Lecture 14 (03/27/18). Channels. Decoding. Preview of the Capacity Theorem.

A. Barg

The concept of a communication channel in information theory is an abstraction for transmitting digital (and analog) information from the sender to the recipient over a noisy medium. Examples of physical channels are wireless links, cable communications (optical, coaxial, etc.), writing onto digital media (flash, magnetic), and many more.

Let $\mathcal{X}, \mathcal{Y}$ be finite sets. A mapping $W : \mathcal{X} \to \mathcal{Y}$ is called *stochastic* if the image of $x \in \mathcal{X}$ is a random variable taking values in $\mathcal{Y}$. Denote $P(x \mapsto y)$ by $W(y|x)$, the probability of y conditional on the given input x.

Definition: A discrete memoryless channel (DMC) is a stochastic mapping $W : \mathcal{X} \to \mathcal{Y}$. We use the letter W to refer both to the channel itself and to the probability distribution $W(y|x)$. The sets $\mathcal{X}$ and $\mathcal{Y}$ are called the input alphabet and the output alphabet of W, respectively. The channel is represented by a stochastic matrix whose rows are labelled by the elements of $\mathcal{X}$ (input letters) and columns by the elements of $\mathcal{Y}$ (output letters). By definition $\sum_{y\in\mathcal{Y}} W(y|x) = 1$ for any $x \in \mathcal{X}$.

Examples. 1. Z-channel (called so because its diagram resembles the letter Z).

$$W = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ 0 & 1 \end{pmatrix}$$

2. Binary symmetric channel (BSC(p)) $W : \{0,1\} \to \{0,1\}$

$$W = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

3. Binary erasure channel (BEC$((p))$ $W : \{0,1\} \to \{0,1,?\}$

$$W = \begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix}$$

There are more examples in the textbook.

Definition: Let $\mathcal{M}$ be a finite set of cardinality M and let $f : \mathcal{M} \to \mathcal{X}^n$ be a mapping. A code $\mathcal{C}$ of length n over the alphabet $\mathcal{X}$ is the image of f in $\mathcal{X}^n$. We say that a message $m \in \mathcal{M}$ is encoded into a codeword $x_m \in \mathcal{C}$ if $f(m) = x_m$. The set of codewords $\{x_1, \dots, x_M\}$ is called a channel code[1]. The number $R = \frac{1}{n}\log M$ is called the rate of the code $\mathcal{C}$. Below we denote general n-vectors by x^n, y^n and keep the above notation for the codewords.

The codewords are "transmitted over the channel". This means the following. The mapping W is extended from $\mathcal{X}$ to $\mathcal{X}^n$ using the memoryless property of W:

$$W^n(y^n|x^n) = \prod_{i=1}^{n} W(y_i^n|x_i^n), \text{where } x^n = (x_1^n, \dots, x_n^n), y^n = (y_1^n, \dots, y_n^n).$$

The result of transmitting the codeword x_m over the channel W is a vector $y^n \in \mathcal{Y}^n$ with probability $W^n(y^n|x_m)$.

Messages are encoded and transmitted as codewords to provide the recipient with the functionality of correcting errors that may occur in the channel. Error correction is performed by a *decoder*, i.e., a mapping $g : \mathcal{Y}^n \to \mathcal{M}$. The decoder is a deterministic mapping constructed so as to minimize the probability of incorrect recovery of transmitted messages.

[1]Sometimes the term code is used to refer to f and then the set of codewords $\mathcal{C}$ is called the codebook.

Optimal decoders. We briefly discuss optimal decoding rules. Let $\Pr(m)$ be a probability distribution on $\mathcal{M}$. Let y^n be the "received vector", i.e., the output of the channel. The posterior probability that the transmitted message was m equals

$$P(m|y^n) = \frac{\Pr(m)W^n(y^n|x_m)}{P(y^n)}, \tag{1}$$

where $P(y^n) = \sum_{m=1}^{M} \Pr(m)W^n(y^n|x_m)$. Assume that $g(y^n) = m$, then the error probability is $p_e = 1 - P(m|y^n)$. To minimize p_e decode y to m such that

$$P(m|y^n) \geq P(m'|y^n) \quad \text{for all } m' \neq m \tag{2}$$

(ties are broken arbitrarily). This rule is called the *maximum aposteriori probability* (MAP) decoder. If $\Pr(m) = 1/M$ is uniform, then the MAP decoder is equivalent to the *maximum likelihood* (ML) decoder g_{ML} given by $g(y^n) = m$ if

$$W^n(y^n|x_m) \geq W^n(y^n|x_{m'}) \quad \text{for all } m' \neq m.$$

To see this, use the Bayes formula (1) in (2). If $\Pr(m) = 1/M$ is not uniform, then the ML decoder is generally suboptimal.

ML and MAP decoders are computationally very hard because of the large search involved in finding $g(y)$.

Preview of the Shannon capacity theorem. The following discussion is informal. It uses the simple case of the BSC to explain the nature of channel capacity in geometric terms. Consider transmission over W=BSC$(p), p < 1/2$. Let $d_H(x^n, y^n) = |\{i : x_i \neq y_i\}|$ be the Hamming distance between the (binary n-dimensional) vectors x^n and y^n.

Let x^n be the transmitted vector and y^n the received vector. The typical value of the distance $d_H(x^n, y^n) \approx np$. In other words, $\Pr\{|d_H(x^n, y^n) - np| \geq n\alpha\}$ is small, where $\alpha > 0$ is a small number. Therefore define the decoder value $g(y^n)$ as follows: if there is a unique codevector $x_m \in \mathcal{C}$ such that $|d_H(x_m, y^n) - np| \leq n\alpha$, then $g(y^n) = x_m$, otherwise put $g(y) = x_1$ (or any other arbitrary codevector). Below we call vectors y^n whose distance from x_m is about np *typical* for x_m.

The number of typical vectors $y^n \in \{0,1\}^n$ for a given x^n is

$$|\{y^n \in \{0,1\}^n : \ |d_H(x^n, y^n) - np| \leq n\lambda\}| = \sum_{i:|i-np|\leq n\lambda} \binom{n}{i}. \tag{3}$$

Lemma 1. *Let* $1 \leq \lambda \leq 1/2$, *then*

$$\frac{1}{n+1} 2^{nh(\lambda)} \leq \sum_{i=0}^{\lambda n} \binom{n}{i} \leq 2^{nh(\lambda)}. \tag{4}$$

Proof :

$$1 = (\lambda + (1-\lambda))^n \geq \sum_{i=0}^{\lambda n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} \geq \sum_{i=0}^{\lambda n} \binom{n}{i} (1-\lambda)^n \Big(\frac{\lambda}{1-\lambda}\Big)^{\lambda n} = 2^{-nh(\lambda)} \sum_{i=0}^{\lambda n} \binom{n}{i}$$

∎

We would like the sets $T_\alpha(x_m) \triangleq \{y^n \in \mathcal{Y}^n : \ |d_H(x_m, y^n) - np| \leq n\alpha\}$ for different x_m to be disjoint. Note that

$$|T_\alpha(\cdot)| = \sum_{i=n(p-\alpha)}^{n(p+\alpha)} \binom{n}{i}.$$

Suppose first that

$$M|T_\alpha(\cdot)| \overset{(4)}{\geq} \frac{1}{n+1} M 2^{nh(p-\alpha)} > 2^n, \tag{5}$$

then a point y^n in the output space is typical on average for an exponentially large number of codevectors, namely for $A = 2^{n(R+h_2(p-\alpha)-o(1))}/2^n = 2^{n(R+h_2(p-\alpha)-1-o(1))} = 2^{n\varepsilon}$ codevectors, where we denoted $\varepsilon = R + h_2(p) - \alpha'' - 1$. This means that a significant proportion of points y^n is typical for exponentially many codewords. In this case decoding with low error probability is impossible (for instance, the maximum error probability is close to 1). We observe that (5) implies the following inequality:

$$R > 1 - h(p) + \alpha''$$

where α'' is small if so is α. Thus if this inequality is true, the error probability is large.

At the same time, if the cumulative volume of the typical sets around the codewords satisfies

$$M|T_\alpha(\cdot)| = 2^{nR}|T_\alpha(\cdot)| \overset{(4)}{\leq} 2^{n(R+h_2(p+\alpha))},$$

i.e., is less than 2^n (the total size of the output space), then there can exist codes in which decoding is correct with large probability. We will show by random choice that such codes indeed exist.

We notice that there is a dividing point $R = 1 - h_2(p)$ between the low and high error probability of decoding. This value of the rate is called the *capacity* of the channel W. We can also "flip" the question by asking

Threshold noise level: We are given a code $\mathcal{C}$ of rate R. Assuming that $\mathcal{C}$ is chosen optimally for transmission over a BSC(p), what is the largest p for which the code can guarantee reliable transmission?

The above argument shows that the maximum p is $p_{th} = h_2^{-1}(1 - R)$ (Plot the function $h_2^{-1}(z)$ to visualize the dependence on the rate).

Lecture 15 (03/29/18). Channel capacity

The following is a formalization of the discussion in the previous section. Consider $BSC(p)$, i.e., a stochastic mapping $W : \mathcal{X} \to \mathcal{Y}, \mathcal{X} = \mathcal{Y} = \{0, 1\}$ such that

$$W(y|x) = (1-p)1_{\{x=y\}} + p1_{\{x\neq y\}}.$$

Let

$$T_\alpha(x^n) = \{y^n : |d_H(x^n, y^n) - np| \leq n\alpha\}, \quad \alpha > 0.$$

Let $\mathcal{C} = \{x_1^n, \ldots, x_M^n\}$ be a code (below we omit the superscript ^{n} from the notation for the codewords). Let

$$D(x_m) \triangleq \{y^n \in \{0,1\}^n : \forall_{m'}\ W^n(y^n|x_m) \geq W^n(y^n|x_{m'})\}$$

be the decision region for the codeword x_m. Denote by

$$\lambda_m = \sum_{y^n \in D(x_m)^c} W^n(y^n|x_m)$$

the error probability of decoding conditional on transmitting the mth codeword and let

$$\lambda_{max}(\mathcal{C}) = \max_m \lambda_m$$

be the maximum error probability of the code $\mathcal{C}$.

Theorem 2. *(Shannon's capacity theorem for the BSC, lower bound) Given $\varepsilon > 0, \gamma > 0, p < 1/2$ and $R \leq 1 - h(p) - \gamma$, there exists $n_0 = n_0(\varepsilon, \gamma)$ such that for any $n \geq n_0$ there exists a code $\mathcal{C} \subset \{0,1\}^n$ of cardinality 2^{Rn}, whose maximum error probability of decoding on a BSC(p) satisfies $\lambda_{\max} \leq \varepsilon$.*

Proof : Let $M = 2^{R'n}$, where $R' = R + \frac{1}{n}$. Choose $\mathcal{C} = \{x_1, \ldots, x_M\} \in \{0,1\}^{Mn}$ by randomly assigning codewords to the messages with uniform probability $\Pr(f(m) = x_m) = 2^{-n}$ independently of each other (below we use the notation x_m for codewords, omitting the superscript x_m^n). Suppose that y^n is the received vector. Let us use the following decoder $g : \mathcal{Y}^n \to \mathcal{M}$: if there is a *unique codeword* $x_m \in \mathcal{C}$ such that $y^n \in T_\alpha(x_m)$, assign $g(y^n) = m$. In all other situations put $g(y^n) = 1$. (This mapping is called the ***typical pairs decoder***).

Let $Z_m = 1_{\{(y^n \in T_\alpha(x_m))\}}, m = 1, \ldots, M$ be the indicator random variable of the event $\{y^n$ is typical for $x^m\}$. Suppose that the transmitted vector is x_1. The probability of error λ_1 satisfies

$$\lambda_1 \leq \Pr\{Z_1 = 0\} + \Pr\{\sum_{m=2}^{M} Z_m \geq 1\}.$$

We have[2]

$$\Pr\{Z_1 = 0\} = \Pr\{y^n \notin T_\alpha(x_1)\} = \Pr\{|d_H(x_1, y^n) - np| > n\alpha\}$$
$$\overset{\text{Chebyshev inequality}}{\leq} \frac{np(1-p)}{(n\alpha)^2} = p(1-p)n^{-\delta}$$

[2]The Chebyshev inequality states that for any random variable X with finite expectation and variance $\mathrm{Var}(X)$ we have

$$\Pr\{|X - EX| \geq a\} \leq \frac{\mathrm{Var}\, X}{a^2}.$$

by taking $\alpha = n^{-(1-\delta)/2}, \delta > 0$. By taking n sufficiently large, we can guarantee that $\Pr\{Z_1 = 0\} \le \beta$, where $\beta > 0$ is arbitrarily small. Next for $m \ge 2$

$$\Pr\{Z_m = 1\} = \frac{|T_\alpha(\cdot)|}{2^n} \overset{(4)}{\le} 2^{-n(1-h(p+\alpha))}.$$

Use the union bound:

$$\Pr\{\sum_{m=2}^{M} Z_m \ge 1\} \le M \Pr\{Z_m = 1\} \le 2^{-n(1-R'-h(p+\alpha))} \le 2^{n(\alpha'-\gamma+\frac{1}{n})},$$

where we write $h(p+\alpha) = h(p)+\alpha'$, and α' is small if α is small (note that $\alpha' > 0$ since $p < 1/2$). By taking a sufficiently large n we can ensure that $\alpha' < \gamma - \frac{1}{n}$, and so $\Pr\{\sum_{m=2}^{M} Z_m \ge 1\} \le \beta$.

Now let use compute the average probability of error over all codeword assignments f :

$$\begin{aligned} P_e = E_F \lambda(\mathcal{C}) &= \frac{1}{M} \sum_{\mathcal{C}} Pr(\mathcal{C}) \sum_{m=1}^{M} \lambda_m(\mathcal{C}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_m(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) \le 2\beta \end{aligned} \tag{6}$$

where $\Pr(\mathcal{C}) = \prod_{m=1}^{M} \Pr\{f(m) = x_m\}$. Here F is the random mapping. Since we go over all the mappings, the sum $\sum_{\mathcal{C}} Pr(\mathcal{C})\lambda_m(\mathcal{C})$ does not depend on m.

By (6) there exists a code $\mathcal{C}^*$ for which the error probability averaged over M codewords satisfies $\lambda(\mathcal{C}) \le 2\beta$. By the Markov inequality, $|\{m \in \mathcal{M} : \lambda_m(\mathcal{C}^*) \ge 2\lambda(\mathcal{C}^*\}| \le M/2$. Thus, there exist at least $M/2$ messages[3] whose codewords in $\mathcal{C}^*$ are decoded with error probability $\lambda_m(\mathcal{C}^*) \le 4\beta$. Denote this set of codewords by $\widetilde{\mathcal{C}}^*$ and take $\beta = \varepsilon/4$. Thus there is a code $\widetilde{\mathcal{C}}^*$ of cardinality $\frac{M}{2} = 2^{n(R'-\frac{1}{n})}$, i.e., of rate R with $\lambda_{\max}(\widetilde{\mathcal{C}}^*) \le \varepsilon$. ∎

Observe that the probability $\lambda_{\max}$ falls as $n^{-\varepsilon}$. By using the optimal (i.e., MAP) decoder together rather than the typical pairs decoder it is possible to show that there exist much better codes for the BSC.

Theorem 3. *For any rate $R, 0 \le R < 1 - h_2(p)$ there exists a sequence of codes $\mathcal{C}_i, i = 1, 2 \ldots$ of growing length n such that*

$$\frac{\log |\mathcal{C}_i|}{n} \to R$$

and

$$\lambda_{\max}(\mathcal{C}_i) \le 2^{-n\{D(\delta(R)\|p))(1-o(1)\}} \tag{7}$$

where $D(x\|y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$, $\delta(R) \triangleq h_2^{-1}(1-R)$, and $o(1) \to 0$ as $n \to \infty$.

We will omit the proof.

Note that the decline rate of the maximum error probability is a much faster (exponential) function of the code length n than in the above argument. We took a loss by using a suboptimal decoder (and a simpler proof).

Finite-length scaling. The efficiency of our transmission design can be measured by the number of messages that can be transmitted reliably. Suppose that the code rate is $R = 1 - h_2(p) - \gamma$, where $\gamma > 0$ is small and p is the transmission probability of the BSC. Suppose moreover that we require

[3]This idea is called *expurgation*.

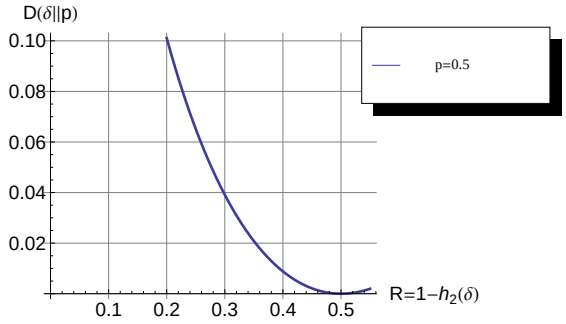
that $\lambda_{\max} = \varepsilon$. We already understand that we will have to choose a sufficiently long code. What is the smallest n that can guarantee this?

As $R \to 1 - h(p)$, we have $\delta(R) \to p$. We have $D(\delta(1-h(p))\|p) = D(p\|p) = 0$,

$$D'_\delta(\delta\|p) = \log \frac{(1-\delta)p}{\delta(1-p)}\Big|_{\delta=p} = 0, \quad D''_\delta(\delta\|p) = \frac{1}{\ln 2}\frac{1}{\delta(1-\delta)}.$$

If $R = 1 - h_2(p) - \gamma$ then $\delta = p + \gamma'$ where γ' is some small number. Expanding D into a power series in the neighborhood of $\delta = p$ we obtain

$$D(\delta\|p) = \frac{1}{2}D''_\delta(\delta\|p)(\delta - p)^2 + o((\delta-p)^2) = O((\delta-p)^2).$$



From (7) we obtain

$$n \geq \frac{\log(1/\varepsilon)}{(\delta - p)^2} \tag{8}$$

(constants omitted). Let us rephrase this by finding how γ (gap to capacity) depends on the code length n.

To answer this, rewrite (8) as follows:

$$\delta - p \geq (\log(1/\varepsilon))^{1/2}\frac{1}{\sqrt{n}}$$

Now substitute $\delta = h^{-1}(1-R)$ to find that $R \leq 1 - h(p - O(n^{-1/2}))$, or

$$R \leq 1 - h(p) - O\Big(\frac{1}{\sqrt{n}}\Big).$$

To conclude:

Proposition 4. *The gap-to-capacity for optimal codes scales as* $n^{-1/2}$.

The outcome of this calculation is called *finite-length scaling* of the code sequence on the BSC. The same order of scaling is true for optimal codes used on any binary-input discrete memoryless channel (DMC).

We return to the textbook, and state the Shannon capacity theorem for a DMC (pp.192–195, 200). Then we consider examples (pp. 187–191), and then (next class) prove the capacity theorem.