EEE 551 Information Theory (Spring 2022)

Chapter 2: Entropy, Relative Entropy, and Mutual Information

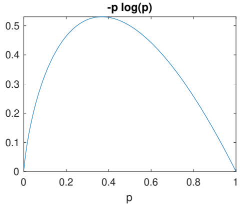
Quick Probability Review and Notation

- Random variable: $X, Y, Z, \ldots$
- Sample value of a random variable: $x, y, z, \ldots$
- Alphabet of a random variable: $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$
- The alphabet does **not** need to consist of only numbers; e.g. $\mathcal{X} = \{a, b, c\}$
- Probability mass function (PMF): $p(x) = p_X(x) = \Pr\{X = x\}$
- Joint PMF: $p(x, y) = p_{X,Y}(x, y) = \Pr\{X = x, Y = y\}$
- Conditional PMF: $p(x|y) = p_{X|Y}(x|y) = \Pr\{X = x|Y = Y\} = \dfrac{p(x, y)}{p(y)}$
- Variables X and Y are independent iff $p(x, y) = p(x)p(y)$, or equivalently $p(x|y) = p(x)$
- Expectation: $\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} p(x)f(x)$

Entropy

- A measure of the "information" or "uncertainty" in a random variable
- Entropy of a discrete random variable X with PMF $p(x)$:

$$
\begin{aligned}
H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\
&= \mathbb{E}\left[\log \frac{1}{p(X)}\right]
\end{aligned}
$$



- $\log \frac{1}{p(x)}$ measures the "surprisingness" of observing $X = x$, so entropy is the "expected surprisingness" of X
- log is typically base 2: entropy is measured in "bits"[1]
- If natural $\log$ (denoted $\ln$), then entropy is measured in "nats"
- By convention, $0 \log 0 = 0$
- We sometimes write

$$H(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log p_n$$

i.e. the entropy of the random variable with distribution $(p_1, p_2, \ldots, p_n)$

[1] Fun fact! Shannon's 1948 paper was one of the first uses of the term "bit" for "binary digit".

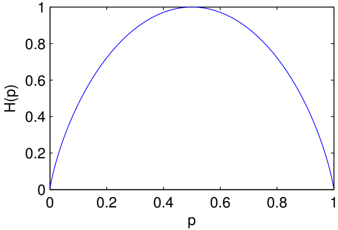
Example 1: Bernoulli random variable

- Let $X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$

  May equivalently write $X \sim \text{Bern}(p)$ (Bernoulli distribution)

$$H(X) = -p \log p - (1-p) \log(1-p)$$

  This quantity can be written $H(p, 1-p)$ or just $H(p)$ (binary entropy function)



- If $p = 0$ or $p = 1$, then $H(X) = 0$ (source is deterministic)
- If $p = 1/2$, then $H(X)$ is maximum (1 bit), since "uncertainty" is largest

Positivity of Entropy

Entropy is non-negative, i.e. $H(X) \geq 0$

Proof:

- Since $0 \leq p(x) \leq 1$, $\log \frac{1}{p(x)} \geq 0$.
- Thus $H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] \geq 0$.

Example 2: Uniform random variable

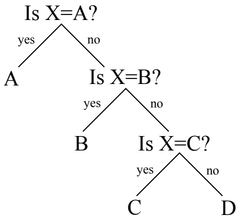
- Alphabet size $|\mathcal{X}| = m$ (could be $\{1, 2, \ldots, m\}$, or any other set with m elements)
- $p(x) = \frac{1}{m}$ for all $x \in \mathcal{X}$
- $H(X) = -\sum_{x \in \mathcal{X}} \frac{1}{m} \log \frac{1}{m} = \sum_{x \in \mathcal{X}} \frac{1}{m} \log m = m \cdot \frac{1}{m} \log m = \log m.$
- For example, if $m = 32$, $H(X) = \log 32 = 5$ bits

Example 3

Let $X = \begin{cases} A & \text{with probability } 1/2 \\ B & \text{with probability } 1/4 \\ C & \text{with probability } 1/8 \\ D & \text{with probability } 1/8 \end{cases}$

$$\begin{aligned} H(X) &= \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 \\ &= \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{8}(3) \\ &= 1.75 \text{ bits} \end{aligned}$$

- Entropy is minimum average number of yes/no questions to determine X
- This is exactly equal for this example; in general it is **approximately** equal[2]



[2] We'll get back to this

Joint Entropy

- Given random variables X, Y with joint PMF $p(x, y)$, the joint entropy is

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x,y)} \\
&= \mathbb{E}\left[\log\frac{1}{p(X,Y)}\right]
\end{aligned}
$$

- Similarly define joint entropy for more random variables: e.g. $H(X, Y, Z)$, $H(X_1, X_2, \ldots, X_n)$

Joint Entropy of Independent Random Variables

- If X and Y are independent, then $H(X,Y) = H(X) + H(Y)$

  Proof:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\
&= -\sum_{x,y} p(x)p(y) \log \left[p(x)p(y)\right] \\
&= -\sum_{x,y} p(x)p(y) \big(\log[p(x)] + \log[p(y)] \big) \\
&= -\sum_{x,y} p(x)p(y) \log p(x) - \sum_{x,y} p(x)p(y) \log p(y) \\
&= -\sum_{x} p(x) \log p(x) - \sum_{y} p(y) \log p(y) \\
&= H(X) + H(Y)
\end{aligned}
$$

- Similarly, if $X_1, X_2, \ldots, X_n$ are mutually independent, then

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

Conditional Entropy

- Given random variables X, Y with joint PMF $p(x,y)$, conditional PMF $p(y|x) = \dfrac{p(x,y)}{p(x)}$, the entropy of Y conditioned on the event that $X = x$ is

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

- The conditional entropy of Y given X is the above quantity averaged over X:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x,y} p(x,y)\log p(y|x) \\ &= \mathbb{E}\left[\log\frac{1}{p(Y|X)}\right]. \end{aligned}$$

- $H(Y|X) \geq 0$.

Chain Rule

- $H(X,Y) = H(X) + H(Y|X)$
- i.e. the uncertainty of X and Y is equal to the uncertainty of X plus the uncertainty of Y given X
- **Proof:**

$$\begin{aligned}
H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\
&= -\sum_{x,y} p(x,y) \log \left[p(x) \cdot p(y|x)\right] \\
&= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\
&= -\sum_{x} p(x) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

- Consequence: $H(Y|X) = H(X,Y) - H(X)$

Further Forms of the Chain Rule

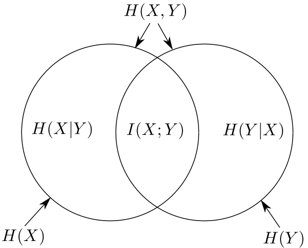
- $H(X,Y) = H(Y) + H(X|Y)$
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- $$\begin{aligned} H(X_1, X_2, \ldots, X_n|Z) &= H(X_1|Z) + H(X_2|X_1, Z) + \cdots \\ &\qquad + H(X_n|X_1, \ldots, X_{n-1}, Z) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}, Z) \end{aligned}$$

Mutual Information

- Given variables X, Y, **mutual information** is the amount of information in X about Y and vice versa:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \\ &= \mathbb{E}\left[\log\frac{p(X,Y)}{p(X)\,p(Y)}\right] \\ &= H(X)+H(Y)-H(X,Y) \\ &= H(X)-H(X|Y) \\ &= H(Y)-H(Y|X) \end{aligned}$$

- $I(X;Y) = I(Y;X)$
- Can also include multiple variables, e.g. $I(X;Y,Z)$, $I(X,Y;Z_1,\ldots,Z_n)$
- Venn diagram representation:

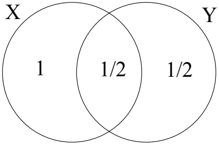


Example 1

$p(x,y)$			
$Y \setminus X$	1	2	3
1	1/4	1/4	0
2	1/4	0	1/4

- $p(x) = [1/2, 1/4, 1/4] \implies H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = \frac{3}{2}$
- $p(y) = [1/2, 1/2] \implies H(Y) = 1$
- $H(X,Y) = 4 \cdot \frac{1}{4}\log 4 = 2$
- $H(Y|X) = H(X,Y) - H(X) = 2 - \frac{3}{2} = \frac{1}{2}$

  Alternatively: $H(Y|X) = \frac{1}{2}H(Y|X=1) + \frac{1}{4}H(Y|X=2) + \frac{1}{4}H(Y|X=3) = \frac{1}{2}$
- $H(X|Y) = H(X,Y) - H(Y) = 2 - 1 = 1$
- $I(X;Y) = H(Y) - H(Y|X) = 1 - \frac{1}{2} = \frac{1}{2}$

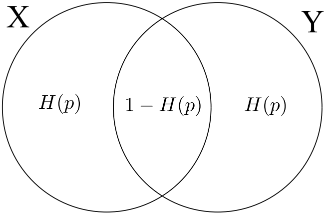


Example 2

- $\mathcal{X}, \mathcal{Y} = \{0, 1\}$, $X \sim \mathsf{Bern}(1/2)$, $p(y|x) = \begin{cases} 1-p & y = x \\ p & y \neq x \end{cases}$
- $H(X) = 1$, $H(Y) = 1$

$$\begin{aligned} H(Y|X) &= \frac{1}{2}H(Y|X=0) + \frac{1}{2}H(Y|X=1) \\ &= \frac{1}{2}H(p) + \frac{1}{2}H(p) \\ &= H(p) \end{aligned}$$

- $H(X, Y) = H(X) + H(Y|X) = 1 + H(p)$
- $I(X; Y) = H(Y) - H(Y|X) = 1 - H(p)$
- $H(X|Y) = H(X) - I(X; Y) = H(p)$



Conditional Mutual Information

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)\,p(y|z)} \\
&= \mathbb{E}\left[\log \frac{p(X,Y|Z)}{p(X|Z)\,p(Y|Z)}\right] \\
&= H(X|Z) - H(X|Y,Z) \\
&= \sum_{z} p(z) I(X;Y|Z=z)
\end{aligned}
$$

Chain rule for mutual information

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

  Proof:
$$
\begin{aligned}
I(X;Y,Z) &= H(Y,Z) - H(Y,Z|X) \\
&= \big[H(Y) + H(Z|Y)\big] - \big[H(Y|X) + H(Z|X,Y)\big] \\
&= \big[H(Y) - H(Y|X)\big] + \big[H(Z|Y) - H(Z|X,Y)\big] \\
&= I(X;Y) + I(X;Z|Y)
\end{aligned}
$$

- In general:

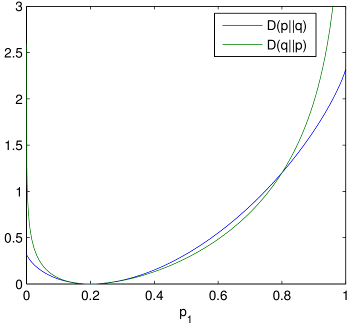
$$
\begin{aligned}
I(X;Y_1,Y_2,\dots,Y_n) &= I(X;Y_1) + I(X;Y_2|Y_1) + \cdots + I(X;Y_n|Y_1,\dots,Y_{n-1}) \\
&= \sum_{i=1}^{n} I(X;Y_i|Y_1,\dots,Y_{i-1})
\end{aligned}
$$

Relative Entropy (or Kullback-Leibler Divergence)

between two distributions $p(x)$ and $q(x)$ on the same alphabet $\mathcal{X}$:

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]$$

- $D(p\|q)$ is roughly a distance between two distributions, but $D(p\|q) \neq D(q\|p)$, and it does not satisfy the triangle inequality
- $D(p\|q) \geq 0$ with equality iff $p = q$ (we'll prove later)
- **Example:** $p = [p_1 \;\; 1 - p_1]$, $q = [0.2 \;\; 0.8]$

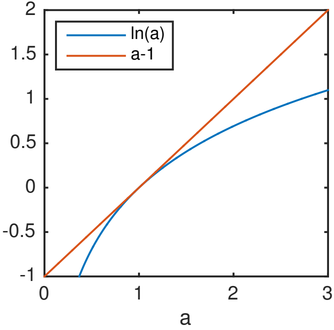


- $I(X;Y) = D\big(p(x,y) \| p(x)\, p(y)\big)$

A Simple Inequality

For $a > 0$, $\quad \ln a \leq a - 1$, $\quad$ with equality iff $a = 1$.

Proof by picture:



Non-negativity of Relative Entropy

For any $p(x)$, $q(x)$, $D(p\|q) \geq 0$ with equality iff $p = q$

Proof:

$$
\begin{aligned}
D(p\|q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= (\log e) \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \\
&= -(\log e) \sum_{x \in \mathcal{X}} p(x) \ln \frac{q(x)}{p(x)} \\
&\geq -(\log e) \sum_{x \in \mathcal{X}} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \\
&= -(\log e) \sum_{x \in \mathcal{X}} [q(x) - p(x)] \\
&= 0.
\end{aligned}
$$

Equality iff $\dfrac{q(x)}{p(x)} = 1$ for all $x \in \mathcal{X}$, i.e. $p = q$.

Non-negativity of Mutual Information

- $I(X;Y) \geq 0$, with equality iff X and Y are independent

  Proof:

  $$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} = D\big(p(x,y) \| p(x)\,p(y)\big) \geq 0$$

  Equality iff $p(x,y) = p(x)\,p(y)$ (i.e. X and Y are independent)

- $I(X;Y|Z) \geq 0$, with equality iff X and Y are independent given Z
  i.e. $p(x,y|z) = p(x|z)p(y|z)$

  Proof:

  $$\begin{aligned} I(X;Y|Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)\,p(y|z)} \\ &= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)\,p(y|z)} \\ &= \sum_{z} p(z) D\big(p(x,y|z) \| p(x|z)\,p(y|z)\big) \geq 0 \end{aligned}$$

Additional Properties of Entropy & Mutual Information

- $H(Y|X) \leq H(Y)$ (i.e. conditioning reduces entropy)
  Proof: $H(Y) - H(Y|X) = I(X;Y) \geq 0$

  Equality iff X and Y are independent
- $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$

  Proof:

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

  Equality iff $X_1, X_2, \ldots, X_n$ all independent
- $H(X) \geq 0$, with equality iff X is deterministic
- $H(Y|X) \geq 0$, with equality iff $Y = g(X)$ for some function g
- $I(X;X) = H(X)$
- $H(X|X) = 0$

Uniform bound

$H(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ is the number of elements in alphabet $\mathcal{X}$

Proof:

Let $p(x)$ be the PMF of X, and $q(x) = \dfrac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$

$$
\begin{aligned}
0 &\leq D(p\|q) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_x p(x) \log \frac{p(x)}{1/|\mathcal{X}|} \\
&= \sum_x p(x) \log p(x) + \sum_x p(x) \log |\mathcal{X}| \\
&= -H(X) + \log |\mathcal{X}|
\end{aligned}
$$

Equality iff $p = q$, i.e. X is uniformly distributed on $\mathcal{X}$

Conditioning DOES NOT Reduce Mutual Information

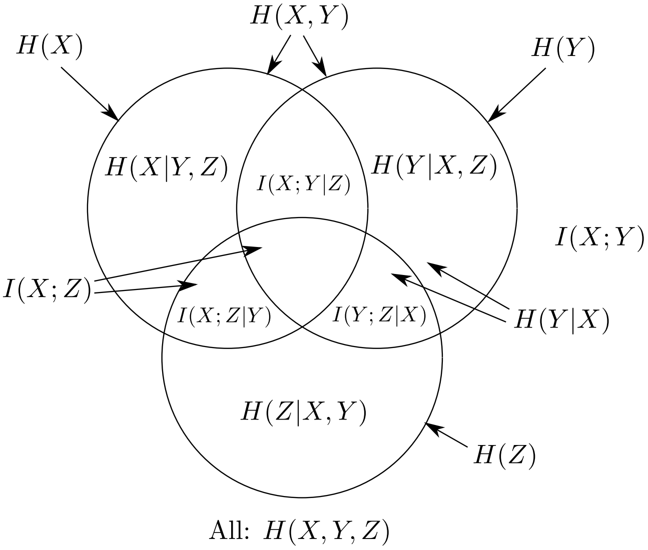
Example: Let $X \sim \text{Bern}(1/2)$ and $Y \sim \text{Bern}(1/2)$ be independent
Let $Z = X \oplus Y$

x	y	z	$p(x,y,z)$
0	0	0	1/4
0	0	1	0
0	1	0	0
0	1	1	1/4
1	0	0	0
1	0	1	1/4
1	1	0	1/4
1	1	1	0

$$I(X;Y) = 0$$
$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 1 - 0 = 1$$

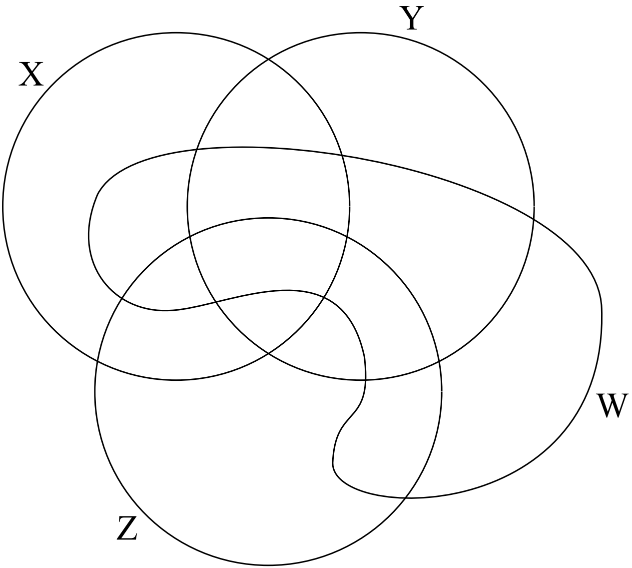
In this case, conditioning **increases** mutual information!

Venn Diagram Representation for 3 Variables



Need to be careful about the center section: it is given by $I(X;Y) - I(X;Y|Z)$ (sometimes written $I(X;Y;Z)$) which can be positive or negative

Venn Diagram Representation for 4 Variables



Markov Chains

- Random variables X, Y, Z form a **Markov chain** denoted $X \to Y \to Z$ if X and Z are conditionally independent given Y, i.e.

$$p(x, z|y) = p(x|y)\, p(z|y)$$

- $X \to Y \to Z$ iff $p(z|x, y) = p(z|y)$:

$$p(z|x, y) = \frac{p(x, y, z)}{p(x, y)} = \frac{p(y)\, p(x|y)\, p(z|y)}{p(x, y)} = \frac{p(x, y)\, p(z|y)}{p(x, y)} = p(z|y)$$

- $X \to Y \to Z$ iff $I(X; Z|Y) = 0$
- $X \to Y \to Z$ iff $Z \to Y \to X$ (sometimes written $X \leftrightarrow Y \leftrightarrow Z$)
- If $Z = f(Y)$, then $X \to Y \to Z$
- $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ if

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_2) \cdots p(x_{n-1}|x_{n-2})\, p(x_n|x_{n-1})$$

- $X \to Y \to Z \to W$ implies $X \to Y \to Z$ and $X \to Y \to W$

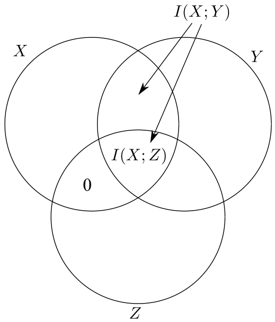
Data Processing Inequality

If $X \to Y \to Z$ is a Markov chain, then $I(X;Z) \leq I(X;Y)$
(i.e. shared information cannot **increase** by processing)

Proof: Assuming $X \to Y \to Z$,

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

Proof by Venn diagram:

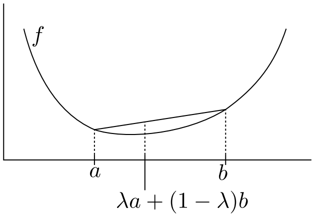


Special case: $I(X;g(Y)) \leq I(X;Y)$ for any function g, since we may take $Z = g(Y)$

Convex and Concave Functions

- A real-valued function $f(a)$ with $a \in \mathbb{R}^n$ is **convex** if for all $0 \le \lambda \le 1$ and a, b

$$f(\lambda a + (1-\lambda)b) \le \lambda f(a) + (1-\lambda)f(b)$$



If inequality is strict for all $0 < \lambda < 1$, then f is **strictly convex**

- A function $f(a)$ is **concave** if for all $0 \le \lambda \le 1$ and a, b

$$f(\lambda a + (1-\lambda)b) \ge \lambda f(a) + (1-\lambda)f(b)$$

- f is convex iff $-f$ is concave
- For scalar a, $f(a)$ is convex iff $f''(a) \ge 0$ for all a, and strictly convex iff $f''(a) > 0$ for all a

Convexity Properties of Entropy and Mutual Information

- $H(X)$ is a concave function of $p(x)$

  Proof:
  Let $f(p) = -p \log p$. f is strictly concave for $p \geq 0$:

$$f'(p) = -\log p - \frac{p \log e}{p} = -\log p - \log e$$
$$f''(p) = -\frac{\log e}{p} < 0.$$

  Thus $H(X) = \sum_x f(p(x))$ is a concave function of the vector $(p(x), x \in \mathcal{X})$

- Since $p(x,y) = p(x)\,p(y|x)$, think of $I(X;Y)$ as a function of $p(x)$ and $p(y|x)$
  - For a fixed $p(y|x)$, $I(X;Y)$ is a concave function of $p(x)$
  - For a fixed $p(x)$, $I(X;Y)$ is a convex function of $p(y|x)$

    (proofs in Cover-Thomas)
- $D(p\|q)$ is convex in the pair (p,q) (proof in Cover-Thomas)

Fano's Inequality

Robert Fano (1917–2016)

- Given $X \to Y \to \hat{X}$, where $\hat{X}$ is an estimate of X using Y
- Let $P_e = \Pr\{X \neq \hat{X}\}$
- Then
$$H(X|Y) \leq H(P_e) + P_e \log\big(|\mathcal{X}| - 1\big)$$

Consequences

- If $P_e = 0$, then $H(X|Y) = 0$ (i.e. X is a function of Y)
- $H(X|Y) \leq 1 + P_e \log|\mathcal{X}|$ (weaker form of Fano's inequality that we will often use)

Proof:

Let $E = \begin{cases} 0, & \text{if } X = \hat{X} \\ 1, & \text{if } X \neq \hat{X} \end{cases}$

$$
\begin{aligned}
H(X|Y) &= H(X) - I(X;Y) \\
&\leq H(X) - I(X;\hat{X}) \\
&= H(X|\hat{X}) \\
&= H(X,E|\hat{X}) - H(E|X,\hat{X}) \\
&= H(X,E|\hat{X}) \\
&= H(E|\hat{X}) + H(X|E,\hat{X}) \\
&\leq H(E) + H(X|E,\hat{X}) \\
&= H(P_e) + \Pr\{E=0\}H(X|\hat{X},E=0) + \Pr\{E=1\}H(X|\hat{X},E=1) \\
&= H(P_e) + P_e H(X|\hat{X},E=1) \\
&\leq H(P_e) + P_e \log(|\mathcal{X}| - 1)
\end{aligned}
$$