Information Theory and Networks

Lecture 24: Channel Capacity

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

October 9, 2013

Part I

Channel Capacity

To make no mistakes is not in the power of man; but from their errors and mistakes the wise and good learn wisdom for the future.

Plutarch

Section 1

Channel Coding Theorem

Capacity

Definition (Operational Channel Capacity)

The highest rate of bits we can send per input symbol, with an arbitrarily low probability of error is called the operational channel capacity.

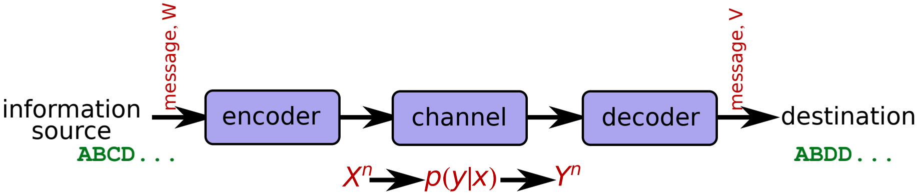
Definition (Information Capacity)

The information capacity of a discrete memoryless channel with inputs $X \in \mathcal{X}$ and outputs $Y \in \mathcal{Y}$, and channel transition matrix $p(Y|X)$ is

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

where $I(X;Y)$ is the mutual information of X and Y.

Digital Communications Channels



Definition (Discrete Channel)

A discrete channel is a system with an input alphabet $\mathcal{X}$, and output alphabet $\mathcal{Y}$, and a probability transition matrix $p(y|x)$ that describes the probability of observing the output symbol $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$.

We denote a Discrete Memoryless Channel (DMC) by the triple $(\mathcal{X}, p(y|x), \mathcal{Y})$.

Digital Communications Channels

We will work with DMC (Discrete Memoryless Channels) with no feedback $(\mathcal{X}, p(y|x), \mathcal{Y})$. Then

Definition

The *n*th extension of a DMC is the channel $(\mathcal{X}^n, p(y^{(n)}|x^{(n)}), \mathcal{Y}^n)$ where

$$p(y_k|x^{(k)}, y^{(k-1)}) = p(y_k|x_k), \text{ for } k = 1, 2, \ldots, n$$

and/or

$$p\left(y^{(n)}\middle|x^{(n)}\right) = \prod_{i=1}^{n} p(y_i|x_i)$$

Channel Codes

Definition (Channel Code)

A (M, n) code for channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of

1. An index set $\{1, 2, \ldots, M\}$
2. An encoding function with block size n

   $$X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$$

   yielding codewords $\{X^n(1), X^n(2), \ldots, X^n(M)\}$, called the codebook.
3. A decoding function

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

   which is a deterministic rule which assigns a guess to each possible received vector.

Errors

Definition

The conditional probability of error given that index i is sent is

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

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

Errors

Definition

The maximal probability of error $\lambda^{(n)}$ for an (M, n) code is defined as

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

and the average probability of error $P_e^{(n)}$ is

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i = P(I \neq g(Y^n))$$

where I is a random index uniformly chosen from $\{1, 2, \ldots, M\}$.

Rate and Capacity

Definition (Rate)

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

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

A rate is said to be achievable if there exists a sequence of $\left(\lceil 2^{nR} \rceil, n\right)$ codes such that the maximal probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$.

Definition (Operational Channel Capacity)

The capacity of a DMC is the supremum of all the achievable rates.

Shannon's Second Theorem

Theorem (Shannon's Channel Coding Theorem)

All rates below capacity $C = \max_{p_X(x)} I(X;Y)$ are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.
Conversely, any sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$ must have $R \leq C$.

Shannon's Second Theorem

Shannon's Channel Coding Theorem.

Full proof [CT91, pp.198-209], but some intuition follows:

1. We want to exploit the law of large numbers for larger blocks to obtain something like convergence to accurate estimates.
2. We can't increase capacity of a memoryless channel by using it multiple times, independently.
3. So there need to be some structure in what we send, and we are achieving this through our set of codewords.
4. By choosing a set of codewords that are reasonable distances apart, we hope that the errors result in sequences that are closer to the real codeword than any other.
5. It turns out random codewords are good enough.

□

Random Codes

1. Fix $p(x)$, and generate a random $(2^{nR}, n)$ code by taking

$$P\left(X^n(i) = x_1 x_2 \cdots x_n\right) = \prod_{k=1}^{n} p\left(x_k\right) \text{ for each } i \in \{1, 2, \ldots, M = 2^{nR}\}$$

2. Write codewords as a $2^{nR} \times n$ matrix, with IID rows

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

3. The probability of a particular code is

$$P(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{k=1}^{n} p(x_k(w))$$

Using Random Codes

To use code $\mathcal{C}$

1. Assume receiver and sender both know the code, and also the transition probabilities $p(y|x)$.
2. Assume message chosen according to uniform distribution

$$P\big(W = w\big) = 2^{-nR}, \;\; \text{for } w = 1, 2, \ldots, 2^{nR}$$

   and the wth codeword $x^n(w)$ is sent.
3. Receiver receives Y^n according to the distribution

$$P\big(y^{(n)}\big|x^{(n)}(w)\big) = \prod_{i=1}^{n} p(y_i|x_i(w))$$

4. Receiver decodes by guessing that w is the input that generates a jointly typical sequence $(x^{(n)}(w), y^{(n)})$.

Joint AEP (see [CT91, Theorem 8.6.1,pp.195-196]

Definition (Jointly Typical)

The set $A_\epsilon^{(n)}$ of jointly typical sequences WRT to $p(x, y)$ is the set of sequences of n pairs (x_i, y_i) with entropies ϵ-close to the true entropy, i.e.,

$$A_\epsilon^{(n)} = \left\{ (x^{(n)}, y^{(n)}) \,\middle|\, d_X < \epsilon, d_Y < \epsilon, d_{X,Y} < \epsilon, \right\}$$

where

$$\begin{aligned} d_X &= \left| -\frac{1}{n} p(x^{(n)}) - H(X) \right| \\ d_X &= \left| -\frac{1}{n} p(y^{(n)}) - H(Y) \right| \\ d_{X,Y} &= \left| -\frac{1}{n} p(x^{(n)}, y^{(n)}) - H(X, Y) \right| \end{aligned}$$

Joint AEP (see [CT91, Theorem 8.6.1,pp.195-196])

Theorem (Joint AEP)

Let $(X^{(n)}, Y^{(n)})$ be sequences of length n drawn IID according to $p(x^{(n)}, y^{(n)}) = \prod_i p(x_i, y_i)$, and choose $A_\epsilon^{(n)}$ to be the set of jointly typical sequences WRT to $p(x, y)$ then

1. $P\left((X^{(n)}, Y^{(n)}) \in A_\epsilon^{(n)}\right) \to 1$ *as* $n \to \infty$
2. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}$
3. *If* $(\tilde{X}^{(n)}, \tilde{Y}^{(n)}) \sim p(x^{(n)})p(y^{(n)})$*, i.e.,* $\tilde{X}^{(n)}$ *and* $\tilde{Y}^{(n)}$ *are independent with the same marginals as* $p(x^{(n)}, y^{(n)})$ *then*

   $$P\left((\tilde{X}^{(n)}, \tilde{Y}^{(n)}) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

   and for sufficiently large n

   $$P\left((\tilde{X}^{(n)}, \tilde{Y}^{(n)}) \in A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Implications of Joint AEP

The jointly typical set has

- about $2^{nH(X)}$ typical X sequences
- about $2^{nH(Y)}$ typical Y sequences
- about $2^{nH(X,Y)}$ jointly typical sequences

So when the two variables are not independent, $H(X, Y) < H(X) + H(Y)$, and hence not all pairs are jointly typical.

For a fixed $Y^{(n)}$ we can consider about $2^{nI(X;Y)}$ such pairs before we are likely to find a jointly typical pair.

That suggests there are about $2^{nI(X;Y)}$ distinguishable signals $X^{(n)}$.

Analysis of Random Codes

Actual assignment algorithm

1. Receiver receives Y^n according to the distribution

$$P\left(y^{(n)}\middle|x^{(n)}(w)\right) = \prod_{i=1}^{n} p(y_i|x_i(w))$$

2. Receiver decodes by guessing that w is the input that generates a jointly typical sequence:
   - If there is one codeword $(x^{(n)}(\hat{w}), y^{(n)}) \in A_\epsilon^{(n)}$, then we decode as $\hat{w}$.
   - If there are two codewords such that $(x^{(n)}(w_i), y^{(n)}) \in A_\epsilon^{(n)}$, then we declare an error event 2.
   - If there is no codeword $(x^{(n)}(w), y^{(n)}) \in A_\epsilon^{(n)}$, then we declare an error event 1.
3. In the 1st case, if $\hat{w} \neq w$ we also declare an error event 2.

Analysis of Random Codes

Probability of errors:

- Probability the jointly typical sequence exists $\rightarrow 1$ as $n \rightarrow \infty$ by the first property of the Joint AEP
  - so probability of type 1 errors $P_1^{error} \rightarrow 0$
- Consider type 2 errors: then for some $i \neq j$

$$\left(x^{(n)}(w_i), y^{(n)}(w_j) \right) \in A_\epsilon^{(n)}$$

  - by the code generation process $x^{(n)}(w_i)$ and $x^{(n)}(w_j)$ are independent
  - hence $x^{(n)}(w_i)$ and $y^{(n)}(w_j)$ are independent
  - by third property of Joint AEP, for independent $x^{(n)}(w_i)$ and $y^{(n)}(w_j)$

$$P\left((x^{(n)}(w_i), y^{(n)}(w_j)) \in A_\epsilon^{(n)} \right) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

Analysis of Random Codes

Probability of errors: consider $w = 1$ WLOG

- There are 2^{nR} codewords, and so $2^{nR} - 1$ possible incorrect codewords, so the chance of a type 2 error is

$$\begin{aligned} P_2^{error} &\leq \left(2^{nR} - 1\right) 2^{-n(I(X;Y) - 3\epsilon)} \\ &\leq 2^{-n(I(X;Y) - 3\epsilon - R)} \end{aligned}$$

- Take rate $R < I(X;Y) - 3\epsilon$, then as $n \to \infty$ we have

$$P_2^{error} \to 0$$

Cons of Random Codes

So why don't we use random codes

1. very large blocks needed for asymptotic results to hold
2. assumes we know $p(y|x)$
3. all codewords must be shared
    1. $2^{nR} \times n$ matrix needs to be shared for large n
4. decoding very inefficient
    1. compute all alternatives and decide which is jointly typical?
    2. or store the mapping, which is impractical for even medium blocks

So these random codes are only really suitable for proofs, but there are other places where random codes are used for real, but we will concentrate on some others.

BTW, [CT91] from 1991, says there are no efficient codes that reach capacity – that's not true anymore, just to give an indication of how recent this all is.

Further reading I

- Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.
- David J. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2011.