Lecture 1:

Entropy, Divergence and Mutual Information

Probability

- A random variable X, takes on values in the set $\mathcal{X}$. The event $\{X = x\}$ is the event that X takes on the particular value $x \in \mathcal{X}$.

- We write $X \sim P_X$ to denote that P_X is the pmf of X, when X is discrete.

- When $|\mathcal{X}| = M$ is finite, the pmf P_X is also represented as the probability vector
$$\mathbf{p} = (p_1, p_2, \ldots, p_M), \qquad p_i = P_X(x_i)$$
where we assume a given (fixed) indexing of the elements of $\mathcal{X}$ with the integers $1, \ldots, M$.

- A probability vector $\mathbf{p}$ has non-negative components that satisfy $\sum_i p_i = 1$, therefore, it is a point in the probability simplex.

- A random sequence, or discrete-time random process is $\{X_i : i = 1, 2, \ldots\}$.

i.i.d. Sequences and Joint pmf

- The sequence is i.i.d., with marginal pmf P_X, if for any $i_1, i_2, \ldots, i_n$ we have

$$\mathbb{P}(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n}) = \prod_{j=1}^{n} P_X(x_{i_j})$$

- We indicate a random n-sequence (random vector) as $X^n = (X_1, \ldots, X_n)$.
- A random n-sequence X^n takes on values in $\mathcal{X}^n$, the set of (row) vectors of length n over $\mathcal{X}$, denoted by $\mathbf{x} = (x_1, \ldots, x_n)$.
- The joint pmf of X^n is denoted by

$$P_{X^n}(\mathbf{x}) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Conditional pmf

- By definition

$$\mathbb{P}(\mathsf{A}|\mathsf{B}) = \frac{\mathbb{P}(\mathsf{A}, \mathsf{B})}{\mathbb{P}(\mathsf{B})}$$

  (defined only if $\mathbb{P}(\mathsf{B}) > 0$).
- Conditional probability mass function of Y given X:

$$P_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x)$$

- Telescopic property of probability

$$\mathbb{P}(X = x, Y = y, Z = z) = \mathbb{P}(X = x)\mathbb{P}(Y = y|X = x)\mathbb{P}(Z = z|X = x, Y = y)$$

  (obviously, this generalizes to random vectors X^n).
- Written in terms of probability mass functions:

$$P_{X,Y,Z}(x, y, z) = P_X(x)P_{Y|X}(y|x)P_{Z|X,Y}(z|x, y)$$

Entropy

Definition 1. *The entropy $H(X)$ of a discrete random variable $X \sim P_X$ over $\mathcal{X}$ is defined by:*

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log(P_X(x)) = -\mathbb{E}\left[\log(P_X(X))\right]$$

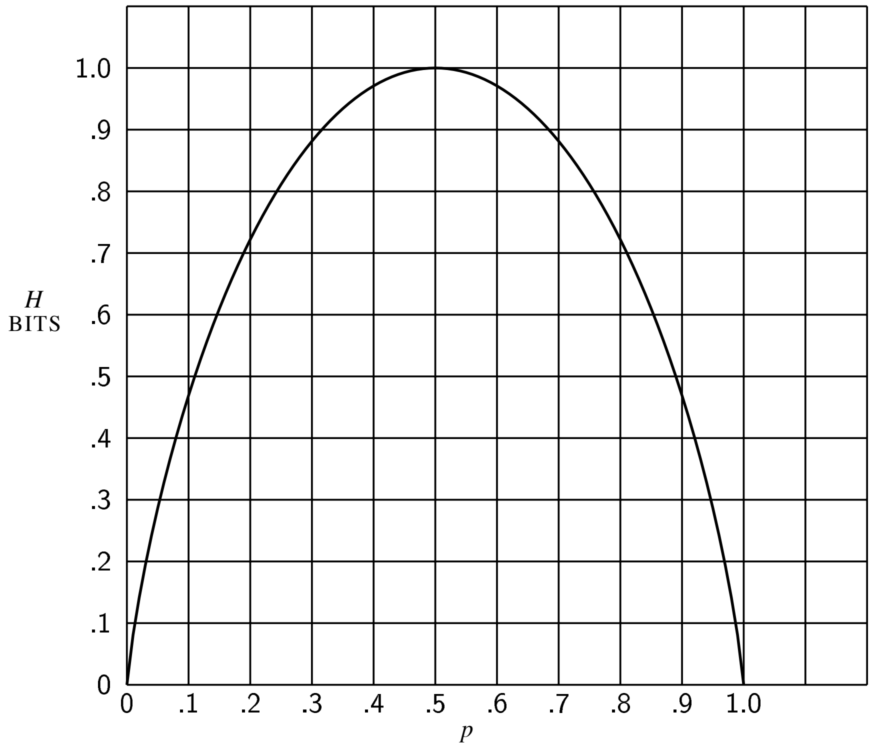
$\Diamond$

Example 1. *Binary entropy function: for $X \sim$ Bernoulli-p, we have*

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} = \mathcal{H}_2(p)$$

More in general, we indicate by $\mathcal{H}(\mathbf{p})$ the entropy function denoted as a function of the probability vector $\mathbf{p}$. $\Diamond$

Binary Entropy Function $\mathcal{H}_2(p)$



Joint and Conditional Entropy

Definition 2. *The joint entropy of a discrete random n-sequence $X^n \sim P_{X^n}$ over $\mathcal{X}$ is:*

$$H(X^n) = -\sum_{\mathbf{x} \in \mathcal{X}^n} P_{X^n}(\mathbf{x}) \log(P_{X^n}(\mathbf{x})) = -\mathbb{E}\left[\log(P_{X^n}(X^n))\right]$$

$\Diamond$

Definition 3. *For two jointly distributed random vectors X^n, Y^m over $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint pmf P_{X^n,Y^m}, the conditional entropy of X^n given Y^m is:*

$$\begin{aligned} H(X^n|Y^m) &= -\sum_{\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^m} P_{X^n,Y^m}(\mathbf{x},\mathbf{y}) \log(P_{X^n|Y^m}(\mathbf{x}|\mathbf{y})) \\ &= -\mathbb{E}\left[\log P_{X^n|Y^m}(X^n|Y^m)\right] \end{aligned}$$

$\Diamond$

Chain Rule for Entropy

Lemma 1. *Chain Rule for Entropy: we have*

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}) \end{aligned}$$

□

- Notice: the chain rule follows from the telescoping property

$$P_{X^n} = P_{X_1} P_{X_2|X_1} P_{X_3|X_1,X_2} \cdots P_{X_n|X_1,\ldots,X_{n-1}}$$

Examples

- "Developing" entropy in different ways:

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

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) = H(Y|Z) + H(X|Y,Z)$$

- Notice that, in general,

$$H(Y|X) \neq H(X|Y)$$

however

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

Divergence (Cross-Entropy)

Definition 4. *Let P_X and Q_X denote two pmfs over $\mathcal{X}$. The divergence (aka, cross-entropy) of P_X and Q_X is given by*

$$D(P_X \| Q_X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)}$$

$\Diamond$

- Also known just as Kullback-Leibler Distance, Information Divergence or Relative Entropy.
- Non-symmetric: $D(P_X \| Q_X) \neq D(Q_X \| P_X)$ in general.
- If for some $x \in \mathcal{X}$ we have $Q_X(x) = 0$ and $P_X(x) > 0$, then $D(P_X \| Q_X) = \infty$.
- It is a "sort of distance" between two pmfs.

Conditional divergence

Definition 5. *Let $P_{Y|X}$ and $Q_{Y|X}$ denote two conditional pmfs for Y given X, and let P_X denote a pmf for X. The conditional divergence of $P_{Y|X}$ and $Q_{Y|X}$ with respect to P_X is given by*

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)}$$

$\Diamond$

Lemma 2. *Chain Rule for Divergence: for two joint pmfs $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X Q_{Y|X}$ we have*

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_X \| Q_X) + D(P_{Y|X} \| Q_{Y|X} | P_X)$$

$\square$

Mutual Information

Definition 6. *Let $X, Y \sim P_{X,Y}$. The mutual information of X and Y is given by*

$$I(X;Y) = D(P_{X,Y} \| P_X P_Y) = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}$$

$\Diamond$

- Mutual information in terms of conditional divergence:

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

- Mutual information as a difference of divergences: let $X, Y \sim P_{X,Y} = P_X P_{Y|X}$ and let Q_Y be an arbitrary marginal pmf for Y, then

$$I(X;Y) = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_Y | P_X)$$

(the upper bound will be clear in a moment ....)

Symmetry of Mutual Information

- From the definition:

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

Mutual Information in terms of Entropies

- It is immediate to see that:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

- If X and Y are independent, then $I(X;Y) = 0$.
- $I(X;X) = H(X)$.

Chain rule for mutual information

Lemma 3. *Chain Rule for Mutual Information: Let X^n and Y be jointly distributed as $P_{X^n,Y}$, then we have*

$$\begin{aligned} I(X^n;Y) &= I(X_1;Y) + I(X_2;Y|X_1) + \cdots + I(X_n;Y|X_1,\ldots,X_{n-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X^{i-1}) \end{aligned}$$

$\square$

Elementary Inequalities

- No general inequality relationship between $I(X;Y|Z)$ and $I(X;Y)$ exists, but there are special cases.

- Special case 1: if $P_{X,Y,Z} = P_X P_Z P_{Y|X,Z}$ then

$$I(X;Y|Z) \geq I(X;Y)$$

- Special case 2: if $P_{X,Y,Z} = P_Z P_{X|Z} P_{Y|X}$, i.e., if $Z \to X \to Y$ (Markov chain), then

$$I(X;Y|Z) \leq I(X;Y)$$

Convex sets

- $\mathcal{R} \subseteq \mathbb{R}^d$ is a convex set if $\mathbf{x}, \mathbf{y} \in \mathcal{R}$ implies that $\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{R}$ for all $\alpha \in [0, 1]$.

- $\mathcal{R}$ is convex if it contains all convex combinations of its points: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{R}$, then

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}_i \in \mathcal{R}, \quad \text{for all coefficients } \alpha_i \geq 0, \ \sum_{i=1}^{n} \alpha_i = 1$$

- The convex hull of a set $\mathcal{S} \subseteq \mathbb{R}^d$ is the smallest convex set that contains $\mathcal{S}$. We write $\mathcal{R} = \mathrm{coh}\mathcal{S}$.

Lemma 4. *Fenchel-Eggleston-Carathéodory: let $\mathcal{S} \subseteq \mathbb{R}^d$ be a connected compact (i.e., closed and bounded) set. Any point in $\mathcal{R} = \mathrm{coh}\mathcal{S}$ can be represented as the convex combination of at most d points in $\mathcal{S}$.* $\square$

Convex Functions

- Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ denote a real-valued function, and define its *epigraph* as the set
$$\mathrm{Epi}(g) = \{(\mathbf{x}, a) : g(\mathbf{x}) \leq a\}$$

- g is a convex function if $\mathrm{Epi}(g)$ is a convex set.

- In simpler terms, g is convex if for any $\mathbf{x}, \mathbf{y}$ in its domain and $\alpha \in [0, 1]$ we have
$$\alpha g(\mathbf{x}) + (1 - \alpha) g(\mathbf{y}) \geq g(\alpha \mathbf{x} + (1 - \alpha)\mathbf{y})$$

- g is called concave if $-g$ is convex.

- If g is twice differentiable, then g is convex iff its Hessian (matrix of second derivatives)
$$\nabla \times \nabla g(\mathbf{x}) = \left[\frac{\partial^2 g(\mathbf{x})}{\partial x_i \partial x_j}\right] \succeq 0, \quad \forall\, \mathbf{x} \in \mathrm{Dom}(g)$$

Jensen's Inequality

Lemma 5. *Jensen's Inequality: Let g denote a convex function over $\mathbb{R}^n$ and let X^n denote a random n-sequence, then*

$$\mathbb{E}\left[g(X^n)\right] \geq g\left(\mathbb{E}[X^n]\right)$$

□

Information Inequality (1)

Theorem 1. *Information Inequality: Let P_X, Q_X be two pmfs defined on $\mathcal{X}$, then*

$$D(P_X \| Q_X) \geq 0$$

with equality iff $P_X(x) = Q_X(x)$ for all $x \in \mathcal{X}$ where they are both non-zero. □

Proof: Define $\mathcal{A} = \{x \in \mathcal{X} : P_X(x) > 0\}$. Then

$$\begin{aligned}
-D(P_X \| Q_X) &= \sum_{x \in \mathcal{A}} P_X(x) \log \frac{Q_X(x)}{P_X(x)} \\
&\leq \log \left(\sum_{x \in \mathcal{A}} P_X(x) \frac{Q_X(x)}{P_X(x)} \right) \quad \text{Jensen's Ineq.} \\
&= \log \sum_{x \in \mathcal{A}} Q_X(x) \leq \log \sum_{x \in \mathcal{X}} Q_X(x) = 0
\end{aligned}$$

■

Information Inequality (2)

Corollary 1.

$$I(X;Y) \geq 0$$

with equality iff X and Y are independent. $\square$

Corollary 2.

$$I(X;Y|Z) \geq 0$$

with equality iff X and Y are conditionally independent given Z. $\square$

Corollary 3. *Conditioning reduces entropy:*

$$H(Y) \geq H(Y|X)$$ *with equality iff X and Y are independent.*

$\square$

Information Inequality (3)

Corollary 4. *The uniform pmf maximizes entropy: for $X \sim P_X$ over the finite set $\mathcal{X}$ of size $|\mathcal{X}|$, we have*

$$H(X) \leq \log |\mathcal{X}|$$

with equality iff X is uniform over $\mathcal{X}$. □

Theorem 2. *Independence bound on joint entropy:*

$$H(X^n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality iff X^n has independent components. □

Convexity Properties

Theorem 3. *Convexity of divergence: Consider $D(P_X \| Q_X)$ as a function of the vector $(\mathbf{p}, \mathbf{q})$, where $\mathbf{p}$ is the probability vector associated with P_X and $\mathbf{q}$ is the probability vector associated with Q_X. Then, $D(P_X \| Q_X)$ is a convex function.* □

Corollary 5. *Concavity of entropy: Consider $H(X)$ as a function $H(\mathbf{p})$ of the probability vector associated with P_X. Then $H(X)$ is a concave function.* □

Corollary 6. *Concavity/Convexity of mutual information: Consider $I(X;Y)$ as a function of $\mathbf{p}$, the probability vector associated with P_X, and of $\mathbf{P}$, the conditional probability matrix associated with $P_{Y|X}$. Then, $I(X;Y)$ is a concave function of $\mathbf{p}$ for any fixed $\mathbf{P}$, and a convex function of $\mathbf{P}$ for any fixed $\mathbf{p}$.* □

Data Processing Inequality

Theorem 4. *Data processing inequality: If $X \to Y \to Z$ (i.e., $P_{X,Y,Z} = P_X P_{Y|X} P_{Z|Y}$), then $I(X;Z) \leq I(Y;Z)$ and $I(X;Z) \leq I(X;Y)$.* □

Proof: Expand $I(X,Y;Z)$ in two ways

$$\begin{aligned} I(X,Y;Z) &= I(X;Z) + I(Y;Z|X) \\ &= I(Y;Z) + I(X;Z|Y) \end{aligned}$$

and notice that $I(X;Z|Y) = 0$ while $I(Y;Z|X) \geq 0$ (operating on $I(X;Y,Z)$ we prove the other inequality). ■

Fano Inequality

Theorem 5. *Fano Inequality: Let $(X, \widehat{X}) \sim P_{X,\widehat{X}}$ be two jointly distributed random variables taking on values in the same alphabet $\mathcal{X}$, and define $P_e = \mathbb{P}(X \neq \widehat{X})$. Then,*

$$H(X|\widehat{X}) \leq \mathcal{H}_2(P_e) + P_e \log |\mathcal{X}| \leq 1 + P_e \log |\mathcal{X}|$$

□

Proof: Define $E = 1\{X \neq \widehat{X}\}$ be the error indicator random variable. Then, we use the chain rule and write:

$$\begin{aligned} H(X, E|\widehat{X}) &= H(X|\widehat{X}) + H(E|X, \widehat{X}) \\ &= H(E|\widehat{X}) + H(X|E, \widehat{X}) \end{aligned}$$

Since $H(E|X, \widehat{X}) = 0$, $H(E|\widehat{X}) \leq H(E) = \mathcal{H}_2(P_e)$ and

$$H(X|E, \widehat{X}) = P_e H(X|E = 1, \widehat{X}) + (1 - P_e) H(X|E = 0, \widehat{X}) \leq P_e \log |\mathcal{X}|$$

the result follows. ■

End of Lecture 1