Quantum information theory (MAT4430) **Spring 2021**

Lecture 1: What is information theory?

Lecturer: Alexander Müller-Hermes

In the first lecture, we will briefly repeat some basic results in discrete probability theory and introduce two key problems in classical information theory:

- **Source coding:** The compression of a discrete memoryless information source.
- **Channel coding:** To send information reliably via a noisy communication channel at rates as high as possible.

We will end with a perspective of what we will study in the rest of the course.

1 Discrete probability theory

To talk about classical information theory, we will need some very basic terminology from finite probability theory. In the following, Σ will denote an alphabet, i.e., a countable set of symbols (a priori without any further structure). Often, our alphabets will be finite, and occasionally we will choose the particular alphabet

$$[d] := \{1, \ldots, d\},$$

or some other set. Let us define the two basic concepts of probability theory:

- A *probability distribution on* Σ is a function $p : \Sigma \to [0,1]$ such that $\sum_{x\in\Sigma} p(x) = 1$. We denote the set of probability distributions on Σ by $\mathcal{P}(\Sigma)$.
- A (discrete) *random variable* X is given by a pair (Σ, p) of an alphabet Σ, not necessarily finite, and a probability distribution $p \in \mathcal{P}(\Sigma)$. We say that X is Σ-valued and distributed according to p, and we write $X \sim p$.

If a random variable X is Σ-valued and distributed according to $p \in \mathcal{P}_d$, then we interpret $p(x)$ for $x \in \Sigma$ as the probability that X takes the value x, and we will write

$$\mathrm{P}(X = x) = p(x).$$

Given a subset $S \subset \Sigma$ we will write

$$\mathrm{P}(X \in S) = \sum_{x\in S} p(x).$$

We will sometimes simplify our language slightly, by not specifying the smallest alphabet Σ of values that a random variable X can take. For example, we will call a random variable $\mathbb{R}$-valued if it takes values on a discrete subset of $\mathbb{R}$, but not all values in $\mathbb{R}$.

Definition 1.1 (Joint and marginal probability distributions)**.** *Consider alphabets Σ_A and Σ_B and the product alphabet $\Sigma_A \times \Sigma_B$. For every distribution $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$, we can define the* marginal distributions *$p_A \in \mathcal{P}(\Sigma_A)$ and $p_B \in \mathcal{P}(\Sigma_B)$ by*

$$p_A(x) = \sum_{y'\in\Sigma_B} p_{AB}(x, y') \qquad \textit{and} \qquad p_B(y) = \sum_{x'\in\Sigma_A} p_{AB}(x', y),$$

for any $x \in \Sigma_A$ and any $y \in \Sigma_B$. The distributions in $\mathcal{P}(\Sigma_A \times \Sigma_B)$ are also called joint distributions*. All of these definitions generalize to more than two alphabets.*

It is common to write (X, Y) for a random variable with values in $\Sigma_A \times \Sigma_B$ distributed according to some probability distribution $p_{AB} \in \Sigma_A \times \Sigma_B$. We will refer to (X, Y) as a *pair of random variables* with values in $\Sigma_A \times \Sigma_B$ and joint distribution p_{AB}. This notation suggests that $X \sim p_A$ and $Y \sim p_B$ can somehow be considered as independent entities, but in general this is not so. In general, the two marginals p_A and p_B do not describe the entire probability distribution p_{AB}, and we refer to aspects not captured by the marginal distributions as *correlations*. There is a case where the two marginals are describing the entire joint distribution:

Definition 1.2 (Independence)**.** *We say that a pair of random variables (X, Y) with values in $\Sigma_A \times \Sigma_B$ distributed according to some probability distribution $p_{AB} \in \Sigma_A \times \Sigma_B$ is* independent *if the probability distribution p_{AB} factorizes as*

$$p_{AB}(x, y) = p_A(x)p_B(y).$$

In this case, the marginals p_A and p_B determine the entire joint distribution p_{AB} and we write $p_{AB} = p_A \times p_B$.

The definition of independence generalizes to general N-tuples of random variables. We will often consider sequences $(X_n)_{n\in\mathbb{N}}$ of random variables with values in Σ that are of independently and identically distributed according to some distribution $p \in \mathcal{P}(\Sigma)$. By this we mean, that for each $N \in \mathbb{N}$ the tuple $(X_1, \ldots, X_N)$ consists of independent random variables with joint distribution $p^{\times N}$, i.e., such that

$$p^{\times N}(x_1, \ldots, x_N) = p(x_1) \cdots p(x_N).$$

Probability theorists developed their own notation to deal with non-independent pairs (or tuples) of random variables. Given a pair (X, Y) of random variables with values in $\Sigma_A \times \Sigma_B$ distributed according to some probability distribution $p_{AB} \in \Sigma_A \times \Sigma_B$, we write

$$\mathrm{P}\left(X = x, Y = y\right) = p_{AB}(x, y).$$

It is common to use this notation quite liberally, and we will write

$$\mathrm{P}\left((X, Y) \in S\right) = \sum_{(x,y)\in S} p_{AB}(x, y),$$

for some set $S \subseteq \Sigma_A \times \Sigma_B$, and if $\Sigma = \Sigma_A = \Sigma_B$ has some additional structure we may even do some arithmetic such as

$$\mathrm{P}\left(X + Y = 3\right) = \mathrm{P}\left(X + Y \in \{(x, y) \in \Sigma \times \Sigma \ : \ x + y = 3\}\right).$$

If we just mention one of the random variables, then we will always mean the marginal distributions, i.e., we have

$$\mathrm{P}\left(X = x\right) = p_A(x) \quad \text{and} \quad \mathrm{P}\left(Y = y\right) = p_B(y).$$

If $p_B(y) \neq 0$, then we write

$$\mathrm{P}\left(X = x | Y = y\right) = \frac{p_{AB}(x, y)}{p_B(y)},$$

which we may abbreviate as $p(x|y)$. We call the probability distribution $p(\cdot|y) \in \mathcal{P}\left(\Sigma_A\right)$ the *conditional distribution* or the *probability distribution of X conditioned to $Y = y$*. This notation can be generalized by setting

$$\mathrm{P}\left(X \in S_A | Y \in S_B\right) = \frac{\sum_{x\in S_A y\in S_B} p_{AB}(x, y)}{\sum_{y'\in S_B} p_B(y')},$$

for subsets $S_A \subseteq \Sigma_A$ and $S_B \subseteq \Sigma_B$, if $\sum_{y' \in S_B} p_B(y') \neq 0$. As before, we may abbreviate the corresponding probability distribution as $p(\cdot|Y \in S_B) \in \mathcal{P}(\Sigma_A)$, which is called the *probability distribution of X conditioned to $Y \in S_B$*. These probabilities satisfy a few properties which can be derived directly from the definitions:

Theorem 1.3 (Properties of joint and conditional probabilities)**.** *Consider a pair of random variables (X, Y) with values in $\Sigma_A \times \Sigma_B$ distributed according to some probability distribution $p_{AB} \in \Sigma_A \times \Sigma_B$. We have:*

- ***Product rule:***

$$P(X \in S_A, Y \in S_B) = P(Y \in S_B) P(X \in S_A | Y \in S_B).$$

- ***Sum rule:***

$$P(X \in S_A) = \sum_{y \in \Sigma_B} P(Y = y) P(X \in S_A | Y = y).$$

- ***Bayes' theorem:***

$$P(Y \in S_B | X \in S_A) = \frac{P(Y \in S_B)\, P(X \in S_A | Y \in S_B)}{P(X \in S_A)}.$$

When talking about random variables attaining values in some countable subset $\Sigma \subset \mathbb{R}$, it will be useful to define the following two functions:

- **Expected value:** $\mathbb{E}[X] = \sum_{x \in \Sigma} p(x) x$.
- **Variance:** $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

In general, neither the expected value nor the variance have to be finite. However, we will often restrict to the case where the probability distributions have finite support, i.e., $p(x) \neq 0$ only for a finite number of $x \in \Sigma$, and in this case the expected value and the variance are finite.

The following lemma can be verified easily:

Lemma 1.4. *Let $\Sigma_A, \Sigma_B \subset \mathbb{R}$ be countable subsets and (X, Y) a pair of random variables with values in $\Sigma_A \times \Sigma_B$. Then, we have*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

If the random variables (X, Y) are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\, \mathbb{E}[Y].$$

We will sometimes need the following elementary inequalities of probability theory:

Theorem 1.5 (Markov's and Chebychev's inequalities)**.** *Let Z denote a random variable with values in a countable subset $\Sigma \subset \mathbb{R}$ and distributed according to $p \in \mathcal{P}(\Sigma)$.*

1. *(Markov's inequality) For every $\epsilon > 0$, we have*

$$P[|Z| \geqslant \epsilon] \leqslant \frac{\mathbb{E}[|Z|]}{\epsilon}.$$

2. *(Chebychev's inequality) For every $\epsilon > 0$, we have*

$$P\left[(Z - \mathbb{E}[Z])^2 \geqslant \epsilon\right] \leqslant \frac{Var[Z]}{\epsilon}.$$

Proof. Both inequalities are trivially satified if $\mathbb{E}\left[|Z|\right] = \infty$ or $\mathrm{Var}\left[Z\right] = \infty$. For the first inequality note that

$$\mathrm{P}\left[|Z| \geqslant \epsilon\right] = \sum_{|z|\geqslant\epsilon} p(z) \leqslant \sum_{|z|\geqslant\epsilon} p(z)\frac{|z|}{\epsilon} \leqslant \frac{\mathbb{E}\left[|Z|\right]}{\epsilon},$$

where we multiplied each summand by a number $|z|/\epsilon \leqslant 1$, and in the last inequality we added more non-negative terms. For the second inequality, we just insert the positive-valued random variable $(Z - \mathbb{E}\left[Z\right])^2$ into the first inequality. □

Using Chebychev's inequality we can prove the following theorem:

Theorem 1.6 (Weak law of large numbers)**.** *Let $\Sigma \subset \mathbb{R}$ be an alphabet. Consider a random variable Y with values in Σ and distributed according to $p \in \mathcal{P}(\Sigma)$ with expected value $\mu = \mathbb{E}\left[Y\right]$ and $Var[Y] < \infty$. If $(Y_n)_{n\in\mathbb{N}}$ are independent random variables identically distributed to Y, then*

$$\lim_{n\to\infty} P\left(\left|\frac{Y_1 + Y_2 + \cdots + Y_n}{n} - \mu\right| \geqslant \epsilon\right) = 0,$$

for every $\epsilon > 0$.

Proof. For any $n \in \mathbb{N}$ consider the random variable

$$Z_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{n},$$

and note that

$$\mathbb{E}\left[Z_n\right] = \mu.$$

Using Chebychev's inequality, we have

$$\mathrm{P}\left(\left|Z_n - \mu\right| \geqslant \epsilon\right) = \mathrm{P}\left[(Z_n - \mu)^2 \geqslant \epsilon^2\right] \leqslant \frac{\mathrm{Var}\left(Z_n\right)}{\epsilon^2}.$$

Finally, note that

$$\begin{aligned}\mathrm{Var}\left(Z_n\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{E}\left[Y_iY_j\right] - \mu^2 &= \frac{1}{n^2}\sum_{i\neq j}^{n}\mathbb{E}\left[Y_i\right]\mathbb{E}\left[Y_j\right] + \frac{1}{n^2}\sum_{i}^{n}\mathbb{E}\left[Y_i^2\right] - \mu^2 \\ &= \frac{1}{n}\left(\mathbb{E}\left[Y^2\right] - \mu^2\right) = \frac{1}{n}\mathrm{Var}\left(Y\right),\end{aligned}$$

and we conclude that

$$\mathrm{P}\left(\left|\frac{Y_1 + Y_2 + \cdots + Y_n}{n} - \mu\right| \geqslant \epsilon\right) \leqslant \frac{\mathrm{Var}\left(Y\right)}{n\epsilon^2} \to 0 \quad \text{as} \quad n \to \infty.$$

□

2 What is information?

Consider the following strings. Which of these contain a high amount of information (intuitively speaking)? Can you rationalize your intuition?

1. 000000000000000000000000000000000000000000000000000000000000000000

2. 0101010101010101010101010101010101010101010101010101010101010101010

3. 3141592653589793238462643383279502884197169399375105820974944592307

4. 1621683129157654671616347956219518784030306919262080790346992725831

5. “By virtue of its innermost intention, and like all questions about language, structuralism escapes the classical history of ideas which already supposes structuralism’s possibility, for the latter naively belongs to the province of language and propounds itself within it.Nevertheless, by virtue of an irreducible region of irreflection and spontaneity within it, by virtue of the essential shadow of the undeclared, the structuralist phenomenon will deserve examination by the historian of ideas. For better or for worse. Everything within this phenomenon that does not in itself transparently belong to the question of the sign will merit this scrutiny; as will everything within it that is methodologically effective, thereby possessing the kind of infallibility now ascribed to sleepwalkers and formerly attributed to instinct, which was said to be as certain as it was blind.”[1]

6. “Preheat oven to 220 degrees C. Melt the butter in a saucepan. Stir in flour to form a paste. Add water, white sugar and brown sugar, and bring to a boil. Reduce temperature and let simmer. Place the bottom crust in your pan. Fill with apples, mounded slightly. Cover with a lattice work crust. Gently pour the sugar and butter liquid over the crust. Pour slowly so that it does not run off. Bake 15 minutes in the preheated oven. Reduce the temperature to 175 degrees C. Continue baking for 35 to 45 minutes, until apples are soft.”[2]

7. “A direct search on the CDC 6600 yielded

   $$27^5 + 84^5 + 110^5 + 133^5 = 144^5$$

   as the smallest instance in which four fifth powers sum to a fifth power. This is a counterexample to a conjecture by Euler that at least n nth powers are required to sum to an nth power, $n > 2$.”[3]

As you may have noticed there might be different possible notions of “information”. Maybe, you had the idea of defining “information” in terms of compressibility such that a string of symbols has a large amount of “information” if it cannot be compressed too much, and it has a low amount of “information” if it can be compressed a lot. This is indeed the intuition behind the notion of “information” that we are going to define. However, to make it precise, we need to think about what it means to compress a bit string. One possible notion based on a model of computation would be as follows:

Definition 2.1 (Kolmogorov complexity)**.** *The Kolmogorov complexity of a string is the length of the shortest program of a Turing machine producing the string when initialized on an empty tape.*

While this notion of complexity is very elegant and has a very general scope, it has a serious disadvantage: The Kolmogorov complexity is uncomputable, i.e., there cannot exist an algorithm to compute it on any kind of computer. We will use a different definition first introduced by the mathematician Claude Shannon. To introduce this definition, we have to (as often in applied mathematics) first properly define the problem.

[1] Jaques Derrida, Writing and difference, Force and Signification

[2] https://www.allrecipes.com/recipe/12682/apple-pie-by-grandma-ople/

[3] Lander, L. J., Parkin, T. R. (1966) Bulletin of the American Mathematical Society, 72(6), 1079.

3 Classical source coding

The first idea is to not focus on the actual content of the string, but rather consider the statistical properties of the symbols appearing in it. To make this precise, we define what we mean by an information source:

Definition 3.1 (Discrete memoryless information sources)**.** *Let Σ denote a finite alphabet. A* discrete memoryless source on Σ *(DMS) is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ that are independently and identically distributed and take values in Σ .*

An example of a discrete memoryless source is the text written by a monkey on a typewriter. The random variables X_i is then the letter the monkey hits at time i. Of course, our definition of an information source is a strong idealization and in english text (not written by a monkey) consecutive symbols are correlated. For instance the combination "ed" will occurr more often than the combination "xz". We could even envision cases, where the probability of later symbols depends on all previous symbols using a kind of memory. Such non-i.i.d. information sources are studied extensively in information theory, but we will restrict ourselves to the simple setting stated above.

Informally, a compression scheme is a pair of an encoding and a decoding function. The encoding function maps blocks of symbols to bit strings of length as short as possible, and the decoding function should reverse this process. The main insight is to not require the decoding to work perfectly, but rather to measure its probability of success.

Definition 3.2 (Compression scheme)**.** *For any $\delta > 0$ and any $n, m \in \mathbb{N}$ an (n, m, δ)-compression scheme for a discrete memoryless source $(X_n)_{n\in\mathbb{N}}$ with distribution $p \in \mathcal{P}(\Sigma)$ on the alphabet Σ is a pair of functions*

$$E : \Sigma^n \to \{0, 1\}^m \quad \textit{and} \quad D : \{0, 1\}^m \to \Sigma^n,$$

such that the success probability satisfies

$$P\Big((D \circ E)(X_1, \dots, X_n) = (X_1, \dots, X_n)\Big) = \sum_{x_1,\dots,x_n\in S} p(x_1)\cdots p(x_n) \geqslant 1 - \delta,$$

where

$$S = \{(x_1, \dots, x_n) \in \Sigma^n \ : \ (D \circ E)(x_1, \dots, x_n) = (x_1, \dots, x_n)\},$$

denotes the set where the compression succeeds.

A good compression scheme will have two properties: The success probability will be high, and it compresses a string into few bits, i.e., the ratio m/n is low. It is intuitively clear, that there should be a tradeoff between the success probability and the compression rate: If we do not compress at all, then the success probability can be 1, but if we want to compress $n > 1$ symbols into $m = 1$ bits, then the success probability will (usually) be small. Shannon's next insight was to consider compression schemes in the asymptotic limit $n \to \infty$. To make this precise, we will define asymptotically achievable rates:

Definition 3.3 (Achievable rates)**.** *A number $R \in \mathbb{R}^+$ is called an* achievable rate *for compression of a discrete memoryless source $(X_n)_{n\in\mathbb{N}}$, if for every $n \in \mathbb{N}$ there exists an (n, m_n, δ_n) compression scheme for $(X_n)_{n\in\mathbb{N}}$ such that*

$$R = \lim_{n\to\infty} \frac{m_n}{n} \qquad \textit{and} \qquad \lim_{n\to\infty} \delta_n = 0.$$

It would be cool if we could find the optimal achievable rate. This is what Shannon did:

Theorem 3.4 (Shannon's source coding[4] theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ denote a discrete memoryless source on the alphabet Σ with distribution $p \in \mathcal{P}(\Sigma)$. The* Shannon entropy *is given by*

$$H(p) = -\sum_{x\in\Sigma} p(x)\log(p(x)). \tag{1}$$

1. *Any rate $R > H(p)$ is achievable for compression of the discrete memoryless source $(X_n)_{n\in\mathbb{N}}$.*

2. *If there is a sequence of (n_k, m_k, δ_k)-compression schemes for the discrete memoryless source $((X_n)_{n\in\mathbb{N}})$ satisfying*

$$\lim_{k\to\infty} n_k = \infty \qquad \text{and} \qquad \lim_{k\to\infty} \frac{m_k}{n_k} = R < H(p),$$

then we have $\lim_{k\to\infty}\delta_k = 1$, i.e., the success probability converges to zero.

Shannon's source coding theorem shows that rates close to $H(p)$ are achievable for compression, and that rates lower than $H(p)$ cannot be achieved with success probability converging to 1. In the following, we will construct a compression scheme that achieves rates arbitrarily close to $H(p)$. This will prove one direction of Theorem 3.4, and for the other direction we have to argue that such a sequence of schemes does not exist.

Let us first get some intuition about how the compression scheme will work: For a discrete memoryless source with some non-trivial and non-uniform distribution $p \in \mathcal{P}(\Sigma)$ not all strings $(x_1, x_2, \ldots, x_n) \in \Sigma^n$ of length n will have the same probability. For example, when you toss a biased coin with the probability for "tails" much larger than the probability for "heads", then it is very unlikely to observe a string of 100 "heads" in a row. Typically, we will observe strings of length n in which each symbol $x \in \Sigma$ occurs approximately $p(x)n$ times. To construct an efficient coding scheme with high success probability it might therefore be enough to focus on such typical strings and ignore the untypical ones. How can we characterize the typical strings? To get some intuition, let us consider a string $(x_1, \ldots, x_n)$ such that each symbol $x \in \Sigma$ occurs approximately $p(x)n$ times. What is the probability of observing such a string? We can compute it as

$$p(x_1)\cdots p(x_n) = \Pi_{x\in\Sigma} p(x)^{\#\{x_i = x\}} \approx \Pi_{x\in\Sigma} p(x)^{p(x)n} = 2^{n\sum_{x\in\Sigma} p(x)\log(p(x))} = 2^{-nH(p)},$$

and magically the Shannon entropy appears in the exponent. Motivated by this intuition, we state the following definition:

Definition 3.5 (Typical strings)**.** *Let Σ be an alphabet and $p \in \mathcal{P}(\Sigma)$ a probability distribution. For $n \in \mathbb{N}$ and $\epsilon > 0$ a string $(x_1, \ldots, x_n) \in \Sigma^n$ is called ϵ-typical for the distribution p if*

$$2^{-n(H(p)+\epsilon)} < p(x_1)\cdots p(x_n) < 2^{-n(H(p)-\epsilon)}.$$

We denote the set of these strings by $\mathcal{T}_{n,\epsilon}(p)$.

How many typical strings are there, and how likely is it that a string obtained from a discrete memoryless source is typical? To answer these questions we will use the weak law of large numbers from probability theory:

Lemma 3.6 (Properties of typical strings)**.** *Let Σ be an alphabet and $p \in \mathcal{P}(\Sigma)$ a probability distribution. We have:*

1. *For any $n \in \mathbb{N}$ and $\epsilon > 0$ we have*

$$|\mathcal{T}_{n,\epsilon}(p)| < 2^{n(H(p)+\epsilon)}.$$

[4] The term *source coding* is synonymous to compression, and it is more common in the literature.

2. For any $\epsilon > 0$ we have

$$\lim_{n\to\infty} P\Big((X_1,\dots,X_n) \in \mathcal{T}_{\epsilon,n}(p)\Big) = 1,$$

where $(X_n)_{n\in\mathbb{N}}$ is a discrete memoryless source distributed according to p.

Proof.

Ad 1.: By Definition 3.5 and the normalization of probability distributions, we have

$$2^{-n(H(p)+\epsilon)}|\mathcal{T}_{n,\epsilon}(p)| < \sum_{(x_1,\dots,x_n)\in\mathcal{T}_{n,\epsilon}(p)} p(x_1)\cdots p(x_n) \leqslant 1.$$

Ad 2.: Consider the function $f : \Sigma \to [0,\infty)$ given by

$$f(x) = \begin{cases} -\log(p(x)), & \text{if } p(x) > 0 \\ 0, & \text{if } p(x) = 0, \end{cases}$$

and the random variable $Z = f(X)$, where X is distributed according to p. Observe, that the expectation value of Z is given by

$$\mu = \mathbb{E}\left(Z\right) = \sum_{x\in\Sigma} p(x)f(x) = H(p).$$

We conclude from the weak law of large numbers that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\frac{f(X_1)+f(X_2)+\cdots+f(X_n)}{n} - H(p)\right| < \epsilon\right) = 1, \qquad (2)$$

whenever $(X_n)_{n\in\mathbb{N}}$ is a discrete memoryless source distributed according to p. Note that

$$\left|\frac{f(x_1)+f(x_2)+\cdots+f(x_n)}{n} - H(p)\right| < \epsilon,$$

holds for a string $(x_1,\dots,x_n) \in \Sigma^n$ if and only if

$$-n(H(p)+\epsilon) < \log\left(p(x_1)\cdots p(x_n)\right) < -n(H(p)-\epsilon).$$

After applying the exponential function, this is equivalent to $(x_1,\dots,x_n) \in \mathcal{T}_{n,\epsilon}(p)$, and we can rewrite (2) into

$$\lim_{n\to\infty} \mathrm{P}\Big((X_1,\dots,X_n) \in \mathcal{T}_{n,\epsilon}(p)\Big) = 1.$$

This finishes the proof.

□

Let us emphasize again the intuition behind Lemma 3.6: There are $2^{n\log(|\Sigma|)}$ many strings of length n in Σ^n, but at most $2^{n(H(p)+\epsilon)}$ of them are ϵ-typical for the distribution p. For large n and if $H(p) < \log(|\Sigma|)$ these are very few strings compared to the total number. Still, when receiving strings of length n from the information source, then we will essentially only get typical strings for large n. Let us exploit this fact, to construct a compression scheme:

Proof of Theorem 3.4.
Direct part. For $\epsilon > 0$ and any $n \in \mathbb{N}$ we will construct an $(n, \lceil n(H(p)+\epsilon)\rceil, \delta_n)$ compression scheme for the discrete memoryless source $(X_n)_{n\in\mathbb{N}}$ over the alphabet Σ distributed according to $p \in \mathcal{P}(\Sigma)$ such that $\delta_n \to 0$ as $n \to \infty$. Since

$$H(p) + \epsilon = \lim_{n\to\infty} \frac{\lceil n(H(p) + \epsilon)\rceil}{n},$$

this shows that the $H(p) + \epsilon$ is an achievable rate.

For $n \in \mathbb{N}$ and $\epsilon > 0$ we will construct a compression scheme which succeeds on all typical strings, i.e., we have $S = \mathcal{T}_{n,\epsilon}(p)$ in the terminology of Definition 3.2. First, we set $m = \lceil n(H(p) + \epsilon)\rceil$ and we choose a bit string $b(x_1, x_2, \ldots, x_n) \in \{0, 1\}^m$ for any typical sequence $(x_1, \ldots, x_n) \in \mathcal{T}_{n,\epsilon}(p)$. The first case of Lemma 3.6 shows that there are enough bit strings of length m to do this. Now, we define an encoding function $E_n : \Sigma^n \to \{0, 1\}^m$ by

$$E_n(x_1, \ldots, x_n) = \begin{cases} b(x_1, \ldots, x_n), & \text{if } (x_1, \ldots, x_n) \in \mathcal{T}_{\epsilon,n}(p) \\ (0, 0, \ldots, 0) & \text{if } (x_1, \ldots, x_n) \notin \mathcal{T}_{n,\epsilon}(p), \end{cases}$$

and a decoding function $D_n : \{0, 1\}^m \to \Sigma^n$ by

$$\begin{aligned} &D_n(b_1, \ldots, b_m) \\ &\quad = \begin{cases} (x_1, \ldots, x_n), & \text{if } (b_1, \ldots, b_m) = b(x_1, x_2, \ldots, x_n) \text{ for some } (x_1, \ldots, x_n) \in \mathcal{T}_{n,\epsilon}(p) \\ (f, f, \ldots, f) & \text{if } (b_1, \ldots, b_m) \neq b(x_1, x_2, \ldots, x_n) \text{ for any } (x_1, \ldots, x_n) \in \mathcal{T}_{n,\epsilon}(p), \end{cases} \end{aligned}$$

for some symbol $f \in \Sigma$ corresponding to a failure. From this construction it follows that

$$(D_n \circ E_n)(x_1, \ldots, x_n) = (x_1, \ldots, x_n),$$

whenever $(x_1, \ldots, x_n) \in \mathcal{T}_{n,\epsilon}(p)$ (and maybe in the additional case where $x_i = f$ for all $i \in \{1, \ldots, n\}$). Therefore, we conclude that the success probability equals

$$\mathrm{P}\Big((D_n \circ E_n)(X_1, \ldots, X_n) = (X_1, \ldots, X_n)\Big) \geqslant \mathrm{P}\Big((X_1, \ldots, X_n) \in \mathcal{T}_{n,\epsilon}(p)\Big) =: 1 - \delta_n.$$

By the second case of Lemma 3.6, we see that $\delta_n \to 1$ as $n \to \infty$. This finishes the proof.

Converse part. Consider a sequence of (n_k, m_k, δ_k)-compression schemes for the discrete memoryless source $(X_n)_{n\in\mathbb{N}}$ such that $\lim_{k\to\infty} n_k = \infty$ and

$$\lim_{k\to\infty} \frac{m_k}{n_k} = R < H(p).$$

For each $k \in \mathbb{N}$ let S_k denote the set of strings on which the (n_k, m_k, δ_k)-compression scheme in the sequence succeeds (see Definition 3.2). We have

$$|S_k| \leqslant 2^{m_k},$$

since encoding more than 2^{m_k} strings into a set with 2^{m_k} elements necessarily leads to a collision. Furthermore, note that for each $k \in \mathbb{N}$ we have

$$S_k \subseteq (S_k \cap \mathcal{T}_{n_k,\epsilon}(p)) \cup (\Sigma^{n_k} \setminus \mathcal{T}_{n_k,\epsilon}(p)),$$

for any $\epsilon > 0$, which implies that

$$\begin{aligned} 1 - \delta_k &= \sum_{(x_1,\ldots,x_{n_k})\in S_k} p(x_1)\cdots p(x_{n_k}) \\ &\leqslant \sum_{(x_1,\ldots,x_{n_k})\in S_k\cap\mathcal{T}_{n_k,\epsilon}(p)} p(x_1)\cdots p(x_{n_k}) + \mathrm{P}\left[(X_1, \ldots, X_{n_k}) \notin \mathcal{T}_{n_k,\epsilon}(p)\right] \\ &\leqslant 2^{-n_k(H(p)-\epsilon)}|S_k| + \mathrm{P}\left((X_1, \ldots, X_{n_k}) \notin \mathcal{T}_{n_k,\epsilon}(p)\right). \end{aligned}$$

Finally, we note that as $k \to \infty$ we have

$$2^{-n_k(H(p)-\epsilon)}|S_k| \leqslant 2^{-n_k(H(p)-\frac{m_k}{n_k}-\epsilon)} \to 0,$$

when we choose $\epsilon < H(p) - R$, and

$$\mathrm{P}\left((X_1, \ldots, X_{n_k}) \notin \mathcal{T}_{n_k,\epsilon}(p)\right) \to 0,$$

by Lemma 3.6. Therefore, the failure probability of the compression scheme satisfies $\delta_k \to 1$ as $k \to \infty$. □

4 Information transmission over noisy channels

Another basic problem of information theory is to determine the maximal rates at which information can be send reliably over noisy channels. Again, this problem was solved by Claude Shannon. In this introduction, we will only state Shannon's channel coding theorem, and we will postpone the proof until later. We start with a definition:

Definition 4.1 (Classical communication channel)**.** *Let Σ_A, Σ_B denote two alphabets. A* communication channel *is given by a function $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ mapping each symbol in Σ_A to a probability distribution over Σ_B.*

To get a concrete picture, envision a noisy telegraph line (not very up-to-date of course), where trying to send the letter "a" might result in the letters "a","b", or "c" to appear at the other end of the line with different probabilities depending on the quirks of the system. Note that in this case $\Sigma_A = \Sigma_B$. How would you send your messages over such a telegraph line? One idea might be to encode your message by adding some redundancy. To stay in the example we might just repeat every symbol five times: If we want to send the symbol "a", then we would input the string "aaaaa" into the telegraph line. Even if some error happens, and the string "caaaba" comes out the other end, the receiver could still guess that probably the symbol "a" was the intended message. This seems to work fine, but is it the best we can do? To quantify what we mean by best, we can again define the information transmission problem similarly to the compression problem from before:

Definition 4.2 (Coding schemes)**.** *For alphabets Σ_A, Σ_B let $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ denote a communication channel. An (n, M, δ)-coding scheme for information transmission over the channel N is a pair of functions*

$$E : \{1, 2, \ldots, M\} \to \Sigma_A^n \quad \textit{and} \quad D : \Sigma_B^n \to \{1, 2, \ldots, M\},$$

such that

$$\min_{i \in \{1,2,\ldots,M\}} P\Big(N^{\times n} \circ E(i) \in D^{-1}(i)\Big) \geqslant 1 - \delta. \tag{3}$$

Here, $N^{\times n}$ is the n-fold direct product of N with itself acting as

$$N^{\times n}(x_1, \ldots, x_n) = (N(x_1), \ldots, N(x_n)),$$

on $(x_1, \ldots, x_n) \in \Sigma_A^n$.

The previous definition might be a bit difficult to parse. It should be read as follows: There are two functions, the encoder E and the decoder D. The encoder E encodes a message (labelled by $1, \ldots, M$) into a string in Σ_A^n of length n. The symbols $E(i)_1, E(i)_2, \ldots$ making up the string corresponding to message i are then send successively through the communication channel leading to a product of probability distributions

$$N^{\times n} \circ E(i) = \Big(N(E(i)_1), N(E(i)_2), \ldots, N(E(i)_n)\Big),$$

which we may interpret as a probability distribution on Σ_B^n. Receiving one possible string in Σ_B^n, the receiver applies the decoding map D thereby obtaining a guess for what the message could be. In (3) the success probability is given by the probability that the string $N^{\times n} \circ E(i)$ is in the preimage of $D^{-1}(i)$. Finally, we consider the minimal probability of success over all messages to be our figure of merit. Note that we made the implicit assumption that consecutive applications of the communication channel are independent from each other leading to the product distribution in (3). This is an idealization, and there are many information theorists studying non-i.i.d. scenarios for channel coding. However, here we focus on the simplest case. As in the case for compression, we can define asymptotically achievable rates:

Definition 4.3 (Achievable rates for channel coding). *A number $R \in \mathbb{R}^+$ is called an* achievable rate *for transmitting information over the communication channel $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ on the alphabets Σ_A and Σ_B, if for every $n \in \mathbb{N}$ there exists an (n, M_n, δ_n) coding scheme such that*

$$R = \lim_{n\to\infty} \frac{\log(M_n)}{n} \quad \textit{and} \quad \lim_{n\to\infty} \delta_n = 0.$$

The following definition is central for information theory and goes back to Shannon:

Definition 4.4 (Capacity of a channel). *The* capacity $C(N)$ *of a communication channel N is the supremum of the achievable rates for transmitting information over it.*

Is it possible to compute the capacity, and does it fully characterize the achievable rates for communication? Yes, both questions where again answered by Claude Shannon. Shannon's channel coding theorem gives a formula for the capacity of a communication channel in terms of the joint probability distributions obtained from "sending" a probability distribution through the channel. We need to introduce another entropic quantity:

Definition 4.5 (Mutual information). *The* mutual information *of a joint probability distribution $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ is given by*

$$I(A : B)_{p_{AB}} = H(p_A) + H(p_B) - H(p_{AB}).$$

The mutual information is never negative (Homework), and it quantifies how close the joint distribution is to the product distribution of its marginals. You can check, that $I(A : B)_{p_{AB}} = 0$ if $p_{AB} = p_A \times p_B$. Consider a communication channel $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ and we write $N(y|x)$ for the probability of obtaining the symbol $y \in \Sigma_B$ at the output of the channel after the symbol $x \in \Sigma_A$ has been send. Note that $\sum_{y\in\Sigma_B} N(y|x) = 1$ for any $x \in \Sigma_A$. Given a probability distribution $p_A \in \mathcal{P}(\Sigma_A)$ we can now define a joint probability distribution $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ by setting

$$p^N_{AB}(x, y) = p_A(x)N(y|x). \tag{4}$$

This joint probability distribution describes the joint probability of inputs and outputs for the communication channel N and it is easy to verify that p_A is a marginal of p^N_{AB}. Finally, we can state the following:

Theorem 4.6 (Shannon's channel coding theorem). *For alphabets Σ_A and Σ_B let $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ denote a communication channel. The capacity of N is given by*

$$C(N) = \sup_{p_A\in\mathcal{P}(\Sigma_A)} I(A : B)_{p^N_{AB}},$$

and a rate R is achievable if and only if

$$R < C(N).$$

We will postpone the proof of this theorem until later, but we still want to point out one remarkable feature about its proof: It is non-constructive! Shannon's proof shows that generating coding schemes at random will almost always achieve rates very close to the capacity. It turned out to be very difficult to construct specific codes with rates close to the capacity for general communication channels. The channel coding theorem was proved in 1948, and it took until 1992 when a family of codes (called turbo codes) where invented achieving rates close to capacity. Then, it took until 2006 when a family of codes (called polar codes) was invented that provably achieved rates arbitrarily close to capacity.

5 What will be the topic of the course?

Information theory really took off after Shannon's paper "A mathematical theory of communication" in 1948 containing all the results we have seen so far. The general theory was intended to describe how any physical systems processes information. However, in the 1920s another fundamental theory took off: Quantum mechanics. From experiments with atomic and subatomic particles it became clear that classical mechanics and statistical physics does not describe nature on its smallest scales. It took until the 1960s and 1970s when physicists and some mathematicians realised that classical information theory itself does not describe how quantum mechanical systems process information. They started an effort to generalize information theory into what we call "quantum information theory" today. This course starts from the fundamentals of quantum mechanics and will give a thorough introduction to modern quantum information theory. In particular, we will answer the following questions:

- What are the fundamental limitations of quantum communication? Why can't we clone quantum states? Why can't we communicate faster than light by exploiting entanglement?
- What is the maximum amount of information that can be stored in a quantum system?
- How can quantum states be compressed?
- How can classical information and quantum information be transmitted over quantum channels?
- How is it possible to transmit information through two channels each having zero capacity?