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

Week 3: Coding against probabilistic errors

Lecturer: João Ribeiro

Recommended reading: These notes are based on Chapters 2 and 3 of the "Elements of Information Theory" book by Cover and Thomas. For an alternative exposition (that also discusses connections to 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

So far we have focused on codes correcting a bounded number of adversarial errors. We have seen some initial constructions of such codes (parity, Hamming, simplex, Hadamard), and also obtained bounds on the optimal rate-distance tradeoff for these codes.

Adversarial error models are often too pessimistic. Therefore, it is also natural to consider models where errors occur probabilistically. This setting was considered already in Shannon's early work [Sha48a, Sha48b]. This seminal work introduced many revolutionary ideas, kickstarting the field of information theory. More than 75 years later, Shannon's paper is still a pleasant read. You can find it here. The Wikipedia page about this work is also worth reading.

We will study coding for noisy channels with "memoryless" errors. These include natural probabilistic variants of errors we have seen so far:

- Channels where each input bit is flipped (from 0 to 1 or vice-versa) independently with some error probability $p \in [0, 1]$. These are called *binary symmetric channels*.
- Channels where each input bit is erased (replaced by a "?") with some erasure probability $p \in [0, 1]$. These are called *binary erasure channels*.

Concretely, we are interested in the probabilistic analog of the rate-distance tradeoff – we want to understand the optimal rate we can achieve while ensuring "reliable communication" through these channels. This is called the *channel capacity*. We will define all of this more formally soon.

In these 2-part notes we will work towards one of the crown jewels of information theory – Shannon's *noisy channel coding theorem*, which Shannon established already in his initial work. This theorem gives an exact characterization of the channel capacity for *discrete memoryless channels*, including the binary symmetric channels and binary erasure channels above. This should be contrasted with our knowledge in the adversarial setting, where we only know relatively loose bounds on the optimal rate for a given minimum distance.

The first part of these notes develops the information-theoretic concepts required to prove the noisy channel coding theorem.

1 Discrete memoryless channels

We begin by formally defining the family of probabilistic channels we will focus on. A *discrete memoryless channel* (DMC) Ch with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ behaves as follows:

- When we send an input $x \in \mathcal{X}$ through Ch, it outputs a sample from some distribution Y_x supported on $\mathcal{Y}$.
- When we send n inputs $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ through Ch, the channel processes these inputs in a memoryless fashion. More precisely, Ch outputs $Y = (Y_1, \ldots, Y_n)$, where the Y_i's are independent and each Y_i is distributed like Y_{x_i}. This means that the probability of observing channel output $y = (y_1, \ldots, y_n)$ on input $x = (x_1, \ldots, x_n)$ is

$$\Pr[Y = y | X = x] = \prod_{i=1}^{n} \Pr[Y_{x_i} = y_i].$$

As mentioned above, the two staple DMCs are the binary symmetric channel (BSC) and the binary erasure channel (BEC).

Formally, the BSC with error probability p has input and output alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and is characterized by the conditional channel output distribution

$$\Pr[Y_x = y] = \begin{cases} 1 - p, & \text{if } y = x, \\ p, & \text{if } y = 1 - x, \end{cases}$$

for $x \in \{0, 1\}$. This channel can be extended to non-binary alphabets. For alphabet size q, on input $x \in \{1, \ldots, q\}$ the channel outputs x with probability $1 - p$ and with probability p outputs a uniformly random sample from $\{1, \ldots, q\} \setminus \{x\}$.

The BEC with error probability p has input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, 1, ?\}$, and is characterized by the conditional channel output distribution

$$\Pr[Y_x = y] = \begin{cases} 1 - p, & \text{if } y = x, \\ p, & \text{if } y =?, \\ 0, & \text{if } y = 1 - x, \end{cases}$$

for $x \in \{0, 1\}$. It can also be easily generalized to non-binary alphabets.

2 Basic information-theoretic concepts

We now develop the basic concepts from information theory that we need to establish the noisy channel coding theorem. Information theory is a broad subject, and these basic concepts play a fundamental role in tackling many other problems as well.

2.1 Entropy

When studying a random variable X it is natural to try and measure the *uncertainty* we have about X. There are several ways of doing this, each with their own advantages and disadvantages. In these notes we introduce one such important measure of uncertainty, called the *Shannon entropy*. Since we will not be seeing other entropy measures in this course, we will often omit the "Shannon" part. For simplicity (and because this is all we need), we will focus on random variables taking values on a finite set. It is possible to extend these notions to more general settings.

Definition 1 (Entropy) *The* (Shannon) entropy *of a random variable X supported on a finite set $\mathcal{X}$, denoted by $H(X)$, is given by*

$$H(X) = -\sum_{x \in \mathcal{X}} \Pr[X = x] \log \Pr[X = x],$$

where we use the convention that $0 \log 0 = 0$.

Let's record some properties of $H(X)$ that are easy to work out. We have $H(X) \geq 0$ for all random variables X. Also, if $\Pr[X = x] = 1$ for some x, then $H(X) = 0$, so singleton distributions have minimum entropy, which makes sense. On the other extreme, if X is uniformly distributed over $\mathcal{X}$, then $H(X) = \log |\mathcal{X}|$. We will prove later on that no random variable supported on $\mathcal{X}$ can have larger entropy, so uniform distributions maximize entropy (which, again, makes sense).

We can also work out the entropy of simple random variables. For example, if X is a Bernoulli random variable with success probability p, then

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

This is the binary entropy function that appeared in our estimate for the volume of a Hamming ball! We plot it below as a function of $p \in [0, 1]$.

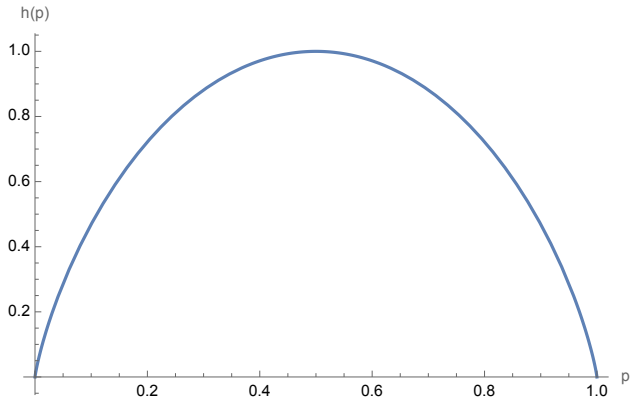

Figure 1: Binary entropy function $h(p)$.

The reader may wonder why we defined $H(X)$ this way. It turns out that if we take an axiomatic perspective and start by writing down reasonable properties that we expect a measure of uncertainty

to satisfy, then we can show that $H(X)$ is (up to a multiplicative constant) the only function satisfying these properties. We will establish some of these properties soon. One intuitive way of thinking about $H(X)$ is that $H(X)$ is the average number of bits you need to describe an outcome of X. We will get some intuition on why this holds soon.

We introduce some additional definitions.

Definition 2 (Joint entropy) *For any two random variables X and Y we define their* joint entropy *as*

$$H(X,Y) = -\sum_{x,y} \Pr[X = x, Y = y] \log \Pr[X = x, Y = y].$$

Definition 3 (Conditional entropy) *For any two random variables X and Y we define the* conditional entropy *of Y given X as*

$$H(Y|X) = \sum_{x \in \mathcal{X}} \Pr[X = x] \cdot H(Y|X = x).$$

The following result gives a useful way of decomposing joint entropies.

Theorem 1 (Chain rule for entropy) *For any random variables X, Y, and Z we have*

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

Proof: We will prove that

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

The argument extends easily to conditioning on some random variable Z.

The desired equality is a consequence of the fact that

$$\log \Pr[X = x, Y = y] = \log(\Pr[X = x] \cdot \Pr[Y = y|X = x]) = \log \Pr[X = x] + \log \Pr[Y = y|X = x].$$

Using this, we can write

$$\begin{aligned}
H(X,Y) &= -\sum_{x,y} \Pr[X = x, Y = y] \log \Pr[X = x, Y = y] \\
&= -\sum_{x,y} \Pr[X = x, Y = y] \log \Pr[X = x] - \sum_{x,y} \Pr[X = x, Y = y] \log \Pr[Y = y|X = x] \\
&= H(X) - \sum_{x} \Pr[X = x] \sum_{y} \Pr[Y = y|X = x] \log \Pr[Y = y|X = x] \\
&= H(X) + H(Y|X).
\end{aligned}$$

■

2.2 Mutual information

Given two random variables X and Y, it is also natural to try and measure the amount of information that learning X reveals about Y. We will define the *mutual information* between X and Y to capture this, but first we need some important definitions.

Definition 4 (Kullback-Leibler divergence) *The* Kullback-Leibler (KL) divergence *between two probability mass functions P and Q supported on a finite set $\mathcal{S}$,*[1] *denoted by* $\mathrm{D}_{\mathrm{KL}}(P\|Q)$*, is given by*

$$\mathrm{D}_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{S}} P(x)\log\frac{P(x)}{Q(x)}.$$

You may think of the KL divergence as measuring how close the two pmfs P and Q are. However, it is not a distance! In particular, KL divergence is not symmetric, and does not satisfy the triangle inequality.

It is also not clear at first sight whether $\mathrm{D}_{\mathrm{KL}}(P\|Q) \geq 0$ always, but this turns out to be true. This is one of the most important inequalities in information theory.

Theorem 2 *We have* $\mathrm{D}_{\mathrm{KL}}(P\|Q) \geq 0$ *for any pmfs P and Q.*

Proof: The desired inequality follows by combining the fact that $x \mapsto \log x$ is a concave[2] function with *Jensen's inequality*: if $f : D \to \mathbb{R}$ is a concave function on a convex domain $D \subseteq \mathbb{R}$ and X is a random variable supported on D, then $\mathrm{E}[f(X)] \leq f(\mathrm{E}[X])$.[3]

[1] A probability mass function (pmf) $P : \mathcal{S} \to [0,1]$ satisfies $P(x) \geq 0$ for all $x \in \mathcal{S}$ and $\sum_{x\in\mathcal{S}} P(x) = 1$. Since we are dealing with finite sets only, we may confuse random variables X and their respective pmfs P satisfying $P(x) = \Pr[X = x]$, which we may also call the *distribution* of X.

[2] A set $D \subseteq \mathbb{R}$ is *convex* if $x_1, x_2 \in D$ implies that $\lambda x_1 + (1-\lambda)x_2 \in D$ for all $\lambda \in [0,1]$. A function $f : D \to \mathbb{R}$ is *concave* on a convex domain $D \subseteq \mathbb{R}$ if for any $x_1, x_2 \in D$ and $\lambda \in [0,1]$ it holds that $f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$. We say that f is *convex* if $-f$ is concave.

[3] We can prove Jensen's inequality for finitely supported random variables by induction on the size of the support. When $\mathcal{X} = \{x_1, x_2\}$ and $\lambda = \Pr[X = x_1]$ we have $\mathrm{E}[f(X)] = \lambda f(x_1) + (1-\lambda)f(x_2) \leq f(\lambda x_1 + (1-\lambda)x_2) = f(\mathrm{E}[X])$. If $\mathcal{X} = \{x_1, \ldots, x_n\}$ for $n \geq 3$, then for $\lambda_i = \Pr[X = x_i]$ and $\gamma = \sum_{j=2}^{n} \lambda_j = 1 - \lambda_1$ we have

$$\mathrm{E}[f(X)] = \lambda_1 f(x_1) + \gamma \sum_{i=2}^{n} \frac{\lambda_i}{\gamma} f(x_i) \leq \lambda_1 f(x_1) + \gamma f\left(\sum_{i=2}^{n} \frac{\lambda_i}{\gamma} x_i\right) \leq f\left(\lambda_1 x_1 + \sum_{i=2}^{n} \lambda_i x_i\right) = f(\mathrm{E}[X]).$$

The first inequality uses the induction hypothesis for supports of size $n-1$.

Let $\mathcal{X} = \{x : P(x) > 0\}$. Then, we have

$$\begin{aligned} -\mathrm{D}_{\mathrm{KL}}(P\|Q) &= \sum_{x\in\mathcal{X}} P(x)\log\frac{Q(x)}{P(x)} \\ &\leq \log\left(\sum_{x\in\mathcal{X}} P(x)\cdot\frac{Q(x)}{P(x)}\right) \\ &= \log\left(\sum_{x\in\mathcal{X}} Q(x)\right) \\ &\leq \log 1 \\ &= 0. \end{aligned}$$

The first inequality follows from Jensen's inequality applied to the logarithm. ■

We define mutual information in terms of KL divergence. Intuitively, it measures how close the joint distribution of X and Y is to the distribution we would get if we assumed X and Y were independent.

Definition 5 (Mutual information) *The* mutual information *between two random variables X and Y, denoted by $I(X;Y)$, is given by*

$$I(X;Y) = \mathrm{D}_{\mathrm{KL}}(P_{XY}\|P_X \times P_Y) = \sum_{x,y}\Pr[X=x,Y=y]\log\frac{\Pr[X=x,Y=y]}{\Pr[X=x]\cdot\Pr[Y=y]},$$

where P_{XY} denotes the joint distribution of X and Y and $P_X \times P_Y$ is the product distribution with P_X and P_Y the pmfs of X and Y, respectively.

Theorem 3 *We have $I(X;Y) = H(Y) - H(Y|X) \geq 0$ and $I(X;Y) = I(Y;X)$.*

Proof: The fact that $I(X;Y) \geq 0$ follows directly from Theorem 2. The fact that $I(X;Y) = I(Y;X)$ follows easily by inspecting the expression for $I(X;Y)$.

Finally, the fact that $I(X;Y) = H(Y) - H(Y|X)$ is a consequence of the fact that

$$\log\frac{\Pr[X=x,Y=y]}{\Pr[X=x]\cdot\Pr[Y=y]} = \log\frac{\Pr[Y=y|X=x]}{\Pr[Y=y]} = -\log\Pr[Y=y] + \log\Pr[Y=y|X=x].$$

Then,

$$\begin{aligned} I(X;Y) &= \sum_{x,y}\Pr[X=x,Y=y]\log\frac{\Pr[X=x,Y=y]}{\Pr[X=x]\cdot\Pr[Y=y]} \\ &= -\sum_{x,y}\Pr[X=x,Y=y]\log\Pr[Y=y] \\ &+ \sum_{x}\Pr[X=x]\sum_{y}\Pr[Y=y|X=x]\log\Pr[Y=y|X=x] \\ &= H(Y) - H(Y|X). \end{aligned}$$

■

The non-negativity of mutual information has as an immediate corollary the useful fact that revealing additional information cannot increase the entropy of a random variable.

Corollary 1 (Conditioning reduces entropy) *For any two random variables X and Y we have*

$$H(Y|X) \leq H(Y).$$

In particular, this means that $H(X,Y) \leq H(X) + H(Y)$.

We can also use the non-negativity of the KL divergence to show that uniform distributions maximize entropy.

Corollary 2 (Uniform distribution has maximum entropy) *For every random variable X supported on $\mathcal{X}$ we have $H(X) \leq \log|\mathcal{X}|$, and this is achieved when X is uniformly distributed over $\mathcal{X}$.*

Proof: Let P be the pmf of X and let U be the uniform distribution over $\mathcal{X}$ (so $U(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$). Then,

$$0 \leq \mathrm{D}_{\mathrm{KL}}(P\|U) = \sum_x P(x) \log \frac{P(x)}{U(x)} = \sum_x P(x)(\log P(x) + \log|\mathcal{X}|) = \log|\mathcal{X}| - H(X).$$

■

We can also define a notion of *conditional* mutual information.

Definition 6 (Conditional mutual information) *The* mutual information between X and Y conditioned on Z *is given by*

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

It is a good exercise to convince yourself that, unlike entropy, mutual information can increase after conditioning!

Mutual information also satisfies a chain rule.

Theorem 4 (Chain rule for mutual information) *For any random variables X, Y, and Z we have*

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

Proof: We have

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

■

2.3 The asymptotic equipartition property and typicality

We now discuss a central property of Shannon entropy, the *asymptotic equipartition property* (AEP) which finds applications to data compression, channel coding, and many other problems. In particular, the AEP will play a key role in the error-correction procedure for the codes we analyze.

Intuitively, the AEP states that if we obtain n samples $x_1, \ldots, x_n$ independently and identically distributed (i.i.d.) according to some random variable with pmf P, then $P(x_1, \ldots, x_n) = \prod_{i=1}^n P(x_i) \approx 2^{-nH(X)}$ with high probability. More precisely, we have the following result, which is a direct consequence of the weak law of large numbers.[4]

Theorem 5 (Asymptotic equipartition property) *For every $\varepsilon, \delta > 0$ there exists n_0 such that for all $n \geq n_0$ and $X_1, \ldots, X_n$ i.i.d. according to X with pmf P we have*

$$\Pr_{(x_1,\ldots,x_n)\sim(X_1,\ldots,X_n)}\left[\left|-\frac{1}{n}\log P(x_1,\ldots,x_n) - H(X)\right| \leq \varepsilon\right] \geq 1-\delta,$$

where $P(x_1, \ldots, x_n) = \prod_{i=1}^n P(x_i)$.

Proof: This is a direct consequence of the weak law of large numbers by noticing that $H(X) = \mathrm{E}[-\log P(X)]$. ∎

Intuitively, we call a sequence $(x_1, \ldots, x_n)$ *typical* if $P(x_1, \ldots, x_n) \approx 2^{-nH(X)}$. We can then talk about the sets of typical sequences.

Definition 7 (Typical sets) *For a random variable X supported on $\mathcal{X}$, a real number $\varepsilon > 0$ and a positive integer n, we define the ε-typical set $A_\varepsilon^{(n)}$ with respect to X with pmf P as*

$$A_\varepsilon^{(n)} = \left\{x \in \mathcal{X}^n : 2^{-n(H(X)+\varepsilon)} \leq P(x_1,\ldots,x_n) \leq 2^{-n(H(X)-\varepsilon)}\right\}.$$

Theorem 6 *The following holds for any $\varepsilon > 0$:*

1. *For all sufficiently large n we have $\Pr[(X_1, \ldots, X_n) \in A_\varepsilon^{(n)}] \geq 1 - \varepsilon$;*
2. *$|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$;*
3. *For all sufficiently large n we have $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$.*

Proof:

1. Follows directly from the asymptotic equipartition property;

[4] The weak law of large numbers states that if $X_1, \ldots, X_n$ are i.i.d. according to X, then their empirical mean converges to the expected value of X in probability. That is, $\Pr_{x_1,\ldots,x_n\sim(X_1,\ldots,X_n)}\left[\left|\frac{1}{n}\sum_{i=1}^n x_i - \mathrm{E}[X]\right| > \varepsilon\right] \to 0$ as $n \to \infty$.

2. Suppose that $|A_\varepsilon^{(n)}| > 2^{n(H(X)+\varepsilon)}$. Then,

$$\Pr[(X_1,\ldots,X_n) \in A_\varepsilon^{(n)}] = \sum_{x \in A_\varepsilon^{(n)}} P(x) \geq |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)} > 1,$$

   a contradiction.

3. Suppose that $|A_\varepsilon^{(n)}| < (1-\varepsilon)2^{n(H(X)-\varepsilon)}$. Then,

$$\Pr[(X_1,\ldots,X_n) \in A_\varepsilon^{(n)}] = \sum_{x \in A_\varepsilon^{(n)}} P(x) \leq |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)} < 1-\varepsilon.$$

   This contradicts the first bullet when n is sufficiently large.

■

In sum, Theorem 6 says that there is a "typical set" of size $\approx 2^{nH(X)}$ such that (i) a tuple $(X_1,\ldots,X_n)$ of n i.i.d. samples from X falls into this set with high probability, and (ii) the elements of this set are nearly equiprobable under the distribution of $(X_1,\ldots,X_n)$.

There is a direct application of this result to data compression. We will not work this out rigorously, but here is the idea. Say that you receive n i.i.d. samples from some random source X. You would like to store these samples using as little memory as possible, in expectation. Roughly speaking, Theorem 6 guarantees that it suffices to use $\approx nH(X)$ bits of memory in expectation, and this turns out to be optimal. The intuition is that we can use bitstrings of length approximately $n(H(X)+\varepsilon)$ to describe the typical sequences (those in $A_\varepsilon^{(n)}$), and longer bitstrings to describe non-typical sequences. Since the n i.i.d. samples will be typical with high probability, the expected description length is close to $\log|A_\varepsilon^{(n)}|$, which is approximately $nH(X)$ for small ε.

3 An information-theoretic argument for bounding the volume of a Hamming ball

We use some of the concepts we have developed so far to upper bound the volume of a Hamming ball. For simplicity we focus on the binary case, but this can be generalized to larger alphabets.

Theorem 7 *For any positive integers n and $r \leq n/2$ it holds that*

$$\mathrm{Vol}_2(n,r) \leq 2^{nh(r/n)},$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

Proof: A good way to think about this is that the amount of information required to describe a uniformly random vector X from the radius-r ball is smaller than the amount of information required to describe n independent coin tosses with success probability r/n, because:

1. coordinates of X cannot be 1 with probability greater than r/n (otherwise the vector would fall outside the ball with some probability);
2. the coordinates in X are correlated. For example, when $X_1 = 1$ the probability that $X_2 = 1$ becomes smaller, because there is an upper bound on the total number of 1s in X.

Now we proceed to the rigorous argument. Let X be uniformly distributed over the Hamming ball of radius r. Then,

$$\log \mathrm{Vol}_2(n, r) = H(X) \leq \sum_{i=1}^{n} H(X_i).$$

Let I denote a uniformly random index between 1 and n. We have

$$H(X_I) \geq H(X_I|I) = \frac{1}{n} \sum_{i=1}^{n} H(X_i) \geq \frac{1}{n} \log \mathrm{Vol}_2(n, r). \tag{1}$$

All that remains is to upper bound $H(X_I)$ appropriately. First, note that X_I is a Bernoulli random variable. Since $w_H(X) \leq r$ always, the probability that $X_I = 1$ is at most $r/n \leq 1/2$. As the entropy of a Bernoulli random variable with success probability p increases with p for $p \in [0, 1/2]$, we conclude that

$$H(X_I) \leq h(p). \tag{2}$$

Combining Equations (1) and (2) yields the desired bound. ∎

4 Notions of channel capacity

4.1 Coding capacity

We begin by defining the notion of *coding capacity*, which is the analog of the rate-distance tradeoff for probabilistic channels. While in the adversarial setting we could demand perfect error-correction, in the probabilistic setting we must allow for a small probability of decoding error. Then, roughly speaking, the coding capacity of a channel is the largest rate that can be achieved while ensuring that the decoding error probability goes to 0 as the block length of the code increases.

More formally, consider a DMC Ch with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. An (n, R, ε)-*code* for Ch is a code $\mathcal{C} \subseteq \mathcal{X}^n$ of rate[5] $\frac{\log |\mathcal{C}|}{n} = R$ for which there exists a decoding function $\mathsf{Dec} : \mathcal{Y}^n \to \mathcal{X}^n$ such that for every codeword $c \in \mathcal{C}$ we have $\Pr[\mathsf{Dec}(Y_c) \neq c] \leq \varepsilon$, where Y_c denotes the channel output on input c.

Definition 8 (Achievable rate) *Given a DMC* Ch*, a real number* $R \geq 0$ *is said to be an* achievable rate *for* Ch *if there exists a family of codes* $(\mathcal{C}_n)_{n \in \mathbb{N}}$ *such that* $\mathcal{C}_n$ *is an* (n, R_n, ε_n)*-code with* $R_n \to R$ *and* $\varepsilon_n \to 0$ *as* $n \to \infty$*.*

[5] Note that here we do not normalize the rate by $\log |\mathcal{X}|$, and so the rate may be larger than 1 if $|\mathcal{X}| > 2$.

The coding capacity of a channel is the largest achievable rate. This matches the intuition of the coding capacity as the optimal rate of reliable information transmission, where "reliable" means "vanishing decoding error probability".

Definition 9 (Coding capacity) *The* coding capacity *of a DMC* Ch*, denoted* $C(\mathsf{Ch})$*, is given by*

$$C(\mathsf{Ch}) = \sup\{R \geq 0 : R \text{ is an achievable rate for } \mathsf{Ch}\}.$$

Given a channel Ch, what is its coding capacity? This is a central problem in channel coding. We can even be more ambitious and ask whether we can get close to this capacity using codes with highly efficient encoding and error-correction algorithms!

4.2 Information capacity

We also introduce another notion of channel capacity, called the *information capacity* of a channel. Roughly speaking, this quantifies, through mutual information, how much information the channel output reveals about the channel input.

Definition 10 (Information capacity) *Given a DMC* Ch *with input alphabet* $\mathcal{X}$*, we define its* information capacity*, denoted* $I(\mathsf{Ch})$*, as*

$$I(\mathsf{Ch}) = \sup_X I(X;Y),$$

where the supremum is taken over all random variables X *supported on* $\mathcal{X}$ *and* Y *is the channel output on input* X*.*

For many DMCs, computing the information capacity is not difficult.

Information capacity of the BSC. As our first example, we compute the information capacity of the BSC with error probability p. For any input distribution X we have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - h(p).$$

To see this, note that for $b \in \{0,1\}$ we have that $(Y|X=b)$ is a Bernoulli random variable with success probability p or $1-p$. Therefore, $H(Y|X=b) = h(p)$ always, and so $H(Y|X) = h(p)$. Since Y is supported on $\{0,1\}$, we know that $H(Y) \leq \log 2 = 1$, so $I(X;Y) \leq 1 - h(p)$ for all X. To see that this bound can be achieved, take X uniformly distributed over $\{0,1\}$, and note that then Y is also uniformly distributed over $\{0,1\}$. This means that $I(X;Y) = H(Y) - h(p) = 1 - h(p)$, and so the information capacity of the BSC with error probability p is $1 - h(p)$.

Information capacity of the BEC. We now compute the information capacity of the BEC with erasure probability p. For any input distribution X we have

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

Note that when $Y = b \neq ?$ we have $H(X|Y = b) = 0$, and $H(X|Y = ?) = H(X)$. Therefore, $H(X|Y) = pH(X)$, and so $I(X;Y) = (1-p)H(X)$ for all X.

Since $H(X) \leq 1$ for all X, we conclude that $I(X;Y) \leq 1 - p$ for all X. Furthermore, this can be achieved by taking X to be uniformly distributed over $\{0,1\}$, and so the information capacity of the BEC with erasure probability p is $1 - p$.

4.3 A preview of the noisy channel coding theorem

In the next notes, we will prove Shannon's noisy channel coding theorem by combining typicality with the probabilistic method.

Theorem 8 (Noisy channel coding theorem) *For any DMC* Ch *it holds that* $C(\mathsf{Ch}) = I(\mathsf{Ch})$.

For example, this means that over the BSC with error probability p we can communicate reliably at any rate $R < 1 - h(p)$, while all codes of rate $R > 1 - h(p)$ will have non-vanishing (actually, as we shall see, quite large) decoding error probability.

4.4 Beyond memoryless channels

In many scenarios errors are not memoryless. There are versions of Shannon's noisy channel coding theorem that apply to channels that are not DMCs. However, while for DMCs the noisy channel coding theorem gives us a nice way of computing the coding capacity, this is not the case more generally, because determining the information capacity for more complicated channels can be very challenging

For example, consider the following (apparently innocent) variant of the BEC, called the *binary deletion channel* (BDC) with deletion probability d: on input $x \in \{0,1\}^n$, the BDC independently deletes each x_i with probability d (and does not replace the deleted symbol by a "?"). If the input is 101010 and the BDC deletes the second, third, and fifth bits, the output of the channel is 100. This is not a DMC, although the channel behavior is memoryless.

We have a version of Shannon's noisy channel coding theorem for the BDC, *but, despite its simplicity, we know very little about the capacity of this channel!* This is one of the major open research directions in information theory. If you would like to learn more, see this survey.

References

[Sha48a] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[Sha48b] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948.