\ 0 & \text{if } \hat{X} = X. \end{cases}$$

Then $P_Z(1) = P_e$ and $H(Z) = H_b(P_e)$. Now we use the chain rule to derive the following:

$$\begin{aligned}
H(X, Z|\hat{X}) &= H(X|\hat{X}) + \underbrace{H(Z|X, \hat{X})}_{=0} \\
&= H(Z|\hat{X}) + H(X|\hat{X}, Z) \\
&\leq H(Z) + H(X|\hat{X}, Z) \\
&= H_b(P_e) + P_Z(0)\underbrace{H(X|\hat{X}, Z=0)}_{=0} + P_Z(1)H(X|\hat{X}, Z=1) \\
&\leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1) && \text{(since } X \neq \hat{X}\text{)} \\
\implies H(X|\hat{X}) &\leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1).
\end{aligned}$$

The upper bound is obtained by upper bounding $H_b(P_e)$ and $\log(|\mathcal{X}| - 1)$. Using the DPI for mutual information, we get the lower bound. □

Theorem 4.7 (***Lossless Source Coding Theorem***)**.** *The optimal rate for lossless source coding of a discrete memoryless source X is*

$$R^* = H(X).$$

Proof Sketch. To prove this theorem, we need to verify the following two statements:

- **Achievability.** For every $R > R^* = H(X)$ there exists a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$.
- **Converse.** For every sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$, the source coding rate $R \geq R^* = H(X)$.

Proof. **(Achievability).** For $\varepsilon > 0$, let $R = (1 + \varepsilon)H(X)$. So for a given n our coding scheme maps n symbols into a $\mathcal{M}$ such that $|\mathcal{M}| = 2^{nR} = 2^{n(1+\varepsilon)H(X)}$. By the upper bound on the cardinality of the typical set in Thm. 3.10 (c), we can uniquely map each typical sequence in $\mathcal{T}_\varepsilon^{(n)}$ to an index in $\mathcal{M}$. Therefore, we define the typicality coding scheme:

- (Encoding). Assign a unique index $m(x^{(n)})$ to each $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}$. Assign $m = 1$ to all $x^{(n)} \notin \mathcal{T}_\varepsilon^{(n)}$.
- (Decoding). Upon receiving the index m, the decoder returns the typical sequence corresponding to m, i.e., return $\hat{x}^{(n)} = x^{(n)}(m)$ for the unique $x^{(n)}(m) \in \mathcal{T}_\varepsilon^{(n)}$.

All typical sequences are recovered error-free. Thus, $P_\varepsilon^{(n)} = \mathbb{P}[X^{(n)} \notin \mathcal{T}_\varepsilon^{(n)}]$ which vanishes to zero as $n \to \infty$ by Thm. 3.10 (b). This completes the achievability proof.
(Converse). Given a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_\varepsilon^{(n)} = 0$, let M be a random variable corresponding to the index generated by the encoder. Note that $X^{(n)} - M - \hat{X}^{(n)}$ forms a Markov chain, by Fano's inequality, we have

$$H(X^{(n)}|M) \leq H(X^{(n)}|\hat{X}^{(n)}) \leq 1 + P_e \log|\mathcal{X}^n| \leq 1 + nP_e \log|\mathcal{X}|.$$

Now consider

$$\begin{aligned}
nR &\geq H(M) \\
&= I(M; X^{(n)}) && \text{(since } H(X^{(n)}|M) = 0\text{)} \\
&= nH(X) - H(X^{(n)}|M) \\
&\geq nH(X) - 1 - nP_e \log|\mathcal{X}| \\
\implies R &\geq H(X) - \frac{1}{n} - P_\varepsilon^{(n)} \log|\mathcal{X}|.
\end{aligned}$$

By taking $n \to \infty$, we conclude that $R \geq H(X)$. □

5 Joint Typicality

Definition 5.1 (***Joint Type***)**.** Let $(x^{(n)}, y^{(n)})$ be 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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$ 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 and similarly for Y, i.e.,

$$\sum_{y \in \mathcal{Y}} \pi(x, y|x^{(n)}, y^{(n)}) = \pi(x|x^{(n)}), \qquad \sum_{x \in \mathcal{X}} \pi(x, y|x^{(n)}, y^{(n)}) = \pi(y|y^{(n)}).$$

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

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

Theorem 5.4 (***Marginal Typicality***)**.** *Let $(x^{(n)}, y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X, Y)$ and $P(x^{(n)}, y^{(n)}) = \prod_{i=1}^n P_{X,Y}(x_i, y_i)$. Then*

(a) $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$ and $y^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(Y)$,

(b) $P(x^{(n)}) \stackrel{\circ}{=} 2^{-nH(X)}$ and $P(y^{(n)}) \stackrel{\circ}{=} 2^{-nH(Y)}$,

(c) $P(x^{(n)}|y^{(n)}) \stackrel{\circ}{=} 2^{-nH(X|Y)}$ and $P(y^{(n)}|x^{(n)}) \stackrel{\circ}{=} 2^{-nH(Y|X)}$,

(d) $P(x^{(n)}, y^{(n)}) \stackrel{\circ}{=} 2^{-nH(X,Y)}$,

(e) $2^{n(1-\varepsilon)H(X,Y)} \leq |\mathcal{T}_\varepsilon^{(n)}(X, Y)| \leq 2^{n(1+\varepsilon)H(X,Y)}$,

where we define $P(\cdot) \stackrel{\circ}{=} 2^{-nH(\cdot)}$ to be $2^{-n(1+\varepsilon)H(\cdot)} \leq P(\cdot) \leq 2^{-n(1-\varepsilon)H(\cdot)}$.

Proof. **(a)** Observe that

$$\begin{aligned} &(1-\varepsilon)P_{X,Y}(x, y) \leq \pi(x, y|x^{(n)}, y^{(n)}) \leq (1+\varepsilon)P_{X,Y}(x, y) \\ \implies &\sum_{y \in \mathcal{Y}} (1-\varepsilon)P_{X,Y}(x, y) \leq \sum_{y \in \mathcal{Y}} \pi(x, y|x^{(n)}, y^{(n)}) \leq \sum_{y \in \mathcal{Y}} (1+\varepsilon)P_{X,Y}(x, y) \\ \implies &(1-\varepsilon)P_X(x) \leq \pi(x|x^{(n)}) \leq (1+\varepsilon)P_X(x), \end{aligned}$$

in other words, $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$. Similarly, $y^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(Y)$.

□

Theorem 5.5 (***Conditional Typicality Lemma***)**.** *Let $(X, Y) \sim P_{X,Y}$. Suppose $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$ and $Y^{(n)} \sim P_{Y^{(n)}|X^{(n)}=x^{(n)}} = \prod_{i=1}^n P_{Y|X=x_i}$. Then, for every $\varepsilon > \varepsilon'$,*

$$\lim_{n \to \infty} \mathbb{P}[(x^{(n)}, Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X, Y)] = 1.$$

Remark 5.6 (***Need for both*** ε ***and*** ε')**.** Let (X, Y) be a pair of independent Bern(1/2) random variables. Let $k = \lfloor (n/2)(1+\varepsilon) \rfloor$ and $x^{(n)}$ be a binary sequence with k ones followed by $(n-k)$ zeros. Then $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$. Let $Y^{(n)}$ be an i.i.d. Bern(1/2) sequence, independent of $x^{(n)}$. Then

$$\mathbb{P}[(x^{(n)}, Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)] \leq \mathbb{P}\left[\sum_{i=1}^{k} Y_i < (k+1)/2\right] \overset{n\to\infty}{\longrightarrow} \frac{1}{2}.$$

Thus, the fact that $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$ and $Y^{(n)} \sim P_{Y^{(n)}|X^{(n)}=x^{(n)}}$ does not necessarily imply that $\mathbb{P}[(x^{(n)}, Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)]$.

Theorem 5.7 (***Conditionally Typical Set***)**.** *Let* $(X, Y) \sim P_{X,Y}$ *and define*

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

Then for every $x^{(n)} \in \mathcal{X}^n$,

$$|\mathcal{T}_\varepsilon^{(n)}(Y|x^{(n)})| \leq 2^{n(H(Y|X)+\delta(\varepsilon))}$$

for some $\delta(\varepsilon) \to 0$ *and* $\varepsilon \to 0$. *Additionally, if* $x^{(n)} \in \mathcal{T}_{\varepsilon'}^{(n)}(X)$ *with* $\varepsilon' < \varepsilon$, *then*

$$|\mathcal{T}_\varepsilon^{(n)}(Y|x^{(n)})| \geq [1 - \mathcal{O}(1/n)]2^{n(H(Y|X)-\delta(\varepsilon))}, \quad n \to \infty.$$

Remark 5.8. The upper bound on the conditionally typical set holds for every $x^{(n)} \in \mathcal{X}^n$ and not just for $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$.

Theorem 5.9 (***Joint Typicality Lemma***)**.** *Let* $(X, Y) \sim P_{X,Y}$. *Suppose* $x^{(n)} \in \mathcal{T}_{\varepsilon'}^{(n)}(X)$ *and* $\bar{Y}^{(n)} \sim P_{\bar{Y}^{(n)}} = \prod_{i=1}^n P_Y$ *(instead of* $\prod_{i=1}^n P_{Y|X=x_i}$*). Then, for every* $\varepsilon > \varepsilon'$,

$$(1 - \mathcal{O}(1/n))2^{-n(I(X;Y)+\delta(\varepsilon))} \leq \mathbb{P}[(x^{(n)}, \bar{Y}^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)] \leq 2^{-n(I(X;Y)-\delta(\varepsilon))},$$

where $\delta(\varepsilon) \to 0$ *as* $\varepsilon \to 0$. *Furthermore, let* $\bar{X}^{(n)} \sim P_{\bar{X}^{(n)}} = \prod_{i=1}^n P_X$ *and* $\bar{Y}^{(n)} \sim P_{\bar{Y}^{(n)}} = \prod_{i=1}^n P_Y$. *Then*

$$\mathbb{P}[(\bar{X}^{(n)}, \bar{Y}^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)] \leq 2^{-n(I(X;Y)-\delta(\varepsilon))},$$

where $\delta(\varepsilon) \to 0$ *as* $\varepsilon \to 0$.

6 Channel Coding

Problem (Refer Fig. 2)**.** The sender encodes the message M into a codeword $X^{(n)}$ and transmits it over the channel in n time instances. Upon receiving the noisy sequence $Y^{(n)}$, the receiver decodes it to obtain the estimate $\hat{M}$ of the message. The channel coding problem is to find the channel capacity, which is the highest rate R such that the probability of decoding error can be made to decay asymptotically to zero with the code block length n.

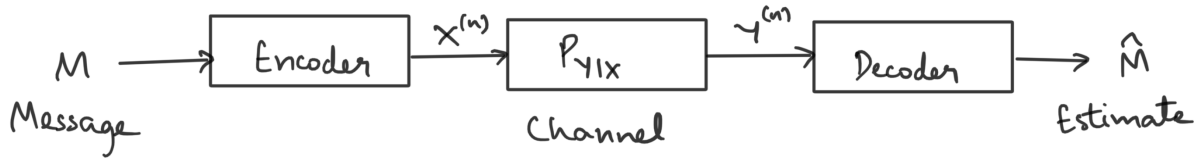


Figure 2: Point-to-point communication system.

Definition 6.1. A ***discrete memoryless channel (DMC)***, denoted by $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ and informally referred to as $P_{Y|X}$, consists of

- a finite input alphabet $\mathcal{X}$,
- a finite output alphabet $\mathcal{Y}$ and
- a collection of conditional probability mass functions $P_{Y|X=x}$ over $\mathcal{Y}$ for every $x \in \mathcal{X}$.

The channel is *stationary* (time-invariant) and *memoryless* in the sense that when it is used n times with message M drawn from an arbitrary set and input $X^{(n)} \in \mathcal{X}^n$, the output $Y_i \in \mathcal{Y}$ at time $i \in [n]$ given $(M, X^{(i)}, Y^{(i-1)})$ is distributed according to

$$P_{Y_i|X^{(i)},Y^{(i-1)},M}(y_i|x^{(i)}, y^{(i-1)}, m) = P_{Y|X}(y_i|x_i), \quad i \in [n].$$

Equivalently, memoryless property induces the following Markov chain:

$$(X^{(i-1)}, Y^{(i-1)}, M) - X_i - Y_i.$$

Definition 6.2. A $(2^{nR}, n)$ ***channel coding scheme*** (simply code) for a DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ consists of

- a message set $[2^{nR}]$,
- an encoder $m : [2^{nR}] \to \mathcal{X}^n$ that assigns a codeword $x^{(n)}(m)$ to each $m \in [2^{nR}]$, and
- a decoder $\hat{m} : \mathcal{Y}^n \to [2^{nR}] \cup \{e\}$ that assigns an estimate $\hat{m} \in [2^{nR}]$ or an error message e to each received sequence $y^{(n)}$.

Definition 6.3. We say that a DMC is used ***without feedback*** if

$$P_{X_i|X^{(i-1)},Y^{(i-1)},M}(x_i|x^{(i-1)}, y^{(i-1)}, m) = P_{X_i|X^{(i-1)}}(x_i|x^{(i-1)}), \quad i \in [n].$$

Lemma 6.4. *If a DMC is used without feedback, then*

$$P_{Y^{(n)}|X^{(n)},M}(y^{(n)}|x^{(n)}, m) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i), \quad \forall\ n \geq 1.$$

Equivalently, no feedback induces $Y_i - (M, X^{(i)}) - X^{(i+1:n)}$.

Proof. Consider the following joint distribution

$$\begin{aligned}
&P_{X^{(n)},Y^{(n)},M}(x^{(n)}, y^{(n)}, m) \\
&\quad = P_M(m) \prod_{i=1}^{n} P_{X_i|X^{(i-1)},Y^{(i-1)},M}(x_i|x^{(i-1)}, y^{(i-1)}, m) P_{Y_i|X^{(i)},Y^{(i-1)},M}(y_i|x^{(i)}, y^{(i-1)}, m) \\
&\quad = P_M(m) \prod_{i=1}^{n} P_{X_i|X^{(i-1)},M}(x_i|x^{(i-1)}, m) P_{Y|X}(y_i|x_i) = P_{X^{(n)},M}(x^{(n)}, m) \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).
\end{aligned}$$

But, $P_{X^{(n)},Y^{(n)},M}(x^{(n)}, y^{(n)}, m) = P_{X^{(n)},M}(x^{(n)}, m) P_{Y^{(n)}|X^{(n)},M}(y^{(n)}|x^{(n)}, m)$. □

Assumption 1. *The message is uniformly distributed over the message set, i.e.,* $P_M(m) = \dfrac{1}{2^{nR}}$.

Definition 6.5. The set $\mathcal{C} = \{x^{(n)}(1), x^{(n)}(2), \ldots, x^{(n)}(2^{nR})\}$, which is the set of all codewords, is referred to as the ***codebook*** associated with the $(2^{nR}, n)$ code. Define $\lambda_m(\mathcal{C})$ to be the conditional probability of error given that message m is sent, i.e.,

$$\lambda_m(\mathcal{C}) = \mathbb{P}[\hat{M} \neq m | M = m] = \sum_{y^{(n)} \in \mathcal{Y}^n} P_{Y^{(n)}|X^{(n)}}(y^{(n)}|x^{(n)}(m)) \mathbb{I}_{[\hat{m}(y^{(n)}) \neq m]}.$$

Then the ***average probability of error*** for a $(2^{nR}, n)$ code is defined as

$$P_e^{(n)}(\mathcal{C}) = \mathbb{P}[\hat{M} \neq M] = \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \lambda_m(\mathcal{C}).$$

A rate R is said to be ***achievable*** if there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$. The ***capacity*** C of a DMC is the supremum of all achievable rates.

Theorem 6.6 (***Channel Coding Theorem***)**.** *Under Assumption 1, the capacity of the DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ without feedback is given by*

$$C = \max_{P_X} I(X;Y).$$

Proof Sketch. To prove this theorem, we need to verify the following two statements:

- **Achievability.** For every $R < C = \max_{P_X} I(X;Y)$ there exists a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$.
- **Converse.** For every sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$, the rate $R \leq C = \max_{P_X} I(X;Y)$.

Proof. **(Achievability).** We use ***random coding***. Fix the probability mass function P_X that attains the information capacity C. Randomly and independently generate 2^{nR} sequences $x^{(n)}(m), m \in [2^{nR}]$, each according to $P_{X^{(n)}} = \prod_{i=1}^n P_X$. The generated sequences constitute a random codebook $\mathcal{C}$ such that

$$\mathbb{P}[\mathcal{C}] = \prod_{m=1}^{2^{nR}} \prod_{i=1}^{n} P_X(x_i(m)).$$

The chosen codebook $\mathcal{C}$ is revealed to both the encoder and the decoder before transmission commences.

- (Encoding). To send a message $m \in [2^{nR}]$, transmit $x^{(n)}(m)$.
- (Decoding). We use ***joint typicality decoding***. Let $y^{(n)}$ be the received sequence. The receiver declares that $\hat{m}(y^{(n)}) \in [2^{nR}]$ is sent if it is the unique message such that

  $$(x^{(n)}(\hat{m}), y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y).$$

  Otherwise - if there is none or more than one such message - it declares an error e.

Assuming the message m is sent, the decoder makes an error if $(x^{(n)}(m), y^{(n)}) \notin \mathcal{T}_\varepsilon^{(n)}(X,Y)$ or if there is another message $m' \neq m$ such that $(x^{(n)}(m'), y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)$. Consider the

average probability of error averaged over M (as defined above) and averaged over the ensemble of codebooks

$$\begin{aligned}
\mathbb{P}[\hat{M} \neq M] &= \mathbb{E}_{\mathcal{C}}[P_e^{(n)}(\mathcal{C})] \\
&= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}} \lambda_m(\mathcal{C})\right] \\
&= \frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}} \mathbb{E}_{\mathcal{C}}\left[\lambda_m(\mathcal{C})\right] \\
&\overset{(*)}{=} \frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}} \mathbb{E}_{\mathcal{C}}\left[\lambda_1(\mathcal{C})\right] \\
&= \mathbb{E}_{\mathcal{C}}[\lambda_1(\mathcal{C})] \\
&= \mathbb{P}[\hat{M} \neq M | M = 1] \\
&= \mathbb{P}[\hat{M} \neq 1 | M = 1],
\end{aligned}$$

where $(*)$ follows from the fact that the conditional probability of error averaged over the ensemble of codebooks does not depend on the message we conditioned on. Hence, we can assume without loss of generality that the message $M = 1$ was sent. The decoder makes an error iff one or both of the following events occur:

$$\begin{aligned}
\mathcal{E}_1 &= \{(X^{(n)}(1), Y^{(n)}) \notin \mathcal{T}_\varepsilon^{(n)}(X,Y)\}, \\
\mathcal{E}_2 &= \{(X^{(m)}(1), Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y) \text{ for some } m \neq 1\},
\end{aligned}$$

Thus, by union bound,

$$\mathbb{P}[\hat{M} \neq M] = \mathbb{P}[\hat{M} \neq 1 | M = 1] = \mathbb{P}[\mathcal{E}_1 \cup \mathcal{E}_1] \leq \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2].$$

We now bound each term. Note that

$$Y^{(n)} \sim \prod_{i=1}^{n} P_{Y|X=x_i(1)}, \text{ i.e., } (X^{(n)}(1), Y^{(n)}) \sim \prod_{i=1}^{n} P_{X,Y}.$$

Then by Theorem 3.10(b), $\mathbb{P}[\mathcal{E}_1]$ decays to zero as $n \to \infty$. For the second term, since for $m \neq 1$,

$$(X^{(n)}(m), Y^{(n)}) \sim \prod_{i=1}^{n} P_X P_Y,$$

by joint typicality lemma, for every $m \neq 1$, we have

$$\mathbb{P}[(X^{(n)}(m), Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)] \leq 2^{-n(I(X;Y)-\delta(\varepsilon))} = 2^{-n(C-\delta(\varepsilon))}.$$

Using the union bound once more, we get

$$\mathbb{P}[\mathcal{E}_2] \leq \sum_{m=2}^{2^{nR}} \mathbb{P}[(X^{(n)}(m), Y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X,Y)] \leq 2^{-n(C-R-\delta(\varepsilon))},$$

which decays to zero as $n \to \infty$, as long as $R < C - \delta(\varepsilon)$.

To summarize, we have just shown that the average error probability, averaged over all codebooks, vanishes if $R < C$. This implies that there must exists a sequence of codes $(2^{nR}, n)$ with $P_e^{(n)} \to 0$ as $n \to \infty$. This proves that $R < C$ is achievable.

(Converse). Given a sequence of $(2^{nR}, n)$ codes with $\lim_{n\to\infty} P_e^{(n)} = 0$, let M be the message sent. Then

$$
\begin{aligned}
nR &= H(M) \\
&= I(M; Y^{(n)}) + H(M|Y^{(n)}) \\
&= I(M; Y^{(n)}) + H(M|Y^{(n)}, \hat{M}) && \text{(since } M \perp\!\!\!\perp \hat{M}|Y^{(n)}\text{)} \\
&\leq I(M; Y^{(n)}) + H(M|\hat{M}) \\
&\leq I(M; Y^{(n)}) + 1 + nRP_e^{(n)} && \text{(by Fano's inequality)} \\
&= \sum_{i=1}^{n} I(M; Y_i|Y^{(i-1)}) + 1 + nRP_e^{(n)} \\
&= \sum_{i=1}^{n} I(M, Y^{(i-1)}; Y_i) - \underbrace{I(Y^{(i-1)}; Y_i)}_{\leq 0} + 1 + nRP_e^{(n)} \\
&\leq \sum_{i=1}^{n} I(M, Y^{(i-1)}; Y_i) + 1 + nRP_e^{(n)} \\
&= \sum_{i=1}^{n} I(X_i, M, Y^{(i-1)}; Y_i) - \underbrace{I(X_i; Y_i|M, Y^{(i-1)})}_{\leq 0} + 1 + nRP_e^{(n)} \\
&\leq \sum_{i=1}^{n} I(X_i, M, Y^{(i-1)}; Y_i) + 1 + nRP_e^{(n)} \\
&= \sum_{i=1}^{n} I(X_i; Y_i) + I(M, Y^{(i-1)}; Y_i|X_i) + 1 + nRP_e^{(n)} \\
&= \sum_{i=1}^{n} I(X_i; Y_i) + 1 + nRP_e^{(n)} && \text{(for a DMC } (M, Y^{(i-1)}) - X_i - Y_i \text{ holds)} \\
&\leq nC + 1 + nRP_e^{(n)} && \text{(where } C = \max_{P_X} I(X;Y)\text{)} \\
\implies R &\leq C - \frac{1}{n} - RP_e^{(n)}.
\end{aligned}
$$

By taking $n \to \infty$, we conclude that $R \leq C$. □

7 Differential Entropy

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.

[4] Abbas El Gamal and and Young-Han Kim. Network information theory. Cambridge university press, 2011.