EE276 Information Theory Lecture 2 - 01/09/2024

Lecture 2: Entropy, Relative Entropy, and Mutual Information

Lecturer: Tsachy Weissman

In this lecture, we will introduce certain key measures of information, that play crucial roles in theoretical and operational characterizations throughout the course. These include the entropy, the mutual information, and the relative entropy. We will also exhibit some key properties exhibited by these information measures.

1 Notation

A quick summary of the notation

1. **Random Variable:** X
2. **Specific Value:** x, x_1, etc.
3. **Alphabet:** $\mathcal{X} = \{x_1, x_2, ..., x_M\}$ (An alphabet of size M)

For discrete random variables, we will write (interchangeably) $P(X = x)$, $P_X(x)$ or most often just, $p(x)$. Similarly, for a pair of random variables X_1, X_2 we write $P(X_1 = x_1, X_2 = x_2) = P_{X_1,X_2}(x_1, x_2) = p(x_1, x_2)$ and $P(X_2 = x_2 \mid X_1 = x_1) = P_{X_2|X_1}(x_2 \mid x_1) = p(x_2 \mid x_1)$.

2 Properties of Random Variables

Fact 1. *Assume X is a nonnegative random variable. Then*

$$\mathbb{E}[X] \geq 0,$$

with equality if and only if $X = 0$ with probability 1 (i.e. X is deterministically 0).

Fact 2. *For random variables X_1 and X_2,*

$$\mathbb{E}[\max\{X_1, X_2\}] \geq \max\{\mathbb{E}[X_1], \mathbb{E}[X_2]\},$$

with equality if and only if $P(X_1 \geq X_2) = 1$ or $P(X_1 \leq X_2) = 1$.

Proof: Consider $X = \max\{X_1, X_2\} - X_1$. Since X is nonnegative random variable, we can apply Fact 1. Similarly, the same holds for $Y = \max\{X_1, X_2\} - X_2$.

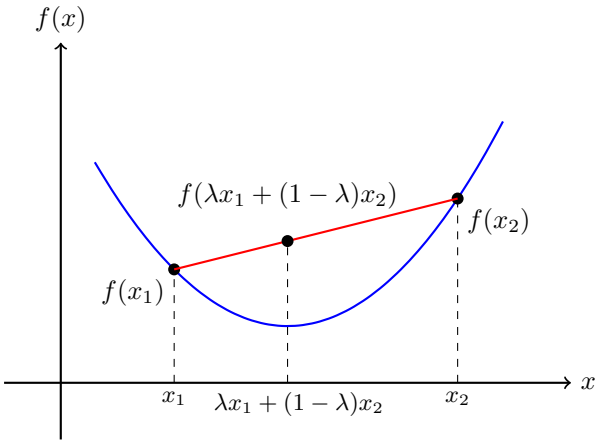
3 Convex Function

Definition 3. *Convex Function: A function f is **convex** if $\forall x_1, x_2$ and $0 \leq \lambda \leq 1$,*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Definition 4. *Strictly Convex Function: A function f is **strictly convex** if $\forall x_1 \neq x_2$ and $0 < \lambda < 1$,*

$$f(\lambda x_1 + (1 - \lambda)x_2) < \lambda f(x_1) + (1 - \lambda)f(x_2).$$



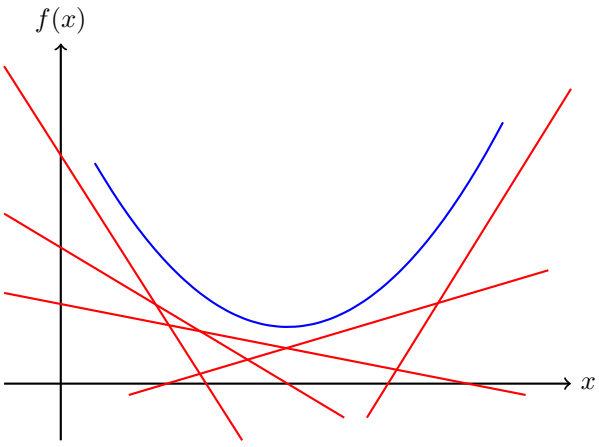
For a convex function, the straight line connecting any two points on the function lies above or on the curve. Note that piecewise linear convex functions cannot be strictly convex.

Fact 5. *Let $\mathcal{L}$ be the set of all affine functions l of the form $l(x) = ax + b$. Given function f, let*

$$\mathcal{L}_f = \{l \in \mathcal{L} : l(x) \leq f(x), \forall x\}.$$

Then f being convex is equivalent to $f(x) = \max\{l(x) : l \in \mathcal{L}_f\}$.

This fact shows that if f is a convex function, then its graph $\{(x, f(x)) : x \in \mathbb{R}\}$ can be seen as an upper bound on the set of affine functions that lie below it.



Lemma 6. ***Jensen's Inequality:*** *Let Q denote a function on a random variable X. Jensen's inequality states that $\forall Q$ that are convex:*

$$\mathbb{E}[Q(X)] \geq Q(\mathbb{E}[X]). \tag{1}$$

Further, if Q is strictly convex, equality holds iff X is deterministic.
Conversely, if Q is a concave *function, then*

$$\mathbb{E}[Q(X)] \leq Q(\mathbb{E}[X]). \tag{2}$$

Proof:

Using Fact 5,

$$Q(X) = \max_{L \in \mathcal{L}_f} L(X). \tag{3}$$

Thus, by linearity:

$$\begin{aligned}
\mathbb{E}[Q(X)] &= \mathbb{E}[\max_{L \in \mathcal{L}} L(X)] \\
&\geq \max_{L \in \mathcal{L}} \mathbb{E}[L(X)] \qquad (\because \text{Fact 2}) \\
&= \max_{L \in \mathcal{L}} L(\mathbb{E}[X]) \qquad (\because L \text{ is affine}) \\
&= Q(\mathbb{E}[X]) \qquad (\because (3))
\end{aligned}$$

The same argument for concave functions can be done using the minimum instead of the maximum.

4 Entropy

Definition 7. *"Surprise" Function:*

$$S(u) \triangleq \log \frac{1}{p(u)} \tag{4}$$

A lower probability of u translates to a greater "surprise" that it occurs.
Note here that we use log to mean $\log_2$ by default, rather than the natural log ln, as is typical in some other contexts. This is true throughout these notes: log is assumed to be $\log_2$ unless otherwise indicated.

Definition 8. ***Entropy****: Let U a discrete random variable taking values in alphabet $\mathcal{U}$. The* ***entropy*** *of U is given by:*

$$H(U) \triangleq \mathbb{E}[S(U)] = \mathbb{E}\left[\log\left(\frac{1}{p(U)}\right)\right] = \mathbb{E}\big[-\log\left(p(U)\right)\big] = -\sum_u p(u) \log p(u) \tag{5}$$

Where U represents all u values possible to the variable.

The entropy is a property of the underlying distribution $P_U(u), u \in \mathcal{U}$ that measures the amount of randomness or surprise in the random variable.

4.1 Properties of Entropy

Suppose $\mathcal{U} = \{u_1, u_2, ..., u_M\}$

1. $H(U) \leq \log M$, with equality iff U is uniformly distributed i.e. $p(u) = \frac{1}{M}\ \forall u$

   Proof:

$$\begin{aligned}
H(U) &= \mathbb{E}\left[\log \frac{1}{p(U)}\right] && (6)\\
&\leq \log \mathbb{E}\left[\frac{1}{p(U)}\right] \text{ (By Jensen's inequality because log is concave)} && (7)\\
&= \log \sum_u p(u) \cdot \frac{1}{p(u)} && (8)\\
&= \log M. && (9)
\end{aligned}$$

   Equality by Jensen's inequality iff $\frac{1}{p(U)}$ is deterministic, iff $p(u) = \frac{1}{M}\ \forall u \in U$

2. $H(U) \geq 0$, with equality iff U is deterministic.

   Proof:

$$H(U) = \mathbb{E}\left[\log \frac{1}{p(U)}\right] \geq 0 \text{ because } \log \frac{1}{p(U)} \geq 0 \tag{10}$$

   The equality occurs iff $\log \frac{1}{p(U)} = 0$ with probability 1, i.e. $p(U) = 1$ with probability 1, so U must be deterministic.

3. For a PMF q define

$$H_q(U) \triangleq \mathbb{E}\left[\log \frac{1}{q(U)}\right] = \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{q(u)}. \tag{11}$$

   Then:

$$H(U) \leq H_q(U), \tag{12}$$

   with equality iff $q = p$.

   Proof:

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{1}{p(u)}\right] - \mathbb{E}\left[\log \frac{1}{q(u)}\right] \tag{13}$$

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{q(u)}{p(u)}\right] \tag{14}$$

$$\leq \log \mathrm{E}\left[\frac{q(u)}{p(u)}\right] \tag{15}$$

$$= \log \sum_{u \in \mathcal{U}} p(u) \frac{q(u)}{p(u)} \tag{16}$$

$$= \log \sum_{u \in \mathcal{U}} q(u) \tag{17}$$

$$= \log 1 \tag{18}$$

$$= 0 \tag{19}$$

   Thus,

$$H(U) - H_q(U) \leq 0.$$

   Equality only holds when $\frac{q(u)}{p(u)}$ is deterministic, which occurs when $q = p$ (distributions are identical). Note that $H_q(U)$ is called "cross entropy" in some places.

Definition 9. ***Relative Entropy***[1] *An measure of distance between probability distributions is relative entropy:*

$$D(p \parallel q) \triangleq H_q(U) - H(U) = \mathbb{E}\left[\log \frac{p(u)}{q(u)}\right] = \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)} \tag{20}$$

Note that by property 3, the relative entropy is always greater than or equal to 0, with equality iff $q = p$. For now, relative entropy can be thought of as a measure of discrepancy between two probability distributions. We will soon see that it is central to information theory.

[1] Some students may be familiar with relative entropy as Kullback-Leibler (KL) divergence

4.

Definition 10. ***Joint Entropy of*** U ***and*** V**.** *For a pair of random variables* (U, V)*, the joint entropy is defined as*

$$H(U,V) = \mathbb{E}[\log \frac{1}{P(U,V)}.$$

5. (Not covered in lecture) If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) \tag{21}$$

Note: $H(X_1, X_2, \ldots, X_n)$ is called the **joint entropy** of $X_1, X_2, \ldots, X_n$.
Proof:

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \mathbb{E}\left[-\log p(x_1, x_2, \ldots, x_n)\right] && (22)\\ &= \mathbb{E}\left[-\log \prod_{i=1}^{n} p(x_i)\right] && (23)\\ &= \mathbb{E}\left[-\sum_{i=1}^{n} \log p(x_i)\right] && (24)\\ &= \sum_{i=1}^{n} \mathbb{E}\left[-\log p(x_i)\right] && (25)\\ &= \sum_{i=1}^{n} H(X_i). && (26) \end{aligned}$$

6.

Definition 11. ***Conditional Entropy of*** X ***given*** Y

$$\begin{aligned} H(X \mid Y) &\triangleq \mathbb{E}\left[\log \frac{1}{p(X \mid Y)}\right] && (27)\\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x \mid y)} && (28)\\ &= \sum_{y} p(y) \left[\sum_{x} p(x \mid y) \log \frac{1}{p(x \mid y)}\right] && (29)\\ &= \sum_{y} p(y) H(X \mid Y = y). && (30) \end{aligned}$$

7. **Chain Rule:**

$$\begin{aligned} H(X,Y) &\triangleq \mathbb{E}\big[\log \frac{1}{P(X,Y)}\big] && (31)\\ &= \mathbb{E}\big[\log \frac{1}{P(Y)P(X \mid Y)}\big] && (32)\\ &= H(Y) + H(X \mid Y) && (33) \end{aligned}$$

We can take this one step further with (5):

$$H(X,Y) = H(Y) + H(X \mid Y) \leq H(X) + H(Y), \tag{34}$$

with equality holding iff X and Y are independent.

8. **Conditioning reduces entropy:** $H(X \mid Y) \leq H(X)$ with equality iff X and Y are independent.
   Proof:

$$H(X) - H(X \mid Y) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] - \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right] \tag{35}$$
$$= \mathbb{E}\left[\log \frac{p(X \mid Y)}{p(X)} \frac{p(Y)}{p(Y)}\right] \tag{36}$$
$$= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \tag{37}$$
$$= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{38}$$
$$= D(P_{x,y} \parallel P_x \times P_y) \tag{39}$$
$$\geq 0 \tag{40}$$

   $D(P_{x,y} \parallel P_x \times P_y) \geq 0$ because relative entropy can never be negative. Equality holds iff $P_{x,y} \equiv P_x \times P_y$, ($X$ and Y are independent).

5 Mutual information

Finally, we introduce the mutual information, which in class was introduce in lecture 3.

Definition 12. ***Mutual information between*** X ***and*** Y

We now define the mutual information between random variables X and Y distributed according to the joint PMF $P(x, y)$:

$$I(X;Y) \triangleq D(P_{x,y} \parallel P_x \times P_y) \tag{41}$$
$$= H(X) - H(X|Y) \tag{42}$$
$$= H(X) + H(Y) - H(X,Y) \tag{43}$$

(May find any of these in the literature) The mutual information tells how helpful one variable is at reducing uncertainty in the other. We'll learn more about this quantity in future lectures and homeworks. For now, we'll note that while relative entropy **is not** symmetric, mutual information **is**.

6 Exercises

1. "Data processing decreases entropy" (note that this statement only applies to deterministic functions)
   $Y = f(X) \Rightarrow H(Y) \leq H(X)$ with equality when f is one-to-one.
   Note: Proof is part of a future homework.

2. "Data processing on side information increases entropy"
   $Y = f(X) \Rightarrow H(Z|X) \leq H(Z|Y)$
   True more generally:
   whenever $Y - X - Z$ (Markov Relation), i.e., $p(Z|X,Y) = p(Z|X)$, then $H(Z|X) \leq H(Z|Y)$
   Note: Proof is part of a future homework.

3. **Definition 13.** ***Conditional mutual information***

$$I(X;Y|Z) \triangleq H(X|Z) - H(X|Y,Z) \tag{44}$$

Show that: $I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2|Y_1)$
Proof:

$$\begin{aligned} I(X;Y_1,Y_2) &= H(X) - H(X|Y_1,Y_2) && (45)\\ &= H(X) - H(X|Y_1,Y_2) - H(X|Y_1) + H(X|Y_1) && (46)\\ &= [H(X) - H(X|Y_1)] + [H(X|Y_1) - H(X|Y_1,Y_2)] && (47)\\ &= I(X;Y_1) + I(X;Y_2|Y_1) && (48) \end{aligned}$$