

Solutions to Problem Set 10

discussed:
15 July 2022

Data Compression With And Without Deep Probabilistic Models

Prof. Robert Bamler, University of Tuebingen

Course materials available at <https://robamler.github.io/teaching/compress22/>

Problem 10.1: Lossy Compression With a Variational Autoencoder (VAE)

The accompanying jupyter notebook contains code for a variational autoencoder that can be used for lossy compression of handwritten digits. Most of it is already implemented, you just have to fill in a few key steps. Follow the instructions in the notebook.

Problem 10.2: Channel Capacity

In this problem, you will calculate the capacity of two toy examples of noisy channels. Then, you will come up with a simple channel coding scheme that is optimal for one of the considered channels.

In the lecture, we discussed the problem of transmitting a bit string $\mathbf{S} \in \{0, 1\}^k$ of length k over a noisy channel $P(\mathbf{Y}|\mathbf{X})$:

$$\begin{array}{ccccccc} \mathbf{S} & \xrightarrow{\text{channel encoder}} & \mathbf{X} & \xrightarrow{\text{channel}} & \mathbf{Y} & \xrightarrow{\text{channel decoder}} & \hat{\mathbf{S}} \\ \cap & P(\mathbf{X}|\mathbf{S}) & \cap & P(\mathbf{Y}|\mathbf{X}) & \cap & P(\hat{\mathbf{S}}|\mathbf{Y}) & \cap \\ \{0, 1\}^k & & \mathbb{X}^n & & \mathbb{Y}^n & & \{0, 1\}^k \end{array}$$

We defined the channel capacity of a memoryless channel $P(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^n P(Y_i|X_i)$:¹

$$C := \max_{P(X_i)} I_P(X_i; Y_i). \quad (1)$$

Here, the maximization is over all possible input distributions $P(X_i)$ (this input distribution is under our control since we are designing the channel encoder $P(\mathbf{X}|\mathbf{S})$ that outputs \mathbf{X}). The channel capacity C is an important property of a channel since it quantifies how many bits we can reliably transmit per channel invocation: the noisy channel coding theorem states that, for long bit strings \mathbf{S} , we can transmit k bits with only n channel invocations at an arbitrarily small probability of failure as long as $\frac{k}{n} < C$.

Let's now actually calculate the capacity of some example channels.

¹More precisely, the maximum ("max") in Eq. 1 should be a supremum ("sup") because the maximum might not exist. We'll favor readability over mathematical rigor here, though.

- (a) **The binary symmetric channel:** Consider a channel $P(Y_i|X_i)$ that maps a binary input $X_i \in \mathbb{X} = \{0, 1\}$ to a binary output $Y_i \in \mathbb{Y} = \{0, 1\}$. The channel flips bits with some probability $f \in [0, 1]$:

$$P(Y_i=y_i | X_i=x_i) = \begin{cases} 1-f & \text{if } y_i = x_i \\ f & \text{if } y_i \neq x_i. \end{cases} \quad (2)$$

Show that the capacity of this channel is $C = 1 - H_2(f)$ where $H_2(f) = -f \log_2 f - (1-f) \log_2 (1-f)$ is the entropy of a Bernoulli distribution with parameter f .

Hint: Write the mutual information as $I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i|X_i)$ (see Eq. 9 on Problem Set 4). To evaluate both entropies, you have to assume some input probability distribution $P(X_i)$. Consider a general input probability distribution $P(X_i=0) = 1 - \alpha$, $P(X_i=1) = \alpha$ and maximize the mutual information over α .

Solution: We express the mutual information as $I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i|X_i)$, where we immediately find $H_P(Y_i|X_i) = \sum_{x_i \in \{0,1\}} P(X_i=x_i) H_P(Y_i | X_i=x_i) = H_2(f)$ since, due to the symmetry of the channel, the conditional entropy of $P(Y_i | X_i=x_i)$ is always $H_2(f)$, independently of x_i .

To obtain $H_P(Y_i)$, we assume a generic input distribution $P(X_i)$ with $P(X_i=0) = 1 - \alpha$ and $P(X_i=1) = \alpha$ and calculate the $P(Y_i)$ by marginalizing over X_i ,

$$\begin{aligned} P(Y_i=y_i) &= \sum_{x_i \in \{0,1\}} P(X_i=x_i) P(Y_i=y_i | X_i=x_i) \\ &= \begin{cases} (1-\alpha)(1-f) + \alpha f = 1 - (\alpha + f - 2\alpha f) & \text{if } y_i = 0; \\ \alpha(1-f) + (1-\alpha)f = \alpha + f - 2\alpha f & \text{if } y_i = 1. \end{cases} \end{aligned}$$

Therefore, we obtain for the mutual information:

$$I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i|X_i) = H_2(\alpha + f - 2\alpha f) - H_2(f).$$

To calculate the channel capacity, we have to maximize $I_P(X_i; Y_i)$ over all input distributions $P(X_i)$, i.e., over all $\alpha \in [0, 1]$. Recalling that the entropy of a coin flip, $H_2(\xi)$, is maximized if the coin is unbiased (i.e., $\xi = \frac{1}{2}$), we solve $\alpha + f - 2\alpha f = \frac{1}{2}$ and for α and obtain that for $\alpha = \frac{1}{2}$ we get $H_P(Y_i) = H_2(\frac{1}{2}) = 1$ and therefore

$$C = \max_{\alpha \in [0,1]} I_P(X_i; Y_i) = 1 - H_2(f).$$

■

- (b) **The noisy parking disc:** (this is a variation of the “noisy typewriter” example from the MacKay book, see link on the course website). Consider a channel $P(Y_i|X_i)$ where both the input X_i and the output Y_i is an integer from one to twelve, i.e., $\mathbb{X} = \mathbb{Y} = \{1, 2, \dots, 12\}$. Picture these twelve numbers arranged in a circle, like they are on an analog clock or a parking disc. The sender points

to number x_i on the circle and the receiver reads off the indicated number as y_i . Unfortunately, the sender has very thick fingers, and therefore the receiver may confuse the indicated number with one of its immediate neighbors. More precisely,

$$P(Y_i = y_i | X_i = x_i) = \begin{cases} \frac{1}{3} & \text{if } y_i \in \{x_i \ominus 1, x_i, x_i \oplus 1\} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where “ \ominus ” and “ \oplus ” denote subtraction and addition that wraps around.

- (i) Show that the channel capacity is $C = 2$ bits.

Hint: write the mutual information again as $I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i | X_i)$. Why does it suffice to maximize only $H_P(Y_i)$? What is the maximum entropy $H_P(Y_i)$ of a random variable $Y_i \in \{1, \dots, 12\}$? Notice that you don't need to find the optimal input distribution $P(X_i)$ to derive the capacity C here.

Solution: We express the mutual information again as $I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i | X_i)$. Like in the case of the binary symmetric channel, the marginal entropy $H_P(Y_i | X_i = x_i)$ is independent of x_i due to the symmetry of the channel: the conditional distribution $P(Y_i | X_i = x_i)$ is always a uniform distribution over three outcomes ($x_i \ominus 1$, x_i , and $x_i \oplus 1$) and therefore has entropy $H_P(Y_i | X_i = x_i) = \log_2 3$, which means that the expected marginal entropy $H_P(Y_i | X_i) = \mathbb{E}_{x_i \sim P(X_i)} [H_P(Y_i | X_i = x_i)] = \log_2 3$. The marginal entropy of the outcome, $H_P(Y_i)$, can again be maximized (with value $H_P(Y_i) = \log_2 12$) if we find an input distribution $P(X_i)$ that makes $P(Y_i)$ a uniform distribution. Due to the symmetry of the channel, this is clearly again achieved by making $P(X_i)$ a uniform distribution. Thus, the channel capacity is

$$\begin{aligned} C &= \max_{P(X_i)} I_P(X_i; Y_i) = \max_{P(X_i)} [H_P(Y_i) - H_P(Y_i | X_i)] \\ &= \log_2 12 - \log_2 3 = \log_2 \frac{12}{3} = \log_2 4 = 2. \end{aligned}$$

■

- (ii) Show that one possible input distribution that maximizes $I_P(X_i; Y_i)$ in Eq. 1 is a uniform distribution, i.e., $P(X_i = x_i) = \frac{1}{12} \forall x_i \in \mathbb{X}$.

Solution: See symmetry argument above. For an explicit calculation that a uniform $P(X_i)$ leads to a uniform $P(Y_i)$, we'd have to marginalize $P(X_i, Y_i)$ over X_i ,

$$\begin{aligned} P(Y_i = y_i) &= \sum_{x_i=1}^{12} P(X_i = x_i) P(Y_i = y_i | X_i = x_i) = \sum_{x_i=y_i \ominus 1}^{y_i \oplus 1} \frac{1}{12} \frac{1}{3} \\ &= 3 \times \frac{1}{12} \frac{1}{3} = \frac{1}{12} \quad \forall y_i \in \{1, \dots, 12\}. \end{aligned}$$

■

- (iii) While a uniform input distribution $P(X_i) = \frac{1}{12}$ does maximize the mutual information $I_P(X_i; Y_i)$, designing a channel code that uses all possible input values $x_i \in \{1, \dots, 12\}$ is somewhat difficult in practice. Luckily, the uniform distribution is not the only input distribution that maximizes the mutual information for the noisy parking disc channel. Can you come up with some very simple channel encoder $P(\mathbf{X}|\mathbf{S})$ and channel decoder $P(\hat{\mathbf{S}}|\mathbf{Y})$ that admits perfect reconstruction of all possible inputs $\mathbf{s} \in \{0, 1\}^k$ and that allows you to transmit exactly 2 bits per channel invocation (i.e., $\frac{k}{n} = C = 2$)?

Hint: You don't need any fancy theorems here. Just think simple.

Note: By coming up with a concrete encoder/decoder pair that reaches the alleged limit of $\frac{k}{n} = C$, you prove here that the noisy channel coding theorem holds for the specific example of the noisy parking disc channel. In the next lecture, we'll prove the theorem for general channels. The proof consists of constructing *random* channel codes that will actually turn out to be quite similar to the channel code that solves the noisy parking disc problem.

Solution: We can completely eliminate any chance of confusion if we use a channel encoder that only ever outputs numbers x_i that are divisible by 3. This leaves 4 possible values for x_i (3, 6, 9, and 12), which we can use to encode two bits of information, and it partitions the output space \mathbb{Y} into 4 disjoint sets that each correspond to a unique input x_i :

$$\begin{aligned} x_i=3 &\Rightarrow y_i \in \{2, 3, 4\}; & x_i=9 &\Rightarrow y_i \in \{8, 9, 10\}; \\ x_i=6 &\Rightarrow y_i \in \{5, 6, 7\}; & x_i=12 &\Rightarrow y_i \in \{11, 12, 1\}. \end{aligned}$$

Since these four subsets of \mathbb{Y} are disjoint, the channel decoder can uniquely identify the input symbol x_i for any y_i . ■

Problem 10.3: Data Processing Inequality

In this problem, you will prove a fundamental theorem of communication systems: the data processing inequality. This inequality will become crucial for the theory of lossy compression.

Consider a sequence of random variables X_i , $i \in \{1, \dots, n\}$ that form a Markov chain $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$, as introduced in the lecture on May 12 (and discussed on Problem Set 5), i.e.,

$$P(X_1, X_2, \dots, X_n) = P(X_1) P(X_2|X_1) P(X_3|X_2) \cdots P(X_n|X_{n-1}) \quad (4)$$

(for a Markov chain $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$).

A Markov chain can model any *memoryless* process, i.e., a process where one chains together stateless stochastic operations and each stochastic operation takes as input *only* the output of the immediately preceding operation. For example, think about kids at a birthday party who play a game of telegraph (German: “Flüsterpost”).

The data processing inequality makes two statements about how information propagates through such a Markov chain:

- Information about *past* items X_i can only *decrease* but never increase along a Markov chain. More formally:

$$I_P(X_i; X_j) \geq I_P(X_i; X_k) \quad \forall i < j < k \quad (\text{for Markov chains}). \quad (5)$$

- Information about *future* items X_k can only *increase* but never decrease along a Markov chain. More formally:

$$I_P(X_i; X_k) \leq I_P(X_j; X_k) \quad \forall i < j < k \quad (\text{for Markov chains}). \quad (6)$$

The following steps guide you through the proofs of Eqs. 5 and 6.

- (a) In order to relate Eqs. 5-6 to their respective verbal statements, recall why the mutual information $I_P(X; Y)$ can be interpreted as a measure of how much information Y gives us about X and vice versa. This was discussed in Problem 4.4 (c) on Problem Set 4 and in the paragraph marked “Interpretation” below it.

Solution: See Problem 4.4 (c) on Problem Set 4. ■

- (b) We’ll first prove Eq. 6:

- (i) Recall that, if we pick three items X_i , X_j , and X_k of a Markov chain in order (i.e., $i < j < k$), then they form a Markov chain $X_i \rightarrow X_j \rightarrow X_k$ (if this is not obvious to you, then refer back to Problem 5.2 (b) (ii) on Problem Set 5). Thus, $P(X_i, X_j, X_k) = P(X_i)P(X_j|X_i)P(X_k|X_j)$.

Solution: The relation $P(X_i, X_j, X_k) = P(X_i)P(X_j|X_i)P(X_k|X_j)$ follows by marginalizing Eq. 4 over all $X_{i'}$ with $i' \notin \{i, j, k\}$, see solutions to Problem 5.2 (b) (ii) on Problem Set 5. ■

- (ii) Use $I_P(X; Y) = H_P(Y) - H_P(Y|X)$ to derive a relation of the form

$$I_P(X_j; X_k) - I_P(X_i; X_k) = \mathbb{E}_P \left[-\log_2 \left(\frac{P(X_?|X_?)}{P(X_?|X_?)} \right) \right] \quad (7)$$

where each “?” is either i , j , or k .

Solution:

$$\begin{aligned} I_P(X_j; X_k) - I_P(X_i; X_k) &= H_P(X_k) - H_P(X_k|X_j) - (H_P(X_k) - H_P(X_k|X_i)) \\ &= H_P(X_k|X_i) - H_P(X_k|X_j) \\ &= \mathbb{E}_P \left[-\log_2 P(X_k|X_i) + \log_2 P(X_k|X_j) \right] \\ &= \mathbb{E}_P \left[-\log_2 \left(\frac{P(X_k|X_i)}{P(X_k|X_j)} \right) \right]. \end{aligned}$$

■

- (iii) Use Jensen's inequality to pull the logarithm in Eq. 7 out of the expectation. Then write out the expectation as a weighted average (using the fact that $P(X_i, X_j, X_k) = P(X_i)P(X_j|X_i)P(X_k|X_j)$) and prove Eq. 6.

Solution: Using Jensen's inequality for the convex function $f(\xi) = -\log_2 \xi$, we obtain the bound

$$\begin{aligned}
I_P(X_j; X_k) - I_P(X_i; X_k) &= \\
&= \mathbb{E}_P \left[-\log_2 \left(\frac{P(X_k|X_i)}{P(X_k|X_j)} \right) \right] \\
&\geq -\log_2 \left(\mathbb{E}_P \left[\frac{P(X_k|X_i)}{P(X_k|X_j)} \right] \right) \\
&= -\log_2 \left(\sum_{x_i, x_j, x_k} P(X_i=x_i) P(X_j=x_j | X_i=x_i) P(X_k=x_k | X_j=x_j) \right. \\
&\quad \left. \times \frac{P(X_k=x_k | X_i=x_i)}{P(X_k=x_k | X_j=x_j)} \right) \\
&= -\log_2 \left(\sum_{x_i, x_j} \left[P(X_i=x_i, X_j=x_j) \underbrace{\sum_{x_k} P(X_k=x_k | X_i=x_i)}_{=1} \right] \right) \\
&= -\log_2(1) = 0.
\end{aligned}$$

Thus, $I_P(X_j; X_k) \geq I_P(X_i; X_k)$ ■

- (c) To prove Eq. 5, recall that, for three random variables X_i , X_j , and X_k , the statement “ $X_i \rightarrow X_j \rightarrow X_k$ is a Markov chain” is equivalent to the statement “ X_i and X_k are conditionally independent given X_j ”. Use the symmetry of conditional independence to argue that $X_k \rightarrow X_j \rightarrow X_i$ is also a Markov chain and therefore Eq. 5 holds.

Solution: As we proved in Problem 5.1. (a) the statement “ $X_i \rightarrow X_j \rightarrow X_k$ is a Markov chain” is equivalent to the statement “ X_i and X_k are conditionally independent given X_j ”, i.e.,

$$P(X_i, X_k|X_j) = P(X_i|X_j) P(X_k|X_j).$$

Clearly, this equation is invariant if we exchange i with k . Therefore, the statement “ $X_i \rightarrow X_j \rightarrow X_k$ is a Markov chain” is equivalent to the statement “ $X_k \rightarrow X_j \rightarrow X_i$ is a Markov chain”. ■

- (d) **What is information?** The data processing inequality can be interpreted as follows: assume we feed some input data X_1 into some (possibly nondeterministic) machine that processes the data and outputs X_2 , and we then feed X_2 (but not X_1) into some other (possibly nondeterministic) machine that outputs X_3 . Using the interpretation of the mutual information reviewed in part (a), the data processing

inequality Eq. 5 then tells us that any information about X_1 that gets destroyed by the first machine cannot be regenerated by the second machine.

Think about what this means for the interpretation of our notion of “information”. How well does our formal notion of the “information content” capture what we would colloquially consider as “information”? For example, think about a cryptographic pipeline $X_1 \rightarrow X_2 \rightarrow X_3$ where X_1 is a clear text message, X_2 is the encrypted representation of X_1 , and X_3 is the decrypted message (thus, $X_3 = X_1$). What does Eq. 5 imply about $I_P(X_1; X_2)$?

Or think about a crime scene, where the perpetrator first destroys as much evidence as they can, and the police then recover some of it. How much information about the crime do the police unveil, according to our very specific notion of information?

These considerations should be a reminder that information theory uses a very specific notion of the term “information”. Any information theoretical statement should always be considered within the context of this specific notion. Like many other technical terms in other branches of science, the term “information” is used in information theory as a *metaphor*. In particular, information theory does not take computational feasibility into account (the absence of a so-called *computational model* is one of the main differences between information theory and cryptography).

Solution: In the cryptography example, we have $X_3 = X_1$ and therefore $I_P(X_1; X_3) = H_P(X_1) + H_P(X_3) - H_P((X_1, X_3)) = H_P(X_1) + H_P(X_1) - H_P(X_1) = H_P(X_1)$. Thus, by the information processing inequality, we have $I_P(X_1; X_2) \geq I_P(X_1; X_3) = H_P(X_1)$ while at the same time $I_P(X_1; X_2) = H_P(X_1) - H_P(X_1|X_2)$. Thus, $H_P(X_1|X_2) = 0$, i.e., any eavesdropper who knows the encrypted representation X_2 has zero uncertainty about the clear text message X_1 . Clearly, this only makes sense if we assume that the eavesdropper can break the code. For public-key cryptography, this would mean that the eavesdropper has to have extremely large computational capabilities.

In the crime scene example, one could formally argue that the police do not unveil any new information because, since the crime was in the past, any information that they find out about the crime must have been there already. Clearly, this does not match our colloquial use of the term “information”. ■

- (e) **Can we exceed the channel capacity?** The precise phrasing of the channel coding theorem that we used in the lecture only states that ratios $\frac{k}{n}$ up to the channel capacity C are achievable. But is it possible to exceed the channel capacity without introducing errors? More precisely, if we transmit k bits using only n invocations of a memoryless channel with capacity C , can we somehow make the transmission error arbitrarily small for all possible input strings even if $\frac{k}{n} > C$?

Hint: This problem is still about the data processing inequality.

Solution: This is not possible. Assume we have a memoryless channel with

capacity C and a channel encoder/decoder pair for bit strings of length k that invokes the channel n times. The original bit string \mathbf{S} , channel encoded representation \mathbf{X} , channel output \mathbf{Y} , and reconstructed bit string $\hat{\mathbf{S}}$ form a Markov chain $\mathbf{S} \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{\mathbf{S}}$. Therefore, by combining both forms of the data processing inequality (Eqs. 5 and 6), we find, for all input distributions $P(\mathbf{S})$,

$$I_P(\mathbf{S}; \hat{\mathbf{S}}) \stackrel{(5)}{\leq} I_P(\mathbf{S}; \mathbf{Y}) \stackrel{(6)}{\leq} I_P(\mathbf{X}; \mathbf{Y}).$$

We can further bound the mutual information between \mathbf{X} and \mathbf{Y} by the sum of the mutual information between X_i and Y_i for each $i \in \{1, \dots, n\}$ by exploiting the assumption that the channel $P(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^n P(Y_i|X_i)$ is memory free, which implies that $H_P(\mathbf{Y}|\mathbf{X}) = \mathbb{E}_P[-\sum_{i=1}^n \log P(Y_i|X_i)] = \sum_{i=1}^n \mathbb{E}_P[-\log P(Y_i|X_i)] = \sum_{i=1}^n H_P(Y_i|X_i)$ and therefore,

$$\begin{aligned} I_P(\mathbf{X}; \mathbf{Y}) &= \underbrace{H_P(\mathbf{Y})}_{\leq \sum_{i=1}^n H_P(Y_i)} - \underbrace{H_P(\mathbf{Y}|\mathbf{X})}_{=\sum_{i=1}^n H_P(Y_i|X_i)} \\ &\leq \sum_{i=1}^n (H_P(Y_i) - H_P(Y_i|X_i)) = \sum_{i=1}^n I_P(X_i; Y_i) \leq nC. \end{aligned}$$

Thus, in total, $I_P(\mathbf{S}; \hat{\mathbf{S}}) \leq nC$ for all input distributions $P(\mathbf{S})$. We now choose $P(\mathbf{S})$ to be the uniform distribution over all bit strings of length k and assume that we can communicate all bit strings without errors. Thus, $\hat{\mathbf{S}} = \mathbf{S}$ and therefore $I_P(\mathbf{S}; \hat{\mathbf{S}}) = H_P(\mathbf{S}) = k$ for uniform $P(\mathbf{S})$. This proves that $k \leq nC$, i.e., we cannot communicate without errors above the channel capacity. \blacksquare