Lecture 1: Information Measures

Course: Network Information Theory (Fall 2016)
Dpt. of Computer Science.
École Normale Supérieure de Lyon (ENS de Lyon)
Samir M. Perlaza

1 Notation

Let X be a random variable taking values in the set $\mathcal{X}$.

- The probability mass function (pmf) of X is $P_X : \mathcal{X} \rightarrow [0,1]$.
- The set of all possible pmfs on $\mathcal{X}$ is $\triangle(\mathcal{X})$, i.e., $P_X \in \triangle(\mathcal{X})$.
- The support of P_X is $\text{supp}(P_X) = \{x \in \mathcal{X} : P_X(x) > 0\}$.

Let Y be a second random variable

- The joint pmf of X and Y is $P_{XY} : \mathcal{X} \times \mathcal{Y} \rightarrow [0,1]$.
- The conditional pmf of Y conditioning on X is $P_{Y|X} : \mathcal{X} \times \mathcal{Y} \rightarrow [0,1]$.

2 Preliminaries

- **Definition 1 (Strict Positiveness)** *The pmf P_X is strictly positive if:*

$$\forall x \in \mathcal{X}, \quad P_X(x) > 0. \tag{1}$$

- **Definition 2 (Absolute Continuity)** *Given two probability measures P and Q defined on a measurable space $(\mathcal{X}, \mathcal{F}(\mathcal{X}))$, P is absolutely continuous with respect to Q ($P << Q$) if*

$$\forall \mathcal{A} \in \mathcal{F}(\mathcal{X}) : Q(\mathcal{A}) = 0 \quad \textit{implies} \quad P(\mathcal{A}) = 0. \tag{2}$$

- If $P << Q$ then supp $(Q) \subseteq$ supp (P)
- If $P << Q$ and $Q << P$ then supp $(Q) =$ supp (P)
- **Definition 3 (Independence)** *The random variables* $X_1, X_2, \ldots, X_n$ *are mutually independent if* $\forall (x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$:

$$P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = \prod_{t=1}^{n} P_{X_t}(x_t). \tag{3}$$

- **Definition 4 (Conditional Independence)** *The random variables* X *and* Z *are mutually independent conditioning on* Y *if* $\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$:

$$P_{XYZ}(x, y, z) = \begin{cases} P_{X|Y}(x|y) P_{Z|Y}(z|y) P_Y(y) & \textit{if} \quad P_Y(y) > 0 \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

- **Definition 5 (Markov Chain)** *The random variables* $X_1, X_2, \ldots, X_n$ *form a Markov chain, notation* $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$, *if* $\forall (x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$:

$$P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) =$$
$$\begin{cases} P_{X_1}(x_1) P_{X_2|X_1}(x_2|x_1) P_{X_3|X_2}(x_3|x_2) \cdots P_{X_n|X_{n-1}}(x_n|x_{n-1}) & \textit{if} \quad \prod_{t=1}^{n} P_{X_t}(x_t) > 0 \\ 0 & \textit{otherwise} \end{cases} \tag{5}$$

- $X \rightarrow Y \rightarrow Z$ implies conditional independence between X and Z given Y.
- $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ implies $X_n \rightarrow X_{n-1} \rightarrow \cdots \rightarrow X_1$.
- **Proposition 1 (Markov Subchains)** *Let* $\mathcal{N} = \{1, 2, \ldots, n\}$ *be a finite set with a partition* $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_m$, *with* $n > 2$ *and* $m > 2$. *Assume that for all* $(r, s) \in \mathcal{N}_i \times \mathcal{N}_j$, *with* $i > j$, *it holds that* $r > s$. *Let also* $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ *form a Markov chain and consider the random variable* $\boldsymbol{Z}_{\mathcal{N}_i} = \left(X_{k_1}, X_{k_2}, \ldots, X_{k_{|\mathcal{N}_i|}}\right)$, *with* $k_\ell \in \mathcal{N}_i$ *for all* $\ell \in \mathcal{N}_i$. *Then, the random variables* $\boldsymbol{Z}_{\mathcal{N}_1}, \boldsymbol{Z}_{\mathcal{N}_2}, \ldots, \boldsymbol{Z}_{\mathcal{N}_m}$, *form the following Markov chain:*

$$\boldsymbol{Z}_{\mathcal{N}_1} \rightarrow \boldsymbol{Z}_{\mathcal{N}_2} \rightarrow \cdots \rightarrow \boldsymbol{Z}_{\mathcal{N}_m}. \tag{6}$$

  Moreover, for a fixed $2 < q \leqslant m$,

$$\boldsymbol{Z}_{\mathcal{N}'_1} \rightarrow \boldsymbol{Z}_{\mathcal{N}'_2} \rightarrow \cdots \rightarrow \boldsymbol{Z}_{\mathcal{N}'_q}, \tag{7}$$

  is also a Markov chain, where $\mathcal{N}'_i \subseteq \mathcal{N}_i$, *for all* $i \in \{1, 2, \ldots, q\}$.

- **Example 1 (Markov Subchains)** *Let* $X_1 \to X_2 \to \cdots \to X_6$ *form a Markov chain. Then, the following are also Markov chains:*

$$X_1 \to X_2 \to X_3 \tag{8}$$
$$X_2 \to X_4 \to X_6 \tag{9}$$
$$(X_1, X_2) \to X_3 \to X_4 \tag{10}$$
$$X_1 \to (X_3, X_4) \to X_6 \tag{11}$$

- **Definition 6 (Variational Distance)** *Let* P *and* Q *be two probability measures on the mesurable space* $(\mathcal{X}, \mathcal{F}(\mathcal{X}))$. *The variational distance between* P *and* Q, *denoted by* $V(P, Q)$ *is*
$$V(P, Q) = 2 \sup_{\mathcal{A} \in \mathcal{F}(\mathcal{X})} |P(\mathcal{A}) - Q(\mathcal{A})| \tag{12}$$

- **Proposition 2** *Let* P, Q *and* S *be three probability measures on the mesurable space* $(\mathcal{X}, \mathcal{F}(\mathcal{X}))$. *Then, the following holds*

$$V(P,Q) \geq 0 \quad \textit{[positiveness]} \tag{13a}$$
$$V(P,Q) = V(Q,P) \quad \textit{[symmetry]} \tag{13b}$$
$$V(P,S) \leqslant V(P,Q) + V(Q,S) \quad \textit{[triangle inequality]} \tag{13c}$$

- **Proposition 3** *Let* P *and* Q *be two probability measures on the measurable space* $(\mathcal{X}, \mathcal{F}(\mathcal{X}))$. *Then,* $V(P, Q) = 0$ *if and only if* P *and* Q *are identical.*

- From Proposition 2 and Proposition 3, total variation is a distance in the formal sense.

- **Proposition 4** *Let* P *and* Q *be two probability measures on the measurable space* $(\mathcal{X}, \mathcal{F}(\mathcal{X}))$, *with* $\mathcal{X}$ *a countable set. Then,*
$$V(P, Q) = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \tag{14}$$

- **Proposition 5** *Let* P_{XY} *and* Q_{XY} *be two joint probability measures on the measurable space* $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}(\mathcal{X} \times \mathcal{Y}))$. *Let* P_X *and* Q_X *be the marginals of* P_{XY} *and* Q_{XY}, *respectively. Then,*
$$V(P_X, Q_X) \leqslant V(P_{XY}, Q_{XY}). \tag{15}$$

- **Proposition 6** *Let P_{XY} and Q_{XY} be two joint probability measures on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}(\mathcal{X} \times \mathcal{Y}))$. Let also P_X and Q_X be the marginals of P_{XY} and Q_{XY}, respectively and assume that*

$$P_{XY} = P_X P_{Y|X}, \text{ and} \tag{16}$$
$$Q_{XY} = Q_X P_{Y|X}. \tag{17}$$

Then,

$$V(P_X, Q_X) = V(P_{XY}, Q_{XY}). \tag{18}$$

- **Definition 7 (KL-Divergence)** *Let P and Q be two probability distributions defined on the set $\mathcal{X}$. Then, the KL-divergence between P and Q, denoted by $D\left(P||Q\right)$, is*

$$D\left(P||Q\right) = \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right). \tag{19}$$

- **Proposition 7** *Let P and Q be two probability distributions defined on the set $\mathcal{X}$. Then,*

$$D\left(P||Q\right) \geqslant 0 \tag{20}$$

- $D\left(P||Q\right) \neq D\left(Q||P\right)$
- If $P \not\ll Q$ then $D\left(P||Q\right) = \infty$
- If $P \ll Q$ then $D\left(P||Q\right) \leqslant \infty$

3 Shannon's Information Measures

- **Definition 8 (Entropy)** *Let $\mathcal{X}$ be a countable set and let also X be a random variable with pmf $P_X : \mathcal{X} \to [0, 1]$. The entropy of X, denoted by $H(X)$, is*

$$H(X) = - \sum_{x \in \mathrm{supp}(P_X)} P_X(x) \log_2 \left(P_X(x)\right). \tag{21}$$

- $H(X) = -\mathbb{E}_X \left[\log_2 \left(P_X(X) \right) \right]$
- $0 \leqslant H(X) \leqslant \log_2 \left(|\mathcal{X}| \right)$
- H(X) is a function of P_X and it is continous in $\triangle\left(\mathcal{X}\right)$ w.r.t. total variation distance.

- $H(f(X)) \leqslant H(X)$, with equality only when f is a one-to-one mapping over $\text{supp}\,(P_X)$

- **Definition 9 (Joint Entropy)** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two countable sets and let also X and Y be two random variables with joint pmf $P_{XY} : \mathcal{X} \times \mathcal{Y} \to [0,1]$. The joint entropy of X and Y, denoted by $H(X,Y)$, is*

$$H(X,Y) = -\sum_{(x,y)\in \text{supp}(P_{XY})} P_{XY}(x,y) \log_2 \left(P_{XY}(x,y)\right). \tag{22}$$

- $H(X,Y) = -\mathbb{E}_{XY}\left[\log_2 \left(P_{XY}(X,Y)\right)\right]$

- **Definition 10 (Conditional Entropy)** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two countable sets and let also X and Y be two random variables with pmf P_X and conditional pmf $P_{Y|X}$. The entropy of X conditioning on Y, denoted by $H(X|Y)$, is*

$$H(X|Y) = -\sum_{y\in \text{supp}(P_Y)} P_Y(y) \sum_{x\in \text{supp}\left(P_{X|Y=y}\right)} P_{X|Y}(x|y) \log_2 \left(P_{X|Y}(x|y)\right). \tag{23}$$

- $H(X|Y) = -\mathbb{E}_{XY}\left[\log_2 \left(P_{X|Y}(X|Y)\right)\right]$
- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- $H(X) \geqslant H(X|Y)$
- $H(X_1,\ldots,X_n) = H(X_1) + H(X_2|X_1) + \sum_{t=3}^{n} H(X_t|X_1,\ldots,X_{t-1})$
- $H(X_1,\ldots,X_n) \leqslant \sum_{t=1}^{n} H(X_t)$
- $H(X_1,\ldots,X_n|Y) = H(X_1|Y) + H(X_2|Y,X_1) + \sum_{t=3}^{n} H(X_t|Y,X_1,\ldots,X_{t-1})$
- **Definition 11 (Mutual Information)** *Let X and Y be two random variables with strictly positive joint pmf P_{XY} and marginal pmfs P_X and P_Y. Then, the mutual information between X and Y, denoted by $I(X;Y)$, is*

$$I(X;Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{XY}(x,y) \log_2 \left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) \tag{24}$$

- $I(X;Y) = \mathbb{E}_{XY}\left[\log_2\left(\frac{P_{Y|X}(Y|X)}{P_Y(Y)}\right)\right] = \mathbb{E}_{XY}\left[\log_2\left(\frac{P_{X|Y}(X|Y)}{P_X(X)}\right)\right]$
- $I(X;Y) \geqslant 0$
- $I(X;Y) = 0$ if and only if X and Y are mutually independent
- $I(X;Y) = I(Y;X)$
- $I(X;X) = H(X)$
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $I(X;Y) = H(X) + H(Y) - H(Y,X)$
- $I(X;Y,Z) \geqslant I(X;Y)$, with equality if and only if $X \to Y \to Z$.
- **Definition 12 (Conditional Mutual Information)** *Let X, Y and Z be three random variables with strictly positive joint pmf P_{XYZ} and conditional pmfs $P_{XY|Z}$, $P_{X|Z}$ and $P_{Y|Z}$. Then, the mutual information between X and Y conditioning on Z, denoted by $I(X,Y|Z)$, is*

$$I(X;Y|Z) = -\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} P_{XYZ}(x,y,z)\log_2\left(\frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right). \tag{25}$$

- $I(X;Y|Z) = 0$ if and only if $X \to Z \to Y$
- $I(X_1,\ldots,X_n;Y) \geqslant 0$
- $I(X_1,\ldots,X_n;Y) = I(X_1;Y) + I(X_2;Y|X_1) + \sum_{t=3}^{n} I(X_t;Y|X_1,X_2,\ldots,X_{t-1})$
- $I(X_1,\ldots,X_n;Y|Z) = I(X_1;Y|Z) + I(X_2;Y|Z,X_1) + \sum_{t=3}^{n} I(X_t;Y|ZX_1,X_2,\ldots,X_{t-1})$
- **Proposition 8** *Let X, Y and Z be three random variables such that $P_{XYZ}(x,y,z) = P_X(x)P_Y(y)P_{Z|XY}(z|x,y)$ for all (x,y,z) in $\operatorname{supp}(P_{XYZ})$. Then,*

$$I(X;Y|Z) \geqslant I(X;Y). \tag{26}$$

- **Proposition 9** *Let X, Y and Z be three random variables such that $Z \to X \to Y$. Then, $I(X;Y|Z) \leqslant I(X;Y)$.*

- **Proposition 10 (Data Processing Inequality)** *Let X, Y and Z be three random variables such that $X \to Y \to Z$. Then, $I(X;Z) \leqslant I(X;Y)$.*

- **Proposition 11 (Fano's Inequality)** *Let X and $\hat{X}$ be two random variables with* $\operatorname{supp}(X) = \operatorname{supp}\left(\hat{X}\right)$. *Let also $E = \mathbb{1}_{\left\{X \neq \hat{X}\right\}}$ be a binary random variable. Then,*

$$H\left(X|\hat{X}\right) \leqslant H(E) + P_E(1)\log_2(|\mathcal{X}| - 1). \tag{27}$$