Solutions to Problem Set 12

discussed:
19 July 2023

Data Compression With And Without Deep Probabilistic Models
Prof. Robert Bamler, University of Tübingen

Course materials available at https://robamler.github.io/teaching/compress23/

Note: This week's tutorial will be replaced by Lucas Theis' talk. The two short problems below are meant to be discussed in small groups in the lecture before the talk.

Problem 12.1: Recovering the Lossless Limit

In the last lecture, we considered a lossless compression pipeline $\mathbf{X} \to \mathbf{S} \to \mathbf{X}'$, and we stated that the expected bit rate is bounded by the rate/distortion curve,

$$\mathbb{E}_P[\text{bit rate}] \geq \mathcal{R}(\mathcal{D}) \qquad \text{with} \qquad \mathcal{R}(\mathcal{D}) := \inf_{\substack{P(\mathbf{X}'|\mathbf{X}):\\ \mathbb{E}_P[d(\mathbf{X},\mathbf{X}')]\leq\mathcal{D}}} I_P(\mathbf{X};\mathbf{X}'). \tag{1}$$

Here, $d(\mathbf{x}, \mathbf{x}') \geq 0$ quantifies how much a reconstruction $\mathbf{x}'$ differs from the original message $\mathbf{x}$, and $\mathcal{D}$ specifies how much distortion we accept in expectation.

What do you get for $\mathcal{R}(\mathcal{D})$ in the limit of lossless compression, $\mathcal{D} = 0$, assuming that $d(\mathbf{x}, \mathbf{x}') = 0$ if and only if $\mathbf{x} = \mathbf{x}'$? Interpret your result.

Solution: From the theory of lossless compression, we expect to find $\mathcal{R}(0) = H_P(\mathbf{X})$. Indeed, for $\mathcal{D} = 0$, the infimum on the right-hand side of Eq. 1 only runs over the single mapping $P(\mathbf{X}' \,|\, \mathbf{X})$ where $\mathbf{X}' = \mathbf{X}$. For this mapping, we find

$$I_P(\mathbf{X};\mathbf{X}') = H_P(\mathbf{X}) - \underbrace{H_P(\mathbf{X} \,|\, \mathbf{X}')}_{=0 \text{ for } \mathbf{X}=\mathbf{X}'} = H_P(\mathbf{X}).$$

■

Problem 12.2: The Noisy Parking Disk

This problem is meant to provide some intuition for the optimal channel coders we constructed in the last lecture. The problem is an adaptation of the "noisy typewriter" example from the MacKay book (see link on the course website).

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

$$C := \sup_{P(X_i)} I_P(X_i; Y_i). \tag{2}$$

Consider a memoryless channel where both the inputs $X_i \in \mathcal{X}$ and the outputs $Y_i \in \mathcal{Y}$ are integers from one to twelve, i.e., $\mathcal{X} = \mathcal{Y} = \{1, 2, \ldots, 12\}$. Picture these twelve numbers

arranged in a circle, like they are on an analog clock or a parking disc. Transmitting a symbol $x_i \in \mathcal{X}$ goes as follows: the sender points to the 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 might confuse the indicated number with one of its immediate neighbors. More precisely,

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

where "$\ominus$" and "$\oplus$" denote subtraction and addition that wraps around the circle.

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

Hint: express the mutual information 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 $Y_i \in \mathcal{Y}$? Notice that you don't need to find the optimal input distribution $P(X_i)$ to derive the capacity C here.

Solution: Expressing the mutual information as $I_P(X_i;Y_i) = H_P(Y_i) - H_P(Y_i|X_i)$, we find that, for the particular channel in Eq. 3, $H_P(Y_i|X_i) = \log_2(3)$ is independent of the input distribution $P(X_i)$. Therefore, maximizing $I_P(X_i;Y_i)$ is equivalent to maximizing $H_P(Y_i)$. We obtain a maximum value of $H_P(Y_i) = \log_2(12)$ if $P(Y_i)$ is a uniform distribution, which is indeed easy to achieve, e.g., by making $P(X_i)$ uniform (see part (b) below). Therefore, we have

$$C = \sup_{P(X_i)} I_P(X_i;Y_i) = \sup_{P(X_i)} \Big[H_P(Y_i) - \log_2(3)\Big] = \log_2(12) - \log_2(3) = \log_2(4) = 2.$$

Thus, according to the channel coding theorem, it should be possible to communicate 2 bits of information per channel invocation. The channel coding theorem only guarantees that this is possible with arbitrarily small probability of error in the limit of long messages. But we'll see in part (c) below that, for this particular channel, we can in fact achieve the channel capacity with zero probability of error and for arbitrarily short (even-length) bit strings. ■

(b) Show that one possible input distribution that maximizes $I_P(X_i;Y_i)$ in Eq. 2 is a uniform distribution, i.e., $P(X_i{=}x_i) = \frac{1}{12}\ \forall x_i \in \mathcal{X}$.

Solution: In principle, there are two ways how one can calculate the $I_P(X_i;Y_i)$ for a given input distribution $P(X_i)$

(i) calculate the marginal output distribution, $P(Y_i) = \sum_{X_i} P(X_i)\, P(Y_i \,|\, X_i)$, and then calculate $I_P(X_i;Y_i) = H_P(Y_i) - H_P(Y_i \,|\, X_i)$; or

(ii) perform Bayesian inference to calculate $P(X_i \,|\, Y_i) = \frac{P(X_i)\, P(Y_i \,|\, X_i)}{\sum_{X_i} P(X_i)\, P(Y_i \,|\, X_i)}$, and then calculate $I_P(X_i;Y_i) = H_P(X_i) - H_P(X_i \,|\, Y_i)$.

In this particular example, both approaches are feasible. We'll use approach (i) here since we already noted that $H_P(Y_i \,|\, X_i) = \log_2(3)$ in part (a) above. Due to the symmetry of the channel, it is easy to see that, for uniform $P(X_i)$, also $P(Y_i)$ is uniform and thus $H_P(Y_i) = \log_2(12)$. Thus, we indeed have $I_P(X_i; Y_i) = H_P(Y_i) - H_P(Y_i \,|\, X_i) = \log_2(12) - \log_2(3) = 2 = C$. ■

(c) While a uniform input distribution $P(X_i{=}x_i) = \frac{1}{12}\ \forall x_i \in \mathcal{X}$ does maximize the mutual information $I_P(X_i; Y_i)$, designing a channel code that uses all possible input values $x_i \in \mathcal{X}$ 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 $P(\mathbf{S}' \,|\, \mathbf{Y})$ that admit perfect reconstruction of all possible inputs $\mathbf{s} \in \{0,1\}^k$, and that allow you to transmit exactly 2 bits per channel invocation?

Hint: You don't need any fancy theorems here. Just think simple: how can you avoid ambiguities on the receiver side given the specific form of the channel in Eq. 3?

Solution: The simplest solution is to partition the symbol space $\mathcal{X}$ into 4 subsets of three consecutive numbers each, e.g., $\mathcal{X} = \{1,2,3\} \cup \{4,5,6\} \cup \{7,8,9\} \cup \{10,11,12\}$. If we send the center digit of any of these subsets (i.e., 2, 5, 8, or 11) over the channel, then the channel output is guaranteed to be from the same subset, and thus the decoder can uniquely recover the input. Thus, we propose the following deterministic channel encoder and decoder for bit strings $\mathbf{s}$ of length 2: the encoder maps $\mathbf{s} \in \{0,1\}^2$ to $\mathcal{C}(\mathbf{s})$ where

$$\mathcal{C}(\text{“00”}) = 2; \quad \mathcal{C}(\text{“01”}) = 5; \quad \mathcal{C}(\text{“10”}) = 8; \quad \mathcal{C}(\text{“11”}) = 11.$$

The decoder receives $y_i \sim P(Y_i \,|\, X_i{=}\mathcal{C}(\mathbf{s}))$ and maps it to the bit string $\mathcal{C}^{-1}(y_i)$,

$$\mathcal{C}^{-1}(1) = \mathcal{C}^{-1}(2) = \mathcal{C}^{-1}(3) = \text{“00”}; \qquad \mathcal{C}^{-1}(7) = \mathcal{C}^{-1}(8) = \mathcal{C}^{-1}(9) = \text{“10”};$$
$$\mathcal{C}^{-1}(4) = \mathcal{C}^{-1}(5) = \mathcal{C}^{-1}(6) = \text{“01”}; \quad \mathcal{C}^{-1}(10) = \mathcal{C}^{-1}(11) = \mathcal{C}^{-1}(12) = \text{“11”}.$$

If we want to transmit longer bit strings then we simply use these channel coders and the channel multiple times. ■