An Introduction to Information Theory

Guangyue Han

The University of Hong Kong

July, 2018@HKU

Outline of the Talk

Outline of the Talk

- Fundamentals of Information Theory
- Research Directions: Memory Channels
- Research Directions: Continuous-Time Information Theory

Fundamentals of Information Theory

The Birth of Information Theory

The Birth of Information Theory

The Birth of Information Theory

C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379-423,623-656, 1948.

Classical Information Theory

Classical Information Theory

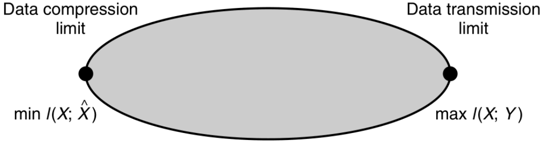

FIGURE 1.2. Information theory as the extreme points of communication theory.

Information Theory Nowadays

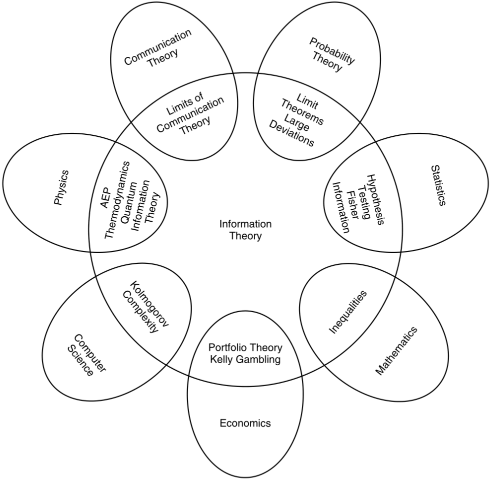

FIGURE 1.1. Relationship of information theory to other fields.

Entropy: Definition

Entropy: Definition

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

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

Entropy: Definition

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

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

Let

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

Then,

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

Entropy: Measure of Uncertainty

Entropy: Measure of Uncertainty

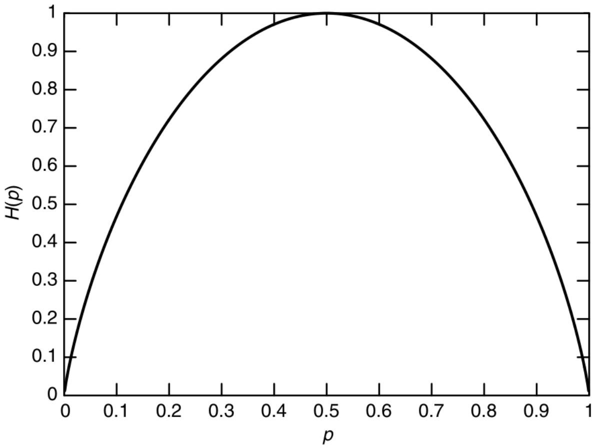

FIGURE 2.1. $H(p)$ vs. p.

Joint Entropy and Conditional Entropy

Joint Entropy and Conditional Entropy

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

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

Joint Entropy and Conditional Entropy

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

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

If $(X, Y) \sim p(x, y)$, the **conditional entropy** $H(Y|X)$ is defined as

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

Joint Entropy and Conditional Entropy

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

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

If $(X, Y) \sim p(x, y)$, the **conditional entropy** $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \end{aligned}$$

Joint Entropy and Conditional Entropy

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

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

If $(X, Y) \sim p(x, y)$, the **conditional entropy** $H(Y|X)$ is defined as

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

All Entropies Together

All Entropies Together

Chain Rule

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

All Entropies Together

Chain Rule

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

Proof.

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

All Entropies Together

Chain Rule

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

Proof.

$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \end{aligned}$$

All Entropies Together

Chain Rule

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

Proof.

$$\begin{aligned} H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x, y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x)p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x) \end{aligned}$$

All Entropies Together

Chain Rule

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

Proof.

$$\begin{aligned} H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x, y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x)p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x) \end{aligned}$$

All Entropies Together

Chain Rule

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

Proof.

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X).
\end{aligned}$$

Mutual Information

Mutual Information

Original Definition

The mutual information $I(X;Y)$ between two discrete random variables X, Y with joint distribution $p(x,y)$ is defined as

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

Mutual Information

Original Definition

The mutual information $I(X;Y)$ between two discrete random variables X, Y with joint distribution $p(x,y)$ is defined as

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

Alternative Definitions

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

Mutual Information

Original Definition

The mutual information $I(X;Y)$ between two discrete random variables X, Y with joint distribution $p(x,y)$ is defined as

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

Alternative Definitions

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

Mutual Information

Original Definition

The mutual information $I(X;Y)$ between two discrete random variables X, Y with joint distribution $p(x, y)$ is defined as

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

Alternative Definitions

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

Mutual Information and Entropy

Mutual Information and Entropy

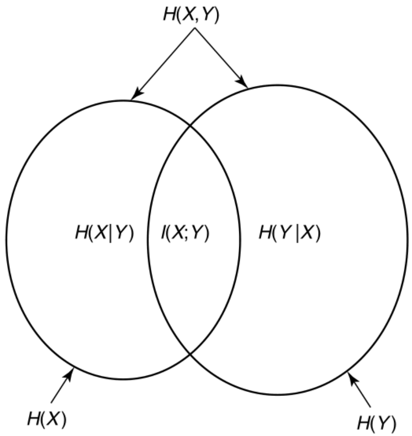

FIGURE 2.2. Relationship between entropy and mutual information.

Asymptotic Equipartition Property Theorem

Asymptotic Equipartition Property Theorem

AEP Theorem

If $X_1, X_2, \ldots,$ are i.i.d $\sim p(x)$, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}$$

Asymptotic Equipartition Property Theorem

AEP Theorem

If $X_1, X_2, \ldots,$ are i.i.d $\sim p(x)$, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}$$

Proof.

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i)$$

□

Asymptotic Equipartition Property Theorem

AEP Theorem

If $X_1, X_2, \ldots,$ are i.i.d $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}$$

Proof.

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \to -\mathbb{E}[\log p(X)]$$

□

Asymptotic Equipartition Property Theorem

AEP Theorem

If $X_1, X_2, \ldots,$ are i.i.d $\sim p(x)$, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}$$

Proof.

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \to -\mathbb{E}[\log p(X)] = H(X).$$

□

The Shannon-McMillan-Breiman Theorem

The Shannon-McMillan-Breiman Theorem

The Shannon-McMillan-Breiman Theorem

Let $\{X_n\}$ be a finite-valued stationary ergodic process. Then, with probability 1,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X),$$

The Shannon-McMillan-Breiman Theorem

The Shannon-McMillan-Breiman Theorem

Let $\{X_n\}$ be a finite-valued stationary ergodic process. Then, with probability 1,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X),$$

where $H(X)$ here denotes the entropy rate of the process $\{X_n\}$, namely,

$$H(X) = \lim_{n\to\infty} H(X_1, X_2, \ldots, X_n)/n.$$

The Shannon-McMillan-Breiman Theorem

The Shannon-McMillan-Breiman Theorem

Let $\{X_n\}$ be a finite-valued stationary ergodic process. Then, with probability 1,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X),$$

where $H(X)$ here denotes the entropy rate of the process $\{X_n\}$, namely,

$$H(X) = \lim_{n\to\infty} H(X_1, X_2, \ldots, X_n)/n.$$

Proof.

There are many. The simplest is the **sandwich** argument by Algoet and Cover [1988]. □

Typical Set: Definition and Properties

Typical Set: Definition and Properties

Definition

The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$

Typical Set: Definition and Properties

Definition

The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$

Properties

Typical Set: Definition and Properties

Definition

The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$

Properties

- $(1-\varepsilon)2^{n(H(X)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ for n sufficiently large.

Typical Set: Definition and Properties

Definition

The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$H(X) - \epsilon \le -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \le H(X) + \epsilon.$$

Properties

- $(1-\varepsilon)2^{n(H(X)-\epsilon)} \le |A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$ for n sufficiently large.
- $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for n sufficiently large.

Typical Set: A Pictorial Description

Typical Set: A Pictorial Description

Consider all the instances $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ of i.i.d. $(X_1, X_2, \cdots, X_n)$ with distribution $p(x)$.

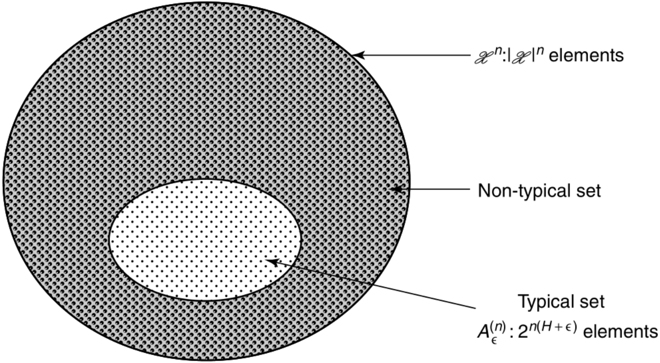

FIGURE 3.1. Typical sets and source coding.

Source Coding (Data Compression)

Source Coding (Data Compression)

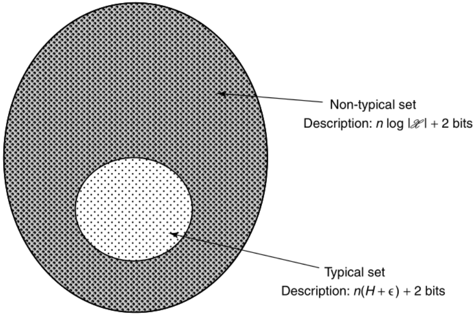

FIGURE 3.2. Source code using the typical set.

Source Coding (Data Compression)

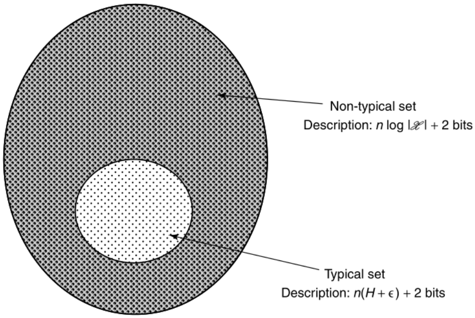

FIGURE 3.2. Source code using the typical set.

- Represent each typical sequence with about $nH(X)$ bits.

Source Coding (Data Compression)

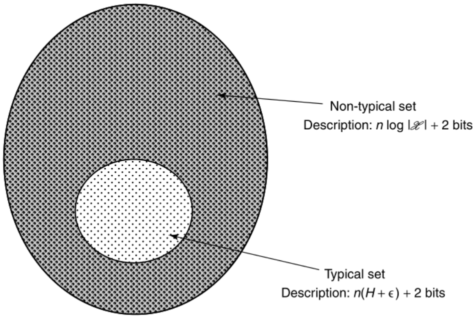

FIGURE 3.2. Source code using the typical set.

- Represent each typical sequence with about $nH(X)$ bits.
- Represent each non-typical sequence with about $n \log |\mathcal{X}|$ bits.

Source Coding (Data Compression)

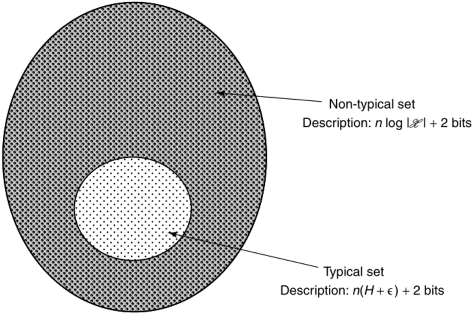

FIGURE 3.2. Source code using the typical set.

- Represent each typical sequence with about $nH(X)$ bits.
- Represent each non-typical sequence with about $n \log |\mathcal{X}|$ bits.
- Then we have a one-to-one and easily decodable code.

Shannon's Source Coding Theorem

Shannon's Source Coding Theorem

The average bits needed is

Shannon's Source Coding Theorem

The average bits needed is

$$\mathbb{E}[l(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n)$$

Shannon's Source Coding Theorem

The average bits needed is

$$\mathbb{E}[l(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n)$$

$$= \sum_{x_1, \ldots, x_n \in A_\epsilon^{(n)}} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n) + \sum_{x_1, \ldots, x_n \notin A_\epsilon^{(n)}} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n)$$

Shannon's Source Coding Theorem

The average bits needed is

$$\mathbb{E}[l(X_1,\dots,X_n)] = \sum_{x_1,\dots,x_n} p(x_1,\dots,x_n) l(x_1,\dots,x_n)$$

$$= \sum_{x_1,\dots,x_n \in A_\epsilon^{(n)}} p(x_1,\dots,x_n) l(x_1,\dots,x_n) + \sum_{x_1,\dots,x_n \notin A_\epsilon^{(n)}} p(x_1,\dots,x_n) l(x_1,\dots,x_n)$$

$$= \sum_{x_1,\dots,x_n \in A_\epsilon^{(n)}} p(x_1,\dots,x_n) nH(X) + \sum_{x_1,\dots,x_n \notin A_\epsilon^{(n)}} p(x_1,\dots,x_n) n \log|\mathcal{X}|$$

Shannon's Source Coding Theorem

The average bits needed is

$$\mathbb{E}[l(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n)$$

$$= \sum_{x_1, \ldots, x_n \in A_\epsilon^{(n)}} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n) + \sum_{x_1, \ldots, x_n \notin A_\epsilon^{(n)}} p(x_1, \ldots, x_n) l(x_1, \ldots, x_n)$$

$$= \sum_{x_1, \ldots, x_n \in A_\epsilon^{(n)}} p(x_1, \ldots, x_n) nH(X) + \sum_{x_1, \ldots, x_n \notin A_\epsilon^{(n)}} p(x_1, \ldots, x_n) n \log |\mathcal{X}| \approx nH(X).$$

Shannon's Source Coding Theorem

The average bits needed is

$$\mathbb{E}[l(X_1,\ldots,X_n)] = \sum_{x_1,\ldots,x_n} p(x_1,\ldots,x_n) l(x_1,\ldots,x_n)$$

$$= \sum_{x_1,\ldots,x_n \in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) l(x_1,\ldots,x_n) + \sum_{x_1,\ldots,x_n \notin A_\epsilon^{(n)}} p(x_1,\ldots,x_n) l(x_1,\ldots,x_n)$$

$$= \sum_{x_1,\ldots,x_n \in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) nH(X) + \sum_{x_1,\ldots,x_n \notin A_\epsilon^{(n)}} p(x_1,\ldots,x_n) n \log|\mathcal{X}| \approx nH(X).$$

Source Coding Theorem

For any information source distributed according to $X_1, X_2, \cdots \sim p(x)$, the compression rate is always greater than $H(X)$, but it can be arbitrarily close to $H(X)$.

Communication Channel: Definition

Communication Channel: Definition

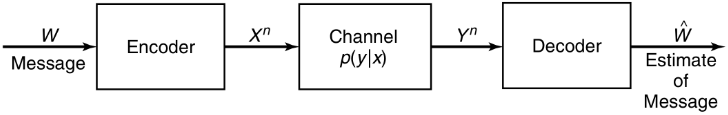

FIGURE 7.8. Communication channel.

Communication Channel: Definition

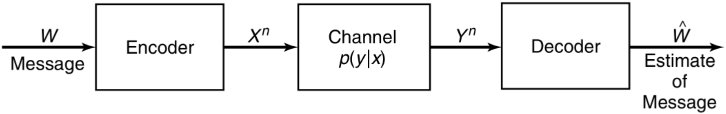

FIGURE 7.8. Communication channel.

- A message W results in channel inputs $X_1(W), \ldots, X_n(W)$;

Communication Channel: Definition

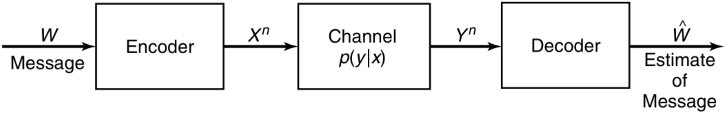

FIGURE 7.8. Communication channel.

- A message W results in channel inputs $X_1(W), \ldots, X_n(W)$;
- And they are received as a random sequence

$$Y_1, \ldots, Y_n \sim p(y_1, \ldots, y_n | x_1, \ldots, x_n).$$

Communication Channel: Definition

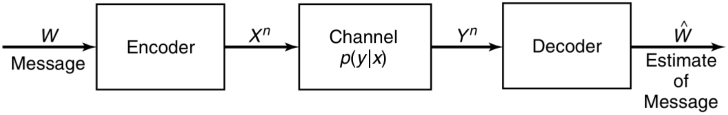

FIGURE 7.8. Communication channel.

- A message W results in channel inputs $X_1(W), \dots, X_n(W)$;
- And they are received as a random sequence
$$Y_1, \dots, Y_n \sim p(y_1, \dots, y_n | x_1, \dots, x_n).$$
- The receiver then guesses the index W by an appropriate decoding rule $\hat{W} = g(Y_1, \dots, Y_n)$.

Communication Channel: Definition

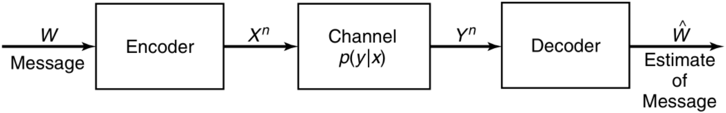

FIGURE 7.8. Communication channel.

- A message W results in channel inputs $X_1(W), \ldots, X_n(W)$;
- And they are received as a random sequence
$$Y_1, \ldots, Y_n \sim p(y_1, \ldots, y_n | x_1, \ldots, x_n).$$
- The receiver then guesses the index W by an appropriate decoding rule $\hat{W} = g(Y_1, \ldots, Y_n)$.
- The receiver makes an error if $\hat{W}$ is not the same as W that was transmitted.

Communication Channel: An Example

Communication Channel: An Example

Binary Symmetric Channel

$$p(Y=0|X=0)=1-p, \qquad p(Y=1|X=0)=p,$$
$$p(Y=0|X=1)=p, \qquad p(Y=1|X=1)=1-p.$$

Communication Channel: An Example

Binary Symmetric Channel

$$p(Y=0|X=0)=1-p, \qquad p(Y=1|X=0)=p,$$
$$p(Y=0|X=1)=p, \qquad p(Y=1|X=1)=1-p.$$

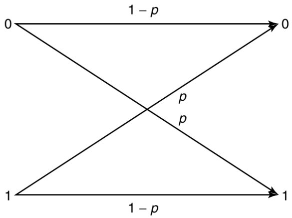

FIGURE 7.5. Binary symmetric channel. $C = 1 - H(p)$ bits.

Tradeoff between Speed and Reliability

Tradeoff between Speed and Reliability

Speed

To transmit 1: we transmit 1. It is likely that we receive 0. Note that the transmission rate is 1.

Tradeoff between Speed and Reliability

Speed

To transmit 1: we transmit 1. It is likely that we receive 0. Note that the transmission rate is 1.

Reliability

To transmit 1: we transmit 11111. Though it is likely that we receive something else, such as 11011, but more likely than not, we can correct the possible error. Note that the transmission rate is however $1/5$.

Shannon's Channel Coding Theorem: Statement

Shannon's Channel Coding Theorem: Statement

Channel Coding Theorem

For any discrete memoryless channel, asymptotically perfect transmission rate **below** the **capacity**

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

is always possible, but is not possible **above** the capacity.

Shannon's Channel Coding Theorem: Proof

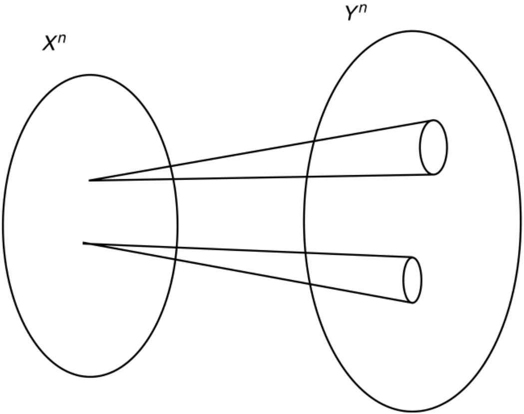

FIGURE 7.7. Channels after n uses.

Shannon's Channel Coding Theorem: Proof

Shannon's Channel Coding Theorem: Proof

- For each typical input n-sequence, there are approximately $2^{nH(Y|X)}$ possible typical output sequences, all of them equally likely.

Shannon's Channel Coding Theorem: Proof

- For each typical input n-sequence, there are approximately $2^{nH(Y|X)}$ possible typical output sequences, all of them equally likely.
- We wish to ensure that no two X input sequences produce the same Y output sequence. Otherwise, we will not be able to decide which X sequence was sent.

Shannon's Channel Coding Theorem: Proof

- For each typical input n-sequence, there are approximately $2^{nH(Y|X)}$ possible typical output sequences, all of them equally likely.
- We wish to ensure that no two X input sequences produce the same Y output sequence. Otherwise, we will not be able to decide which X sequence was sent.
- The total number of possible typical Y sequences is approximately $2^{nH(Y)}$. This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences.

Shannon's Channel Coding Theorem: Proof

- For each typical input n-sequence, there are approximately $2^{nH(Y|X)}$ possible typical output sequences, all of them equally likely.
- We wish to ensure that no two X input sequences produce the same Y output sequence. Otherwise, we will not be able to decide which X sequence was sent.
- The total number of possible typical Y sequences is approximately $2^{nH(Y)}$. This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences.
- The total number of disjoint sets is less than or equal to $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Hence, we can send at most approximately $2^{nI(X;Y)}$ distinguishable sequences of length n.

Capacity of Binary Symmetric Channels

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Proof.

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

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \end{aligned}$$

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) H(p) \end{aligned}$$

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) H(p) \\ &= H(Y) - H(p) \end{aligned}$$

Capacity of Binary Symmetric Channels

The **capacity** of a binary symmetric channel with crossover probability p is $C = 1 - H(p)$, where

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

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) H(p) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p). \end{aligned}$$

Capacity of Additive White Gaussian Channels

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0, 1)$, is $C = \frac{1}{2}\log(1 + P)$.

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0,1)$, is $C = \frac{1}{2}\log(1+P)$.

Proof.

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

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0,1)$, is $C = \frac{1}{2}\log(1+P)$.

Proof.

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

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0, 1)$, is $C = \frac{1}{2} \log(1 + P)$.

Proof.

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

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0,1)$, is $C = \frac{1}{2}\log(1+P)$.

Proof.

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

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \leq P$ and $Z \sim N(0, 1)$, is $C = \frac{1}{2}\log(1 + P)$.

Proof.

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - H(X+Z|X) \\
&= H(Y) - H(Z|X) \\
&= H(Y) - H(Z) \\
&\leq \frac{1}{2}\log 2\pi e(1+P) - \frac{1}{2}\log 2\pi e
\end{aligned}
$$

□

Capacity of Additive White Gaussian Channels

The capacity of an additive white Gaussian channel $Y = X + Z$, where $\mathbb{E}[X^2] \le P$ and $Z \sim N(0,1)$, is $C = \frac{1}{2}\log(1+P)$.

Proof.

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - H(X+Z|X) \\
&= H(Y) - H(Z|X) \\
&= H(Y) - H(Z) \\
&\le \frac{1}{2}\log 2\pi e(1+P) - \frac{1}{2}\log 2\pi e \\
&= \frac{1}{2}\log(1+P).
\end{aligned}
$$

□

Memory Channels

Memoryless Channels

Memoryless Channels

- Channel transitions are characterized by time-invariant transition probabilities $\{p(y|x)\}$.

Memoryless Channels

- Channel transitions are characterized by time-invariant transition probabilities $\{p(y|x)\}$.
- Channel inputs are independent and identically distributed.

Memoryless Channels

- Channel transitions are characterized by time-invariant transition probabilities $\{p(y|x)\}$.
- Channel inputs are independent and identically distributed.
- Representative examples include (memoryless) binary symmetric channels and additive white Gaussian channels.

Capacity of Memoryless Channels

Capacity of Memoryless Channels

Shannon's channel coding theorem

$$
\begin{aligned}
C &= \sup_{p(x)} I(X;Y) \\
&= \sup_{p(x)} -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.
\end{aligned}
$$

Capacity of Memoryless Channels

Shannon's channel coding theorem

$$
\begin{aligned}
C &= \sup_{p(x)} I(X;Y) \\
&= \sup_{p(x)} -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.
\end{aligned}
$$

The Blahut-Arimoto algorithm (BAA)

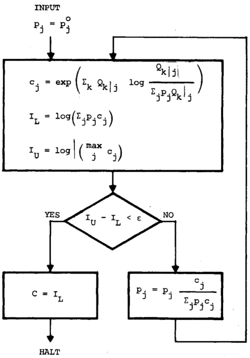

Fig. 1. Capacity algorithm.

Memory Channels

Memory Channels

- Channel transitions are characterized by probabilities
$$\{p(y_i|x_1, \ldots, x_i, y_1, \ldots, y_{i-1}, s_i)\},$$
where channel outputs are possibly dependent on previous and current channel inputs and previous outputs and current channel state; for example, inter-symbol interference channels, flash memory channels, Gilbert-Elliot channels.

Memory Channels

- Channel transitions are characterized by probabilities
$$\{p(y_i|x_1, \ldots, x_i, y_1, \ldots, y_{i-1}, s_i)\},$$
where channel outputs are possibly dependent on previous and current channel inputs and previous outputs and current channel state; for example, inter-symbol interference channels, flash memory channels, Gilbert-Elliot channels.
- Channel inputs may have to satisfy certain constraints which necessitate dependence among channel inputs; for example, (d, k)-RLL constraints, more generally, finite-type constraints.

Memory Channels

- Channel transitions are characterized by probabilities
$$\{p(y_i|x_1, \ldots, x_i, y_1, \ldots, y_{i-1}, s_i)\},$$
where channel outputs are possibly dependent on previous and current channel inputs and previous outputs and current channel state; for example, inter-symbol interference channels, flash memory channels, Gilbert-Elliot channels.
- Channel inputs may have to satisfy certain constraints which necessitate dependence among channel inputs; for example, (d, k)-RLL constraints, more generally, finite-type constraints.
- Such channels are widely used in a variety of real-life applications, including magnetic and optical recording, **solid state drives**, communications over band-limited channels with inter-symbol interference.

Capacity of Memory Channels

Capacity of Memory Channels

Despite a great deal of efforts by Zehavi and Wolf [1988], Mushkin and Bar-David [1989], Shamai and Kofman [1990], Goldsmith and Varaiya [1996], Arnold, Loeliger, Vontobel, Kavcic and Zeng [2006], Holliday, Goldsmith, and Glynn [2006], Vontobel, Kavcic, Arnold and Loeliger [2008], Pfister [2011], Permuter, Asnani and Weissman [2013], Han [2015], ...

Capacity of Memory Channels

???

Capacity of Memory Channels

Capacity of Memory Channels

Shannon's channel coding theorem

$$
\begin{aligned}
C &= \sup_{p(x)} I(X;Y) \\
&= \sup_{p(x)} \lim_{n\to\infty} -\frac{1}{n} \sum_{x_1^n, y_1^n} p(x_1^n, y_1^n) \log \frac{p(x_1^n, y_1^n)}{p(x_1^n)p(y_1^n)}.
\end{aligned}
$$

Capacity of Memory Channels

Shannon's channel coding theorem

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

$$= \sup_{p(x)} \lim_{n\to\infty} -\frac{1}{n} \sum_{x_1^n, y_1^n} p(x_1^n, y_1^n) \log \frac{p(x_1^n, y_1^n)}{p(x_1^n)p(y_1^n)}.$$

The Generalized Blahut-Arimoto algorithm (GBAA) by Vontobel, Kavcic, Arnold and Loeliger [2008]

Algorithm 45 (Generalized BAA): Let $\mathcal{Q} = \mathcal{Q}(\mathcal{B})$ be a given FSMS manifold and let W be the channel law of a given FSMC. Let $\{Q_{ij}^{\langle 0\rangle}\} \in \mathrm{relint}(\mathcal{Q})$ be some initial (freely chosen) FSMS process. For iterations $r = 0, 1, 2, \ldots$, perform alternatively the following two steps.

- **First Step:** For each $(i,j) \in \mathcal{B}$ calculate $T_{ij}^{\langle r\rangle} \triangleq T_{ij}(Q_{ij}^{\langle r\rangle}, W)$ according to Definition 41. The values $T_{ij}^{\langle r\rangle}$ can be approximated by the procedure given in Section V-C.
- **Second Step:** The new FSMS process $\{Q_{ij}^{\langle r+1\rangle}\}$ is chosen to maximize $\Psi(Q_{ij}^{\langle r\rangle}, Q_{ij}, W)$, i.e.,

$$\left\{Q_{ij}^{\langle r+1\rangle}\right\} \triangleq \arg \max_{\{Q_{ij}\}\in\mathcal{Q}} \Psi\left(Q_{ij}^{\langle r\rangle}, Q_{ij}, W\right)$$

  and is calculated according to the algorithm in Lemma 44 with inputs $\{\tilde{Q}_{ij}\} \triangleq \{Q_{ij}^{\langle r\rangle}\}$ and W and output $\{Q_{ij}^{\langle r+1\rangle}\} \triangleq \{Q_{ij}^{*}\}$.

Convergence of the GBAA

Convergence of the GBAA

The GBAA will converge if the following conjecture is true.

Convergence of the GBAA

The GBAA will converge if the following conjecture is true.

Concavity Conjecture [Vontobel *et al.* 2008]

$I(X;Y)$ and $H(X|Y)$ are both concave with respect to a chosen parameterization.

Convergence of the GBAA

The GBAA will converge if the following conjecture is true.

Concavity Conjecture [Vontobel *et al.* 2008]

$I(X;Y)$ and $H(X|Y)$ are both concave with respect to a chosen parameterization.

Unfortunately, the concavity conjecture is **not** true in general [Li and Han, 2013].

A Randomized Algorithm [Han 2015]

A Randomized Algorithm [Han 2015]

With appropriately chosen step sizes $a_n = 1/n^a$, $a > 0$,

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n),$$

where

A Randomized Algorithm [Han 2015]

With appropriately chosen step sizes $a_n = 1/n^a$, $a > 0$,

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n),$$

where

- θ_0 is randomly selected from the parameter space Θ;

A Randomized Algorithm [Han 2015]

With appropriately chosen step sizes $a_n = 1/n^a$, $a > 0$,

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n),$$

where

- θ_0 is randomly selected from the parameter space Θ;
- $g_{n^b}(\theta)$ is a simulator for $I'(X(\theta); Y(\theta))$;

A Randomized Algorithm [Han 2015]

With appropriately chosen step sizes $a_n = 1/n^a$, $a > 0$,

$$\theta_{n+1} = \theta_n + a_n g_{n^b}(\theta_n),$$

where

- θ_0 is randomly selected from the parameter space Θ;
- $g_{n^b}(\theta)$ is a simulator for $I'(X(\theta); Y(\theta))$;
- $$0 < \beta < \alpha < 1/3, \quad b > 0, \quad 2a + b - 3b\beta > 1,$$

  here, α, β are some "hidden" parameters involved in the definition of $g_{n^b}(\theta)$.

Our Simulator of $I'(X;Y)$

Our Simulator of $I'(X; Y)$

Define

$$q = q(n) \triangleq n^{\beta}, \quad p = p(n) \triangleq n^{\alpha}, \quad k = k(n) \triangleq n/(n^{\alpha} + n^{\beta}).$$

Our Simulator of $I'(X;Y)$

Define

$$q = q(n) \triangleq n^{\beta}, \quad p = p(n) \triangleq n^{\alpha}, \quad k = k(n) \triangleq n/(n^{\alpha} + n^{\beta}).$$

For any j with $iq + (i-1)p + 1 \leq j \leq iq + ip$, define

$$W_j = -\left(\frac{p'(Y_{j-\lfloor q/2 \rfloor})}{p(Y_{j-\lfloor q/2 \rfloor})} + \cdots + \frac{p'(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})}{p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})} \right) \log p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1}),$$

Our Simulator of $I'(X;Y)$

Define

$$q = q(n) \triangleq n^{\beta}, \quad p = p(n) \triangleq n^{\alpha}, \quad k = k(n) \triangleq n/(n^{\alpha} + n^{\beta}).$$

For any j with $iq + (i-1)p + 1 \leq j \leq iq + ip$, define

$$W_j = -\left(\frac{p'(Y_{j-\lfloor q/2 \rfloor})}{p(Y_{j-\lfloor q/2 \rfloor})} + \cdots + \frac{p'(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})}{p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})} \right) \log p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1}),$$

and furthermore

$$\zeta_i \triangleq W_{iq+(i-1)p+1} + \cdots + W_{iq+ip}, \quad S_n \triangleq \sum_{i=1}^{k(n)} \zeta_i.$$

Our Simulator of $I'(X;Y)$

Define

$$q = q(n) \triangleq n^{\beta}, \quad p = p(n) \triangleq n^{\alpha}, \quad k = k(n) \triangleq n/(n^{\alpha} + n^{\beta}).$$

For any j with $iq + (i-1)p + 1 \leq j \leq iq + ip$, define

$$W_j = -\left(\frac{p'(Y_{j-\lfloor q/2 \rfloor})}{p(Y_{j-\lfloor q/2 \rfloor})} + \cdots + \frac{p'(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})}{p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1})}\right) \log p(Y_j | Y_{j-\lfloor q/2 \rfloor}^{j-1}),$$

and furthermore

$$\zeta_i \triangleq W_{iq+(i-1)p+1} + \cdots + W_{iq+ip}, \quad S_n \triangleq \textstyle\sum_{i=1}^{k(n)} \zeta_i.$$

Our simulator for $I'(X;Y)$:

$$g_n(X_1^n, Y_1^n) = H'(X_2 | X_1) + S_n(Y_1^n)/(kp) - S_n(X_1^n, Y_1^n)/(kp).$$

Convergence of Our Algorithm

Convergence of Our Algorithm

Convergence and convergence rate with concavity

If $I(X;Y)$ is concave with respect to θ, then θ_n converges to the unique capacity achieving distribution θ^* almost surely. And for any τ with $2a + b - 3b\beta - 2\tau > 1$, we have

$$|\theta_n - \theta^*| = \tilde{O}(n^{-\tau}).$$

The Ideas for the Proofs

The Ideas for the Proofs

Analyticity result [Han, Marcus, 2006]

The entropy rate of hidden Markov chains is analytic.

Refinements of the Shannon-MaMillan-Breiman theorem [Han, 2012]

Limit theorems for the sample entropy of hidden Markov chains hold.

The Ideas for the Proofs

Analyticity result [Han, Marcus, 2006]

The entropy rate of hidden Markov chains is analytic.

Refinements of the Shannon-MaMillan-Breiman theorem [Han, 2012]

Limit theorems for the sample entropy of hidden Markov chains hold.

The analyticity result states that $I(X;Y) = H(X) + H(Y) - H(X,Y)$ is a "nicely behaved" function.

The Ideas for the Proofs

Analyticity result [Han, Marcus, 2006]

The entropy rate of hidden Markov chains is analytic.

Refinements of the Shannon-MaMillan-Breiman theorem [Han, 2012]

Limit theorems for the sample entropy of hidden Markov chains hold.

The analyticity result states that $I(X;Y) = H(X) + H(Y) - H(X,Y)$ is a "nicely behaved" function.

The refinement results confirm that using Monte Carlo simulations, $I(X;Y)$ and its derivatives can be "well-approximated".

Continuous-Time Information Theory

Continuous-Time Gaussian Non-Feedback Channels

Consider the following continuous-time Gaussian channel:

$$Y(t) = \sqrt{snr} \int_0^t X(s) ds + B(t),\ t \in [0, T],$$

where $\{B(t)\}$ is the standard Brownian motion.

Continuous-Time Gaussian Non-Feedback Channels

Continuous-Time Gaussian Non-Feedback Channels

Theorem (Ducan 1970)

The following I-CMMSE relationship holds:

$$I(X_0^T; Y_0^T) = \frac{1}{2}\mathbb{E}\int_0^T (X(s) - \mathbb{E}[X(s)|Y_0^s])^2\, ds.$$

Continuous-Time Gaussian Non-Feedback Channels

Theorem (Ducan 1970)

The following I-CMMSE relationship holds:

$$I(X_0^T; Y_0^T) = \frac{1}{2}\mathbb{E}\int_0^T (X(s) - \mathbb{E}[X(s)|Y_0^s])^2\, ds.$$

Theorem (Guo, Shamai and Verdu 2005)

The following I-MMSE relationship holds:

$$\frac{d}{dsnr} I(X_0^T; Y_0^T) = \frac{1}{2}\mathbb{E}\int_0^T (X(s) - \mathbb{E}[X(s)|Y_0^T])^2 ds.$$

Continuous-Time Gaussian Feedback Channels

Consider the following continuous-time Gaussian feedback channel:

$$Y(t) = \sqrt{snr} \int_0^t X(s, M, Y_0^s) ds + B(t), \ t \in [0, T],$$

where $\{B(t)\}$ is the standard Brownian motion.

Continuous-Time Gaussian Feedback Channels

Continuous-Time Gaussian Feedback Channels

Theorem (Kadota, Zakai and Ziv 1971)

The following I-CMMSE relationship:

$$I(M; Y_0^T) = \frac{1}{2}\mathbb{E}\int_0^T (X(s, M, Y_0^s) - \mathbb{E}[X(s, M, Y_0^s)|Y_0^s])^2\, ds.$$

Continuous-Time Gaussian Feedback Channels

Theorem (Kadota, Zakai and Ziv 1971)

The following I-CMMSE relationship:

$$I(M; Y_0^T) = \frac{1}{2}\mathbb{E}\int_0^T (X(s, M, Y_0^s) - \mathbb{E}[X(s, M, Y_0^s)|Y_0^s])^2\, ds.$$

Theorem (Han and Song 2016)

The following I-MMSE relationship holds:

$$\frac{d}{dsnr} I(M; Y_0^T) = \frac{1}{2}\int_0^T \mathbb{E}\left[\left(X(s) - \mathbb{E}[X(s)|Y_0^T]\right)^2\right] ds$$

$$+snr \int_0^T \mathbb{E}\left[\left(X(s) - \mathbb{E}\left[X(s)|\, Y_0^T\right]\right) \frac{d}{dsnr} X(s)\right] ds.$$

Capacity of Continuous-Time Gaussian Channels

Capacity of Continuous-Time Gaussian Channels

For either the following continuous-time Gaussian channel:

$$Y(t) = \sqrt{snr} \int_0^t X(s) ds + B(t), \; t \in [0, T],$$

or the following continuous-time Gaussian feedback channel:

$$Y(t) = \sqrt{snr} \int_0^t X(s, M, Y_0^s) ds + B(t), \; t \in [0, T],$$

the capacity is $\mathbf{P}/\mathbf{2}$.

Thank you!