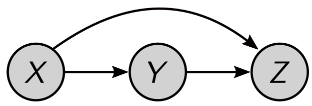
Problem Set 5

published: 17 May 2023
discussion: 24 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/

Problem 5.1: Conditional Independence

In last week's lecture, we learned that every probability distribution P satisfies the so-called chain rule of probability theory. For example, for any three random variables X, Y, and Z, we can always factorize their joint probability distribution as follows (see illustration on the right),



$$P(X, Y, Z) = P(X)\, P(Y \mid X)\, P(Z \mid X, Y). \tag{1}$$

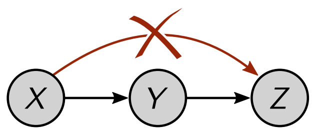
We then introduced the concept of *conditional (statistical) independence* between two random variables X and Z given a third random variable Y, which is defined analogously to the ordinary (i.e., unconditional) statistical independence as follows,

$$X \text{ and } Z \text{ are conditionally independent given } Y \;:\Leftrightarrow\; P(X, Z \mid Y) = P(X \mid Y)\, P(Z \mid Y). \tag{2}$$

(a) Show that conditional independence between X and Z given Y means that, once you know the value of Y, learning about the value of X would not provide any additional information about Z, i.e.,

$$X \text{ and } Z \text{ are cond. independ. given } Y \quad \Leftrightarrow \quad P(Z \mid X, Y) = P(Z \mid Y). \tag{3}$$

Remark: Eq. 3 implies that, if and only if X and Z are conditionally independent given Y, then the chain rule from Eq. 1 simplifies as follows (see illustration on the right),



$$X \text{ and } Z \text{ are cond. indep. given } Y \Leftrightarrow P(X, Y, Z) = P(X)\, P(Y|X)\, P(Z|Y). \tag{4}$$

We refer to the property expressed by Eq. 4 also by saying that X, Y, and Z form a *Markov chain* $X \to Y \to Z$. A Markov chain can be interpreted as a memoryless stochastic process: if you want to draw a random sample from a Markov chain, then you can proceed as follows: first, draw a random sample $x \sim P(X)$, then draw $y \sim P(Y \mid X{=}x)$, and finally draw $z \sim P(Z \mid Y{=}y)$. Notice that, once you've drawn y, you no longer need to keep x in memory because you won't need it for drawing z.

Markov chains play an important role in information theory since communication pipelines can typically be modeled as chains of memoryless stages, where each stage transforms the communicated data into some new representation. We'll meet Markov chains again when we discuss channel coding and lossy compression, and you'll prove an important bound on how information propagates along a Markov chain—the so-called data processing inequality—on Problem Set 10.

Comparison to ordinary independence: we now show that conditional independence is neither a stronger nor a weaker property than ordinary statistical independence.

(b) Show that two random variables X and Z can be statistically independent even if they are *not* conditionally independent given some third random variable Y.

Hint: Consider our Simplified Game of Monopoly. You already showed in Problem 4.1 (b) that X_{red} and X_{blue} are statistically independent. Now show that X_{red} and X_{blue} are, however, *not* conditionally independent given X_{sum}.

(c) Show that two random variables X and Z can be conditionally independent given some third random variable Y even if X and Z are *not* statistically independent.

Hint: Any (nontrivial) Markov process $X \to Y \to Z$ will do: conditioning on Y "cuts" the dependency between X and Z. For example, consider a sequence of three coin tosses and let X, Y, and Z be the number of times that the coin comes up "heads" in the first, the first two, and all three tosses, respectively. Find an expression for $P(Z \,|\, X, Y)$ without being overly formal (think about the experimental setup and the interpretation of conditional probability rather than its formal mathematical definition). Then convince yourself that X and Z are conditionally independent given Y by Eq. 3. Show by providing a counter example that, without conditioning on Y, then X and Z are *not* statistically independent.

Problem 5.2: Expressiveness of Probabilistic Models

In the lecture, we introduced various model architectures to efficiently approximate complicated probability distributions. Let us now analyze how expressive each of these architectures is. In particular, we analyze whether each of the proposed architecture can model *correlations* between symbols in a message, i.e., the fact that, in messages that appear in the real world, symbols are typically *not* statistically independent. All models below describe a message $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ where each symbol X_i, $i \in \{1, 2, \ldots, k\}$ is modeled as a random variable with values from some discrete alphabet $\mathfrak{X}$.

The four parts (a)-(d) of this problem can be solved independently. So don't give up if you have troubles solving one of the parts.

(a) **Fully factorized models:** before we look at more complicated model architectures below, let's consider the most trivial model architecture, which assumes that all symbols X_i, $i \in \{1, 2, \ldots, k\}$ are statistically independent. Such a model is often called "fully factorized" because the joint probability distribution $P(\mathbf{X})$ of the message $\mathbf{X}$ can be written as a product of the marginal distributions:

$$P_{\text{model}}(\mathbf{X}) = \prod_{i=1}^{k} P_{\text{model}}(X_i). \tag{5}$$

Here, we reinstated the subscript "model" because we want to search for the best model, $P^*_{\text{model}}(\mathbf{X})$, that can be written in the form of Eq. 5 and that best approximates some data distribution $P_{\text{data}}(\mathbf{X})$, which is typically *not* fully factorized.

(i) Consider the cross entropy $H\big(P_{\text{data}}(\mathbf{X}), P_{\text{model}}(\mathbf{X})\big)$. Convince yourself that, for a model of the form of Eq. 5 (warning: but not for more general models),

$$H\big(P_{\text{data}}(\mathbf{X}), P_{\text{model}}(\mathbf{X})\big) = \sum_{i=1}^{k} H\big(P_{\text{data}}(X_i), P_{\text{model}}(X_i)\big) \quad \text{(if Eq. 5 holds)} \tag{6}$$

where $P_{\text{data}}(X_i)$ is the marginal distribution of X_i under P_{data} (i.e., the distribution that you obtain if you *marginalize* $P_{\text{data}}(\mathbf{X})$ over all X_j with $j \neq i$).

(ii) Argue that the right-hand side of Eq. 6 is minimized by setting $P^*_{\text{model}}(X_i) = P_{\text{data}}(X_i)$ for all i. Thus, within the class of fully factorized models (Eq. 5), the best approximation $P^*_{\text{model}}(\mathbf{X})$ of an arbitrary distribution $P_{\text{data}}(\mathbf{X})$ is the product of the marginals, $P^*_{\text{model}}(\mathbf{X}) = \prod_{i=1}^{k} P_{\text{data}}(X_i)$.

Hint: what is the cross entropy $H(P, P)$ of a distribution with itself, and why is it smaller or equal than any $H(P, Q)$ for all other distributions $Q \neq P$?

(iii) Convince yourself that, for this optimal fully factorized model, the cross entropy (and thus the expected bit rate) is the sum of the marginal entropies of all symbols under the data distribution,

$$H\big(P_{\text{data}}(\mathbf{X}), P^*_{\text{model}}(\mathbf{X})\big) = \sum_{i=1}^{k} H_{P_{\text{data}}}(X_i) \qquad \text{(if Eq. 5 holds).} \tag{7}$$

(b) **Markov Chains:** as discussed in the lecture, a Markov chain models the creation of a sequence of symbols $X_1, X_2, \ldots, X_k$ as a memoryless stochastic process, i.e.,

$$P(\mathbf{X}) = P(X_1) \prod_{i=2}^{k} P(X_i \,|\, X_{i-1}) \tag{8}$$

where, from here on, we drop the subscript "model" for simplicity.

(i) Show that, although each symbol X_i is conditioned only on its immediately preceding symbol X_{i-1} (for $i > 1$) and not on any earlier symbols, a Markov chain can still model correlations between *any* symbols, not just nearest neighbors. More specifically, show that there exists a model of the form of Eq. 8 where two symbols X_i and X_j are *not* statistically independent for at least *some* $i, j \in \{1, \ldots, k\}$ with $j \geq i + 2$.

Hint: For example, you could consider the Markov chain over the alphabet $\mathfrak{X} = \{0, 1\}$ with $P(X_1{=}0) = P(X_1{=}1) = \frac{1}{2}$ and

$$P(X_i \,|\, X_{i-1}) = \begin{cases} 0.99 & \text{if } X_i = X_{i-1}; \\ 0.01 & \text{if } X_i \neq X_{i-1}. \end{cases} \tag{9}$$

Describe in words what a random sample $\mathbf{x} \sim P(\mathbf{X})$ from this model would typically look like. Then convince yourself (either by explicit calculation or

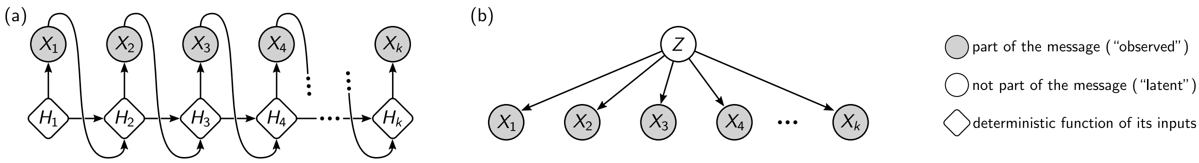


Figure 1: (a) autoregressive model, see Problem 5.2 (c); (b) latent variable model, see Problem 5.2 (d)

by less formal and more intuitive arguments) that all marginal probabilities are $P(X_i\!=\!0) = P(X_i\!=\!1) = \frac{1}{2}\ \forall i$ by symmetry but that, e.g., the conditional probability $P(X_j\!=\!1\,|\,X_i\!=\!1) > \frac{1}{2}$ for at least *some* non-neighboring i, j (it turns out to be true for *all* i, j, but this is more difficult to show formally).

(ii) Now show that, although a Markov chain can model symbols that are not statistically independent, any two symbols X_i and X_l with $l \geq i + 2$ are *conditionally* independent given any X_j with $i < j < l$.

Hint: write out the joint probability of all symbols *up to* X_l as follows,

$$P(\mathbf{X}) = \underbrace{\left(P(X_1)\prod_{\alpha=2}^{i}P(X_\alpha|X_{\alpha-1})\right)}_{=P(X_1,\ldots,X_i)}\underbrace{\left(\prod_{\alpha=i+1}^{j}P(X_\alpha|X_{\alpha-1})\right)}_{=P(X_{i+1},\ldots,X_j|X_i)}\underbrace{\left(\prod_{\alpha=j+1}^{l}P(X_\alpha|X_{\alpha-1})\right)}_{=P(X_{j+1},\ldots,X_l|X_j)}. \tag{10}$$

What do you get if you now marginalize both sides over all symbols except X_i, X_j, and X_l? Compare the result to Eq. 4.

(c) **Autoregressive models:** Figure 1 (a) illustrates an autoregressive model like the one you've used in Problem 3.2. The figure is a graphical representation of the following factorization of the joint probability distribution,

$$P(\mathbf{X}) = \prod_{i=1}^{k} P(X_i\,|\,H_i) \qquad \text{with} \qquad H_1 = \text{fixed};\ H_{i+1} = f(H_i, X_i) \tag{11}$$

where f is some deterministic function (e.g., a neural network). Show that autoregressive models are more powerful than Markov chains in that they can model probability distributions where two symbols X_i and X_l are *not* conditionally independent given some third symbol X_j with $i < j < l$.

Hint: For example, you could consider a toy autoregressive model over the alphabet $\mathfrak{X} = \{0, 1\}$ with $H_1 = 0$ and $H_{i+1} = f(H_i, X_i) = (H_i + X_i) \bmod 10$. Thus, the hidden state H_i counts how many "1" symbols have appeared before symbol X_i (modulo 10 so that the hidden states don't grow out of bounds). Now you could make the probability of "1" symbols depend on H_i, e.g., by setting $P(X_i\!=\!1\,|\,H_i) = \frac{H_i+1}{10}$ and $P(X_i\!=\!0\,|\,H_i) = 1 - \frac{H_i+1}{10}$. Then, consider the first three symbols X_1, X_2, and X_3 (the statement is also true for other triples of symbols, but the calculations are more tedious). Show by explicit calculation that

$$P(X_3\!=\!1 \,|\, X_1\!=\!1, X_2\!=\!1) \neq P(X_3\!=\!1 \,|\, X_2\!=\!1), \tag{12}$$

i.e., that even this simple toy model already violates the right-hand side of Eq. 3. The value of the left-hand side of Eq. 12 follows directly from unrolling the model but calculating the right-hand side takes a few more steps. Before you do these calculations, test your understanding by reasoning in words whether you expect the left-hand side of Eq. 12 to be smaller or larger than the right-hand side.

(d) **Latent variable models:** Figure 1 (b) illustrates a latent variable model. You'll learn how to use latent variable models for effective data compression with the so-called bits-back trick in Lecture 7. But let's first prove here that latent variable models can in fact capture correlations between symbols.

The illustration in Figure 1 (b) is a pictorial representation of the following factorization of a joint probability distribution over symbols $\mathbf{X} = (X_1, \ldots, X_k)$ and a (usually multidimensional) so-called *latent* variable Z,

$$P(\mathbf{X}, Z) = P(Z) \prod_{i=1}^{k} P(X_i \,|\, Z). \tag{13}$$

Here $P(Z)$ is called the "prior distribution" and $P(X_i \,|\, Z)$ is called the "likelihood". At a first glance, the model architecture in Eq. 13 might look like it couldn't possibly capture any correlations between different symbols X_i because the part of Eq. 13 that describes symbols is fully factorized (similar to the model in Eq. 5). However, this impression is deceptive because the symbols X_i are only *conditionally independent given the latent* Z. However, Z is not part of the message. The probabilistic model of the message is the *marginal* distribution of $\mathbf{X}$,

$$P(\mathbf{X}) = \begin{cases} \sum_Z P(\mathbf{X}, Z) & \text{for discrete } Z; \\ \int P(\mathbf{X}, Z)\, dZ & \text{for continuous } Z. \end{cases} \tag{14}$$

Show that the marginal distribution in Eq. 14 can indeed describe correlations between symbols, i.e., a distribution of this form can model data sources where any two symbols X_i and X_l are *not* statistically independent, and are also *not* conditionally independent given any different third symbol X_j.

Hint: You could consider, e.g., a toy model over the alphabet $\mathfrak{X} = \{0, 1\}$ with $k = 3$, boolean $Z \in \{0, 1\}$, and with a likelihood $P(X_i \,|\, Z)$ that is the same for all i. Come up with some explicit probabilities for $P(Z\!=\!z)$ and $P(X_i\!=\!x_i \,|\, Z\!=\!z)$ for all $z, x_i \in \{0, 1\}$. Then show first that $P(X_1\!=\!x_1, X_3\!=\!x_3) \neq P(X_1\!=\!x_1)\, P(X_3\!=\!x_3)$ and finally that $P(X_3\!=\!x_3 \,|\, X_1\!=\!x_1, X_2\!=\!x_2) \neq P(X_3\!=\!x_3 \,|\, X_2\!=\!x_2)$ in your model for some $x_1, x_2, x_3 \in \{0, 1\}$ of your choice. Try to explain your findings in words too: why does knowing the value of, e.g., X_1 influence the probability distribution over X_3?