ECE 534: recap

Generic communication block diagram

- Source
- Encoder
- Channel
- Decoder
- Destination
- Noise

- Remove redundancy
- Controlled adding of redundancy
- Decode signals, detect/correct errors
- Restore source
Efficient, reliable communications

- Source coding
  - Source = random variable
  - Ultimate data compression limit is the source’s entropy $H$

- Channel coding
  - Channel = conditional distributions
  - Ultimate transmission rate is the channel capacity $C$

Reliable communication possible $\iff H < C$
Topics outline

- Entropy and mutual information
- Asymptotic Equipartition Property (AEP)
- Data compression
- Channel capacity
- Differential entropy
- Gaussian channel
- Rate-distortion theory
- Examples from statistics, universal source coding
- Network information theory

Entropy and mutual information

- Entropy of a random variable $X$ with distribution $p(x)$

  $\text{prob}(x)$
  
  $\begin{align*}
  0.2 & \quad 0.5 & \quad 0.3 \\
  1 & \quad 3 & \quad 7 \\
  \Rightarrow & \quad H(X) = ?
  \end{align*}$

- Mutual information between two random variables $X$ and $Y$ with joint distribution $p(x,y)$

  $\begin{align*}
  p(x,y) \\
  \Rightarrow & \quad I(X,Y) = ?
  \end{align*}$
Entropy of a random variable

(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

Entropy of a uniform distribution

- Let X be uniformly distributed over 8 outcomes. What is the entropy of X?

\[
H(X) = \sum_{x=1}^{8} p(x) \log_2(p(x)) = - \sum_{x=1}^{8} \frac{1}{8} \log_2 \left( \frac{1}{8} \right) = \log_2(8) = 3 \text{ (bits)}
\]

- This is the number of bits needed to describe X!

- By extension, for a discrete random variable taking on K outcomes, the **maximal entropy** is attained by a uniform distribution and is equal to the number of bits needed to describe K:

\[
H(X) = \log_2(K)
\]
Entropy of a non-uniform distribution

- Suppose X represents the outcome of a horse race with 8 horses, which win with probabilities \( \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64} \)

\[
H(X) = -\frac{1}{2} \log_2 \left( \frac{1}{2} \right) - \frac{1}{4} \log_2 \left( \frac{1}{4} \right) - \frac{1}{8} \log_2 \left( \frac{1}{8} \right) - \frac{1}{16} \log_2 \left( \frac{1}{16} \right) - 4 \frac{1}{64} \log_2 \left( \frac{1}{64} \right)
\]

= 2 (bits)

- 8 outcomes, 3 bits? But on average can represent with 2 bits!

\[
\begin{align*}
(000, 001, 010, 011, 100, 101, 110, 111) & \quad \text{3 bits} \\
(0, 10, 110, 1110, 111100, 111101, 111110, 111111) & \quad \text{2 bits (on average!)}
\end{align*}
\]

Mutual information between 2 random variables:

\[
I(X; Y) = \sum p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
\]

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

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

(A) I(X;Y) is the reduction in the uncertainty about X due to knowledge of Y

(B) if X, Y are independent I(X;Y) = 0

(C) if X=Y then I(X;Y) = H(X)

(D) I(X;Y) is non-negative
Chapter 2: entropy, divergence and mutual information

**SUMMARY**

**Definition** The entropy $H(X)$ of a discrete random variable $X$ is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$  \hfill (2.156)

**Properties of $H$**

1. $H(X) \geq 0$.
2. $H_b(X) = (\log_b a) H_a(X)$.
3. (Conditioning reduces entropy) For any two random variables, $X$ and $Y$, we have

$$H(X|Y) \leq H(X)$$  \hfill (2.157)

with equality if and only if $X$ and $Y$ are independent.
4. $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$, with equality if and only if the $X_i$ are independent.
5. $H(X) \leq \log |\mathcal{X}|$, with equality if and only if $X$ is distributed uniformly over $\mathcal{X}$.
6. $H(p)$ is concave in $p$. 

*Proof:* We have

$$H(X) - D(p||r) = \sum_{x} p(x) \log p(x) + \sum_{x} p(x) \log r(x)p(x).$$

(2.151)

$$= \sum_{x} p(x) \log r(x)p(x).$$

(2.152)

$$\leq \sum_{x} p(x) 2 \log r(x).$$

(2.153)

$$= \sum_{x} p(x) r(x).$$

(2.154)

where the inequality follows from Jensen's inequality and the convexity of the function $f(y) = 2y^2$.

\[\square\]
Chapter 2: entropy, divergence and mutual information

**Definition** The relative entropy $D(p \parallel q)$ of the probability mass function $p$ with respect to the probability mass function $q$ is defined by

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2.158)$$

**Definition** The mutual information between two random variables $X$ and $Y$ is defined as

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2.159)$$

**Alternative expressions**

$$H(X) = E_p \log \frac{1}{p(X)} \quad (2.160)$$

$$H(X, Y) = E_p \log \frac{1}{p(X, Y)} \quad (2.161)$$

$$H(X|Y) = E_p \log \frac{1}{p(X|Y)} \quad (2.162)$$

$$I(X; Y) = E_p \log \frac{p(X, Y)}{p(X)p(Y)} \quad (2.163)$$

$$D(p||q) = E_p \log \frac{p(X)}{q(X)} \quad (2.164)$$

---

**Chapter 2: entropy, divergence and mutual information**

**Properties of $D$ and $I$**

1. $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$.
2. $D(p \parallel q) \geq 0$ with equality if and only if $p(x) = q(x)$, for all $x \in \mathcal{X}$.
3. $I(X; Y) = D(p(x, y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$ (i.e., $X$ and $Y$ are independent).
4. If $|\mathcal{X}| = m$, and $u$ is the uniform distribution over $\mathcal{X}$, then $D(p \parallel u) = \log m - H(p)$.
5. $D(p||q)$ is convex in the pair $(p, q)$.

**Chain rules**

Entropy: $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n H(X_i|X_{i-1}, \ldots, X_1)$.

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

Relative entropy: $D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

**Jensen’s inequality.** If $f$ is a convex function, then $Ef(X) \geq f(EX)$. 

[Cover+Thomas pg.41-43]
Chapter 2: entropy, divergence and mutual information

Log sum inequality. For $n$ positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.165}
$$

with equality if and only if $\frac{a_i}{b_i}$ is constant.

Data-processing inequality. If $X \rightarrow Y \rightarrow Z$ forms a Markov chain, $I(X; Y) \geq I(X; Z)$.

Sufficient statistic. $T(X)$ is sufficient relative to $\{f_\theta(x)\}$ if and only if $I(\theta; X) = I(\theta; T(X))$ for all distributions on $\theta$.

Fano’s inequality. Let $P_e = \Pr\{\hat{X}(Y) \neq X\}$. Then

$$
H(P_e) + P_e \log |X| \geq H(X|Y). \tag{2.166}
$$

Inequality. If $X$ and $X'$ are independent and identically distributed, then

$$
\Pr(X = X') \geq 2^{-H(X)}, \tag{2.167}
$$

Asymptotic Equipartition Property (AEP)

- Information theoretic analogy of the “law of large numbers”: terribly useful!

- Law of large numbers:

  For i.i.d. $X_i$, 
  $$
  \frac{1}{n} \sum_{i=1}^{n} X_i \approx E[X]
  $$

- Asymptotic equipartition theory:

  For i.i.d. $X_i$, 
  $$
  \frac{1}{n} \log_2 \left( \frac{1}{p(X_1, X_2, \cdots X_n)} \right) \approx H(X)
  $$

- Meaning: all events are equally surprising! Can talk about “typical sets”.
Chapter 3: The Asymptotic Equipartition Property

**AEP.** “Almost all events are almost equally surprising.” Specifically, if \( X_1, X_2, \ldots \) are i.i.d. \( \sim p(x) \), then
\[
\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}
\] (3.28)

**Definition.** The typical set \( A^{(n)}_\epsilon \) is the set of sequences \( x_1, x_2, \ldots, x_n \) satisfying
\[
2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.
\] (3.29)

**Properties of the typical set**
1. If \( (x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon \), then \( p(x_1, x_2, \ldots, x_n) = 2^{-n(H(X)+\epsilon)} \).
2. \( \Pr \{ A^{(n)}_\epsilon \} > 1 - \epsilon \) for \( n \) sufficiently large.
3. \( |A^{(n)}_\epsilon| \leq 2^{n(H(X)+\epsilon)} \), where \( |A| \) denotes the number of elements in set \( A \).

**Definition.** \( a_n \equiv b_n \) means that \( \frac{1}{n} \log \frac{a_n}{b_n} \to 0 \) as \( n \to \infty \).

**Smallest probable set.** Let \( X_1, X_2, \ldots, X_n \) be i.i.d. \( \sim p(x) \), and for \( \delta < \frac{1}{2} \), let \( B^{(n)}_\delta \subset A^n \) be the smallest set such that \( \Pr \{ B^{(n)}_\delta \} \geq 1 - \delta \). Then
\[
|B^{(n)}_\delta| \geq 2^n H.
\] (3.30)

Chapter 4: Entropy Rate

**Entropy rate.** Two definitions of entropy rate for a stochastic process are
\[
H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n),
\] (4.76)
\[
H'(X) = \lim_{n \to \infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1).
\] (4.77)

For a stationary stochastic process,
\[
H(X) = H'(X).
\] (4.78)

**Entropy rate of a stationary Markov chain**
\[
H(X) = - \sum_{ij} \mu_i P_{ij} \log P_{ij}.
\] (4.79)

**Second law of thermodynamics.** For a Markov chain:
1. Relative entropy \( D(\mu_n||\mu_s) \) decreases with time
2. Relative entropy \( D(\mu_n||\mu) \) between a distribution and the stationary distribution decreases with time.
3. Entropy \( H(X_n) \) increases if the stationary distribution is uniform.
Chapter 4: Entropy Rate

4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.

5. The conditional entropy $H(X_0|X_n)$ of the initial condition $X_0$ increases for any Markov chain.

Functions of a Markov chain. If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain and $Y_i = \phi(X_i)$, then

$$H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \leq H(Y) \leq H(Y_n|Y_{n-1}, \ldots, Y_1)$$

(4.80)

and

$$\lim_{n \to \infty} H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = H(Y) = \lim_{n \to \infty} H(Y_n|Y_{n-1}, \ldots, Y_1).$$

(4.81)

Data compression

• Given a source $X$ with distribution $p(x)$, what is the fundamental limit of compression of this source’s information?

$$H(X) = -\sum_x p(x) \log_2(p(x))$$

• Can we construct good codes to achieve this limit?
Chapter 5: Data Compression

**Kraft inequality.** Instantaneous codes $\Leftrightarrow \sum D^{-l_i} \leq 1$.

**McMillan inequality.** Uniquely decodable codes $\Leftrightarrow \sum D^{-l_i} \leq 1$.

**Entropy bound on data compression**

$$L \triangleq \sum p_i l_i \geq H_D(X). \quad (5.136)$$

**Shannon code**

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil \quad (5.137)$$

$$H_D(X) \leq L < H_D(X) + 1. \quad (5.138)$$

**Huffman code**

$$L^* = \min_{\sum D^{-l_i} \leq 1} \sum p_i l_i \quad (5.139)$$

$$H_D(X) \leq L^* < H_D(X) + 1. \quad (5.140)$$

Chapter 5: Data Compression

**Wrong code.** $X \sim p(x), l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil, L = \sum p(x)l(x)$:

$$H(p) + D(p||q) \leq L < H(p) + D(p||q) + 1. \quad (5.141)$$

**Stochastic processes**

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq L_n < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}. \quad (5.142)$$

**Stationary processes**

$$L_n \rightarrow H(X). \quad (5.143)$$

**Competitive optimality.** Shannon code $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil$ versus any other code $l'(x)$:

$$\Pr (l(X) \geq l'(X) + c) \leq \frac{1}{2^{c-1}}. \quad (5.144)$$
Channel capacity

- Information channel capacity:

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

- Operational channel capacity:

\[
\text{Highest rate (bits/channel use) that can communicate at reliably}
\]

- Channel coding theorem says: information capacity = operational capacity

Capacity in general

- Main idea was to reduce the rate (from a 27-letter input per channel use to a 9-letter input per channel use) so as to produce

Non-overlapping outputs!
Chapter 7: Channel Capacity

**Channel capacity.** The logarithm of the number of distinguishable inputs is given by
\[ C = \max_{p(x)} I(X; Y). \]

**Examples**
- Binary symmetric channel: \( C = 1 - H(p) \).
- Binary erasure channel: \( C = 1 - \epsilon \).
- Symmetric channel: \( C = \log |\mathcal{Y}| - H(\text{row of transition matrix}) \).

**Properties of \( C \)**
1. \( 0 \leq C \leq \min\{\log |\mathcal{X}|, \log |\mathcal{Y}|\} \).
2. \( I(X; Y) \) is a continuous concave function of \( p(x) \).

**Joint typicality.** The set \( A_n^{(a)} \) of jointly typical sequences \( \{(x^n, y^n)\} \) with respect to the distribution \( p(x, y) \) is given by
\[
A_n^{(a)} = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \right. \\
\left. \quad \quad \text{and} \quad \left| \frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \right. \\
\left. \quad \quad \text{and} \quad \left| \frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \right\},
\]
where \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \).

---

**Joint AEP.** Let \( (X^n, Y^n) \) be sequences of length \( n \) drawn i.i.d. according to \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \). Then:
1. \( \Pr((X^n, Y^n) \in A_n^{(a)}) \to 1 \) as \( n \to \infty \).
2. \( |A_n^{(a)}| \leq 2^{n(H(X,Y)+\epsilon)} \).
3. If \( (\tilde{X}^n, \tilde{Y}^n) \sim p(x^n) p(y^n) \), then \( \Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A_n^{(a)} \right) \leq 2^{-n(I(X;Y)-3\epsilon)} \).

**Channel coding theorem.** All rates below capacity \( C \) are achievable, and all rates above capacity are not; that is, for all rates \( R < C \), there exists a sequence of \( (2^{nR}, n) \) codes with probability of error \( \lambda^{(n)} \to 0 \). Conversely, for rates \( R > C \), \( \lambda^{(n)} \) is bounded away from 0.

**Feedback capacity.** Feedback does not increase capacity for discrete memoryless channels (i.e., \( C_{FB} = C \)).

**Source–channel theorem.** A stochastic process with entropy rate \( H \) cannot be sent reliably over a discrete memoryless channel if \( H > C \). Conversely, if the process satisfies the AEP, the source can be transmitted reliably if \( H < C \).
Differential entropy (continuous random variables)

- entropy defined for PMFs (discrete variables):
  \[
  H(X) = - \sum_x p(x) \log_2(p(x))
  \]

- differential entropy defined for distributions (continuous):
  \[
  h(X) = - \int f(x) \log(x) \, dx
  \]

Chapter 8: Differential Entropy

- \( h(X) = h(f) = - \int_S f(x) \log f(x) \, dx \) (8.81)
- \( f(X^n) \equiv 2^{-nh(X)} \) (8.82)
- \( \text{Vol}(A^{(n)}_\epsilon) \equiv 2^{nh(X)} \) (8.83)
- \( H([X]_{2^{-n}}) \approx h(X) + n \) (8.84)
- \( h(\mathcal{N}(0, \sigma^2)) = \frac{1}{2} \log 2\pi e\sigma^2 \) (8.85)
- \( h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| \) (8.86)
- \( D(f||g) = \int f \log \frac{f}{g} \geq 0 \) (8.87)
- \( h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n h(X_i|X_1, X_2, \ldots, X_{i-1}) \) (8.88)
Chapter 8: Differential Entropy

\[
\begin{align*}
    h(X|Y) & \leq h(X). \\
    h(aX) & = h(X) + \log |a|. \\
    I(X; Y) & = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \geq 0. \\
    \max_{E[X^{|K}]} h(X) & = \frac{1}{2} \log (2\pi e)^n |K|. \\
    E(X - \hat{X}(Y))^2 & \geq \frac{1}{2\pi e} 2^{h(X|Y)}.
\end{align*}
\]

$2^n h(X)$ is the effective alphabet size for a discrete random variable. $2^{nh(X)}$ is the effective support set size for a continuous random variable. $2^C$ is the effective alphabet size of a channel of capacity $C$.

Entropy of a Gaussian random variable

- differential entropies of Gaussian distributions:

\[
\begin{align*}
    h(\mathcal{N}(0, \sigma^2)) & = \frac{1}{2} \log \left(2\pi e \sigma^2\right) \\
    h(\mathcal{N}_n(\mu, K)) & = \frac{1}{2} \log \left((2\pi e)^n |K|\right)
\end{align*}
\]
Entropy maximization

- **Uniform** distribution maximizes entropy for a given # outcomes

\[
\max_{X:|X|=K} H(X) = \log_2(K)
\]

- **Gaussian** maximizes entropy for a given covariance constraint

\[
\max_{E[XX^T]=K} h(X) = \frac{1}{2} \log ((2\pi e)^n |K|)
\]

AWGN channel capacity

\[
C = \frac{1}{2} \log \left( \frac{|h|^2 P + P_N}{P_N} \right)
\]

\[
= \frac{1}{2} \log (1 + SNR) \quad \text{(bits/channel use)}
\]

What about bits/second and bandwidth of the channel?

\[
C = W \log_2 \left(1 + \frac{P}{WN_0}\right) \quad \text{(bits/second)}
\]
Chapter 9: Gaussian Channel

**Maximum entropy.** \( \max_{EX\sim\mathcal{N}} h(X) = \frac{1}{2} \log 2\pi e \alpha. \)

**Gaussian channel.** \( Y_i = X_i + Z_i; Z_i \sim \mathcal{N}(0, N); \) power constraint \( \frac{1}{n} \sum_{i=1}^{n} X_i^2 \leq P; \) and
\[
C = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right) \text{ bits per transmission.} \tag{9.163}
\]

**Bandlimited additive white Gaussian noise channel.** Bandwidth \( W; \) two-sided power spectral density \( N_0/2; \) signal power \( P; \) and
\[
C = W \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bits per second.} \tag{9.164}
\]

**Water-filling (k parallel Gaussian channels).** \( Y_j = X_j + Z_j, j = 1, 2, \ldots, k; Z_j \sim \mathcal{N}(0, N_j); \sum_{j=1}^{k} X_j^2 \leq P; \) and
\[
C = \frac{1}{2} \sum_{i=1}^{k} \log \left( 1 + \frac{(v - N_i)^+}{N_i} \right), \tag{9.165}
\]
where \( v \) is chosen so that \( \sum_{i}(v - N_i)^+ = nP. \)

---

Chapter 9: Gaussian Channel

**Additive nonwhite Gaussian noise channel.** \( Y_i = X_i + Z_i; Z \sim \mathcal{N}(0, K_Z); \) and
\[
C = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{(v - \lambda_i)^+}{\lambda_i} \right), \tag{9.166}
\]
where \( \lambda_1, \lambda_2, \ldots, \lambda_n \) are the eigenvalues of \( K_Z \) and \( v \) is chosen so that \( \sum_{i}(v - \lambda_i)^+ = P. \)

**Capacity without feedback**
\[
C_n = \max_{tr(K_X) \leq P} \frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|}, \tag{9.167}
\]

**Capacity with feedback**
\[
C_{n, FB} = \max_{tr(K_X) \leq P} \frac{1}{2n} \log \frac{|K_X + Z|}{|K_Z|}. \tag{9.168}
\]

**Feedback bounds**
\[
C_{n, FB} \leq C_n + \frac{1}{2}, \tag{9.169}
\]
\[
C_{n, FB} \leq 2C_n. \tag{9.170}
\]
Rate-distortion theory

- How to represent a continuous random variable using a finite representation?
- How to best compress in a lossy fashion?
- Minimize rate subject to a give distortion (or minimize distortion for a given rate)

\[ R(D) = \min_{p(\hat{x}|x); \sum_{(x, \hat{x})} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \]

where the minimization is over all conditional distributions \( p(\hat{x}|x) \) for which the joint distribution \( p(x, \hat{x}) = p(x)p(\hat{x}|x) \) satisfies the expected distortion constraint.

Rate distortion theorem. If \( R > R(D) \), there exists a sequence of codes \( \hat{X}^n(X^n) \) with the number of codewords \( |\hat{X}^n(\cdot)| \leq 2^{nR} \) with \( Ed(X^n, \hat{X}^n(X^n)) \rightarrow D \). If \( R < R(D) \), no such codes exist.
Chapter 10: Rate-Distortion Theory

**Bernoulli source.** For a Bernoulli source with Hamming distortion,

\[ R(D) = H(p) - H(D). \]  \hspace{1cm} (10.149)

**Gaussian source.** For a Gaussian source with squared-error distortion,

\[ R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}. \]  \hspace{1cm} (10.150)

**Source–channel separation.** A source with rate distortion \( R(D) \) can be sent over a channel of capacity \( C \) and recovered with distortion \( D \) if and only if \( R(D) < C \).

**Multivariate Gaussian source.** The rate distortion function for a multivariate normal vector with Euclidean mean-squared-error distortion is given by reverse water-filling on the eigenvalues.

---

Network information theory

- **Point to point capacity**

- **Multi-user capacity region**
Multiple-access channel. The capacity of a multiple-access channel \((\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})\) is the closure of the convex hull of all \((R_1, R_2)\) satisfying

\[
R_1 < I(X_1; Y|X_2), \\
R_2 < I(X_2; Y|X_1), \\
R_1 + R_2 < I(X_1, X_2; Y)
\]

for some distribution \(p_1(x_1)p_2(x_2)\) on \(\mathcal{X}_1 \times \mathcal{X}_2\).

The capacity region of the \(m\)-user multiple-access channel is the closure of the convex hull of the rate vectors satisfying

\[
R(S) \leq I(X(S); Y|X(S')) \quad \text{for all } S \subseteq \{1, 2, \ldots, m\}
\]

for some product distribution \(p_1(x_1)p_2(x_2)\cdots p_m(x_m)\).

---

Gaussian multiple-access channel. The capacity region of a two-user Gaussian multiple-access channel is

\[
R_1 \leq C \left( \frac{P_1}{N} \right), \\
R_2 \leq C \left( \frac{P_2}{N} \right), \\
R_1 + R_2 \leq C \left( \frac{P_1 + P_2}{N} \right)
\]

where

\[
C(x) = \frac{1}{2} \log(1 + x).
\]

Slepian–Wolf coding. Correlated sources \(X\) and \(Y\) can be described separately at rates \(R_1\) and \(R_2\) and recovered with arbitrarily low probability of error by a common decoder if and only if

\[
R_1 \geq H(X|Y), \\
R_2 \geq H(Y|X), \\
R_1 + R_2 \geq H(X, Y).
\]
Chapter 15: Network Information Theory

**Broadcast channels.** The capacity region of the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$ is the convex hull of the closure of all $(R_1, R_2)$ satisfying

\begin{align}
R_2 &\leq I(U; Y_2), \\
R_1 &\leq I(X; Y_1 | U)
\end{align}

(15.357)

(15.358)

for some joint distribution $p(u) p(x | u) p(y_1, y_2 | x)$.

**Relay channel.** The capacity $C$ of the physically degraded relay channel $p(y, y_1 | x, x_1)$ is given by

\[
C = \sup_{p(x, x_1)} \min \{ I(X, X_1; Y), I(X; Y_1 | X_1) \},
\]

(15.359)

where the supremum is over all joint distributions on $\mathcal{X} \times \mathcal{X}_1$.

---

Chapter 15: Network Information Theory

**Source coding with side information.** Let $(X, Y) \sim p(x, y)$. If $Y$ is encoded at rate $R_2$ and $X$ is encoded at rate $R_1$, we can recover $X$ with an arbitrarily small probability of error iff

\begin{align}
R_1 &\geq H(X | U), \\
R_2 &\geq I(Y; U)
\end{align}

(15.360)

(15.361)

for some distribution $p(y, u)$ such that $X \rightarrow Y \rightarrow U$.

**Rate distortion with side information.** Let $(X, Y) \sim p(x, y)$. The rate distortion function with side information is given by

\[
R_f(D) = \min_{p(w|x), f: Y \rightarrow W} \min_{f: Y \rightarrow W} I(X; W) - I(Y; W),
\]

(15.362)

where the minimization is over all functions $f$ and conditional distributions $p(w | x)$, $|W| \leq |X| + 1$, such that

\[
\sum_x \sum_w \sum_y p(x, y) p(w | x) d(x, f(y, w)) \leq D.
\]

(15.363)
Applications in other areas

- Gambling - skipped Ch.6
- Computer science (Kolmogorov Complexity) - skipped Ch.14
- Physics (thermodynamics) Ch. 16
- Philosophy of Science (Occam’s Razor) - related to Ch. 14
- Economics (investment, portfolio theory) - skipped Ch.16
- Biology (genetics, bio-statistics, neuroscience)
- Statistics - very powerful theory - skipped Ch.11
- Universal source coding - skipped Ch.13

Research in information theory

- http://www.isit2010.info/
- http://arxiv.org/list/cs.IT/recent
- http://www.itsoc.org/
- excellent tool/framework to merge with other fields
Practicalities

• Final exam Monday December 6, 3:30-5:30pm in Lincoln Hall 305

• 3 crib sheets permitted

• Will cover the whole semester

• 2.1 -- 2.10, 3.1 -- 3.3, 4.1 -- 4.2, 5.1 -- 5.10, 7.1 -- 7.13, 8.1 -- 8.6, 9.1 -- 9.6, 10.1 -- 10.6, 15.1 -- 15.7, 15.10

• 5 questions in 2 hours - so more time than midterms

Advertisement

• Next semester I teach ECE531: Detection and Estimation

![Diagram of Detection and Estimation](image)

Detected / classify which "hypothesis" happened

Estimate unknown parameters from noisy observations

Examples?
Evaluations

Thanks!