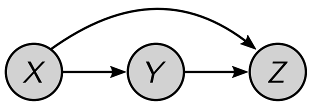
Solutions to Problem Set 5

discussed:
27 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/

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 also illustration on the right),

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

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

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

(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.,

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

Solution: Eq. 3 follows by direct calculation,

$$\begin{aligned} & X \text{ and } Y \text{ are conditionally independent given } Z \\ \Leftrightarrow\ & P(X, Z \,|\, Y) = P(X \,|\, Y)\, P(Z \,|\, Y) \\ \Leftrightarrow\ & P(Z \,|\, Y) = \frac{P(X, Z \,|\, Y)}{P(X \,|\, Y)} = \frac{P(X, Y, Z)}{P(Y)} \frac{P(Y)}{P(X, Y)} = \frac{P(X, Y, Z)}{P(X, Y)} = P(Z \,|\, X, Y). \end{aligned}$$

■

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 also illustration on the right),

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

If three random variables X, Y, and Z satisfy the right-hand side of Eq. 4, then we say that they form a *Markov chain* $X \to Y \to Z$. A Markov chain can be interpreted as a memoryless process: if you want to generate a random sample from a Markov chain, then you can do so by *ancestral sampling*: you first generate some random sample $x \sim P(X)$, then you generate some $y \sim P(Y \,|\, X = x)$, and finally you generate some $z \sim P(Z \,|\, Y = y)$. Notice that, once you've generated the random sample y, you no longer need to keep x in memory because you won't need it for generating z. Later in this course, you will prove the important *data processing inequality*, which states that information about the initial random variable X can never increase along a Markov chain, i.e., $I_P(X;Z) \leq I_P(X;Y)$. The information processing inequality has far reaching consequences, e.g., on how information can propagate along a deep neural network.

(b) Show that conditional independence is neither a strictly stronger nor a strictly weaker property than normal (i.e., unconditional) independence. Thus,

(i) 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}.

Solution: By definition, conditional independence holds if and only if the two probability distributions on the left and right-hand sides of Eq. 2 are equal. Two probability distributions are equal if they assign the same probabilities to all possible inputs. Thus, in order to show that X_{red} and X_{blue} are *not* conditionally independent given X_{sum}, we only have to find a single triple of values x_{red}, x_{blue}, and x_{sum} for which

$$P(X_{\text{red}}\!=\!x_{\text{red}}, X_{\text{blue}}\!=\!x_{\text{blue}} \,|\, X_{\text{sum}}\!=\!x_{\text{sum}}) \\ \neq P(X_{\text{red}}\!=\!x_{\text{red}} \,|\, X_{\text{sum}}\!=\!x_{\text{sum}})\, P(X_{\text{blue}}\!=\!x_{\text{blue}} \,|\, X_{\text{sum}}\!=\!x_{\text{sum}}).$$

You can easily find many examples for x_{red}, x_{blue}, and x_{sum} for which this is the case. For example, we have

$$P(X_{\text{red}}\!=\!1, X_{\text{blue}}\!=\!1 \,|\, X_{\text{sum}}\!=\!3) = 0$$

but, according to our example in the lecture notes for Lecture 4,

$$P(X_{\text{red}}\!=\!1 \,|\, X_{\text{sum}}\!=\!3)\, P(X_{\text{blue}}\!=\!1 \,|\, X_{\text{sum}}\!=\!3) = \frac{1}{2} \times \frac{1}{2} \neq 0.$$

■

(ii) show that two random variables X and Z can be conditionally independent given some third random variable Y even if X and Z by themselves are not statistically independent.

Hint: Almost any Markov process will do. For example, you could consider a sequence of three independent coin tosses and let $C_i \in \{0,1\}$ be the result of the i^{th} coin toss. Assume that the coin is bent (why is this necessary?), i.e., $P(C_i = 1) = \alpha$ and $P(C_i = 0) = 1 - \alpha$ with some $\alpha \neq \frac{1}{2}$. Then define $X := C_1$, $Y := (X + C_2) \bmod 2$, and $Z := (Y + C_3) \bmod 2$. Thus, X and Z are clearly conditionally independent given Y because they satisfy Eq. 3. Show by explicit calculation that X and Z, however, do *not* satisfy normal statistical independence.

Solution: We chose $\alpha = 0.1$ in the suggested toy model and we show that X and Z are not statistically independent by showing that $P(X\!=\!1, Y\!=\!1) \neq P(X\!=\!1)\,P(Z\!=\!1)$. We can easily compute $P(X\!=\!1) = P(C_1 = 1) = \alpha = 0.1$. To obtain the other two probabilities, we have to add up the probabilities of all possible values of C_1, C_2, and C_3 that are consistent with the corresponding requirement, i.e.,

$$\begin{aligned}
P(X\!=\!1, Z\!=\!1) &= P(C_1\!=\!1, C_2\!=\!0, C_3\!=\!0) + P(C_1\!=\!1, C_2\!=\!1, C_3\!=\!1) \\
&= \alpha(1-\alpha)^2 + \alpha^3 = 0.082
\end{aligned}$$

and

$$\begin{aligned}
\underbrace{P(X\!=\!1)}_{=\alpha}\,P(Z\!=\!1) &= \alpha\big[P(C_1\!=\!1, C_2\!=\!0, C_3\!=\!0) + P(C_1\!=\!0, C_2\!=\!1, C_3\!=\!0) \\
&\qquad + P(C_1\!=\!0, C_2\!=\!0, C_3\!=\!1) + P(C_1\!=\!1, C_2\!=\!1, C_3\!=\!1)\big] \\
&= 3 \times \alpha^2(1-\alpha)^2 + \alpha^4 = 0.0244 \neq P(X\!=\!1, Z\!=\!1).
\end{aligned}$$

Note: Had we set $\alpha = \frac{1}{2}$ then we would have ended up with a model where X and Z are "accidentally" statistically independent since $P(X\!=\!x, Z\!=\!z) = \frac{1}{4}$ and $P(X = x) = P(Z = z) = \frac{1}{2}$ for all $x, z \in \{0,1\}$. That doesn't invalidate our claim, though, since we're only claiming that there *exist* models where two random variables that are not statistically independent by themselves are conditionally independent given a third random variable. We don't claim that this the case in all models. ■

Problem 5.2: Correlated Symbols in Various Probabilistic Model Architectures

In last week's lecture, we introduced various architectures for models of complicated probability distributions. We will make these architectures more concrete and discuss which compression method works best with which model architecture in upcoming lectures. But before we do this, let's analyze how capable each model architecture actually is. In particular, we analyze whether each of the proposed probabilistic models can capture *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 trouble with some part.

(a) **Fully factorized models:** before we look at more complicated model architectures below, let's consider the most trivial class of models which assume 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 explicitly reinstated the subscript "model" because we now want to search for the best model, $P^*_{\text{model}}(\mathbf{X})$, in the form of Eq. 5 that approximates some data distribution $P_{\text{data}}(\mathbf{X})$, which is typically *not* fully factorized.

Thus, 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),

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

where, following our usual notation, $P_{\text{data}}(X_i)$ is the marginal distribution of symbol X_i under P_{data} (i.e., the distribution that you obtain if you *marginalize* $P(\mathbf{X})$ over all X_j with $j \neq i$). Then 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. (*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 $Q \neq P$?)

Thus, within the class of fully factorized models, the best approximation $P^*_{\text{model}}(\mathbf{X})$ of an arbitrary distribution $P_{\text{data}}(\mathbf{X})$ is the product of the marginals, $P^*_{\text{model}}(\mathbf{X}) = \prod_{i=1}^{k} P_{\text{data}}(X_i)$. Convince yourself that, for this model, the cross entropy is the sum of the marginal entropies of each symbol under the data distribution,

$$H\big(P_{\text{data}}(\mathbf{X}), P^*_{\text{model}}(\mathbf{X})\big) = \sum_{i=1}^{k} H_{P_{\text{data}}}[X_i] \qquad \text{(for optimal fully factorized model).} \tag{7}$$

Solution: To prove Eq. 6, we simply write out the cross entropy on the left-hand side, use linearity of the expectation, and then marginalize each term over all X_j

with $j \neq i$. For your reference, the following calculation is very detailed; you weren't really expected to write down every individual step like this:

$$
\begin{aligned}
H\big(P_{\text{data}}(\mathbf{X}), P_{\text{model}}(\mathbf{X})\big) &= \mathbb{E}_{P_{\text{data}}(\mathbf{X})}\big[-\log_2 P_{\text{model}}(\mathbf{X})\big] \\
&= \mathbb{E}_{P_{\text{data}}(\mathbf{X})}\left[-\sum_{i=1}^{k} \log_2 P_{\text{model}}(X_i)\right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{P_{\text{data}}(\mathbf{X})}\big[-\log_2 P_{\text{model}}(X_i)\big] \\
&\overset{(*)}{=} -\sum_{i=1}^{k}\left(\sum_{(X_1,\ldots,X_k)\in\mathfrak{X}^k} P_{\text{data}}(X_1,\ldots,X_k) \times \log_2 P_{\text{model}}(X_i)\right) \\
&\overset{(\triangle)}{=} -\sum_{i=1}^{k}\left(\sum_{X_i\in\mathfrak{X}} P_{\text{data}}(X_i) \times \log_2 P_{\text{model}}(X_i)\right) \\
&= \sum_{i=1}^{k} \mathbb{E}_{P_{\text{data}}(X_i)}\big[-\log_2 P_{\text{model}}(X_i)\big] \\
&= \sum_{i=1}^{k} H\big(P_{\text{data}}(X_i), P_{\text{model}}(X_i)\big)
\end{aligned}
$$

Where, in the equality marked with "$(*)$", we explicitly write out the expectation over $\mathbf{X} = (X_1, \ldots, X_k)$, and in the equality marked with "$(\triangle)$", we marginalize over all X_j with $j \neq i$.

Now, to show that setting $P^*_{\text{model}}(X_i) := P_{\text{data}}(X_i)$ minimizes each term on the right-hand side of Eq. 6 over $P_{\text{model}}(X_i)$, we note that the cross entropy of a distribution with itself is just the normal entropy, $H(P,P) = H[P]$. Thus, choosing any other $P'_{\text{model}}(X_i) \neq P_{\text{data}}(X_i)$ would increase the cross entropy by

$$
H\big(P_{\text{data}}(X_i), P'_{\text{model}}(X_i)\big) - H\big[P_{\text{data}}(X_i)\big] = D_{\text{KL}}\big(P_{\text{data}}(X_i)\,||\,P'_{\text{nodel}}(X_i)\big) \geq 0.
$$

Inserting $P^*_{\text{model}}(X_i) := P_{\text{data}}(X_i)$ into the right-hand side of Eq. 6 and using $H(P,P) = H[P]$ leads to Eq. 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, even if they are not nearest neighbors. I.e., show that there exists a model of the form of Eq. 8 where all pairs of symbols X_i and X_j with $i \neq j$ are *not* statistically independent.

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

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

Then convince yourself (either by explicit calculation or simply by reasoning about what this Markov process models) 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 for some $i \neq j$ (it actually turns out that $P(X_j{=}1 \,|\, X_i{=}1) > 1\ \forall j \geq i$ but this is more difficult to show).

Solution: The probabilities in Eq. 9 were deliberately chosen so dramatic to make the interpretation of the model stand out: the model describes sequences of bits, where one typically has long runs of identical bits before the bit flips. Therefore, while we have $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$ for each individual bit X_i by symmetry, any two bits X_i and X_j are more likely to be equal than unequal, especially if $|i - j|$ is not too large. It's easy to show this for symbols that are not too far away from each other:

$$P(X_j{=}1 \,|\, X_i{=}1) \geq P(X_{i+1}{=}1, X_{i+2}{=}1, \ldots, X_j{=}1 \,|\, X_i{=}1) = 0.99^{j-i}$$

which is larger than $P(X_j = 1) = \frac{1}{2}$ as long as $j - i \leq 68$. Therefore, X_i and X_j are not statistically independent in this case.

Note: The equation above only states a *lower bound* on $P(X_j{=}1 \,|\, X_i{=}1)$, but that's enough to prove that there exist some pairs of symbols that are not statistically independent. From our interpretation of Eq. 9, we'd expect that *no* pairs of symbols are statistically independent in this model; they only become *close* to being independent with growing distance $\delta := j - i$ (i.e., $\lim_{\delta\to\infty} I_P(X_i; X_{i+\delta}) = 0$). This is in fact true: using the so-called transfer matrix method, which is out of scope for this problem set, one finds that $P(X_j{=}1 \,|\, X_i{=}1) = \frac{1}{2}\big(1 + 0.98^{j-i}\big) > \frac{1}{2}\ \forall j - i \geq 0$.) ■

(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 $i < j < l$.

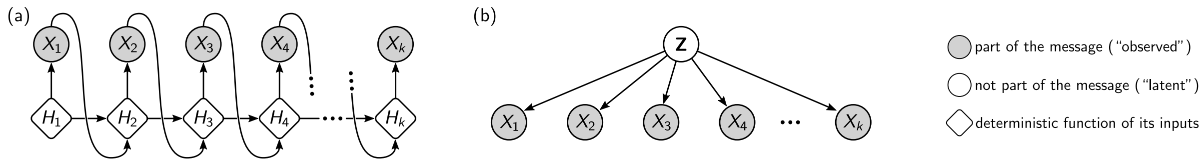


Figure 1: (a) autoregressive model, see Problem 5.2 (c); (b) latent variable model, see Problem 5.2 (d)

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

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

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.

Solution: Marginalizing both sides of Eq. 10 over all symbols except X_i, X_j, and X_l results in

$$P(X_i, X_j, X_l) = P(X_i)\, P(X_j \,|\, X_i)\, P(X_l \,|\, X_j)$$

which is precisely of the form of 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,

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

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 the model definition 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.

Solution: Since every "1"-bit increases the probability of subsequent "1"-bits, we expect the left-hand side of Eq. 12 to be larger than the right-hand side. Let's check this by explicit calculation.

To evaluate the left-hand side of Eq. 12, we can simply unroll the autoregressive model until the point where it models the symbol X_3. Since H_i counts how many "1" symbols have appeared before symbol X_i (modulo 10), we get $H_3 = 2$ and therefore $P(X_3{=}1 \,|\, X_1{=}1, X_2{=}1) = \frac{3}{10} = 0.3$.

To evaluate the right-hand side of Eq. 12, we explicitly write out the conditional probability and then express both enumerator and denominator as a marginalization over X_1,

$$\begin{aligned} P(X_3{=}1 \,|\, X_2{=}1) &= \frac{P(X_2{=}1, X_3{=}1)}{P(X_2{=}1)} = \frac{\sum_{x_1 \in \mathfrak{X}} P(X_1{=}x_1, X_2{=}1, X_3{=}1)}{\sum_{x_1 \in \mathfrak{X}} P(X_1{=}x_1, X_2{=}1)} \\ &= \frac{\frac{9}{10}\frac{1}{10}\frac{2}{10} + \frac{1}{10}\frac{2}{10}\frac{3}{10}}{\frac{9}{10}\frac{1}{10} + \frac{1}{10}\frac{2}{10}} = \frac{24}{110} \approx 0.218 \end{aligned}$$

which is indeed smaller than the left-hand side, as we expected. ■

(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 the next lecture. 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 $\mathbf{X}$,

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

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 $k = 3$, scalar $Z \in \{0, 1\}$, 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?

Solution: Let's use a uniform prior $P(Z\!=\!0) = P(Z\!=\!1) = \frac{1}{2}$ for simplicity and let's model a generative process where all symbols X_i are very likely to be equal to the latent variable Z, i.e.,

$$P(X_i\!=\!x_i \,|\, Z\!=\!z) = \begin{cases} \alpha & \text{if } x_i = z \\ 1 - \alpha & \text{if } x_i \neq z \end{cases}$$

for some $\alpha > \frac{1}{2}$. Since both the prior probability and the likelihood remain unchanged if we simultaneously flip the latent bit Z and all symbols X_i, each individual symbol X_i is either "0" or "1" with equal probability, $P(X_i\!=\!0) = P(X_i\!=\!1) = \frac{1}{2}$. However, we expect that most symbols are equal to Z and thus, *even if we don't know* Z, we can predict that most symbols are probably equal to each other. Or, put in different words, if we know, e.g., that $X_1 = 1$, then the most probable explanation for this is that $Z = 1$, which would then also make it probable that $X_3 = 1$. Conversely, if we know that $X_1 = 0$, then the most probable explanation for this is that $Z = 0$, which would then also make it probable that $X_3 = 0$. Thus, $P(X_3 \,|\, X_1)$ depends on X_1 and therefore X_1 and X_3 are not statistically independent. If we now also know the value of X_2 then we have even more data to reason about Z and, consequently, the probability of X_3 changes again.

More formally, we have, e.g.,

$$P(X_i\!=\!1) = \sum_{z\in\{0,1\}} P(Z\!=\!z, X_i\!=\!1) = \sum_{z\in\{0,1\}} P(Z\!=\!z)\, P(X_i\!=\!1 \,|\, Z\!=\!z)$$
$$= \frac{1}{2}\times\alpha + \frac{1}{2}\times(1-\alpha) = \frac{1}{2}$$

and therefore

$$P(X_1\!=\!1)\, P(X_3\!=\!1) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$$

whereas, for $i \neq j$ we have, according to our model in Eq. 13,

$$\begin{aligned}
P(X_i\!=\!1, X_j\!=\!1) &= \sum_{z\in\{0,1\}} P(Z\!=\!z, X_i\!=\!1, X_j\!=\!1) \\
&= \sum_{z\in\{0,1\}} P(Z\!=\!z)\, P(X_i\!=\!1 \,|\, Z\!=\!z)\, P(X_j\!=\!1 \,|\, Z\!=\!z) \\
&= \frac{1}{2}\big[(1-\alpha)^2 + \alpha^2\big] = \frac{1}{2} - \alpha + \alpha^2 = \frac{1}{4} + \left(\alpha - \frac{1}{2}\right)^2 \\
&> \frac{1}{4} \qquad \forall \alpha \neq \frac{1}{2}
\end{aligned}$$

which proves that X_i and X_j are not statistically independent.

To prove that X_1 and X_3 are also not conditionally independent given X_2 we show that they don't form a Markov chain, i.e., we use Eq. 3 and show that $P(X_3\!=\!1 \,|\, X_1\!=\!1, X_2\!=\!1) \neq P(X_3\!=\!1 \,|\, X_2\!=\!1)$. For simplicity, we chose a concrete value of $\alpha = 0.9$ here. We obtain the right-hand side by combining the above results,

$$P(X_3\!=\!1 \,|\, X_2\!=\!1) = \frac{P(X_2\!=\!1, X_3\!=\!1)}{P(X_2\!=\!1)} = \frac{\frac{1}{2}\big[(1-\alpha)^2 + \alpha^2\big]}{\frac{1}{2}} = 0.82$$

whereas, for the left-hand side,

$$\begin{aligned}
P(X_3\!=\!1 \,|\, X_1\!=\!1, X_2\!=\!1) &= \frac{P(X_1\!=\!1, X_2\!=\!1, X_3\!=\!1)}{P(X_1\!=\!1, X_2\!=\!1)} \overset{(*)}{=} \frac{\frac{1}{2}\big[(1-\alpha)^3 + \alpha^3\big]}{\frac{1}{2}\big[(1-\alpha)^2 + \alpha^2\big]} \\
&= \frac{0.73}{0.82} \approx 0.89 > P(X_3\!=\!1 \,|\, X_2\!=\!1)
\end{aligned}$$

where the equality marked with “(∗)” expresses both the enumerator and the denominator again as a marginalization over $Z \in \{0, 1\}$. ■

No programming problem this week; next week's problem set will be mostly programming however. You'll improve our autoregressive compression method for natural language by replacing the Huffman coder with a range coder, and you'll implement and empirically analyze the bits-back trick for a toy latent variable model.