Lecture-03: Data Processing

1 Data Processing

The definitions of entropy, mutual information, and divergence all extend naturally to any finite number of random variables by treating multiple random variables as a single random vector. However, there are a few new concepts that can only be defined in terms of three random variables. Let \( X, Y, \) and \( Z \) be random variables with joint distribution \( p_{X,Y,Z}(x,y,z) \).

**Definition 1.1.** For three r.v. \( (X,Y,Z) \sim p_{X,Y,Z}(x,y,z) \) defined on \( X \times Y \times Z \), the conditional mutual information (in bits) between \( X \) and \( Y \) given \( Z \) is denoted

\[
I(X;Y|Z) \triangleq \sum_{(x,y,z) \in X \times Y \times Z} p_{X,Y,Z}(x,y,z) \log_2 \frac{p_{X,Y|Z}(x,y,z)}{p_{X|Z}(x,z)p_{Y|Z}(y,z)} = \mathbb{E} \left[ \log_2 \frac{p_{X,Y|Z}(X,Y,Z)}{p_{X|Z}(X,Z)p_{Y|Z}(Y,Z)} \right].
\]

From this, we see that \( I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \). Thus the conditioning is simply inherited by each entropy in the standard decomposition.

**Definition 1.2.** Three r.v. \( (X,Y,Z) \sim p_{X,Y,Z}(x,y,z) \) form a Markov chain \( X - Y - Z \) if

\[
p_{X,Y,Z}(x,y,z) = p_X(x)p_Y(y|x)p_Z(z|y).
\]

This is clearly the same as \( p_{Z|X,Y}(z|x,y) = p_{Z|Y}(z|y) \) for all \( x,y,z \), which is equivalent to the condition that \( X \) and \( Z \) are conditionally independent given \( Y \).

**Lemma 1.3.** Properties of mutual information for three random variables:

1. (chain rule of mutual information) \( I(X;Y,Z) = I(X;Y) + I(X;Z|Y) \).

   **Proof.** This follows from the expectation of the decomposition

   \[
   \log_2 \frac{p_{X,Y,Z}(X,Y,Z)}{p_X(x)p_Y(Y)Z(Z)} = \log_2 \frac{p_{X,Y}(X,Y)p_{Z|X,Y}(Z|X,Y)}{p_X(x)p_Y(Y)p_{Z|Y}(Z|Y)} = \log_2 \frac{p_{X,Y|(X,Z|Y)}}{p_X(x)p_Y(Y)} + \log_2 \frac{p_{X,Z|Y}(X,Z|Y)}{p_{Z|Y}(Z|Y)p_{X|Y}(X|Y)}.
   \]

2. (non-negativity of conditional mutual information) \( I(X;Y|Z) \geq 0 \) with equality iff \( X \) and \( Y \) are conditionally independent given \( Z \).

   **Proof.** First, we observe that

   \[
   I(X;Y|Z) = \sum_z p_Z(z) \mathbb{D}(p_{X,Y|Z=z} \parallel p_{X|Z=z}p_{Y|Z=z}).
   \]

   Each term in this sum is non-negative and equal to zero iff \( p_{X,Y|Z=z}(x,y) = p_{X|Z=z}(x)p_{Y|Z=z}(y) \) for all \( x,y \). Thus, the overall sum is zero iff the condition holds for all \( x,y,z \) (i.e., \( X \) and \( Y \) are conditionally independent given \( Z \)).

**Theorem 1.4 (Data Processing Inequality).** If three r.v. \( (X,Y,Z) \sim p_{X,Y,Z}(x,y,z) \) form a Markov chain \( X - Y - Z \), then \( I(X;Z) \leq I(X;Y) \). For example, if \( Z = f(Y) \) is a function of \( Y \), then \( X - Y - Z \) form a Markov chain.
Proof. Applying the chain rule of mutual information in the two possible orders gives
\[ I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y). \]
Since \( X - Y - Z \) form a Markov chain, \( X \) and \( Z \) are conditionally independent and \( I(X;Z|Y) = 0 \). Thus, we have
\[ I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z). \]
If \( Z = f(Y) \), then \( p_{Z|X,Y}(z|x,y) = \mathbb{1}_{\{z=f(y)\}} = p_{Z|Y}(z,y) \) and \( X - Y - Z \) form a Markov chain. \( \square \)

**Example 1.5.** A system has a random state \( X \) and an experiment with outcome \( Y \) is performed to measure that state. Is it possible that additional processing can produce a new output \( Z = f(Y) \) such that \( H(X|Z) < H(X|Y) \)?

**Theorem 1.6 (Fano’s Inequality).** Let the r.v. \( Y \) be an observation of the r.v. \( X \) and \( \hat{X} = f(Y) \) be an estimate of \( X \). Then, the error probability \( P_e = P(\hat{X} \neq X) \) satisfies
\[ H(P_e) + P_e \log_2(|X| - 1) \geq H(X|Y). \]

*Proof.* Let \( E = \mathbb{1}_{\{\hat{X} = X\}} \) be an indicator r.v. for the error event. Expanding the conditional entropy \( H(E, X|\hat{X}) \) in two ways gives
\[ H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}). \]
Now \( H(X|\hat{X}) \geq H(X|Y) \) by data processing inequality, since \( X - Y - \hat{X} \) form a Markov chain, and \( H(E|X, \hat{X}) = 0 \) since \( E = \mathbb{1}_{\{\hat{X} \neq X\}} \). Further, \( H(E|\hat{X}) \leq H(E) \leq H(P_e) \) since the conditioning reduces entropy. In addition, \( H(X|E = 0, \hat{X}) = 0 \), and we can write
\[ H(X|E = 1, \hat{X}) \leq H(X \neq \hat{X}) \leq \log_2(|X| - 1). \]
This implies that \( H(X|E, \hat{X}) \leq P_e \log_2(|X| - 1) \). Rearranging these terms gives the stated result. \( \square \)

## 2 Sequences of random variables

Let \( (X_t : t \in \mathbb{N}) \) be a random process where each random variable lies in \( \mathcal{X} \). The joint probability distribution of the first \( N \) random variables is denoted \( P_N(x_1, \ldots, x_N) \). Let \( [N] \triangleq \{1, 2, \ldots, N\}, \mathcal{A} \subseteq [N] \), and \( \bar{A} = [N] \setminus \mathcal{A} \) be sets of indices. We will denote subvectors with indices in \( \mathcal{A} \) and \( \bar{A} \) by
\[ x_A = (x_t : t \in \mathcal{A}), \quad x_{\bar{A}} = (x_t : t \in \bar{A}). \]
The marginal distribution of variables in \( \mathcal{A} \) is given by summing over all variables in \( \bar{A} \):
\[ P_{\mathcal{A}}(x_{\mathcal{A}}) = \sum_{x_{\bar{A}}} P_N(x_1, \ldots, x_N). \]

**Definition 2.1.** The entropy rate of a random process is defined to be
\[ h_X \triangleq \lim_{N \to \infty} \frac{1}{N} H(X_1, X_2, \ldots, X_N), \]
if the limit exists.

**Example 2.2.** If the random variables are drawn i.i.d. according to \( p(x) \), then
\[ P_N(x_1, \ldots, x_N) = \prod_{t=1}^N p(x_t). \]
In this case, \( H(X_1, \ldots, X_N) = NH(p) \) and the entropy rate is \( h_X = H(p) \).
Example 2.3. If the random variables form a homogenous Markov chain, then

$$P_N(x_1, \ldots, x_N) = p_1(x_1) \prod_{t=1}^{N} w(x_t \to x_{t+1}),$$

where $p_1(x)$ is the distribution of the initial state and $w(x \to x') = p_{X_{t+1} \mid X_t}(x' \mid x)$ defines the transition probabilities of the chain. In this case, the entropy rate is given by

$$h_X = \lim_{N \to \infty} \frac{1}{N} H(X_1, X_2, \ldots, X_N) = \lim_{N \to \infty} \frac{1}{N} \left( H(X_1) + \sum_{t=1}^{N-1} H(X_{t+1} \mid X_t) \right)$$

$$= \sum_{x \in \mathcal{X}} \left( \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N-1} p_t(x) \sum_{x' \in \mathcal{X}} w(x \to x') \log_2 \frac{1}{w(x \to x')} \right)$$

$$= \sum_{x \in \mathcal{X}} p^*(x) \sum_{x' \in \mathcal{X}} w(x \to x') \log_2 \frac{1}{w(x \to x')}',$$

where the last step assumes that $w(x \to x')$ was chosen so that the limiting occupancy distribution $p^*(x) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N-1} p_t(x)$ exists and is independent of the initial state distribution.