Chapter 2: Entropy and Mutual Information

Chapter 2 outline

• Definitions
• Entropy
• Joint entropy, conditional entropy
• Relative entropy, mutual information
• Chain rules
• Jensen’s inequality
• Log-sum inequality
• Data processing inequality
• Fano’s inequality
Definitions

A discrete random variable $X$ takes on values $x$ from the discrete alphabet $\mathcal{X}$.

The probability mass function (pmf) is described by

$$p_X(x) = p(x) = \Pr\{X = x\}, \text{ for } x \in \mathcal{X}.$$ 

The joint pmf of two random variables $X$ and $Y$ taking on values in alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively is described by

$$p_{X,Y}(x, y) = p(x, y) = \Pr\{X = x, Y = y\}, \text{ for } x, y \in \mathcal{X} \times \mathcal{Y}.$$ 

If $p_X(X = x) > 0$, the conditional probability that the outcome $Y = y$ given that $X = x$ is defined as

$$p_{Y|X}(y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$
Definitions

The events $X = x$ and $Y = y$ are statistically independent if $p(x, y) = p(x)p(y)$.

The random variables $X$ and $Y$ defined over the alphabets $\mathcal{X}$ and $\mathcal{Y}$, resp. are statistically independent if $p_{X,Y}(x, y) = p_X(x)p_Y(y), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$.

The variables $X_1, X_2, \cdots X_N$ are called independent if for all $(x_1, x_2, \cdots, x_N) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \mathcal{X}_N$ we have

$$p(x_1, x_2, \cdots x_N) = \prod_{i=1}^{N} p_{X_i}(x_i).$$

They are furthermore called identically distributed if all variables $X_i$ have the same distribution $p_X(x)$.

Entropy

- Intuitive notions?

- 2 ways of defining entropy of a random variable:
  - axiomatic definition (want a measure with certain properties...)
  - just define and then justify definition by showing it arises as answer to a number of natural questions

**Definition:** The entropy $H(X)$ of a discrete random variable $X$ with pmf $p_X(x)$ is given by

$$H(X) = -\sum_x p_X(x) \log p_X(x) = -E_{p_X(x)}[\log p_X(X)].$$
Order these in terms of entropy
Entropy examples 1

- What’s the entropy of a uniform discrete random variable taking on K values?

- What’s the entropy of a random variable with

  \[ X = [♠, ♦, ♣, ♦], \quad p_X = [1/2; 1/4; 1/8; 1/8] \]

- What’s the entropy of a deterministic random variable?

Entropy: example 2

**Example 2.12.** The entropy of a randomly selected letter in an English document is about 4.11 bits, assuming its probability is as given in table 2.9. We obtain this number by averaging \( \log 1/p_i \) (shown in the fourth column) under the probability distribution \( p_i \) (shown in the third column).
Entropy: example 3

- Bernoulli random variable takes on heads (0) with probability p and tails with probability 1-p. Its entropy is defined as

\[ H(p) := -p \log_2(p) - (1 - p) \log_2(1 - p) \]

Suppose that we wish to determine the value of \( X \) with the minimum number of binary questions. An efficient first question is “Is \( X = a \)?” This splits the probability in half. If the answer to the first question is no, the second question can be “Is \( X = b \)?” The third question can be “Is \( X = c \)?” The resulting expected number of binary questions required is 1.75. This turns out to be the minimum expected number of binary questions required to determine the value of \( X \). In Chapter 5 we show that the minimum expected number of binary questions required to determine \( X \) lies between \( H(X) \) and \( H(X) + 1 \).

2.2 JOINT ENTROPY AND CONDITIONAL ENTROPY

We defined the entropy of a single random variable in Section 2.1. We now extend the definition to a pair of random variables. There is nothing really new in this definition because \((X, Y)\) can be considered to be a single vector-valued random variable.

Definition

The joint entropy \( H(X, Y) \) of a pair of discrete random variables \((X, Y)\) with a joint distribution \( p(x, y) \) is defined as

\[ H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y), \quad (2.8) \]

Entropy

The entropy \( H(X) = -\sum_x p(x) \log p(x) \) has the following properties:

- \( H(X) \geq 0 \), entropy is always non-negative. \( H(X) = 0 \) iff \( X \) is deterministic (0 log(0) = 0).
- \( H(X) \leq \log(|X|) \). \( H(X) = \log(|X|) \) iff \( X \) has uniform distribution over \( X \).
- Since \( H_b(X) = \log_b(a) H_a(X) \), we don’t need to specify the base of the logarithm (bits vs. nat).

Moving on to multiple RVs
Joint entropy and conditional entropy

**Definition:** Joint entropy of a pair of two discrete random variables $X$ and $Y$ is:

$$H(X,Y) := -E_{p(x,y)}[\log p(X,Y)]$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$$

**Definition:** The conditional entropy of $Y$ given a random variable $X$ (average over $X$) is:

$$H(Y|X) := E_{p(x)}[H(Y|X = x)] = \sum_{x \in \mathcal{X}} p(x)H(Y|X = x)$$

$$= -E_{p(x)}E_{p(y|x)}[\log p(Y|X)]$$

$$= -E_{p(x,y)}[\log p(Y|X)] = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x)$$

**Note:** $H(X|Y) \neq H(Y|X)$. 

---

Joint entropy and conditional entropy

- Natural definitions, since....

**Theorem:** Chain rule

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

**Corollary:**

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$
Joint/conditional entropy examples

\[ p(x, y) \begin{array}{c|c|c} \hline x = 0 & y = 0 & y = 1 \\ \hline \frac{1}{2} & \frac{1}{4} \\ \hline x = 1 & 0 & \frac{1}{4} \end{array} \]

\[ \begin{align*}
H(X,Y) &= \\
H(X|Y) &= \\
H(Y|X) &= \\
H(X) &= \\
H(Y) &= 
\end{align*} \]

Entropy is central because...

(A) entropy is the measure of **average uncertainty** in the random variable

(B) entropy is the **average number of bits** needed to describe the random variable

(C) entropy is a lower bound on the **average length of the shortest description** of the random variable

(D) entropy is measured in bits?

(E) \[ H(X) = - \sum_x p(x) \log_2(p(x)) \]

(F) entropy of a deterministic value is 0
Mutual information

- Entropy $H(X)$ is the uncertainty ("self-information") of a single random variable.
- Conditional entropy $H(X|Y)$ is the entropy of one random variable conditional upon knowledge of another.
- The average amount of decrease of the randomness of $X$ by observing $Y$ is the average information that $Y$ gives us about $X$.

**Definition:** The mutual information $I(X;Y)$ between the random variables $X$ and $Y$ is given by

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

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

$$= E_{p(x,y)} \left[ \log_2 \frac{p(X,Y)}{p(X)p(Y)} \right]$$

At the heart of information theory because...

- Information channel capacity:

$$C = \max_{p(x)} I(X;Y)$$

- Operational channel capacity:

Highest rate (bits/channel use) that can communicate at reliably

- Channel coding theorem says: information capacity = operational capacity
Mutual information example

<table>
<thead>
<tr>
<th>( p(x, y) )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
<th>( X ) or ( Y )</th>
<th>( p(x) )</th>
<th>( p(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>1/2</td>
<td>1/4</td>
<td>0</td>
<td>3/4</td>
<td>1/2</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>0</td>
<td>1/4</td>
<td>1</td>
<td>1/4</td>
<td>1/2</td>
</tr>
</tbody>
</table>

Divergence (relative entropy, K-L distance)

**Definition:** Relative entropy, divergence or Kullback-Leibler distance between two distributions, \( P \) and \( Q \), on the same alphabet, is

\[
D(p \parallel q) := E_p \left[ \log \frac{p(x)}{q(x)} \right] = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}
\]

(Note: we use the convention \( 0 \log 0 = 0 \) and \( 0 \log \frac{0}{q} = p \log \frac{p}{0} = \infty \).)

- \( D(p \parallel q) \) is in a sense a measure of the “distance” between the two distributions.
- If \( P = Q \) then \( D(p \parallel q) = 0 \).
- Note \( D(p \parallel q) \) is not a true distance.

\[
D(\blacklozenge, \blacklozenge) = 0.2075 \quad D(\blacklozenge, \lozenge) = 0.1887
\]
K-L divergence example

- $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$
- $P = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$
- $Q = [1/10, 1/10, 1/10, 1/10, 1/10, 1/2]$
- $D(p \parallel q) = ?$ and $D(q \parallel p) = ?$

Mutual information as divergence

**Definition:** The mutual information $I(X; Y)$ between the random variables $X$ and $Y$ is given by

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

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \left[ \log_2 \frac{p(x, y)}{p(x)p(y)} \right]$$

- Can we express mutual information in terms of the K-L divergence?

$$I(X; Y) = D(p(x, y) \parallel p(x)p(y))$$
Mutual information and entropy

Theorem: Relationship between mutual information and entropy.

\[ I(X; Y) = H(X) - H(X|Y) \]

\[ I(X; Y) = H(Y) - H(Y|X) \]

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) \]

\[ I(X; Y) = I(Y; X) \quad \text{(symmetry)} \]

\[ I(X; X) = H(X) \quad \text{("self-information")} \]

"Two’s company, three’s a crowd"

Chain rule for entropy

Theorem: (Chain rule for entropy): \((X_1, X_2, ..., X_n) \sim p(x_1, x_2, ..., x_n)\)

\[ H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1) \]

\[ H(X_1, X_2, X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_1, X_3) - H(X_2, X_3) + H(X_1, X_2, X_3) \]
Conditional mutual information

**Definition:** The conditional mutual information between $X$ and $Y$ given $Z$ is

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

$$= E_{p(x, y, z)} \log \frac{p(X, Y | Z)}{p(X | Z) p(Y | Z)}$$

Chain rule for mutual information

**Theorem:** (Chain rule for mutual information)

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, ..., X_1)$$

**Chain rule for relative entropy in book pg. 24**
What is the grey region?

Another disclaimer....
Convex and concave functions

- A **convex function** $f$ on an interval $[a, b]$ is one for which every chord lies (on or above) the function on that interval.

  \[ f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v), \quad \forall u, v \in [a, b], \ 0 < \lambda < 1 \]

- A function $f$ is **concave** if $-f$ is convex.

*Theorem:* If the function $f$ has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.
Jensen’s inequality

**Theorem:** (Jensen’s inequality) If $f$ is convex, then

$$E[f(X)] \geq f(E[X]).$$

If $f$ is strictly convex, the equality implies $X = E[X]$ with probability 1.

Jensen’s inequality consequences

- **Theorem:** (Information inequality) $D(p \parallel q) \geq 0$, with equality iff $p = q$.
- **Corollary:** (Nonnegativity of mutual information) $I(X; Y) \geq 0$ with equality iff $X$ and $Y$ are independent.
- **Theorem:** (Conditioning reduces entropy) $H(X|Y) \leq H(X)$ with equality iff $X$ and $Y$ are independent.
- **Theorem:** $H(X) \leq \log |\mathcal{X}|$ with equality iff $X$ has a uniform distribution over $\mathcal{X}$.
- **Theorem:** (Independence bound on entropy) $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality iff $X_i$ are independent.
Log-sum inequality

**Theorem:** (Log sum inequality) For nonnegative $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}
$$

with equality iff $a_i/b_i = \text{const}$.

Convention: $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$.

Log-sum inequality consequences

- **Theorem:** (Convexity of relative entropy) $D(p \parallel q)$ is convex in the pair $(p, q)$, so that for pmf’s $(p_1, q_1)$ and $(p_2, q_2)$, we have for all $0 \leq \lambda \leq 1$:

$$
D(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \\
\leq \lambda D(p_1 \parallel q_1) + (1 - \lambda)D(p_2 \parallel q_2)
$$

- **Theorem:** Concavity of entropy For $X \sim p(x)$, we have that

$$
H(p) := H_p(X) \text{ is a concave function of } p(x).
$$

- **Theorem:** (Concavity of the mutual information in $p(x)$) Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$.

- **Theorem:** (Convexity of the mutual information in $p(y|x)$) Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Then, $I(X; Y)$ is a convex function of $p(y|x)$ for fixed $p(x)$. 
Markov chains

Definition: X, Y, Z form a Markov chain in that order (X → Y → Z) iff

\[ p(x, y, z) = p(x)p(y|x)p(z|y) \equiv p(z|y, x) = p(z|y) \]

- X → Y → Z iff X and Z are conditionally independent given Y
- X → Y → Z ⇒ Z → Y → X. Thus, we can write X ↔ Y ↔ Z.

Data-processing inequality

Theorem: (Data-processing inequality) If X → Y → Z, then \( I(X; Y) \geq I(X; Z) \), with equality iff \( I(X; Y|Z) = 0 \).

Corollary: If Z = g(Y), then \( I(X; Y) \geq I(X; g(Y)) \).

Corollary: If X → Y → Z, then I(X; Y) ≥ I(X; Y|Z).
If X → Y → Z, then I(X; Y) ≥ I(X; Y|Z).
Markov chain questions

If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Y|Z)$.

What if $X, Y, Z$ do not form a Markov chain, can $I(X; Y|Z) \geq I(X; Y)$?

If $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5 \rightarrow X_6$, then Mutual Information increases as you get closer together:

$$I(X_1; X_2) \geq I(X_1; X_4) \geq I(X_1; X_5) \geq I(X_1; X_6).$$

Consequences on sufficient statistics

- Consider a family of probability distributions $\{f_\theta(x)\}$ indexed by $\theta$. If $X \sim f(x | \theta)$ for fixed $\theta$ and $T(X)$ is any statistic (i.e., function of the sample $X$), then we have
  $$\theta \rightarrow X \rightarrow T(X).$$

- The data processing inequality in turn implies
  $$I(\theta; X) \geq I(\theta; T(X))$$
  for any distribution on $\theta$.

- Is it possible to choose a statistic that preserves all of the information in $X$ about $\theta$?
Consequences on sufficient statistics

- Consider a family of probability distributions \( \{f_\theta(x)\} \) indexed by \( \theta \).
  If \( X \sim f(\theta \mid \theta) \) for fixed \( \theta \) and \( T(X) \) is any statistic (i.e., function of the sample \( X \)), then we have
  \[ \theta \rightarrow X \rightarrow T(X) \]
- The data processing inequality in turn implies
  \[ I(\theta; X) \geq I(\theta; T(X)) \]
  for any distribution on \( \theta \).
- Is it possible to choose a statistic that preserves all of the information in \( X \) about \( \theta \)?

**Definition: Sufficient Statistic** A function \( T(X) \) is said to be a sufficient statistic relative to the family \( \{f_\theta(x)\} \) if the conditional distribution of \( X \), given \( T(X) = t \), is independent of \( \theta \) for any distribution on \( \theta \) (Fisher-Neyman):

\[
f_\theta(x) = f(x \mid t) f_\theta(t) \quad \Rightarrow \quad \theta \rightarrow T(X) \rightarrow X \quad \Rightarrow \quad I(\theta; T(X)) \geq I(\theta; X)
\]

Hence, \( I(\theta; X) = I(\theta; T(X)) \) for a sufficient statistic.

Example of a sufficient statistic

- If \( X_1, \ldots, X_n \) are independent Bernoulli-distributed random variables with expected value \( p \), then the sum \( T(X) = \sum_{i=1}^{n} X_i \) is a sufficient statistic for \( p \).
- Proof: The joint probability distribution
  \[
p(x_1, \ldots, x_n) = p^{T(x)} (1 - p)^{n-T(x)}
\]
  which satisfies the factorization criterion, with \( f(x \mid t) = 1 \) being just a constant.
- Note that the unknown parameter \( p \) interacts with the data \( X \) only via the statistic \( T(X) \).
Fano’s inequality

Theorem: Fano’s inequality
For any estimator $\hat{X} : X \rightarrow Y \rightarrow \hat{X}$, with $P_e = \Pr\{X \neq \hat{X}\}$, we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

This implies $1 + P_e \log |\mathcal{X}| \geq H(X|Y)$ or $P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}$.

- Fano’s inequality says that the probability of error cannot be too small if $H(X|Y)$ is large i.e., correct estimation only happens when the residual randomness of $X$ is small after the observation of $Y$.

Fano’s inequality consequences

- Corollary: Let $p = \Pr\{X \neq Y\}$. Then,

$$H(p) + p \log |\mathcal{X}| \geq H(X|Y).$$

- Corollary: Let $P_e = \Pr\{X \neq \hat{X}\}$, and constrain $\hat{X} : Y \rightarrow \mathcal{X}$; then

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

- Fano’s bound is a loose bound, but sufficient for many cases of interest ($P_e$ is small and $|\mathcal{X}|$ is quite large).

- Suppose no observation $Y$ so that $X$ must simply be guessed, and order $X \in \{1, 2, \ldots, m\}$ such that $p_1 \geq p_2 \geq \cdots \geq p_m$. Then $\hat{X} = 1$ is the optimal estimate of $X$, with $P_e = 1 - p_1$, and Fano’s inequality becomes

$$H(P_e) + P_e \log(m - 1) \geq H(X).$$

The pmf $(p_1, p_2, \ldots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \ldots, \frac{P_e}{m-1}\right)$ achieves this bound with equality.