Information Theory

Lecture 1

- Course introduction
- Entropy, relative entropy and mutual information: Cover & Thomas (CT) 2.1–5
- Important inequalities: CT2.6–8, 2.10

Information Theory

- Founded by Claude Shannon in 1948.
  - C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379-423, 623-656, 1948
  - "*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*"
- Information theory is concerned with
  - *communication*, *information*, *entropy*, *coding*, *achievable performance*, *performance bounds*, *limits*, *inequalities*,. . .

Shannon's Coding Theorems

- Two *source coding theorems*
  - Discrete sources
  - Analog sources
- The *channel coding theorem*
- The *joint source–channel coding theorem*

Noiseless Coding of Discrete Sources

- A *discrete source* $\mathcal{S}$ (finite number of possible values per output sample) that produces raw data at a rate of R bits per symbol.
- The source has *entropy* $H(\mathcal{S}) \leq R$.
- **Result** (CT5): $\mathcal{S}$ can be *coded* into an alternative, but equivalent, representation at $H(\mathcal{S})$ bits per symbol. The original representation can be recovered *without errors*. This is *impossible* at rates lower than $H(\mathcal{S})$.
- Hence, $H(\mathcal{S})$ is a measure of the "real" information content in the output of $\mathcal{S}$. The coding process removes all that is *redundant*.

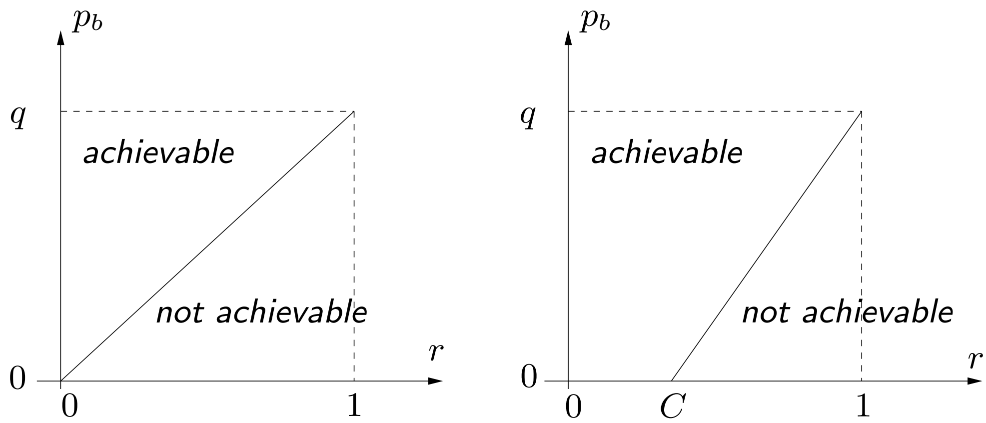
Coding of Analog Sources

- A discrete-time analog source $\mathcal{S}$ (e.g., a sampled speech signal).
- For storage or transmission the source needs to be *coded* ("quantized") into a discrete representation $\hat{\mathcal{S}}$, at R bits per source sample. This process is generally *irreversible*...
- A measure $d(\mathcal{S}, \hat{\mathcal{S}}) \geq 0$ of the *distortion* induced by the coding.
- A function $D_{\mathcal{S}}(R)$, the *distortion-rate function* of the source.
- **Result** (CT10): There exists a way of coding $\mathcal{S}$ into $\hat{\mathcal{S}}$ at rate R (bits per sample), with $d(\mathcal{S}, \hat{\mathcal{S}}) = D_{\mathcal{S}}(R)$. At rate R it is *impossible* to achieve a *lower* distortion than $D_{\mathcal{S}}(R)$.

Channel Coding

- Consider transmitting a stream of *information bits* $b \in \{0, 1\}$ over a binary *channel* with bit-error probability q and *capacity* $C = C(q)$.
- A *channel code* takes a block of k information bits, b, and maps these into a new block of $n > k$ coded bits, c, hence introducing *redundancy*. The "information content" per coded bit is $r = k/n$.
- The coded bits, c, are transmitted and a *decoder* at the receiver produces estimates $\hat{b}$ of the original information bits.
- Overall error probability $p_b = \Pr(\hat{b} \neq b)$.
- **Result** (CT7): As long as $r < C$, a code exists that can achieve $p_b \to 0$. At rates $r > C$ this is *impossible*. Hence, C is a measure of the "quality" or "noisiness" of the channel.

Achievable Rates



before Shannon   after Shannon

The left plot illustrates the rates believed to be achievable before 1948. The right plot shows the rates Shannon proved were achievable. Shannon's remarkable result is that, at a particular channel bit-error rate q, *all rates* below the *channel capacity* $C(q)$ are achievable with $p_b \to 0$.

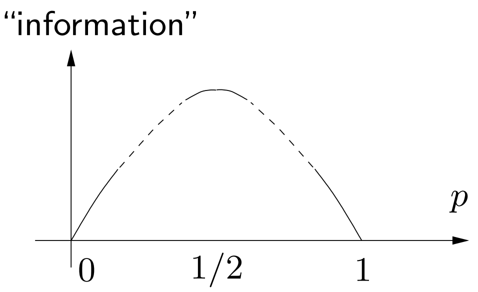
Course Outline

- 1–2: Introduction to Information Theory
  - Entropy, mutual information, inequalities,...
- 3: Data compression
  - Huffman, Shannon-Fano, arithmetic, Lemper-Ziv,...
- 4–5: Channel capacity and coding
  - Block channel coding, discrete and Gaussian channels,...
- 6–8: Linear block codes (book by Roth)
  - G and H matrices, finite fields, cyclic codes and polynomials over finite fields, BCH and Reed-Solomon codes,...
- 9–11: More channel capacity
  - Error exponents, non-stationary and/or non-ergodic channels,...

Senior undergraduate version: 1–8; Ph.D. student version: 1–11

Entropy and Information

- Consider a binary random variable $X \in \{0,1\}$ and let $p = \Pr(X=1)$.
- *Before* we observe the value of X there is a certain amount of *uncertainty* about its value. *After* getting to know the value of X, we gain *information*. **Uncertainty** $\leftrightarrow$ **Information**
- The *average* amount of uncertainty lost = information gained, over a large number of observations, should behave like

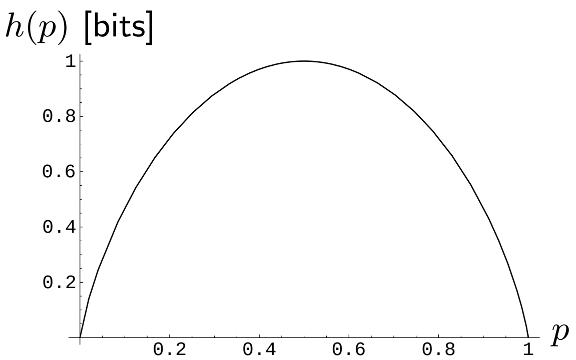


- Define the *entropy* $H(X)$ of the binary variable X as

$$H(X) = \Pr(X=1)\cdot\log\frac{1}{\Pr(X=1)} + \Pr(X=0)\cdot\log\frac{1}{\Pr(X=0)} =$$
$$= -p\cdot\log p - (1-p)\cdot\log(1-p) \triangleq h(p)$$

where $h(x)$ is the *binary entropy function*.

- $\log = \log_2$: unit = *bits*; $\log = \log_e = \ln$: unit = *nats*



- Entropy for a general discrete variable X with alphabet $\mathcal{X}$ and pmf $p(x) \triangleq \Pr(X = x),\ \forall x \in \mathcal{X}$

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

- $H(X)$ = *the average amount of uncertainty removed when observing the value of* X = *the information obtained when observing* X
- It holds that $0 \leq H(X) \leq \log |\mathcal{X}|$
- Entropy for an n-tuple $\mathbf{X} = (X_1, \ldots, X_n)$

$$H(\mathbf{X}) = H(X_1, \ldots, X_n) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

- *Conditional entropy* of Y given $X = x$

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

  - $H(Y|X = x)$ = *the average information obtained when observing* Y *when it is already known that* $X = x$
- *Conditional entropy* of Y given X (on the average)

$$H(Y|X) \triangleq \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

  - Define $g(x) = H(Y|X = x)$. Then $H(Y|X) = Eg(X)$.
- *Chain rule*

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

(c.f., $p(x, y) = p(y|x)p(x)$)

- *Relative entropy* between the pmf's $p(\cdot)$ and $q(\cdot)$

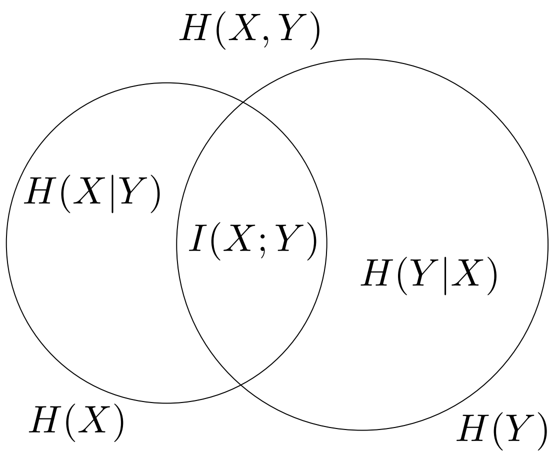
$$D(p\|q) \triangleq \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

  - Measures the "*distance*" between $p(\cdot)$ and $q(\cdot)$. If $X \sim p(x)$ and $Y \sim q(y)$ then a *low* $D(p\|q)$ means that X and Y are *close*, in the sense that their "statistical structure" is similar.

- *Mutual information*

$$\begin{aligned} I(X;Y) &\triangleq D\big(p(x,y)\|p(x)p(y)\big) \\ &= \sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

  - $I(X;Y)$ = *the average information about X obtained when observing Y* (and vice versa).

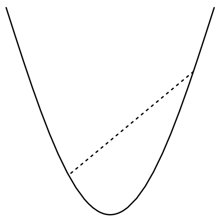


$$I(X;Y) = I(Y;X)$$
$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$I(X;X) = H(X)$$
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Inequalities

- Jensen's inequality
  - based on *convexity*
  - *application:* general purpose inequality
- Log sum inequality
  - based on *Jensen's inequality*
  - *application:* convexity as a function of distribution
- Data processing inequality
  - based on *Markov property*
  - *application:* cannot generate "extrinsic" information
- Fano's inequality
  - based on *conditional entropy*
  - *application:* lower bound on error probability

Convex Functions



$f : \mathcal{D}_f \subset \mathbb{R}^n \to \mathbb{R}$

- *convex*
  $\mathcal{D}_f$ is convex[1] and for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}_f,\ \lambda \in [0,1]$
  $$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \le \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$$
- *strictly convex*
  strict inequality for $\mathbf{x} \neq \mathbf{y},\ \lambda \in (0,1)$
- *(strictly) concave*
  $-f$ (strictly) convex

[1] $\mathbf{x}, \mathbf{y} \in \mathcal{D}_f,\ \lambda \in [0,1] \implies \lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \mathcal{D}_f$

Jensen's Inequality

- For f convex and a random $\mathbf{X} \in \mathbb{R}^n$,

$$f(E[\mathbf{X}]) \leq E[f(\mathbf{X})]$$

- Reverse inequality for f concave
- For f strictly convex (or strictly concave),

$$f(E[\mathbf{X}]) = E[f(\mathbf{X})] \implies \Pr(\mathbf{X} = E[\mathbf{X}]) = 1$$

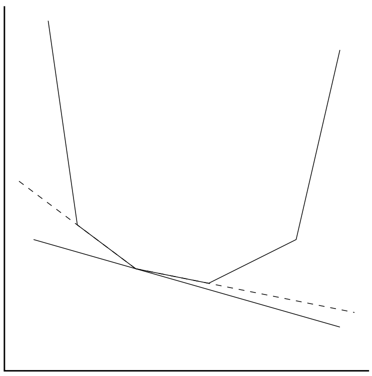
Quick Proof of Jensen's Inequality

Supporting hyperplane characterization of convexity: For f convex and any $\mathbf{x}_0 \in \mathcal{D}_f$ there exists a $\mathbf{n}_0$ such that for all $\mathbf{x} \in \mathcal{D}_f$

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{n_0} \cdot (\mathbf{x} - \mathbf{x}_0)$$

Let $\mathbf{x}_0 = E[\mathbf{X}]$ and take expectations

$$E[f(\mathbf{X})] \geq f(E[\mathbf{X}]) + \mathbf{n}_0 \cdot E\big[(\mathbf{X} - \mathrm{E}[\mathbf{X}])\big]$$



Applications of Jensen's Inequality

- Uniform distribution maximizes entropy ($f(x) = \log x$ concave)

$$H(X) = E \log \frac{1}{p(X)} \le \log \left[E \frac{1}{p(X)} \right] = \log |\mathcal{X}|$$

with equality iff $\frac{1}{p(X)} = \text{constant}$ w.p. 1

- Information Inequality ($f(x) = x \log x$ convex)

$$D(p\|q) = E_q \frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)} \ge E_q \left(\frac{p(X)}{q(X)} \right) \log E_q \frac{p(X)}{q(X)} = 0$$

with equality iff $\frac{q(X)}{p(X)} = \text{constant}$ w.p. 1 (i.e. $p \equiv q$)

- Non-negativity of mutual information

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

with equality iff X and Y independent

- Conditioning reduces entropy

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

with equality iff X and Y independent

- Independence bound on entropy

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

with equality iff X_i independent

similar inequalities hold with extra conditioning

The Log Sum Inequality

For non-negative $a_1, a_2, \ldots, a_n$ 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 $\frac{a_i}{b_i} = \text{constant}$

Applications

- $D(p\|q)$ is convex in the pair (p, q)
- $H(p)$ is concave in p
- $I(X;Y)$ is concave in $p(x)$ for fixed $p(y|x)$
- $I(X;Y)$ is convex in $p(y|x)$ for fixed $p(x)$

Markov Property

- Given the Present, the Past and the Future are independent
- Formally, $X \to Y \to Z$ *Markov* if

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

- Symmetric! $X \to Y \to Z \implies Z \to Y \to X$

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

- Conditional independence

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

- In particular, $X \to Y \to f(Y)$

Data Processing Inequality

$$X \to Y \to Z \implies I(X;Z) \le I(X;Y)$$

In particular,

$$I(X;f(Y)) \le I(X;Y)$$

$\Rightarrow$ No clever manipulation of the data can extract additional information that is not already present in the data itself.

Proof of the Data Processing Inequality

Using the chain rule, expand in two different ways

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + \underbrace{I(X;Y|Z)}_{\ge 0} \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

Corollary

$$X \to Y \to Z \implies I(X;Y|Z) \le I(X;Y)$$

Caution: this last inequality need not hold in general

Fano's Inequality

- Consider the following estimation problem (discrete RV's):
  X random variable of interest
  Y observed random variable
  $\hat{X} = f(Y)$ estimate of X based on Y
- Define the probability of error as

$$P_e = \Pr(\hat{X} \neq X)$$

- Fano's inequality lower bounds P_e

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

$[h(x) = -x \log x - (1-x)\log(1-x)]$

Proof of Fano's Inequality

- Define an indicator random variable for the error event

$$E = \begin{cases} 1, & \hat{X} \neq X \\ 0, & \hat{X} = X \end{cases} ; \qquad \Pr(E=1) = 1 - \Pr(E=0) = P_e$$

- Using the chain rule, expand in two different ways

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

$$H(X|E,Y) = P_e\, H(X|Y, E=1) + (1-P_e)\, \underbrace{H(X|Y,E=0)}_{=0}$$