Introduction to Coding Theory **IST-UL, 2025/2026**

Week 4: Coding against probabilistic errors

Lecturer: João Ribeiro

Recommended reading: These notes are based on Chapter 7 of the "Elements of Information Theory" book by Cover and Thomas. For an alternative exposition (that also discusses inference and learning) see MacKay's book. For a more general and deeper treatment of information theory, check out this recent book by Polyanskiy and Wu.

Introduction

In this second part we use the information-theoretic concepts we developed to prove the noisy channel coding theorem.

Recall from the previous notes that Shannon's noisy channel coding theorem states that the coding capacity $C(\mathsf{Ch})$ of a DMC Ch always equals its information capacity $I(\mathsf{Ch})$. This means that for any $R < I(\mathsf{Ch})$ there exist families of codes of rate R that achieve vanishing decoding error probability over Ch, while any family of codes of rate $R > I(\mathsf{Ch})$ has decoding error probability bounded away from 0 (in fact, the decoding error probability will always be very large in this case).

For many DMCs, this gives an easy way of determining their exact coding capacity by solving a simple maximization problem.

1 Coding capacity $\leq$ information capacity

Our first goal is to show that the information capacity is always an upper bound on the coding capacity.

Fix a DMC Ch with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Let $(\mathcal{C}_n)_{n\in\mathbb{N}}$ be a family of (n, R_n, ε_n)-codes for Ch with rate[1] $R_n \to R$ and $\varepsilon_n \to 0$ as $n \to \infty$. We will show that $R \leq I(\mathsf{Ch})$ for any such family of codes.

Let X^n be uniformly distributed over $\mathcal{C}_n$, and let Y^n be the corresponding channel output. Then, $H(X^n) = \log|\mathcal{C}_n|$, and we have

$$I(X^n; Y^n) = H(X^n) - H(X^n|Y^n) = \log|\mathcal{C}_n| - H(X^n|Y^n). \tag{1}$$

[1] Note that here we define the rate as $R_n = \frac{\log|\mathcal{C}_n|}{n}$, so that $|\mathcal{C}_n| = 2^{R_n n}$.

Also,

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X^n, Y_1, \ldots, Y_{i-1}) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} (H(Y_i) - H(Y_i|X_i)) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \\ &\leq n \cdot \max_X I(X;Y) \\ &= n \cdot I(\mathsf{Ch}) \end{aligned} \tag{2}$$

The second inequality uses the fact that Y_i is conditionally independent of $(X_j, Y_j)_{j \neq i}$ given X_i, since Ch is a DMC.

Putting Equations (1) and (2) together yields

$$I(\mathsf{Ch}) \geq \frac{\log |\mathcal{C}_n|}{n} - \frac{1}{n} H(X^n|Y^n). \tag{3}$$

We know that $\frac{\log |\mathcal{C}_n|}{n} \to R$ as $n \to \infty$, so this would give the desired result $I(\mathsf{Ch}) \geq R$ if we could show that, under the assumption of vanishing decoding error probability, we have $\frac{1}{n} H(X^n|Y^n) \to 0$ as $n \to \infty$. This is a consequence of *Fano's inequality*. Intuitively, it states that if there is a way of predicting X^n from Y^n with high accuracy, then $H(X^n|Y^n)$ is small.

Lemma 1 (Fano's inequality) *Fix any two random variables X and Y, with X supported on $\mathcal{X}$. Suppose that there is an "estimator" f such that $\Pr[f(Y) = X] \leq \varepsilon$. Then,*

$$H(X|Y) \leq h(\varepsilon) + \varepsilon \log |\mathcal{X}|.$$

Proof: Let E be a random variable that is 1 when $f(Y) = X$ and 0 otherwise. Note that $\Pr[E = 0] \leq \varepsilon$, and that Y completely determines X when $E = 1$. Therefore,

$$\begin{aligned} H(X, E|Y) &= H(E|Y) + H(X|E, Y) \\ &\leq H(E) + \Pr[E = 1] \cdot 0 + \Pr[E = 0] \cdot H(X|Y, E = 0) \\ &\leq h(\varepsilon) + \varepsilon \log |\mathcal{X}|. \end{aligned}$$

Since $H(X|Y) \leq H(X, E|Y)$, the desired statement follows. ∎

We now use Fano's inequality to simplify Equation (3). Since $\mathcal{C}_n$ is an (n, R_n, ε_n)-code, there is a decoder Dec such that $\Pr[\mathsf{Dec}(Y^n) \neq X^n] \leq \varepsilon_n$. Therefore,

$$\frac{1}{n} H(X^n|Y^n) \leq \frac{1}{n} (h(\varepsilon_n) + \varepsilon_n \log |\mathcal{X}|^n) \leq \frac{1}{n} + \varepsilon_n \log |\mathcal{X}|.$$

Combining this with Equation (3), we get that

$$I(\mathsf{Ch}) \geq \frac{\log |\mathcal{C}_n|}{n} - \frac{1}{n} - \varepsilon_n \log |\mathcal{X}| \to R + 0 = R$$

when $n \to \infty$.

2 Coding capacity $\geq$ information capacity

It remains to show that any rate $R < I(\mathsf{Ch})$ is achievable. Recall that this means we need to show the existence of a family of codes $(\mathcal{C}_n)_{n\in\mathbb{N}}$ where each $\mathcal{C}_n$ is an (n, R_n, ε_n)-code with $R_n \to R$ and $\varepsilon_n \to 0$ as $n \to \infty$. We will establish the existence of this family of codes through the probabilistic method.

Fix an input distribution X supported on $\mathcal{X}$ with pmf P, and let Y denote the corresponding output distribution with respect to the DMC Ch.

Fix also a rate $R = I(X;Y) - \delta$ for some $\delta > 0$. For a given $n \in \mathbb{N}$, we construct a "preliminary code" $\mathcal{C}'$ of block length n and size 2^{Rn} with encoding and decoding functions $(\mathsf{Enc}, \mathsf{Dec})$ as follows.

Encoding function. This is where the probabilistic method comes in. For each message $i \in \{1, \ldots, 2^{Rn}\}$, we independently sample the corresponding codeword $\mathsf{Enc}(i)$ from $\mathcal{X}^n$ according to the product distribution $P(x) = \prod_{i=1}^n P(x_i)$ (in other words, coordinates of codewords are sampled i.i.d. according to X). Note that there is a positive probability that we assign the same codeword to two different messages, and so $\mathcal{C}'$ may not be a code in the usual sense. We will handle that later – for now we see $\mathcal{C}'$ as a multiset and allow different messages to be mapped to the same codeword.

Decoding function. We will now define a decoder Dec for $\mathcal{C}'$ and analyze its *average* decoding error probability. The decoder will be based on notions of typicality that we have seen before. Fix some $\gamma > 0$ to be set later (think of γ as being very small and the block length n of $\mathcal{C}'$ as being large). To simplify our discussion, we will say that a vector $z^n \in \mathcal{Z}^n$ is γ-*typical* with respect to a random variable Z supported on $\mathcal{Z}$ if

$$2^{-n(H(Z)+\gamma)} \leq \prod_{i=1}^n \Pr[Z = z_i^n] \leq 2^{-n(H(Z)-\gamma)}.$$

We have already shown that:

- For sufficiently large n a vector z^n sampled according to the product distribution $P(z^n) = \prod_{i=1}^n \Pr[Z = z_i^n]$ will be γ-typical with probability at least $1 - \gamma$;
- There are at most $2^{n(H(Z)+\gamma)}$ γ-typical sequences with respect to Z;
- For sufficiently large n there are at least $(1-\gamma)2^{n(H(Z)-\gamma)}$ γ-typical sequences with respect to Z.

Given a channel output $y^n \in \mathcal{Y}^n$, the decoder checks if there exists a unique message $i \in \{1, \ldots, 2^{Rn}\}$ for which the following properties hold:

1. $\mathsf{Enc}(i)$ is γ-typical[2] with respect to X;
2. y^n is γ-typical with respect to Y;
3. $(\mathsf{Enc}(i), y^n)$ is γ-typical with respect to the joint distribution of (X, Y).

If this is the case, then the decoder outputs $\mathsf{Dec}(y^n) = i$. Otherwise, it gives up and outputs $\mathsf{Dec}(y^n) = 0$.

Analyzing the average decoding error probability. As mentioned above, instead of analyzing the maximal decoding error probability (which is what we ultimately care about), we will instead start by analyzing the *average* decoding error probability of $\mathcal{C}'$. We denote this quantity by $\lambda(\mathcal{C}')$, and it is defined as

$$\lambda(\mathcal{C}') = \frac{1}{2^{Rn}} \sum_{i=1}^{2^{Rn}} \Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(i)}) \neq i].$$

We would like to show that there exists a $\mathcal{C}'$ of size 2^{Rn} for which $\lambda(\mathcal{C}')$ is small. To that end, we look at the expected average decoding error probability $\mathrm{E}_{\mathcal{C}'}[\lambda(\mathcal{C}')]$, where the expectation is taken over the sampling of $\mathcal{C}'$ as defined above. We can simplify the expectation as

$$\begin{aligned}
\mathrm{E}_{\mathcal{C}'}[\lambda(\mathcal{C}')] &= \mathrm{E}_{\mathcal{C}'}\left[\frac{1}{2^{Rn}} \sum_{i=1}^{2^{Rn}} \Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(i)}) \neq i]\right] \\
&= \frac{1}{2^{Rn}} \sum_{i=1}^{2^{Rn}} \mathrm{E}_{\mathcal{C}'}\left[\Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(i)}) \neq i]\right] \\
&= \mathrm{E}_{\mathcal{C}'}\left[\Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(1)}) \neq 1]\right],
\end{aligned}$$

where for $x \in \mathcal{X}^n$ we define $Y_x = (Y_{x_1}, \ldots, Y_{x_n})$. The second equality follows by linearity of expectation. The third equality follows by symmetry, since the codewords of $\mathcal{C}'$ are sampled independently and according to the same distribution, so $\mathrm{E}_{\mathcal{C}'}\left[\Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(i)}) \neq i]\right]$ does not depend on i.

Let y^n denote the channel output when sending $\mathsf{Enc}(1)$. There are two possible sources of error for the decoder:

- **Event 1:** The correct input $\mathsf{Enc}(1)$ does not satisfy properties 1–3 above together with y^n;
- **Event 2:** There is $i \neq 1$ such that the pair $(\mathsf{Enc}(i), y^n(1))$ satisfies properties 1–3 above.

We will show that the probability of either of these events occurring is small.

[2] We say that x^n is with respect to X if $2^{-n(H(X)+\gamma)} \leq \prod_{i=1}^{n} \Pr[X = x_i] \leq 2^{-n(H(X)-\gamma)}$.

Upper bounding the probability of the first decoding error event. We begin by upper bounding the probability of the first decoding error event.

Lemma 2 *For all sufficiently large n, the probability of the first decoding error event is at most 3γ.*

Proof: This follows easily from the properties of typical sequences we worked out in the previous notes and recollected above.

Note that the coordinates of $\mathsf{Enc}(1)$ are sampled i.i.d. according to X, and so, for large enough n, $\mathsf{Enc}(1)$ is γ-typical with respect to X with probability at least $1-\gamma$. Furthermore, since Ch is a DMC, the coordinates of y^n are i.i.d. according to Y (the channel output distribution on input X). Therefore, for large enough n, y^n is γ-typical with respect to Y with probability at least $1-\gamma$. Lastly, the pairs $(\mathsf{Enc}(1)_i, y_i^n)_{i=1,\ldots,n}$ are also i.i.d. according to the joint distribution (X,Y), and so $(\mathsf{Enc}(1), y^n)$ is γ-typical with respect to (X,Y) with probability at least $1-\gamma$ when n is sufficiently large.

Consequently, applying the union bound gives that the probability that at least one of the properties fails to hold is at most 3γ. ■

Upper bounding the probability of the second decoding error event. We now proceed to upper bound the probability of the second decoding error event. Fix any message $i \neq 1$. Recall that the coordinates $\mathsf{Enc}(i)$ are sampled i.i.d. according to X, and that $\mathsf{Enc}(i)$ is sampled independently of $\mathsf{Enc}(1)$. Therefore, if y^n denotes the channel output on input $\mathsf{Enc}(1)$, then the coordinates of $(\mathsf{Enc}(i), y^n)$ are i.i.d. according to the product distribution of X and Y. That is, we have $P(\mathsf{Enc}(i), y^n) = \prod_{j=1}^n \Pr[X = \mathsf{Enc}(i)_j] \cdot \Pr[Y = y_j^n]$.

Given this, what is the probability that $(\mathsf{Enc}(i), y^n)$ satisfy properties 1–3 above? The following lemma bounds that probability.

Lemma 3 $(\mathsf{Enc}(i), y^n)$ *satisfy properties* 1–3 *above with probability at most* $2^{-n(I(X;Y)-3\gamma)}$.

Proof: Recall that there are at most $2^{n(H(X,Y)+\gamma)}$ γ-typical sequences with respect to the joint distribution of X and Y. If $(\mathsf{Enc}(i), y^n)$ must satisfy properties 1 and 2, the probability that they land in this typical set is at most

$$2^{n(H(X,Y)+\gamma)} \cdot 2^{-n(H(X)-\gamma)} \cdot 2^{-n(H(Y)-\gamma)} = 2^{-n(I(X;Y)-3\gamma)}.$$

Here, we used the fact that

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

■

An appropriate upper bound on the probability of the second decoding error event follows easily.

Corollary 1 *The probability of the second decoding error event is at most* $2^{-n(\delta-3\gamma)}$.

Proof: By a union bound over the $2^{Rn}-1$ choices for $i \neq 1$ and Lemma 3, the probability that there exists $i \neq 1$ for which the second decoding error event occurs is at most

$$2^{Rn} \cdot 2^{-n(I(X;Y)-3\gamma)} = 2^{-n(I(X;Y)-R-3\gamma)} = 2^{-n(\delta-3\gamma)}.$$

■

Upper bounding the expected average decoding error probability. We upper bound the average decoding error probability by combining Lemma 2 and Corollary 1. Note that if neither Event 1 nor Event 2 occur, then Dec outputs the correct message. Therefore,

$$\mathrm{E}_{\mathcal{C}'}[\lambda(\mathcal{C}')] = \mathrm{E}_{\mathcal{C}'}[\Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(1)}) \neq 1]] \leq \mathrm{E}_{\mathcal{C}'}[\Pr[\text{Event 1}] + \Pr[\text{Event 2}]] \leq 3\gamma + 2^{-n(\delta-3\gamma)}.$$

If $\gamma < \delta/3$ and n is sufficiently large then $2^{-n(\delta-3\gamma)} \leq \gamma$, in which case we get $\mathrm{E}_{\mathcal{C}'}[\lambda(\mathcal{C}')] \leq 4\gamma$. Therefore, we get the following.

Theorem 1 *Let Ch be a DMC with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Fix an input distribution X over $\mathcal{X}$, and let Y be the corresponding channel output distribution according to Ch. Fix any $R < I(X;Y)$ and $0 < \gamma < (I(X;Y)-R)/3$. Then, for all sufficiently large n there exists a multiset $\mathcal{C}'$ in $\mathcal{X}^n$ of size 2^{Rn} with $\lambda(\mathcal{C}') \leq 4\gamma$.*

From average to maximal decoding error probability. Theorem 1 is close to showing what we want: the existence of codes with large block length of rate $\approx I(X;Y)$ and small *maximal* decoding error probability. But we face two problems. First, in Theorem 1 we allow mapping different messages to the same codeword (and so see the code as a multiset). Second, we only bound the average decoding error probability instead of the maximal decoding error probability.

It turns out that both of these problems can be dealt with using just one simple trick. Let $\mathcal{C}'$ be the multiset guaranteed by Theorem 1 for given R and $\gamma < 1/8$ with associated encoding and decoding functions $(\mathsf{Enc}, \mathsf{Dec})$. For each message $i \in \{1, \ldots, 2^{Rn}\}$, define its associated decoding error probability

$$\varepsilon_i = \Pr[\mathsf{Dec}(Y_{\mathsf{Enc}(i)}) \neq i].$$

Note that

$$\lambda(\mathcal{C}') = \frac{1}{2^{Rn}} \sum_{i=1}^{2^{Rn}} \varepsilon_i \leq 4\gamma.$$

This implies that for at least half of the messages i we have $\varepsilon_i \leq 8\gamma$. Otherwise,

$$\lambda(\mathcal{C}') > \frac{1}{2} \cdot 8\gamma = 4\gamma,$$

a contradiction.

Consider $\mathcal{C}$ obtained by taking only the codewords $\mathsf{Enc}(i)$ of $\mathcal{C}'$ with $\varepsilon_i \leq 8\gamma$ (this is called *code expurgation*). The block length of $\mathcal{C}$ is n and its rate is at least

$$\frac{\log|\mathcal{C}|}{n} \geq \frac{\log(|\mathcal{C}'|/2)}{n} = R - \frac{1}{n} \to R$$

when $n \to \infty$. Furthermore, by definition, its maximal decoding error probability is at most 8γ, and γ can be made arbitrarily small if we increase the block length sufficiently. Finally, since $8\gamma < 1$, we get that $\mathcal{C}$ maps different messages to different codewords. Indeed, if i and j are messages such that $\mathsf{Enc}(i) = \mathsf{Enc}(j)$, then $\varepsilon_i = \varepsilon_j = 1$, and so both codewords are thrown out when going from $\mathcal{C}'$ to $\mathcal{C}$.

Summarizing, we conclude that for any input distribution X, rate $R < I(X;Y)$, and $\gamma > 0$, there exists a family of codes $(\mathcal{C}_n)_{n\in\mathbb{N}}$ of rate $R_n \geq R - 1/n \to R$ with maximal decoding error probability 8γ for all sufficiently large n. Since $\gamma > 0$ can be made arbitrarily small by increasing the block length n and since X is an arbitrary input distribution, this concludes the proof of the noisy channel coding theorem.