TUTORIAL III

1 Data processing inequality for mutual information

Recall that:

$$H(X|Y) \stackrel{\text{def}}{=} \sum_{y \in A_Y} P_Y(y) H(X|Y=y); \quad H(X,Y) = H(X) + H(Y|X); \quad \text{and} \quad I(X;Y) \stackrel{\text{def}}{=} H(X) - H(X|Y)$$

0. We know that more information cannot increase uncertainty in the sense that $H(X|Y) \leq H(X)$. Show that this is not true if we do not take the average of Y, i.e. give an example of a pair of random variables (X,Y) such that $H(X|Y=y) > H(X)$ for some y.

We define the conditional mutual information:

$$I(X;Y|Z) \stackrel{\text{def}}{=} H(X|Z) - H(X|Y,Z)$$

If X and Z are conditionally independent given Y (i.e. $\mathbf{P}_{Z|Y,X} = \mathbf{P}_{Z|Y}$), we will use the notation $X \to Y \to Z$ (this notation is motivated by the theory of Markov chains). Notice that $X \to Y \to Z$ implies $Z \to Y \to X$ since $\mathbf{P}_{Z|Y,X} = \mathbf{P}_{Z|Y} \Rightarrow \mathbf{P}_{X|Y,Z} = \mathbf{P}_{X|Y}$.

1. Show that $I(X;Y|Z)$ is the average over Z of $I(X;Y)$, i.e.: $I(X;Y|Z) = \sum_z \mathbf{P}(Z=z) I(X|Z=z; Y|Z=z)$.
2. Show that $I(X;(Y,Z)) = I(X;Z) + I(X;Y|Z)$
3. For any $X \to Y \to Z$, show that the conditional mutual information $I(X;Z|Y)$ is 0.
4. Using question 2 and 3, show the data processing inequality: $I(X;Y) \geq I(X;Z)$ for any $X \to Y \to Z$.
5. Show that for any function g, we have $I(X;Y) \geq I(X;g(Y))$.

2 Typical sets

Let $X^n = X_1 \dots X_n$ be independent and identically distributed bits with $X_1 \sim \text{Ber}(p)$, i.e., $P_{X_1}(0) = 1-p$ and $P_{X_1}(1) = p$ (assume that $0 < p < 1/2$) and let $\delta > 0$ with $p + \delta \leq 1/2$.

1. Recall that $h_2(p) = -p \log_2 p - (1-p)\log_2(1-p)$. Show that for $k \leq n/2$ the following inequality holds:
$$1 + \binom{n}{1} + \cdots + \binom{n}{k} \leq 2^{h_2(k/n)n} \,. \tag{1}$$
2. Using the previous inequality, show that there exists a set $S_\delta \subseteq \{0,1\}^n$ with $|S_\delta| \leq 2^{n \cdot h_2(p+\delta)}$ satisfying the property that
$$\lim_{n \to \infty} \mathbf{P}[X^n \in S_\delta] = 1 \,. \tag{2}$$

3 Fixed-length almost lossless compressor: source coding theorem

Recall that a *fixed length compressor* for source $Y \in \mathcal{Y}$ of length ℓ is a function $C : \mathcal{Y} \to \{0,1\}^\ell$. It has error probability at most δ if there exists a decompressor $D : \{0,1\}^\ell \to \mathcal{Y}$ such that $\mathbf{P}[D(C(Y)) = Y] \geq 1 - \delta$. Let define

$$\ell^{opt}(Y, \delta) = \min\{\ell : \text{ there exists a length } \ell \text{ compressor for } Y \text{ with error probability } \delta\}.$$

We will prove what is usually called *Shannon Source Coding Theorem*: Let $X^n = X_1 \dots X_n$ be a sequence of independent and distributed as $X \in \mathcal{X}$. For any $\delta \in (0,1)$,

$$\lim_{n\to\infty} \frac{\ell^{opt}(X^n, \delta)}{n} = H(X). \tag{4}$$

We will first show the lower bound (also called converse) and then give another proof for the upper bound.

3.1 Converse of Shannon source coding theorem

In this section, we will show that for any $\varepsilon > 0$ and for large enough n, we have $H(X) - 2\varepsilon \leq \ell^{opt}(X^n, \delta)/n$.

Recall that we showed in class that $\ell^{opt}(Y, \delta) = \min\{\lceil \log |S| \rceil : \mathbf{P}\{Y \in S\} \geq 1 - \delta\}$.

1. Prove that we are done if we can show that: for every set $S \subseteq \mathcal{X}^n$ such that $\mathbf{P}[X^n \in S] \geq 1 - \delta$ we have $|S| \geq 2^{n(H(X)-2\varepsilon)}$.
2. Now suppose that there is $S \subseteq \mathcal{X}^n$ such that $|S| \leq 2^{\ell n}$ for some ℓ and $\mathbf{P}[X^n \in S] \geq 1 - \delta$. Prove that

$$\mathbf{P}[X^n \in S] \leq \mathbf{P}\left[-\sum_{i=1}^{n} \log P_X(X_i) \leq \ell n + \varepsilon n \right] + 2^{-\varepsilon n}.$$

3. Deduce that under the assumption of Question 2, $\ell \geq H(X) - 2\varepsilon$, and so the lower bound holds.

3.2 Achievability using random coding

Recall that in order to prove achievability of the source coding theorem, we chose the set S of correctly encoded symbols to be the set of $x^n \in \mathcal{X}^n$ such that $P_{X^n}(x^n) \geq 2^{-n(H(X)-\varepsilon)}$. We will now show a similar result by choosing the set S at random. In fact, we start by considering a general source (i.e., not necessarily iid) and derive an upper bound on the probability of error and will give us the desired result in the special case of an iid source.

Our objective is to show that for any source X and any integer $l \geq 0$, there exists a compressor with error probability

$$\delta \leq \mathbf{P}[-\log_2(P_X(X)) > l - \tau] + 2^{-\tau}, \ \ \forall \tau > 0. \tag{5}$$

1. Let $\tau > 0$, X be a random variable and C be a length l compressor. Let x_0 be a fixed letter of $\mathcal{X}$. Define $D = \{0,1\}^l \to \mathcal{X}$ by

$$D(y) = \begin{cases} x, & \text{if } \exists! x \in \mathcal{X} \text{ s.t. } C(x) = y \text{ and } -\log_2(P_X(x)) \leq l - \tau \\ x_0, & \text{otherwise} \end{cases} \tag{6}$$

Define also

$$J(x, C) = \{x' \in \mathcal{X} : C(x) = C(x'), x \neq x', \text{ and } -\log_2(P_X(x')) \leq l - \tau\}$$

Show that

$$\mathbf{P}[D(C(X)) \neq X] \leq \mathbf{P}[-\log_2(P_X(X)) > l - \tau] + \mathbf{P}[J(X, C) \neq \emptyset]$$

2. Let C be a random length-l compressor, that is for each $x \in \mathcal{X}$, $C(x)$ is a random bit string of length l, with each bit choosen independantly and uniformly from $\{0, 1\}$. Show that

$$\mathbb{E}_C[\mathbf{P}[J(X, C) \neq \emptyset]] \leq 2^{-\tau}$$

   where we compute the mean on the randomness of C but not on X.

3. Prove Eq. (5)

4. Can you build from the proof a length l compressor with error $\delta \leq \mathbf{P}[-\log_2(P_X(X)) > l - \tau] + 2^{-\tau}$?

5. Use Eq. (5) to give a proof of the upper bound in Shannon source coding theorem (4).