Solutions to Problem Set 4

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

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

Notes:

- This problem set is deliberately kept short with no coding problem because we'll use most of the time in the tutorial on May 17 for a final matchmaking round for the group projects.
- Problem 4.2 introduces the concepts of joint and conditional entropies. You will derive important and non-obvious relations between these concepts that will be instrumental throughout the rest of the course. These relations will guide both the choice of machine-learning models as well as the design of coding algorithms, and they will be critical for the theories of lossy compression and noisy communication.

Problem 4.1: Joint and Conditional Information Content

On the last problem set, we analyzed the bit rate of a lossless compression algorithm that is optimal w.r.t. some probabilistic model p_{model} of the data source. We saw that, up to usually negligible rounding effects, the bit rate $R(\mathbf{x})$ for any message $\mathbf{x}$ under such a code is given by the *information content* of $\mathbf{x}$ under the model, i.e., $R(\mathbf{x}) \simeq -\log p_{\text{model}}(\mathbf{x})$. Thus, we can understand the information content as the number of bits that are essential to a message, regardless of the representation in which the message is actually stored or transmitted.

Now that we have a more formal notion of probability theory, we can analyze how each symbol in the message contributes to the information content. In the general case, it is best to regard the symbols in a message as *random variables* $X_1, X_2, \ldots, X_k$. But for this problem, we'll just look at two random variables X and Y. The generalization to more than two random variables is straight forward. In this problem, we assume that X and Y are both *discrete* random variables, since the information content is not defined for continuous random variables. We denote the probabilistic model simply as P (dropping the subsript "model").

(a) **Joint Information Content:** The joint information content for X and Y is the information content for the tuple (X, Y), which can be considered as a random variable in itself $(\omega \mapsto (X(\omega), Y(\omega)))$. Thus, for some specific values x and y, the information content is

$$\begin{aligned} -\log P\big((X,Y) = (x,y)\big) &= -\log P(X = x, Y = y) \\ &= -\log P\big(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\}\big). \end{aligned} \tag{1}$$

Show that the joint information content of $(X, Y) = (x, y)$ is not smaller than the information content of $X = x$ alone and not smaller than the information content of $Y = y$ alone (the information content of $X = x$ is $-\log P(X = x) = -\log P\big(\{\omega \in \Omega : X(\omega) = x\}\big)$). Thus, adding an additional symbol can't decrease the information content.

Solution: It suffices to show that the joint information content of $(X, Y) = (x, y)$ is not smaller than the information content of $X = x$. The comparison to the information content of $Y = y$ follows by symmetry if we swap the names for X and Y. Thus, we have to show that $-\log P\big((X, Y) = (x, y)\big) \geq -\log P(X = x)$. Since the logarithm is a strictly monotonically increasing function, this is equivalent to showing that $P\big((X, Y) = (x, y)\big) \leq P(X = x)$. We have

$$\begin{aligned}
P(X = x) &= P\big(\{\omega \in \Omega : X(\omega) = x\}\big) \\
&= P\big(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\} \cup \{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) \neq y\}\big) \\
&= P\big(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\big) + P\big(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) \neq y\}\big) \\
&= P\big((X, Y) = (x, y)\big) + \underbrace{P(X = x, Y \neq y)}_{\geq 0} \geq P\big((X, Y) = (x, y)\big)
\end{aligned}$$

where, in the step from the second to the third line, we used the fact that the union on the second line is between disjoint sets and the axiom that the probability of the union of disjoint events is the sum of the probabilities of the events. ■

(b) **Marginal and Conditional Information Content:** The information content of $X = x$ alone, $-\log P(X = x)$, is also called the *marginal* information content. We further define the *conditional* information content of $Y = y$ given $X = x$ as $-\log P(Y = y \,|\, X = x)$. Using the definition of the conditional probability from the lecture, $P(Y = y \,|\, X = x) := P(X = x, Y = y)/P(X = x)$, derive the chain rule of information content, which states that the joint information content of $(X, Y) = (x, y)$ is the sum of the marginal information content of $X = x$ and the conditional information content of $Y = y$ given $X = x$. What does this mean for the bit rate of a compression method that uses an autoregressive model?

Solution: The chain rule follows immediately from the definition of the conditional probability and the properties of the logarithm:

$$\begin{aligned}
P(Y = y \,|\, X = x) &= \frac{P(X = x, Y = y)}{P(X = x)} \\
\Rightarrow P(X = x, Y = y) &= P(X = x)\, P(Y = y \,|\, X = x) \\
\Rightarrow -\log P(X = x, Y = y) &= \big(-\log P(X = x)\big) + \big(-\log P(Y = y \,|\, X = x)\big)
\end{aligned}$$

This means that, when we have an autoregressive model of some data source, we can compress messages from that data source by compressing one symbol at a time. Adding up the optimal bit rates for each symbol leads to the optimal bit rate of the message, with no overhead. (Note however, that this assumes that the bit rate

for each symbol is really just the information content, which is not the case in a symbol code because symbol codes are restricted to integer bit rates per symbol. We will learn about compression methods that effectively admit non-integer bit rates per symbol later in the course.) ■

(c) **Nonadditivity of Marginal Information Content:** In general, the joint information content of $(X, Y) = (x, y)$ is *not* the sum of the two marginal information contents of $X = x$ and $Y = y$. In fact, show (by providing a simple example for both cases) that the sum of the two marginal information contents can be both lower and higher than the joint information content.

Solution: Consider two binary random variables X and Y whose probability distribution is given in the following table (the center 2×2 block of the table shows the joint probabilities $P(X = x, Y = y)$ while the last row and column show the marginal probabilities $P(X = x)$ and $P(Y = y)$, respectively):

$P(X = x, Y = y)$	$\downarrow x = 0 \downarrow$	$\downarrow x = 1 \downarrow$	$\downarrow P(Y = y) \downarrow$
$y = 0 \rightarrow$	0.49	0.01	0.5
$y = 1 \rightarrow$	0.01	0.49	0.5
$P(X = x) \rightarrow$	0.5	0.5	

The marginal information content of both $X = x$ and $Y = y$ is one bit for all $x, y \in \{0, 1\}$ because all marginal probabilities are $P(X = x) = P(Y = y) = \frac{1}{2}$. Thus, the sum of the two marginal information contents is always

$$-\log_2 P(X = x) - \log_2 P(Y = y) = 2 \text{ bit} \qquad \forall x, y \in \{0, 1\}.$$

However, the joint information content can be both lower and higher than 2 bit. For $x = y$, the joint probability $P(X = x, Y = y) = 0.49$ is just slightly below $\frac{1}{2}$, and thus the joint information content is just slightly above one bit ($-\log_2 0.49 \approx 1.03$ bit < 2 bit). By contrast, for $x \neq y$, the joint probability $P(X = x, Y = y) = 0.01$ is very low, and thus the joint information content is much higher than 2 bit ($-\log_2 0.01 \approx 6.64$ bit > 2 bit). ■

(d) **Information Content for Statistically Independent Random Variables:** Now assume that X and Y are statistically independent, i.e., $P(X, Y) = P(X)P(Y)$ (which is shorthand for $P(X = x, Y = y) = P(X = x)P(Y = y)\ \forall x, y$). How does the statement from part (c) change in this case.

Solution: For statistically independent random variables, the joint information content is the sum of the marginal information contents:

$$\begin{aligned} & P(X, Y) = P(X)\, P(Y) \qquad \text{(for } X, Y \text{ statistically independent)} \\ \Rightarrow\ & -\log P(X, Y) = -\log P(X) - \log P(Y). \end{aligned}$$

■

Problem 4.2: Joint and Conditional Entropy

As we defined in the lecture, the entropy $H_P(X)$ of a random variable X is its *expected information content*, i.e., $H_P(X) = \mathbb{E}_P[-\log P(X)]$. Similar to Problem 4.1, let's now understand how entropies of two random variables X and Y interact. We will again assume that X and Y are *discrete* random variables since the entropy is not defined for continuous random variables (only a *differential* entropy is defined for these).

(a) **Joint Entropy:** The joint entropy of X and Y is the entropy of the tuple (X, Y). We will explicitly denote it as $H_P\big((X,Y)\big)$ (with double braces) to highlight the distinction from the cross entropy.[1] Show, by applying what you've shown in Problem 4.1 (a), that $H_P\big((X,Y)\big) \geq H_P(X)$ and that $H_P\big((X,Y)\big) \geq H_P(Y)$.

Solution: The entropy is the expected information content, and the act of taking an expectation value preserves semi-inequalities like "$\geq$". Thus, since the joint information content is not smaller than either one of the marginal information contents, the joint entropy is not smaller than either of the marginal entropies. ■

(b) **Marginal and Conditional Entropy:** The entropy of X alone, $H_P(X)$, is also called the *marginal* entropy. We further define two kinds of conditional entropies:

- $H_P(Y|X=x)$ denotes the conditional entropy of Y if we know that X takes a specific value x. In other words, $H_P(Y|X=x)$ is the entropy of the distribution $P(Y|X=x)$, interpreted as a distribution over values of Y. It is thus given by

$$\begin{aligned} H_P(Y|X=x) &= \mathbb{E}_P\big[-\log P(Y|X=x) \,\big|\, X=x\big] \\ &= -\sum_y P(Y=y \,|\, X=x)\ \log P(Y=y \,|\, X=x) \end{aligned} \tag{2}$$

Show (by providing an example for both cases) that $H_P(Y|X=x)$ can be both larger and smaller than $H_P(Y)$.

Solution: Consider two binary random variables X and Y with the following joint and marginal distributions:

$P(X=x, Y=y)$	$\downarrow x=0 \downarrow$	$\downarrow x=1 \downarrow$	$\downarrow P(Y=y) \downarrow$
$y=0 \rightarrow$	$1/4$	$3/8$	$5/8$
$y=1 \rightarrow$	$1/4$	$1/8$	$3/8$
$P(X=x) \rightarrow$	$1/2$	$1/2$	

We can calculate the marginal entropy of Y by looking at the last column, and we obtain $H_P(Y) \approx 0.95$ bit. Further, by normalizing the colums in the center 2×2 block, we obtain the following conditional probabilities $P(Y = y \,|\, X = x)$:

[1] This is not really standard notation, the literature is inconsistent in the notation here; you may find $H(X,Y)$ to refer to either the cross entropy or the joint entropy in the literature.

$P(Y=y \mid X=x)$	$\downarrow x=0\downarrow$	$\downarrow x=1\downarrow$
$y=0\rightarrow$	$1/2$	$3/4$
$y=1\rightarrow$	$1/2$	$1/4$

Therefore, we have $H_P(Y \mid X=0) = 1$ bit $> H_P(Y)$, and $H_P(Y \mid X=1) \approx 0.81$ bit $< H_P(Y)$. ■

- $H_P(Y|X)$ denotes the expectation value of $H_P(Y|X=x)$, where the expectation is taken over x. Thus,

$$\begin{aligned} H_P(Y|X) &= \sum_x P(X=x)\, H_P(Y|X=x) \qquad (3)\\ &= -\sum_x P(X=x) \left[\sum_y P(Y=y \mid X=x)\, \log P(Y=y \mid X=x) \right] \\ &= -\sum_{x,y} P(X=x, Y=y)\, \log P(Y=y \mid X=x) \end{aligned}$$

Derive the chain rule of the entropy:

$$H_P\big((X,Y)\big) = H_P(X) + H_P(Y|X) = H_P(Y) + H_P(X|Y). \qquad (4)$$

We will show on the next problem set that $H_P(Y|X) \leq H_P(Y)$ (the gap between the two is called the "mutual information between X and Y"). Therefore, while conditioning on a *specific* $X=x$ may increase the conditional entropy $H_P(Y|X=x)$ (see above), *in expectation*, conditioning can only decrease the entropy (or keep it unchanged at worst).

Solution:

$$\begin{aligned} H_P(X) + H_P(Y|X) &= \mathbb{E}_P\big[-\log P(X)\big] + \mathbb{E}_P\big[-\log P(Y|X)\big] \\ &= \mathbb{E}_P\big[-\log P(X) - \log P(Y|X)\big] \\ &= \mathbb{E}_P\big[-\log P(X,Y)\big] \\ &= H_P\big((X,Y)\big) \end{aligned}$$

The second equality follows from symmetry by swapping the names of X and Y. ■

(c) **Entropy of Statistically Independent Random Variables:** What is the joint entropy $H_P\big((X,Y)\big)$ and the marginal entropy $H_P(Y|X=x)$ and $H_P(Y|X)$ if the two random variables X and Y are statistically independent, i.e., if $P(X,Y) = P(X)P(Y)$?

Solution: For statistically independent random variables, the conditional probability is equal to the marginal probability:

$$P(Y|X) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)\,P(Y)}{P(Y)} = P(X). \qquad \text{(for } X,\, Y \text{ stat. indep.)}$$

Therefore, we have $H_P(Y|X = x) = H_P(Y|X) = H_P(Y)$. By inserting this into Eq. 4, we find $H_P\big((X, Y)\big) = H_P(X) + H_P(Y)$, i.e., for statistically independent variables, the entropy is additive. ■