Solutions to Problem Set 6

Data Compression With Deep Probabilistic Models
Prof. Robert Bamler, University of Tuebingen

Course material available at https://robamler.github.io/teaching/compress21/

Problem 6.1: Understanding Information Content

This problem was triggered by the results of the interactive part of the last lecture, which indicated that there seems to be a severe lack of understanding about the general concepts of lossless compression and information content among a significant part of the class. This problem invites you to rethink these concepts by applying them to the absolute simplest conceivable setup.

You should also see this problem as a blueprint of how to learn *any* new topic (not just in this course). If there's something you feel you don't quite understand yet, then I encourage you to use the same strategies that you would use if you were debugging some code that doesn't work: reduce the problem to its absolute simplest form, understand that, and then gradually build up from that.

Problem Setup. Consider a data source that generates messages x. Different from the problems so far, we don't care what these messages are—they could be sequences of symbols but they don't have to be. All we care about is that the messages come from a *finite* set $\mathbb{X}$, and that they are uniformly distributed, i.e., $P(X = x) = \frac{1}{|\mathbb{X}|}\ \forall x \in \mathbb{X}$.

(a) What is the information content of any message $x \in \mathbb{X}$?

Solution: Since the probability of all messages $x \in \mathbb{X}$ is the same, the information content is also the same for all $x \in \mathbb{X}$:

$$-\log_2 P(X = x) = -\log_2 \frac{1}{|\mathbb{X}|} = \log_2 |\mathbb{X}| \qquad \forall x \in \mathbb{X}.$$

■

(b) If you were to write out the size $|\mathbb{X}| \in \mathbb{N}$ of the set $\mathbb{X}$ in binary, how many bits would you need? Express the number of bits as a mathematical function of $|\mathbb{X}|$ and compare it to your result from part (a).

Solution: We have:

- if $|\mathbb{X}|$ is a power of 2, then its binary representation is $|\mathbb{X}| = (100\ldots0)_2$, and so it contains $\log_2 |\mathbb{X}| + 1$ bits;
- if $|\mathbb{X}|$ is not a power of 2, then the length of its binary representation is the same as for the next *lower* power of 2.

Thus, in any case, the length of the binary representation of $|\mathbb{X}|$ is $\lfloor \log_2 |\mathbb{X}| \rfloor + 1$. ■

(c) Consider the following method of mapping messages $x \in \mathbb{X}$ to a bit string: let f be an arbitrary *bijective* function from $\mathbb{X}$ onto the set $\{0, \ldots, |\mathbb{X}| - 1\}$. Then, to turn $x \in \mathbb{X}$ into a bit string, we simply write out the binary expansion of the integer $f(x)$.

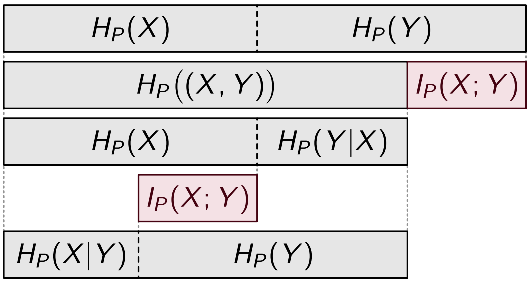
Using your result from part (b), derive an upper bound on the bit rate $R(x)\ \forall x \in \mathbb{X}$ when using this method. Then use your result from part (a) to show that this method achieves the theoretical lower bound on the expected bit rate for lossless compression (within less than one bit of overhead), i.e., that the method is optimal.

Solution: For all $x \in \mathbb{X}$, we have $x \leq |\mathbb{X}| - 1$, and thus length of the binary representation of x is $\lfloor \log_2(|\mathbb{X}| - 1) \rfloor + 1 = \lceil \log_2 |\mathbb{X}| \rceil$, i.e., the information content rounded up to the nearest integer. Thus, this method is optimal within an overhead of at most one bit (per message). ■

(d) Now forget about everything you've learned in this course so far and argue, without using any equations, why the lossless compression method from part (c) is obviously optimal.

Solution: Since all $x \in \mathbb{X}$ have the same probability, the expected bit rate is just the (equally weighted) average of the bit rates for all $x \in \mathbb{X}$. To minimize the average bit rate, we thus have to map from $\mathbb{X}$ to as short bit strings as possible. However, a lossless compression method must map to bit strings *injectively*, so the best thing one can do is to map bijectively to the from $\mathbb{X}$ to the set of $|\mathbb{X}|$ shortest bit strings. This is exactly what the above method does. ■

Problem 6.2: Mutual Information



This problem picks up and completes the discussion on Problem Set 4 by introducing the concept of the *mutual information* $I_P(X;Y)$ between two random variables X and Y. The figure on the right summarizes all relations between the joint entropy $H_P\big((X,Y)\big)$, the marginal entropies $H_P(X)$ and $H_P(Y)$, the conditional entropies $H_P(Y|X)$ and $H_P(X|Y)$, and the mutual information $I_P(X;Y)$ (red boxes).

We consider a probabilistic model P and two discrete random variables X and Y. On Problem Set 4, we derived the chain rule of the entropy (see above figure),

$$H_P\big((X,Y)\big) = H_P(X) + H_P(Y|X) = H_P(Y) + H_P(X|Y). \tag{1}$$

Eq. 1 directly corresponds to how compression with autoregressive models works (see Problem 3.2 on Problem Set 3). To encode both X and Y with an autoregressive compressor, one first encodes X into $H_P(X)$ bits (in expectation) and transmit the encoded bit string. Then, one exploits the fact that the receiver already knows X and one encodes Y into only $H_P(Y|X)$ bits in expectation. According to Eq. 1, this procedure encodes both X and Y into $H_P\big((X,Y)\big)$ bits in total (in expectation), i.e., it reaches optimal compression performance (apart from any overhead due to a suboptimal coding algorithm, e.g., if restricting oneself to a symbol code).

We now ask ourselves the question: how many bits did we save by exploiting the fact that the receiver already knew X at the point when we encoded Y? Or, equivalently: how many additional bits would we have had to transmit had we treated X and Y separately and encoded them into $H_P(X)$ and $H_P(Y)$ bits (in expectation), respectively (see first two rows in the above figure)?

The overhead of treating X and Y separately instead of encoding the tuple (X,Y) jointly is called the *mutual information* $I_P(X;Y)$,

$$I_P(X;Y) := H_P(X) + H_P(Y) - H_P\big((X,Y)\big). \tag{2}$$

Remark: You may assume in this problem that all random variables are discrete. However, it is interesting to note that, unlike the problems on Problem Set 4, the derivations in this problem actually generalize to the continuous case when we interpret H as the *differential* entropy, which we will define later in the course. While an individual differential entropy cannot be interpreted as a number of bits, *differences* between differential entropies can.

(a) **Symmetry of the Mutal Information:** Convince yourself that the mutual information is symmetric, i.e., $I_P(X;Y) = I_P(Y;X)$.

Solution: Follows directly from Eq. 2. ■

(b) **Interpretation of the Mutual Information:** Show that the mutual information can also be expressed as follows (see also last three rows in above figure),

$$\begin{aligned} I_P(X;Y) &= H_P(X) - H_P(X|Y) \\ &= H_P(Y) - H_P(Y|X). \end{aligned} \tag{3}$$

It is important to understand how these realtions can be interpreted: the interpretation of the first line in Eq. 3 is that $I_P(X;Y)$ quantifies *how much we learn about X if someone tells us the value of Y* (i.e., how much knowing Y reduces the entropy of X, in expectation). The second line in Eq. 3 expresses the same with the roles of X and Y swapped.

Solution: See Problem 4.2 (b) on Problem Set 4, where we derived the chain rule of entropy, $H_P\big((X,Y)\big) = H_P(X) + H_P(Y|X) = H_P(Y) + H_P(X|Y)$. Inserting this into Eq. 2 leads to Eq. 3. ■

(c) **Nonnegativity of the Mutual Information:** Convince yourself that the mutual information can be expressed as follows,

$$I_P(X;Y) = \mathbb{E}_P\left[\log\frac{P(X,Y)}{P(X)P(Y)}\right]. \tag{4}$$

Then show that $I_P(X;Y) \geq 0$.

Hint: the proof is essentially identical to the proof that $D_{\mathrm{KL}} \geq 0$, see Problem 3.1 (c) of Problem Set 3. In fact, $I_P(X;Y) = D_{\mathrm{KL}}\big(P(X,Y)\,\big|\big|\,P(X)P(Y)\big)$.

Solution: Eq. 4 follows directly from the definition of the mutual information in Eq. 2. We can then use Jensen's inequality (see Problem 3.1 (c) on Problem Set 3) for the convex function $f(\xi) = -\log\xi$,

$$\begin{aligned} I_P(X;Y) &= \mathbb{E}_P\left[-\log\frac{P(X)P(Y)}{P(X,Y)}\right] \geq -\log\left(E_P\left[\frac{P(X)P(Y)}{P(X,Y)}\right]\right) \\ &= -\log\left(\sum_{x,y} P(X=x)\,P(Y=y)\right) = -\log 1 = 0 \end{aligned}$$

■

(d) **Chain Rule of Mutual Information**: Now consider three random variables X, Y, and Z. Show that

$$I_P(X;Y,Z) = I_P(X;Y) + I_P(X;Z\,|\,Y). \tag{5}$$

Here, the notation $I_P(X;Y,Z)$ denotes the mutual information between X and the tuple (Y,Z) (i.e., we could write it more explicitly as $I_P\big(X;(Y,Z)\big)$). On the right-hand side, the conditional mutual information $I_P(X;Z\,|\,Y)$ is understood to condition *everything* on Y (and then averaged over it), i.e.,

$$I_P(X;Z\,|\,Y) := H_P(X|Y) + H_P(Z|Y) - H_P\big((X,Z)\,\big|\,Y\big). \tag{6}$$

Hint: write $I_P(X;Y,Z)$ in the form of Eq. 4, then use properties of the logarithm.

Solution: We start from

$$I_P(X;Y,Z) = \mathbb{E}_P\left[\log\frac{P(X,Y,Z)}{P(X)\,P(Y,Z)}\right].$$

Inserting the relations $P(X,Y,Z) = P(Y)P(X,Z|Y)$ and $P(Y,Z) = P(Y)P(X|Y)$, and multiplying both the enumerator and the denominator with $P(X,Y)$ leads to

$$\begin{aligned} I_P(X;Y,Z) &= \mathbb{E}_P\left[\log\frac{P(Y)\,P(X,Z|Y)\,P(X,Y)}{P(X)\,P(Y)\,P(Z|Y)\,P(X,Y)}\right] \\ &= \mathbb{E}_P\left[\log\frac{P(X,Y)}{P(X)\,P(Y)} + \log\frac{P(Y)\,P(X,Z|Y)}{P(Z|Y)\,P(X,Y)}\right] \\ &= \mathbb{E}_P\left[\log\frac{P(X,Y)}{P(X)\,P(Y)}\right] + \mathbb{E}_P\left[\log\frac{P(X,Z|Y)}{P(Z|Y)\,P(X|Y)}\right] \\ &= I_P(X;Y) + I_P(X;Z|Y). \end{aligned}$$

■

Problem 6.3: Conditional Independence and Data Processing Inequality

In this problem, you will prove a fundamental theorem of communication systems: the data processing inequality. Consider three random variables X, Y, and Z. In the general case, one can use the chain rule of probability theory to factorize $P(X, Y, Z)$ as follows,

$$P(X, Y, Z) = P(X)\, P(Y|X)\, P(Z|X, Y). \tag{7}$$

Eq. 7 is a general statement about any probability distribution P because it follows directly from the definition of a conditional probability. Thus, the factorization on the right-hand side of Eq. 7 can capture arbitrary dependencies between X, Y, and Z. For example, if you were to use this factorization in a compression method that encodes first X, then Y, and finally Z, then the encoded values of *both* X and Y will have an influence on the probabilistic model $P(Z|X, Y)$ that you will use to encode Z.

In many practical situations, dependencies between random variables are more constrained. Imagine, for example, a game of telegraph (German: "Flüsterpost") with three players, Alice, Bob, and Charley. Alice thinks of a word X and whispers it into Bob's ear. Bob comprehends a word Y, which may or may not be different from X depending on how well Alice and Bob communicated. Bob then whispers Y into Charley's ear, who ultimately comprehends a word Z, which he says out lout.

Surely the final output Z depends on the initial input X, but the dependency is only indirect through Y. If Bob tells us the intermediate word Y then we can make some probabilistic prediction about Z, and it would be irrelevant for this prediction whether or not we *also* knew X. In such a situation, we say that X and Z are *conditionally independent given* Y. More formally,

$$X \text{ and } Z \text{ are conditionally independent given } Y \quad :\Leftrightarrow \quad P(Z|X, Y) = P(Z|Y). \tag{8}$$

If we assume conditional independence between X and Z given Y, then the factorization in Eq. 7 simplifies,

$$P(X, Y, Z) = P(X)\, P(Y|X)\, P(Z|Y) \qquad (\text{if } X,\ Z \text{ cond. indep. given } Y). \tag{9}$$

Notice that, on the right-hand side of Eq. 9, each random variable in the sequence X, Y, Z is conditioned only on its immediate predecessor but not on any others. We say that X, Y, and Z form a *Markov chain* $X \to Y \to Z$. As an example of a Markov chain (apart from the game of telegraph), consider your WiFi connection followed by your telephone line; or consider X, Y, and Z being three different layers in a deep neural network, assuming that the network has no skip connections (but it may have stochastic connections such as dropout).

(a) **Symmetry of Conditional Independence:** Show that an equivalent way of characterizing conditional independence is as follows,

$$X \text{ and } Z \text{ are cond. indep. given } Y \quad \Leftrightarrow \quad P(X, Z \,|\, Y) = P(X|Y)\, P(Z|Y). \tag{10}$$

This formulation highlights the analogy to regular (i.e., unconditional) statistical independence (which we would have if $P(X, Z) = P(X)P(Z)$). It also makes it obvious that conditional independence is symmetric: X and Z are conditionally independent given Y iff Z and X are conditionally independent given Y. Therefore, we can also characterize conditional independence by swapping X and Y in Eq. 8:

$$X \text{ and } Z \text{ are cond. indep. given } Y \qquad \Leftrightarrow \qquad P(X \,|\, Y, Z) = P(X|Y). \tag{11}$$

Solution: We want to show that $P(Z|X, Y) = P(Z|Y) \Leftrightarrow P(X, Z|Y) = P(X|Y)\, P(Z|Y)$. We will show the two directions of the inequality separately:

- "⇒": Assuming $P(Z|X, Y) = P(Z|Y)$, we have

$$P(X, Z|Y) = \frac{P(X, Y)\, P(Z|X, Y)}{P(Y)} = \frac{P(X, Y)\, P(Z|Y)}{P(Y)} = P(X|Y)\, P(Z|Y).$$

- "⇐": Assuming $P(X, Z|Y) = P(X|Y)\, P(Z|Y)$, we have

$$P(Z|X, Y) = \frac{P(Y)\, P(X, Z|Y)}{P(X, Y)} = \frac{P(Y)\, P(X|Y)\, P(Z|Y)}{P(X, Y)} = P(Z|Y).$$

■

(b) **Data Processing Inequality:** Assume that X, Y, and Z form a Markov Chain $X \to Y \to Z$ (i.e., X and Z are conditionally independent given Y). Show that they satisfy the important data processing inequality,

$$I_P(X; Y) \geq I_P(X; Z) \qquad \text{(for a Markov chain } X \to Y \to Z\text{)}. \tag{12}$$

In other words, Y reveals at least as much about X as Z does.

Hint: Consider the difference, $I_P(X; Y) - I_P(X; Z)$, and write out both information contents in the form of Eq. 3. Then convince yourself that $H_P(X|Y) = H_P(X \,|\, Y, Z)$ for a Markov Chain $X \to Y \to Z$, and use this to show that

$$I_P(X; Y) - I_P(X; Z) = I_P(X; Y \,|\, Z) \geq 0 \tag{13}$$

where the proof that $I_P(X; Y \,|\, Z) \geq 0$ is analogous to Problem 6.2 (c).

Solution: Since X and Z are conditionally independent given Y, we have $P(X|Y) = P(X|Y,Z)$ and thus $H_P(X|Y) = \mathbb{E}[-\log P(X|Y)] = \mathbb{E}[-\log P(X|Y,Z)] = H_P(X|Y,Z)$. Thus, we find:

$$\begin{aligned} I_P(X;Y) - I_P(X;Z) &= H_P(X) - H_P(X|Y) - \big(H_P(X) - H_P(X|Z)\big) \\ &= H_P(X|Z) - H_P(X|Y) \\ &= H_P(X|Z) - H_P(X|Y,Z) \\ &= I_P(X;Y\,|\,Z) \geq 0. \end{aligned}$$

■

(c) **Data Processing Inequality, Alternative Form:** Use the symmetry of the mutual information and of conditional independence to derive from Eq. 12 that, for any Markov chain $X \to Y \to Z$,

$$I_P(Y;Z) \geq I_P(X;Z) \qquad \text{(for a Markov chain } X \to Y \to Z\text{).} \tag{14}$$

In other words, Y reveals at least as much about Z as X does.

Solution: The important part here is to think about and understand the provided interpretation of Eq. 14. The proof is simple: due to the symmetry of conditional independence, we can exchange X and Z in Eq. 12, which leads to

$$I_P(Z;Y) \geq I_P(Z;X) \qquad \text{(for a Markov chain } X \to Y \to Z\text{).}$$

We can then use the the symmetry of the mutual information to arrive at Eq. 14. ■

(d) **What is Information?** The data processing inequality can be interpreted as follows. Assume we feed some input data X into some (possibly nondeterministic) machine that processes the data and outputs Y, and we then feed Y (but not X) into some other (possibly nondeterministic) machine that outputs Z. Using the interpretation of the mutual information from Problem 6.2 (b), the data processing inequality Eq. 12 then tells us that the second machine, which processes Y to Z, can only *destroy* any information that's left about X, it cannot (re-)generate any information about X.

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 "information"?

For example, think about a cryptographic setup where X is a clear text message, Y is the encrypted representation of X, and Z is the decrypted message (thus, $Z = X$). What does Eq. 12 imply in this case about $I_P(X;Y)$? Or think about a crime scene, where the perpetrator first destroys as much evidence as they can, and the police then try to recover 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 of “information”, which can sometimes be misleading. 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 cryptographic setup, we have $X = Z$, which means $H_P(X) = H_P(Z) = H_P\big((X, Z)\big)$ and thus

$$I_P(X; Z) = H_P(X) + H_P(Z) - H_P\big((X, Z)\big) = H_P(X).$$

The data processing inequality $I_P(X; Y) \geq I_P(X; Z)$ then tells us that $I_P(X; Y) \geq H_P(X)$, and therefore $I_P(X; Y) = H_P(X)$ since the mutual information cannot be larger than either of the marginal entropies (since $H_P(X) - I_P(X; Y) = H_P(X|Y)$, which is nonnegative for discrete random variables). Thus, we conclude that $I_P(X; Z) = I_P(X; Y)$.

According to our interpretation from Problem 6.2 (b), this would mean that the encrypted representation Y of the message tells us just as much about the clear text X as the reconstruction Z. This may be true in a very strict sense (we may *in theory* be able to crack the encryption using brute-force) but it is certainly not a very useful statement for practical purposes. Thus, a purely information theoretical notion of “information” is not sufficient for a meaningful discussion of cryptography. Cryptography requires some additional assumption, either a computational model or a model of a secure communication channel.

A similar statement could be made about the crime scene: in a strictly information theoretical sense, the process of evidence gathering does not generate any new information since all the information must have been already there to begin with. Yet, I wouldn’t recommend telling a police person that you think they don’t do anything. ■