Solutions to Problem Set 4

discussed:
17 May 2023

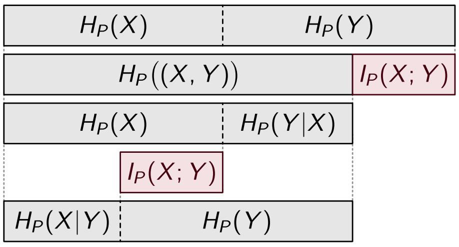
Data Compression With And Without Deep Probabilistic Models
Prof. Robert Bamler, University of Tübingen

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

Note on the Length of This Problem Set

Each question in Problems 4.2-4.4 below can be answered with a one-sentence argument or a single line of calculations (except for two questions ones marked with an asterisk ("*")). So don't try to be overly formal; our goal here is to find *concise* arguments that will help you get an intuition for several important information-theoretical concepts. But pay attention to details: some relations are surprisingly subtle.

How to Use This Problem Set to Study for the Exam



In Problems 4.2-4.4 below, you will derive several important information-theoretical relations, which are summarized in the figure on the right.[1]

When you first solve this problem set, you should use it 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 and the figure on the right as a self-contained reference sheet of important information-theoretical relations.

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

[1] adapted from the book "Information Theory, Inference, and Learning Algorithms" by David MacKay.

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).

Solution: $P(\Omega) = |\Omega|/|\Omega| = 1$ and $P(\emptyset) = |\emptyset|/|\Omega| = 0$ follow trivially from the definition of P. For countable additivity, at most $|\Omega| = 9$ of the events E_i in a countably infinite union $\bigcup_{i=1}^{\infty} E_i$ can be nonempty because otherwise the E_i's couldn't be pairwise disjoint. Thus,

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i:E_i\neq\emptyset} E_i\right) = \frac{|\cup_{i:E_i\neq\emptyset} E_i|}{|\Omega|} \overset{(*)}{=} \frac{\sum_{i:E_i\neq\emptyset} |E_i|}{|\Omega|} \\ &= \sum_{i:E_i\neq\emptyset} P(E_i) \overset{(\blacktriangle)}{=} \sum_{i=1}^{\infty} P(E_i) \end{aligned}$$

where the equality marked "$(*)$" holds because the number of elements in a union of a finite number of finite and pairwise disjoint sets is the sum of the number of elements in each set; and the equality marked "$(\blacktriangle)$" holds because $P(\emptyset) = 0$. ■

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

Solution: We have:

$$\begin{aligned} P(X_{\text{red}} = a) &= \frac{1}{3} \quad \forall a \in \{1,2,3\}; \\ P(X_{\text{blue}} = b) &= \frac{1}{3} \quad \forall b \in \{1,2,3\}; \\ \text{and} \qquad P(X_{\text{red}} = a, X_{\text{blue}} = b) &= \frac{1}{9} \quad \forall a,b \in \{1,2,3\}. \end{aligned}$$

Thus, $P(X_{\text{red}} = a, X_{\text{blue}} = b) = P(X_{\text{red}} = a)\, P(X_{\text{blue}} = b)\ \forall a,b \in \{1,2,3\}$. ■

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

Solution: To disprove statistical independence, it suffices to find a single case (a, s) for which $P(X_{\text{red}} = a, X_{\text{sum}} = s) \neq P(X_{\text{red}} = a)\, P(X_{\text{sum}} = s)$. This is the case, e.g., for $a = 1, s = 3$:

$$P(X_{\text{red}} = 1) = \frac{1}{3}; \qquad \text{and} \qquad P(X_{\text{sum}} = 3) = \frac{\left|\{(1,2),(2,1)\}\right|}{9} = \frac{2}{9};$$

but

$$P(X_{\text{red}} = 1, X_{\text{sum}} = 3) = \frac{\left|\{(1,2)\}\right|}{9} = \frac{1}{9} \neq \frac{1}{3} \times \frac{2}{9}.$$

■

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") w.r.t. a probability distribution P. We further discussed in Lecture 2 that the information content of a given (long) *message* is the bit rate (up to tiny corrections) that one would obtain when compressing the message with a lossless code that is optimal for the model P. In this problem, you'll analyze how many bits each symbol in the message contributes to the information content (and therefore the bit rate) of the full message.

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

(a) **Joint Information Content:** In the notation introduced in the lecture, the *joint information content* of the statement "$X = x$ and $Y = y$" or, equivalently, the information content of the statement "$(X, Y) = (x, y)$", can be written as follows,

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

Argue why the joint information content of "$(X, Y) = (x, y)$" is not smaller than the information content of "$X = x$" and not smaller than the information content of "$Y = y$" (*hint:* the information content of "$X = x$" is $-\log_2 P(X=x) = -\log_2 P\big(\{\omega \in \Omega : X(\omega) = x\}\big)$; identify a superset-subset relationship).

Solution: We showed in the lecture that $P(E_1) \leq P(E_2)$ for events E_1, E_2 with $E_1 \subseteq E_2$. Thus, for $E_1 := \{\omega \in \Omega : X(\omega) = x \;\wedge\; Y(\omega) = y\}$ and $E_2 := \{\omega \in \Omega : X(\omega) = x\}$, we have $P(X=x, Y=y) = P(E_1) \leq P(E_2) = P(X=x)$ and therefore, for the information contents: $-\log_2 P(X=x, Y=y) \geq -\log_2 P(X=x)$. ■

(b) **Marginal and Conditional Information Content:** We refer to the information content of "$X=x$" alone, $-\log_2 P(X=x)$, as the *marginal* information content. We further define the *conditional* information content of "$Y=y$" *given* $X=x$ as $-\log_2 P(Y=y \,|\, X=x)$, where $P(Y=y \,|\, X=x) := P(X=x, Y=y)/P(X=x)$ as defined in the lecture. Show the chain rule of information content, which states:

> 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 imply for lossless compression? 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 y, respectively.

Solution: The claim follows directly from the definition of the information content and the definition of conditional probability given above:

$$\begin{aligned} -\log_2 P(X\!=\!x, Y\!=\!y) &= -\log_2 \big[P(X\!=\!x)\, P(Y\!=\!y \,|\, X\!=\!x)\big] \\ &= -\log_2 P(X\!=\!x) - \log_2 P(Y\!=\!y \,|\, X\!=\!x). \end{aligned}$$

Thus if one wants to encode the tuple (x, y), one could encode x using a code that is optimized for the model $P(X)$, and then encode y using a code that is optimized for the model $P(Y \,|\, X\!=\!x)$, as we did last week with our autoregressive model in Problem 3.2. ■

(c*) **Nonadditivity of Marginal Information Content:** In Problem 2.3 (b) of Problem Set 2, you showed (although using different notation) that *if X and Y are statistically independent*, then the joint information content of "$(X, Y) = (x, y)$" is the sum of the two marginal information contents of "$X\!=\!x$" and "$Y\!=\!y$". 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 "$(X, Y) = (x, y)$".

For both cases (i) and (ii), use the chain rule of information content from part (b) to relate the marginal information content $-\log_2 P(Y\!=\!y)$ to the conditional information content $-\log_2 P(Y\!=\!y \,|\, X\!=\!x)$. Does conditioning on $X = x$ increase or reduce the information content in each of the two cases?

Note: You will show below that one of these cases (i) or (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?

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):

$P(X\!=\!x, Y\!=\!y)$	$\downarrow x\!=\!0 \downarrow$	$\downarrow x\!=\!1 \downarrow$	$\downarrow P(Y\!=\!y) \downarrow$
$y\!=\!0$ →	0.49	0.01	0.5
$y\!=\!0$ →	0.01	0.49	0.5
$P(X\!=\!x)$ →	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 ($-\log_2 0.49 \approx 1.03\,\text{bit} < 2\,\text{bit}$). 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 ($-\log_2 0.01 \approx 6.64\,\text{bit} > 2\,\text{bit}$). ■

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., $H_P(X) = \mathbb{E}_P[-\log_2 P(X)]$. Analogous to Problem 4.2, where we analyzed interactions between information contents of two random variables X and Y, let's now analyze interactions between their entropies. 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.[2] Argue, by applying what you've shown in Problem 4.2 (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 (i.e., calculating a weighted average) 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. ■

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*) $H_P(Y \,|\, X = x)$ 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

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

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

[2] 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.

Note: In Problem 4.4 (c) below, you will show that, *in expectation over* X, the conditional entropy $H_P(Y \,|\, X)$ (see Eq. 3 below) cannot be larger than the marginal entropy $H_P(Y)$. Thus, conditioning on some $X = x$ *typically* reduces the entropy of Y, but there may be some specific values of x where the opposite is the case.

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

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

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

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

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

(c) The notation $H_P(Y \,|\, X)$ denotes the *expected conditional entropy*, i.e., the expectation value of $H_P(Y \,|\, X=x)$ from part (b), where the expectation is taken over x:

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

Use the chain rule of information content from Problem 4.2 (b) to show the chain rule of the entropy (visualized in the lower parts of the figure on page 1):

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

Solution:

$$\begin{aligned} H_P(X) + H_P(Y|X) &= \mathbb{E}_P\big[-\log_2 P(X)\big] + \mathbb{E}_P\big[-\log_2 P(Y|X)\big] \\ &= \mathbb{E}_P\big[-\log_2 P(X) - \log_2 P(Y|X)\big] \\ &= \mathbb{E}_P\big[-\log_2 P(X,Y)\big] \\ &= H_P\big((X,Y)\big) \end{aligned}$$

The second equality in Eq. 4 follows from symmetry by swapping the names of X and Y. ■

(d) What are the joint entropy $H_P\big((X,Y)\big)$ and the two types of conditional entropy, $H_P(Y \mid X\!=\!x)$ and $H_P(Y \mid X)$, 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:

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

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

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.

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

This is an important result as it implies that modeling symbols in a message independently leads to suboptimal compression performance. As discussed in the lecture, one should instead consider a probabilistic model of the entire message.

To show Eq. 5, we define the *mutual information $I_P(X;Y)$ between X and Y*,

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

See the first two rows of the figure on page 1. We then show that $I_P(X;Y) \geq 0$:

(a) 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. 2 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 (b).

Solution: Eq. 7 follows directly from Eq. 6, the definition of the entropy, and properties of the logarithm. One possibly non-obvious step is that an expectation over a marginal distribution like $P(X)$ can also be expressed as an expectation over the joint distribution $P(X,Y)$. For example,

$$\begin{aligned}
H_P[X] &= \mathbb{E}_{P(X)}\big[-\log_2 P(X)\big] \\
&= -\sum_x P(X\!=\!x)\log_2 P(X) \\
&= -\sum_x \Big(\sum_y P(X\!=\!x, X\!=\!y)\Big)\log_2 P(X) \\
&= -\sum_{x,y} P(X\!=\!x, X\!=\!y)\log_2 P(X) \\
&= \mathbb{E}_{P(X,Y)}\big[-\log_2 P(X)\big].
\end{aligned}$$

This is why the lecture notes and problem sets will often just use the shorter notation $\mathbb{E}_P[\,\cdot\,]$ with only subscript "P".

From Eq. 7 and Eq. 3 on last week's problem set, we find that

$$I_P(X;Y) = D_{\mathrm{KL}}\big(P(X,Y)\,\big\|\,P(X)P(Y)\big) \geq 0$$

where the notation $P(X)P(Y)$ denotes the probability distribution (more precisely, the "probability mass function") that assigns to each pair (x,y) the probability $P(X=x)\,P(Y=y)$. The above identification of mutual information with a KL-divergence admits a direct interpretation: the mutual information is the expected overhead (in bitrate) if we compress data from some arbitrary true data distribution $P(X,Y)$ with the probabilistic model $P(X)P(Y)$, i.e., with a model that assumes (possibly wrongfully) that X and Y are statistically independent (you will show on next week's problem set that, within all models that assume statistical independence, the model $P(X)P(Y)$ that is a product of the marginals of the true probability distribution is the optimal one). ■

While we're at it, let's show two more important properties of the mutual information:

(b) **Mutual information is symmetric:** convince yourself that $I_P(X;Y) = I_P(Y;X)$.

Solution: Eq. 6 is clearly invariant under swapping X with Y. ■

(c) **Mutual information measures "Information Gain":** combine Eqs. 4 and 6 to show that the mutual information can also be expressed as follows (illustrated in the last three rows of the figure on page 1),

$$\begin{aligned}
I_P(X;Y) &= H_P(X) - H_P(X\,|\,Y) && (8) \\
&= H_P(Y) - H_P(Y\,|\,X). && (9)
\end{aligned}$$

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).

Solution: Combining Eqs. 4 and 6 leads to Eq. 9:

$$\begin{aligned} I_P(X;Y) &\overset{(6)}{=} H_P(X) + H_P(Y) - H_P\big((X,Y)\big) \\ &\overset{(4)}{=} H_P(X) + H_P(Y) - \big(H_P(X) + H_P(Y|X)\big) \\ &= H_P(Y) - H_P(Y|X). \end{aligned}$$

The relation in Eq. 8 follows similarly, or by using symmetry of $I_P(X;Y)$. ■

Interpretation of Eqs. 8-9: By the source coding theorem, the entropy $H_P(X)$ measures the expected number of bits that someone needs to tell us in order to communicate the value of X. Thus, we can interpret entropy as "amount of uncertainty" or "lack of information" that the receiver has before the communication takes place. 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 information we *gain* about X), in expectation, if someone tells us the value of Y. In fact, $I_P(X;Y)$ is also called "information gain" in some contexts. This interpretation will become helpful when we discuss lossy compression. Analogously, according to Eq. 9, $I_P(X;Y)$ also measures how much information we gain about Y, in expectation, if someone tells us the value of X.

(d) **Mutual information quantifies the degree of statistical dependency:** what is the mutual information $I_P(X;Y)$ if X and Y are statistically independent? Interpret this also in words using the above interpretation of mutual information: 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?

Solution: If X, Y are statistically independent, i.e., $P(X,Y) = P(X)P(Y)$ then

$$I_P(X;Y) = D_{\mathrm{KL}}\big(P(X,Y) \,\|\, P(X)P(Y)\big) = D_{\mathrm{KL}}\big(P(X)P(Y) \,\|\, P(X)P(Y)\big) = 0.$$

This is consistent with our above interpretation of mutual information: if X and Y are statistically independent then knowing X tells us nothing about Y and vice versa. ■