Information Theory: Principles and Applications

Tiago T. V. Vinhoza

April 16, 2010

Why the channel capacity is important?

- Shannon proved that the channel capacity is the maximum number of bits that can be reliably transmitted over the channel.
- Reliably = probability of error can be made arbitrarily small.
- Channel coding theorem.

Intuitive idea of channel capacity as a fundamental limit

- Basic idea: For large block lengths, every channel looks like the noisy typewriter channel shown last class.
- Channel has a subset of inputs that produce disjoint sequences at the output.
- Tipicality arguments.

Intuitive idea of channel capacity as a fundamental limit

- For each typical input sequence of length n, there are approximately $2^{nH(Y|X)}$ possible Y sequences.
- Desirable: No two different X sequences produce the same Y output sequence.
- Total number of typical Y sequences is approx. $2^{nH(Y)}$.
- So that total number of disjoint sets is less or equal to $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$
- At most $2^{nI(X;Y)}$ distinguishable sequences of length n can be sent.

Formalizing the ideas

- Message W is drawn from the index set $\{1, 2, \ldots, M\}$.
- Encoded signal $\mathbf{X}^n(W)$.
- The encoded signal passes through the channel and is received as the sequence $\mathbf{Y}^n$. The channel is described by a transition probability matrix $P(\mathbf{Y}^n|\mathbf{X}^n)$.
- Receiver guesses index W using a decoding rule $\widehat{W} = g(\mathbf{Y}^n)$.
- If $\widehat{W} \neq W$ then an error occurs.

Definition: Discrete Channel

- A discrete channel consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability distributions $p_{Y|X=x}(y)$ one for each $x \in \mathcal{X}$.

$$(\mathcal{X}, p_{Y|X=x}(y), \mathcal{Y})$$

Definition: Extension of the Discrete Memoryless Channel

- The n-th extension of the discrete memoryless channel is the channel $(\mathcal{X}^n, p_{\mathbf{Y}^n|\mathbf{X}^n=\mathbf{x}^n}(\mathbf{y}^n), \mathcal{Y}^n)$ where

$$P(Y_k|\mathbf{X}^k, \mathbf{Y}^{k-1}) = P(Y_k|X_k), \quad k = 1, 2, \ldots, n.$$

- If the channel is used without feedback, that is, the inputs do not depend on the past outputs then the channel transition function for the discrete memoryless channel is

$$p_{\mathbf{Y}^n|\mathbf{X}^n=\mathbf{x}^n}(\mathbf{y}^n) = \prod_{i=1}^{n} p_{Y_i|X_i=x_i}(y_i)$$

Definition: Code for a channel

- An (M, n) code for the channel $(\mathcal{X}, p_{Y|X=x}(y), \mathcal{Y})$ consists of the following:
  - An index set $\{1, 2, \ldots, M\}$.
  - An encoding function $\mathbf{X}^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n$, that generates codewords $\mathbf{X}^n(1), \ldots, \mathbf{X}^n(M)$.
  - A decoding function

    $$g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}.$$

    which assigns a guess to each possible received vector

Definition: Rate of a code

- The rate R of an (M, n) code is:

$$R = \frac{\log M}{n} \text{ bits per transmission.}$$

- A rate R is *achievable* if there exists a sequence of $(2^{\lceil nR \rceil}, n)$ codes such that the maximal probability of error goes to zero as n goes to infinity.
- The *capacity* of a discrete memoryless channel is the supremum of all achievable rates.
- Rates less than capacity yield arbitrarily small probability of error for sufficiently large n.

Definition: Probability of Error

- Conditional probability of error given that index i was sent:

$$\lambda_i = P(g(\mathbf{Y}^n) \neq i | \mathbf{X}^n = \mathbf{X}^n(i)) = \sum_{\mathbf{y}^n} p_{\mathbf{Y}^n|\mathbf{X}^n=\mathbf{x}^n(i)}(\mathbf{y}^n) I(g(\mathbf{y}^n) \neq i)$$

  where $I(\cdot)$ is the indicator function.

- Maximal probability of error:

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots M\}} \lambda_i$$

- Average probability of error for an (M, n) code:

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

Definition: Joint Typical Sequences

- The decoding procedure employed in the proofs will decode a channel output $\mathbf{Y}^n$ as the i-th index if the codeword $\mathbf{X}^n(i)$ is jointly-typical with the received sequence $\mathbf{Y}^n$.
- The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(\mathbf{x}^n, \mathbf{y}^n)\}$ with respect to their joint distribution is the set of n-sequences with sample entropy ϵ-close to the true entropies.

$$
\begin{aligned}
A_\epsilon^{(n)} &= \Bigg\{ (\mathbf{x}^n, \mathbf{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} - H(X) \right| < \epsilon \\
&\quad \left| -\frac{\log p_{\mathbf{Y}^n}(\mathbf{y}^n)}{n} - H(Y) \right| < \epsilon \\
&\quad \left| -\frac{\log p_{\mathbf{X}^n\mathbf{Y}^n}(\mathbf{x}^n, \mathbf{y}^n)}{n} - H(X,Y) \right| < \epsilon \Bigg\}
\end{aligned}
$$

Joint AEP

- Let $(\mathbf{X}^n, \mathbf{Y}^n)$ be sequences of length n drawn i.i.d. according to the joint distribution $p_{\mathbf{X}^n\mathbf{Y}^n}(\mathbf{x}^n, \mathbf{y}^n) = \prod_{i=1}^{n} p_{X_iY_i}(x_i, y_i)$ then
  - $P((\mathbf{X}^n, \mathbf{Y}^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
  - $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$ and $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$
  - If $(\widetilde{\mathbf{X}}^n, \widetilde{\mathbf{Y}}^n)$ are independent and have the same marginals as $\mathbf{X}^n$ and $\mathbf{Y}^n$

$$
\begin{aligned}
P((\widetilde{\mathbf{X}}^n, \widetilde{\mathbf{Y}}^n) \in A_\epsilon^{(n)}) &= \sum_{(\mathbf{x}^n, \mathbf{y}^n) \in A_\epsilon^{(n)}} p_{\mathbf{X}^n}(\mathbf{x}^n) p_{\mathbf{Y}^n}(\mathbf{y}^n) \\
&\leq 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\
&= 2^{-n(I(X,Y)-3\epsilon)}
\end{aligned}
$$

Joint AEP

- There are about $2^{nH(X)}$ typical X sequences, and about $2^{nH(Y)}$ typical Y sequences. However, since there are only $2^{nH(X,Y)}$ jointly typical sequences, not all pairs $(\mathbf{X}^n, \mathbf{Y}^n)$ with $\mathbf{X}^n$ and $\mathbf{Y}^n$ being typical are jointly typical.
- The probability that any randomly chosen pair is jointly typical is about $2^{-nI(X;Y)}$. So, for a fixed $\mathbf{Y}^n$ sequence, we can consider $2^{nI(X;Y)}$ of pairs before we come across a jointly typical pair. This suggests that there are about $2^{nI(X;Y)}$ distinguishable sequences $\mathbf{X}^n$

Channel Coding Theorem

- Achievability: *Consider a discrete memoryless channel with capacity C. All rates $R < C$ are achievable. Specifically, for every rate $R < C$ there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error arbitrarily small.*
- Converse: *Consider a discrete memoryless channel with capacity C. For any sequence of $(2^{nR}, n)$ codes with maximum probability of error as small as we want, then $R < C$.*

Channel Coding Theorem: Achievability

- Generate an $(2^{nR}, n)$ code at random accordion to the distribution $p_X(x)$, that is, generate 2^{nR} codewords according to the probability distribution $p_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} p_{X_i}(x_i)$.
- Exhibit the 2^{nR} codewords as the rows of the matrix

$$\mathbf{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix}$$

- Reveal the code to transmitter and receiver (they both know the channel transition matrix $P(Y|X)$, too.

Channel Coding Theorem: Achievability

- Message W is chosen according to a uniform distribution, that is, $P(W = w) = 2^{-nR}$, for $w = 1, 2, \ldots, 2^{nR}$.
- The codeword $\mathbf{X}^n(w)$, corresponding to the w-th row of matrix $\mathbf{C}$ is sent over the channel.
- The receiver gets sequence $\mathbf{Y}^n$ according to the distribution $p_{\mathbf{Y}^n|\mathbf{X}^n=\mathbf{x}^n(w)}(\mathbf{y}^n) = \prod_{i=1}^{n} p_{Y_i|X_i=x_i(w)}(y_i)$
- Receiver guesses message using typical set decoding
- Receiver declares that index i was sent if
  - $(\mathbf{X}^n(i), \mathbf{Y}^n)$ are jointly typical.
  - there is no other index j such that $(\mathbf{X}^n(j), \mathbf{Y}^n)$ are jointly typical.
- Otherwise the receive declares and error.

Channel Coding Theorem: Achievability

- Analysis of the error probability
  - Instead of calculating the probability of error for a single code, we compute the average over all codes generated at random according to the probability distribution $P(\mathbf{C})$
  - Two types of error events: The output $\mathbf{Y}^n$ is not jointly typical with the transmitted codeword or there is another codeword with is also jointly typical with $\mathbf{Y}^n$.
  - The probability that the transmitted codeword and the received sequence are jointly typical goes to one as shown by the AEP.
  - For the rival codewords, the probability that any one of them is jointly typical with the received sequence is about $2^{-nI(X;Y)}$, so we can use $2^{nI(X;Y)}$ codewords and have a small error probability.

Channel Coding Theorem: Achievability

- Calculating the average error probability

$$
\begin{aligned}
P(\mathcal{E}) &= \sum_{\mathbf{C}} P(\mathbf{C}) P_e^{(n)}(\mathbf{C}) \\
&= \sum_{\mathbf{C}} P(\mathbf{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathbf{C}) \\
&= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathbf{C}} P(\mathbf{C}) \lambda_w(\mathbf{C})
\end{aligned}
$$

- By the symmetry of the code construction, the average probability of error over all codes does not depend on the particular index that was sent.

$$\sum_{\mathbf{C}} P(\mathbf{C}) \lambda_w(\mathbf{C}) \text{ is not a function of } w.$$

Channel Coding Theorem: Achievability

- Assuming WLOG that $W = 1$ was sent.

$$P(\mathcal{E}) = \sum_{\mathbf{C}} P(\mathbf{C})\lambda_1(\mathbf{C}) = P(\mathcal{E}|W = 1)$$

- Defining the events

$$E_i = \{(\mathbf{X}^n(i), \mathbf{Y}^n) \text{ is in } A_\epsilon^{(n)}\}, \quad i = 1, 2, \ldots, 2^{nR}$$

- The error events in our case are
  - $\overline{E_1}$, that is, the complement of E_1 occurs. This means that $\mathbf{Y}^n$ and $\mathbf{X}^n(1)$ are not jointly typical.
  - E_2 or E_3 or ... $E_{2^{nR}}$ occurs. This means that a wrong codeword is jointly typical with $\mathbf{Y}^n$.

Channel Coding Theorem: Achievability

- Evaluating

$$P(\mathcal{E}|W=1) = P(\overline{E_1} \cup E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}})$$
$$\leq P(\overline{E_1}) + \sum_{i=2}^{2^{nR}} P(E_i)$$

- The inequality is due to the union bound.
- By the joint AEP, $P(\overline{E_1}) < \epsilon$ for sufficiently large n.
- As $\mathbf{X}^n(1)$ and $\mathbf{X}^n(i)$ are independent (code generation procedure), it follows that $\mathbf{Y}^n$ and $\mathbf{X}^n(i)$ are also independent if $i \neq 1$. Hence, from the joint AEP

$$P(E_i) \leq 2^{-n(I(X;Y)-3\epsilon)} \quad \text{if } i \neq 1.$$

Channel Coding Theorem: Achievability

- Evaluating

$$\begin{aligned} P(\mathcal{E}|W=1) &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + (2^{nR})2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{n3\epsilon})2^{-n(I(X;Y)-R)} \\ &\leq 2\epsilon \end{aligned}$$

- if n is sufficiently large and $R < I(X;Y) - 3\epsilon$

Channel Coding Theorem: Achievability

- Evaluating

$$\begin{aligned} P(\mathcal{E}|W=1) &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + (2^{nR})2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{n3\epsilon})2^{-n(I(X;Y)-R)} \\ &\leq 2\epsilon \end{aligned}$$

- if n is sufficiently large and $R < I(X;Y) - 3\epsilon$

Channel Coding Theorem: Achievability

- If $R < I(X;Y)$, we can choose ϵ and n so that the average probability of error over all codebooks is less than 2ϵ.
- If the input distribution $p_X(x)$ is the one that achieves the channel capacity C, then the achievability condition is replaced by $R < C$.
- If the average probability of error over all codebooks is less than 2ϵ, than there exists at least one codebook $\mathbf{C}^*$ with an average probability of error $P_e^{(n)} \leq 2\epsilon$.

$$2\epsilon \geq \frac{1}{2^{nR}} \sum_i \lambda_i(\mathbf{C}^*) = P_e^{(n)}$$

- This implies that at least half of the indices i and their codewords have $\lambda_i < 4\epsilon$. Using only this best half of codewords we have 2^{nR-1} codewords and the new rate $R' = R - 1/n \approx R$ for large n.

Channel Coding Theorem: Converse

- We have now to show that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.
- If the maximal error probability goes to zero, the average error probability, $P_e^{(n)}$, also goes to zero. For each n, let W be drawn from a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$. Since W is uniform, $P_e^{(n)} = P(\widehat{W} \neq W)$
- We will resort to the Fano's Inequality to prove the converse.

Channel Coding Theorem: Converse

- Proving the converse:

$$\begin{aligned} nR &= H(W) = H(W|\mathbf{Y}^n) + I(W;\mathbf{Y}^n) \\ &\leq H(W|\mathbf{Y}^n) + I(\mathbf{X}^n(W);\mathbf{Y}^n) \\ &\leq 1 + P_e^{(n)} nR + I(\mathbf{X}^n(W);\mathbf{Y}^n) \\ &\leq 1 + P_e^{(n)} nR + nC \end{aligned}$$

- Dividing by n, and rewriting we get

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

Channel Coding Theorem

- The theorem shows that good codes exist with exponentially small probability of error for long block lengths.
- No systematic way of constructing such codes is provided.
- Random Codes: No structure $\longrightarrow$ Lookup table decoding $\longrightarrow$ Huge table size for large blocks.
- 1950's: Coding theorists started searching for good codes and to devise efficient implementations.

Error Correcting Codes

- Error control coding: Addition of redundancy in a smart way to combat errors induced by the channel.
- Error detection
- Error correction

Error Correcting Codes

- Block Codes
- Linear Codes: Encoding and Decoding
- Hamming Codes

Joint Source Channel Coding

- The source coding theorem states that for data compression $R > H$.
- The channel coding theorem states that for data transmission $R < C$.
- Is the condition $H < C$ necessary and sufficient for sending a source over a channel?

Joint Source Channel Coding

- Consider a source modeled by a finite alphabet stochastic process $V^n = V_1, V_2, \ldots, V_n$,with entropy rate $H(\mathcal{V})$ that satisfies the AEP.
- Achievability: *The source can be sent reliably over a discrete memoryless channel with capacity C if $H(\mathcal{V}) < C$*
- Converse: *The source cannot be sent reliably over a discrete memoryless channel with capacity C if $H(\mathcal{V}) > C$*

Joint Source Channel Coding

- The source channel separation theorem shows that it is possible to design the source code and the channel code separately and combine the results to achieve optimal performance.
- Asymptotic optimality: For finite block length, the probability of error can be reduced by using joint source-channel coding.
- That separation, however, fails for some multiuser channels.

Next Steps

- Lossy Source Coding
- Multiple-user Information Theory