Problem Set 4

published: 12 May 2022
discussion: 20 May 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/

How to Use This Problem Set

This problem set discusses several important information theoretical concepts: conditional information content, conditional, joint, and marginal entropies, and mutual information. While the definition of each individual concept may seem simple, some of their properties that you will prove on this problem set are surprisingly subtle.

You should use this problem set now as an opportunity to recap and expand on the content of the lecture; later, you'll be able to refer back to this problem set as a self-contained reference sheet of important information theoretical relations.

All Problems on this Problem set are designed so that **each question can be answered with either a one-sentence argument or a single line of calculations**. The only exceptions are the two questions marked with an asterisk ("*"), which each require you to come up with a simple example probability distribution.

Problem 4.1: Statistical Independence

In the lecture, we formalized a probabilistic model of our Simplified Game of Monopoly (which consists of throwing two fair three-sided dice—one red die and one blue die—and then recording their sum). For completeness, here's the model:

- sample space: $\Omega = \big\{(a, b) \quad \text{where} \quad a, b \in \{1, 2, 3\}\big\}$
- sigma algebra: $\Sigma = 2^\Omega := \big\{\text{all subsets of } \Omega \text{ (including } \emptyset \text{ and } \Omega)\big\}$
- probability measure P: for all $E \in \Sigma$, let $P(E) := |E|/|\Omega| = |E|/9$

We further defined three random variables, i.e., functions from Ω to $\mathbb{R}$:

- total value: $X_{\text{sum}}\big((a, b)\big) = a + b$
- value of the red die: $X_{\text{red}}\big((a, b)\big) = a$
- value of the blue die: $X_{\text{blue}}\big((a, b)\big) = b$

Now, verify the following claims from the lecture:

(a) Convince yourself that P is a valid probability measure (i.e., $P(\Omega) = 1$, $P(\emptyset) = 0$, and P satisfies countable additivity).

(b) Show that X_{red} and X_{blue} are statistically independent.

(c) Show that X_{red} and X_{sum} are *not* statistically independent.

Problem 4.2: Joint and Conditional Information Content

In the lecture, we identified the quantity "$-\log_2 P(X=x)$" as the information content of the statement "$X = x$" (meaning "the random variable X has value x") under a probabilistic model P. As you've shown in Problem 2.4 on the last problem set, the information content *of a long message* essentially measures (up to tiny corrections) the total bit rate of the message assuming that one uses a lossless code that is optimal for the model P. In this problem, you'll see in which sense precisely the information content *of an individual symbol* can or cannot be interpreted as the individual symbol's contribution to this total bitrate.

For this problem, we'll just look at *two* random variables X and Y. The generalization to more than two random variables is analogous. We further assume that X and Y are both *discrete* random variables since we didn't define information content for continuous random variables.

(a) **Joint Information Content:** The *joint information content* of the statement "$X = x$ and $Y = y$" or, equivalently, the information content of the statement "$(X, Y) = (x, y)$", is

$$\begin{aligned} -\log_2 P\big((X,Y)=(x,y)\big) &= -\log_2 P(X=x, Y=y) \\ &= -\log_2 P\big(\{\omega \in \Omega : X(\omega) = x \ \wedge \ Y(\omega) = y\}\big). \end{aligned} \quad (1)$$

Argue why 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 (*hint:* use the fact that the information content of "$X = x$" is $-\log_2 P(X=x) = -\log_2 P\big(\{\omega \in \Omega : X(\omega) = x\}\big)$ and identify a superset-subset relationship).

(b) **Marginal and Conditional Information Content:** The information content of "$X = x$" alone, $-\log_2 P(X = x)$, is also called *marginal* information content. We further define the *conditional* information content of "$Y = y$" *given* $X = x$ as $-\log_2 P(Y = y \,|\, X = x)$. Using the definition of 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 $X = x$.

Interpret this finding in words: if you want to compress the two symbols x and y in an optimal way, and you want to encode one after the other, what probabilistic model should you use for encoding x and for encoding y, respectively.

(c*) **Nonadditivity of Marginal Information Content:** In Problem 2.3 (b) of the last problem set, you showed (although using different notation) that the joint information content of "$(X, Y) = (x, y)$" is the sum of the two marginal information

contents of "$X=x$" and "$Y=y$" *if* X *and* Y *are statistically independent*. However, this statement is not necessarily true if X and Y are *not* statistically independent.

Provide examples of simple probabilistic models

(i) where the sum of the two marginal information contents of "$X = x$" and "$Y = y$" for some x and y is *larger* than the joint information content of "$(X, Y) = (x, y)$"; and

(ii) where the sum of the two marginal information contents of "$X = x$" and "$Y = y$" for some x and y is *smaller* than the joint information content of "$(X, Y) = (x, y)$".

Using your result from part (b), relate the marginal information content of "$Y=y$" and the conditional information content of "$Y = y$" given $X = x$ to each other for both cases (i) and (ii). Does conditioning on $X = x$ increase or reduce the information content in each of the two cases?

Note: You will see below that one of these cases (i) and (ii) can be regarded as the "typical" case whereas the other one is somewhat of an exception. Using your intuition about information content, can you guess which case is the typical one?

Problem 4.3: Joint and Conditional Entropy

In the lecture, we defined the entropy $H_P(X)$ of a random variable X as its expected information content, i.e., $H_P(X) = \mathbb{E}_P[-\log_2 P(X)]$. Similar to Problem 4.2, 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 entropy is not defined for continuous random variables (only a so-called differential entropy is defined for these).

(a) **Joint Entropy:** The joint entropy of X and Y is simply the entropy of the tuple (X, Y) (interpreted as a random variable that maps $\omega \mapsto \big(X(\omega), Y(\omega)\big)$). We will explicitly denote the joint entroy as $H_P\big((X, Y)\big)$ (with double braces) to highlight the distinction from the cross entropy.[1] Argue, by applying what you've shown in Problem 3.3 (a), that $H_P\big((X, Y)\big) \geq H_P(X)$ and that $H_P\big((X, Y)\big) \geq H_P(Y)$.

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:

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

$$\begin{aligned} H_P(Y \mid X=x) &= \mathbb{E}_{P(Y|X=x)}\left[-\log_2 P(Y \mid X=x)\right] \\ &= -\sum_y P(Y=y \mid X=x)\ \log_2 P(Y=y \mid X=x). \end{aligned} \tag{2}$$

[1] This is not really standard notation. In the literature, you may find the notation "$H(X, Y)$" used for either the cross entropy or the joint entropy, depending on context.

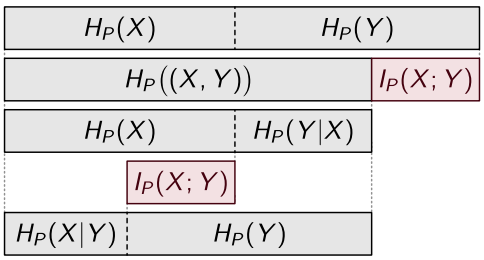

Figure 1: Interplay between marginal entropies ($H_P(X)$ and $H_P(Y)$), joint entropy $H_P\big((X;Y)\big)$, conditional entropies ($H_P(X\,|\,Y)$ and $H_P(Y\,|\,X)$), and mutual information $I_P(X;Y)$ of two arbitrary (discrete) random variables X and Y. Figure adapted from book "Information Theory, Inference, and Learning Algorithms" by David MacKay.

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

Note: In Problem 4.4 below, you will show that, *in expectation over* X, the conditional entropy $H_P(Y\,|\,X)$ (see Eq. 3 below) can never be larger than the marginal entropy $H_P(Y)$. Thus, we can say that conditioning on some $X=x$ *typically* reduces the entropy of Y, but it is possible that certain specific values of x exist for which conditioning on $X=x$ increases the entropy of Y.

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

$$\begin{aligned} H_P(Y\,|\,X) &= \sum_x P(X=x)\, H_P(Y\,|\,X=x) \\ &= -\sum_x P(X=x) \sum_y P(Y=y\,|\,X=x)\, \log_2 P(Y=y\,|\,X=x) \\ &= -\sum_{x,y} P(X=x, Y=y)\, \log_2 P(Y=y\,|\,X=x) \\ &\equiv \mathbb{E}_P\big[-\log_2 P(Y\,|\,X)\big]. \end{aligned} \tag{3}$$

Derive the chain rule of the entropy (visualized in the lower parts of Figure 1):

$$H_P\big((X,Y)\big) = H_P(X) + H_P(Y\,|\,X) = H_P(Y) + H_P(X\,|\,Y). \tag{4}$$

(d) What are the joint entropy $H_P\big((X,Y)\big)$ and the two types of conditional 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 $P(X,Y) = P(X)\,P(Y)$?

Problem 4.4: Mutual Information and Subadditivity of Entropies

We now show that entropies of two random variables X and Y are subadditive, i.e.

$$H_P\big((X,Y)\big) \leq H_P(X) + H_P(Y). \tag{5}$$

To show this, we define the *mutual information* $I_P(X;Y)$ *between* X *and* Y,

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

as illustrated in the first two rows of Figure 1. We then show that $I_P(X;Y) \geq 0$.

(a) **Symmetry of the Mutal Information:** Convince yourself that the mutual information is symmetric, i.e., $I_P(X;Y) = I_P(Y;X)$. (This is not really relevant for the proof of $I_P(X;Y) \geq 0$ but still important to know in general.)

(b) Convince yourself that the mutual information can be expressed as follows,

$$I_P(X;Y) = \mathbb{E}_P\left[\log_2 \frac{P(X,Y)}{P(X)\,P(Y)}\right] \tag{7}$$

Then use Eq. 3 from last week's problem set to express $I_P(X;Y)$ as a Kullback-Leibler divergence between two distributions (which two?). Thus, $I_P(X;Y) \geq 0$ since Kullback-Leibler divergences are nonnegative, as you proved in Problem 3.1.

(c) Combine Eqs. 4 and 6 to show that the mutual information can also be expressed as follows (illustrated in the last three rows of Figure 1),

$$I_P(X;Y) = H_P(X) - H_P(X \,|\, Y) \tag{8}$$
$$= H_P(Y) - H_P(Y \,|\, X). \tag{9}$$

Note: Since $I_P(X;Y) \geq 0$, Eq. 9 implies that $H_P(Y \,|\, X) \leq H(Y)$. Thus, while conditioning on a *specific* $X = x$ may increase the conditional entropy $H_P(Y \,|\, X = x)$ compared to $H_P(Y)$ (see Problem 4.3 (b)), *in expectation*, conditioning can only decrease the entropy (or keep it unchanged at worst).

Interpretation: By the source coding theorem, the entropy $H_P(X)$ measures the expected number of bits that someone needs to tell us before we can be certain about the value of X. Thus, we can interpret entropy as "amount of uncertainty" or "lack of knowledge". Then, the interpretation of Eq. 8 is that the mutual information $I_P(X;Y)$ measures by how much our uncertainty about X decreases (= how much knowledge we gain about X) if someone tells us the value of Y. Analogously, the interpretation of Eq. 9 is that $I_P(X;Y)$ also measures how much we learn about Y if someone tells us the value of X. This interpretation will become helpful when we discuss lossy compression.

(d) What is the mutual information $I_P(X;Y)$ if X and Y are statistically independent? Interpret this also in words: if X and Y are statistically independent (e.g., if they represent the red and the blue die in our Simplified Game of Monopoly), then how much do you learn about X if someone tells you the value of Y, or vice versa?