INFORMATION THEORY
University of Amsterdam, 2019
TEACHERS: Christian Schaffner, Yfke Dulek TAs: Esteban Landerreche, Maximilian Siemers

TEAM NAME: Goppa
STUDENT NAMES: Nikita Tokovenko, Tobias Beers, Robert Jan Schlimbach, Bella Nicholson, Fergus Smiles

Homework set 1

Contents

Problem 1 (7 pt): Two coins and a die

You have two (fair) coins and a (fair) 4-sided die with outcomes $\{1, 2, 3, 4\}$. Let X be the number of heads after flipping the two coins and let Y be the result of rolling the die. Let Z be the average of X and Y.

(2 pt) **a.** What is the distribution P_Z of Z?

Solution: Let us first give the probability distribution for X and Y. X has 3 outcomes based on 4 events: $\{T, T\}$ $(X = 0)$, $\{T, H\}$ $(X = 1)$, $\{H, T\}$ $(X = 1)$ and $\{H, H\}$ $(X = 2)$. Therefore, the probability distribution can be expressed as $P_X(0) = \frac{1}{4}$, $P_X(1) = \frac{2}{4}$ and $P_X(2) = \frac{1}{4}$.

For Y, we have 4 outcomes that happen with the same probability $P_Y(y_i) = \{\frac{1}{4} | y_i \in \{1, 2, 3, 4\}\}$.

We can obtain the event space of Z as the Cartesian product of the event spaces of X and Y: $\mathcal{F}_Z = \{\frac{x+y}{2} | x \in \mathcal{X} \wedge y \in \mathcal{Y}\}$ (see Table 1). All events happen with probability $P_Z(\frac{x+y}{2}) = P_X(x)P_Y(y)$ (see Table 2). Adding probabilities of events with the same value for Z gives us the following probabilities: $P_Z(\frac{1}{2}) = \frac{1}{16}$, $P_Z(1) = \frac{3}{16}$, $P_Z(1\frac{1}{2}) = \frac{4}{16}$, $P_Z(2) = \frac{4}{16}$, $P_Z(2\frac{1}{2}) = \frac{3}{16}$ and $P_Z(3) = \frac{1}{16}$.

		Y			
		1	2	3	4
X	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
	1	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$
	2	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Table 1: values of $P_Z = P_X(x)P_Y(y)$ for all $\{x, y | x \in \mathcal{X}, y \in \mathcal{Y}\}$.

		Y			
		1	2	3	4
X	0	$\frac{1}{2}$	1	$1\frac{1}{2}$	2
	1	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$
	2	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3

Table 2: Values of $Z = \frac{x+y}{2}$ for all $\{x, y | x \in \mathcal{X} \wedge y \in \mathcal{Y}\}$.

□

(3 pt) **b.** Compute the variances of X, Y and Z.

Solution: Note the values for X and P_X, Y and P_Y and Z and P_Z given in 1.a. First, we compute the expected values for X as $E[X] = \sum_i^n P_X(x_i)x_i$. The variance is then given as $\text{var}(X) = E[(X - E[X])^2]$ or $\text{var}(X) = \sum_i^n P_X(x_i)(x_i - E[X])^2$. $E[Y]$, $\text{var}(Y)$, $E[Z]$ and $\text{var}(Z)$ were obtained by similar calculations. This lead to the results of $E[X] = 1$, $\text{var}(X) = \frac{1}{2}$, $E[Y] = 2\frac{1}{2}$, $\text{var}(Y) = 1\frac{1}{4}$, $E[Z] = 1\frac{3}{4}$ and $\text{var}(Z) = \frac{7}{16}$ □

(2 pt) **c.** You play the following game. If $2X \geq Y$, you win X^2 euros and otherwise you lose 1 euro. What is your expected total gain or loss after playing this game 40 times?

Solution: Table 3 demonstrates in which cases the player wins the game. If we compare these outcomes with the probabilities given in Table 2, we note that there are 6 events in which the player wins, with a summed probability of $P(2X \geq Y) = \frac{8}{16} = \frac{1}{2}$. Meaning, the probability of not winning, i.e. loosing the game, is $P(2X < Y) = 1 - \frac{1}{2} = \frac{1}{2}$. The expected gain per win is obtained by evaluating $E[X^2 | 2X \geq Y]$.

		Y			
		1	2	3	4
X	0	F	F	F	F
	1	T	T	F	F
	2	T	T	T	T

Table 3: Boolean values of $2x \geq y$ for all $\{x, y | x \in \mathcal{X} \wedge y \in \mathcal{Y}\}$.

When $2X \geq Y$, $P_X(1) = \frac{1}{2}$ and $P_X(2) = \frac{1}{2}$ (again, please refer tables 2 and 3 for further explanations). Therefore, $E[X^2|2X \geq Y] = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot 2^2 = \frac{1}{2} \cdot 5 = 2\frac{1}{2}$. We also know that $E[-1|2X < Y] = -1$, i.e. in case of loosing the game, the gain is always -1, independently of any variable. The expected gain per game is therefore $E[\text{gain}] = P(2X \geq Y)E[X^2|2X \geq Y] + P(2X < Y)E[-1|2X < Y] = \frac{1}{2} \cdot 2\frac{1}{2} + \frac{1}{2} \cdot (-1) = \frac{3}{4}$. As a result, the expected gain after 40 games is $40 \times \frac{3}{4} = 30$ euros. □

Problem 2 (9 pt): Deriving the weak law of large numbers

(2 pt) **a.** (**Markov's inequality**) For any real non-negative random variable X and any $t > 0$, show that

$$P[X \geq t] \leq \frac{\mathbb{E}[X]}{t}\,.$$

Solution: Since $X > 0$, and X is discrete, we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=0}^{\infty} x_i P(x_i) = \sum_{0 \leq i < t} x_i P(x_i) + \sum_{i=t}^{\infty} x_i P(x_i) \\ &\geq \sum_{i=t}^{\infty} x_i P(x_i) \geq t \sum_{i=t}^{\infty} P(x_i) = tP(X \geq t)\end{aligned}$$

□

(1 pt) **b.** Exhibit a random variable (which can depend on t) that achieves this inequality with equality.

Solution: Consider the binary random variable X with $P_X(0) = \frac{1}{2}$ and $P_X(1) = \frac{1}{2}$. Also consider $t = 1$. We then have $E[X] = \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2}$. Similarly, $P(X \geq t) = P_X(1) = \frac{1}{2}$. Additionally, $\frac{E[X]}{t} = \frac{1}{2} \times \frac{1}{t} = \frac{1}{2}$. Finally, we conclude that equality is achieved as $P(X \geq t) = \frac{1}{2} = \frac{E[X]}{t}$.

An example of equality achieved with $t(X)$ dependent on X is obtained by declaring X as *any* random variable X and setting $t(X) = X$. In this case, $P(X \geq t(X)) = 1$ as X is *always* equal to t. Furthermore, $\frac{E[X]}{E[t(X)]} = \frac{E[X]}{E[X]} = 1$, so we conclude that $P(X \geq t(X)) = \frac{E[X]}{E[t(X)]}$ for all random variables X.

□

(3 pt) **c.** (**Chebyshev's inequality**) Let Y be a random variable with mean μ and variance σ^2. Show that for any $\varepsilon > 0$,

$$P[|Y - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}\,.$$

Hint: Define a random variable $X := (Y - \mu)^2$.

Solution: Let $A \subseteq \Omega = \{\omega \in \Omega : |Y - \mu| \geq \epsilon\} = \{\omega \in \Omega : (Y - \mu)^2 \geq \epsilon^2\}$. Now consider X as defined above: X is a non-negative random variable, hence we can apply Markov's inequality as proven in 2a.

$$P[A] = P[X \geq \epsilon^2] \leq \frac{\mathbb{E}[X]}{\epsilon^2} = \frac{\mathbb{E}[(Y - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

Hence, Chebyshev's inequality has been shown to be derivable from Markov's inequality.

□

(3 pt) **d.** (The weak law of large numbers.) Let $Z_1, Z_2, ..., Z_n$ real i.i.d. random variables with mean $\mu = \mathbb{E}[Z_i]$ and variance $\sigma^2 = \mathbb{E}[(Z_i - \mu)^2] < \infty$. Define the random variables $S_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Show that

$$P[|S_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}\,.$$

Thus, $P[|S_n - \mu| \geq \varepsilon] \to 0$ as $n \to \infty$. This is known as the weak law of large numbers (which we will use heavily in Week 03).

Solution:

First, we begin with Markov's Inequality as given in the first part of this question.

$$P[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Next, suppose that $X = (S_n - \mu)^2$, and $t = \varepsilon^2$. Then,

$$P[(S_n - \mu)^2 \geq \varepsilon^2] \leq \frac{\mathbb{E}[(S_n - \mu)^2]}{\varepsilon^2} \tag{1}$$

In the following steps, we will reformulate the left and right hand sides of the expression shown in Equation 1 separately.

1. *The left-hand side.*

$$P[(S_n - \mu)^2 \geq \varepsilon^2] = P[|S_n - \mu| \geq \varepsilon] \tag{2}$$

2. *The right-hand side.* A pivotal part of our understanding of the right-hand side of Equation 1 is that it can be reformulated using the definition of variance. Below, we prove that it is fair to re-express $\mathbb{E}[(S_n - \mu)^2]$ as $\text{Var}(S_n)$.

$$\begin{aligned}\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_i\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\cdot n\mu\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right)^2\right]\end{aligned}$$

Hence,

$$\frac{\mathbb{E}[(S_n - \mu)^2]}{\varepsilon^2} = \frac{1}{\varepsilon^2}\cdot \text{Var}[S_n] = \frac{1}{\varepsilon^2}\cdot \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]$$

$$= \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[Z_i\right] = \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\cdot n\sigma^2$$

Meaning,

$$\frac{\mathbb{E}[(S_n - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon} \tag{3}$$

If we substitute Equations 2 and 3 into Equation 1, we obtain

$$P[|S_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}$$

, as desired.

□

Problem 3 (3 pt): Multiple-choice test

A multiple-choice exam has 4 choices for each question. A student has studied enough so that the probability she will know the answer to a question is 0.5, the probability that she will be able to eliminate one choice is 0.25, otherwise all 4 choices seem equally plausible. If she knows the answer she will get the question right. If not she has to guess from the 3 or 4 choices.

As the teacher you want the test to measure what the student knows. If the student answers a question correctly, what is the probability she knew the answer? Give your answer with three decimals of precision.

Solution:
In order to calculate the probability that the student answered correctly and knew the solution $P(k|c)$, we must first calculate the probabilities that she selected the correct answer and knew it $P(k, c)$ and the probability of her just being correct $P(c)$.

$$P(k|c) = \frac{P(k, c)}{P(c)} \tag{4}$$

, where we already know $P(k, c) = \frac{1}{2} \cdot 1 = \frac{1}{2}$.

However, to determine $P(c)$, we must also consider the possibilities where she is able to guess the correct solution. To do so, we refer Figure 1 , which provides us a visualization of the probability space.

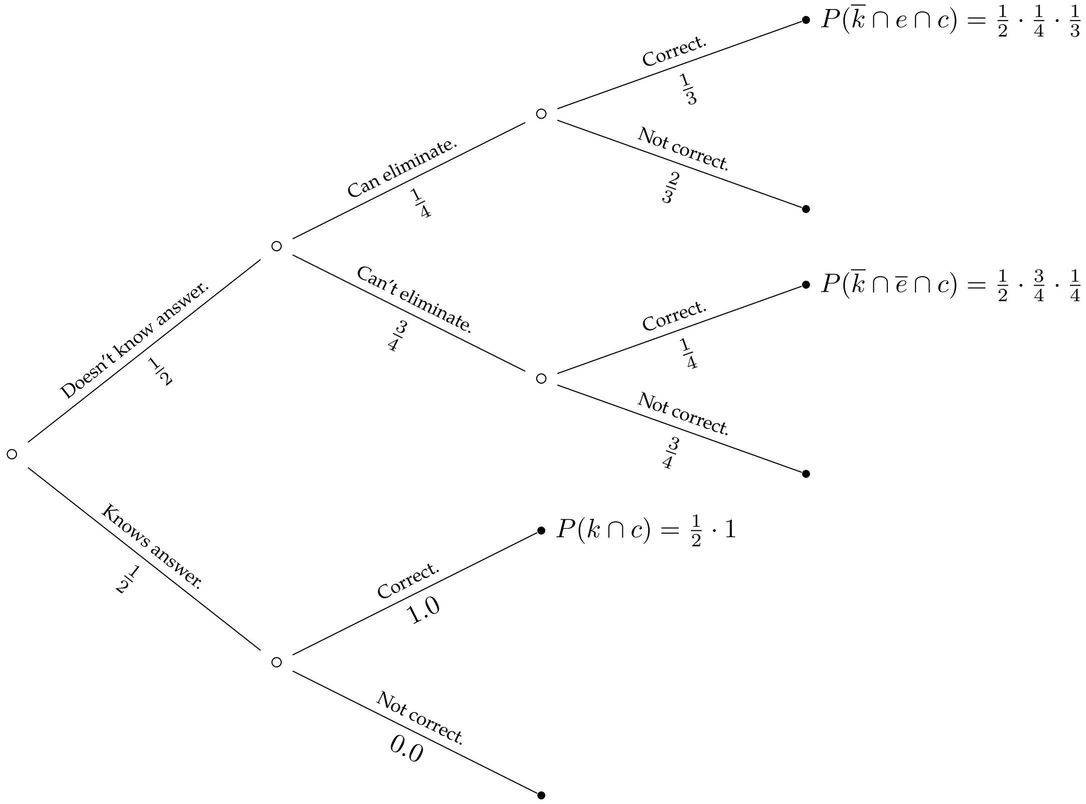

Figure 1: **Probability space visualization**. As we are only interested in the probability of selecting the correct solution, we have omitted all other probability calculations for the sake of legibility.

Thus,

$$P(c) = P(k \cap c) + P(e \cap c) + P(\overline{e} \cap c)$$

$$P(c) = \frac{1}{2} + \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{61}{96}$$

Referring to Equation 4,

$$P(k|c) = \frac{P(k,c)}{P(c)} = \frac{1}{2} \cdot \frac{96}{61} = \frac{48}{61} = 0.787$$

Meaning, if the student selected the correct solution, there is a 0.787 chance that she knew the solution.

□

Problem 4 (8 pt): Computing Variational Distance (programming)

The *total variation distance* between two probability distributions P and Q over the finite alphabet $\mathcal{X}$ is defined as

$$\|P - Q\| := \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$$

This distance measure is symmetric, fulfills the triangle inequality and is normalized, i.e. it is 0 iff $P = Q$ and 1 iff P and Q have disjoint support.

The *collision probability* of a distribution P over finite alphabet $\mathcal{X}$ is defined as

$$Coll(P) := \sum_{x \in \mathcal{X}} P(x)^2$$

In this exercise, we are going to analyze the letter frequencies of *Alice in Wonderland* in five different languages: English, German, Esperanto, Italian and Finnish. You can find all necessary files here: https://github.com/cschaffner/InformationTheory/tree/master/Problems/HW1. Hereby, we are going to consider only the 26 English letters (without space) and ignore that languages like German and Finnish have important other letters such as ä, ö, ü.

For *lang* $\in$ {eng, ger, esp, ita, fin}, let P_{lang} be the frequency distribution of the 26 English letters (without space) of Alice in Wonderland.

(2 pt) **a.** Compute all pairwise variational distances $\|P_{lang} - P_{lang'}\|$ for $lang \neq lang' \in$ {eng, ger, esp, ita, fin}. Which two languages are closest, which two are furthest apart in terms of variational distance?

Note: for this exercise and any future programming exercises, you do not have to submit your code. In the pdf that you hand in, describe in a few sentences which general strategy you used (e.g., what quantities did you compute and in what order?), any choices you made (e.g., how did you treat uppercase letters? How did you deal with edge cases?), and any 'sanity check' computations you may have performed (e.g., did you check what the variational distance between a text file and itself was?) By adding this information, you may still receive partial credit for your approach, even if your final numerical answer is incorrect.

Solution: Our general strategy is as follows: using regular expression, we delete all symbols not from the range [A-Za-z] from the target file. i.e., We exclude all letters and symbols that do not occur in the English alphabet. We do so as a simplifying assumption, as otherwise, we: (**a**) are considering an inconsistent amount of letters per language, which would make the endeavor of calculating variational distance much more difficult, and (**b**) would almost be "cheating" by simply learning the characters unique to a certain language. Thus, defeating the point of this exercise.

Next, we assume the information provided between a capital letter and its lowercase counterpart to be negligible in determining the source language, as uppercase letters are more indicative of a language's grammatical rules than the morphological structure of its words —where we reason the latter will provide us the majority of the information pertinent to this task. The only language that may defy this assumption is German, as —unlike most of languages considered in Table 4 requires the first letter of *every* noun to be upper-cased. Even so, prior (real-world) knowledge suggests the exclusion of uppercase letters is still permissible as the salient letter frequency differences the between English (the only other language shown that also belongs to the Germanic family) are still observed. For example, the letters z and y are swapped on the German keyboard, since z is known to be much more frequent in the German language than in English. Hence, we consider this a fair and safe assumption to make.[1]

After converting the remaining symbols to their respective lowercase forms, we count the number of times they occur in the text. Next, we build distribution by applying softmax activation to every letter count value in the alphabet. For a sanity check, we ensure that all probabilities both non-negative and all together sum to 1.

[1] As none of the team members of Goppa are particularly knowledgeable about the Finnish language, we simply assumed that if the assumptions made hold for all other languages, then they will likely also hold for Finnish.

	eng	esp	fin	ger	ita
eng	0	0.209	0.261	0.144	0.156
esp	0.209	0	0.160	0.237	0.106
fin	0.261	0.160	0	0.302	0.237
ger	0.144	0.237	0.302	0	0.192
ita	0.156	0.106	0.237	0.192	0

Table 4: Pairwise Variational Distances

Table 4 shows pairwise variational distances for every language available. As we can see, the maximal variational distance is obtained for German and Finnish languages, while the minimal variational difference exists between Italian and Spanish. These results as expected given our real world knowledge that Spanish and Italian are the only two languages considered that belong to the Romance language family, and thus share a large number of cognates (along with similar grammatical structures). This, in turn, should yield very similar word frequencies as captured by our metric of variational difference. Alternatively, Finnish comes from the Uralic language family, which is known to be one of the most insular language families in existence. To put it simply, very few languages exist that are closely related to Finnish, and none of which are present here. Meaning, the maximal variation should intuitively exist between Finnish and some other language. The fact that our results corroborate rather than contradict our real world knowledge serves as an additional sanity check.

□

(2 pt) **b.** Compute the five collision probabilities $Coll(P_{lang})$ for $lang \in \{$eng, ger, esp, ita, fin$\}$.
Note: You do not have to submit your code.

Solution:

	eng	esp	fin	ger	ita
$Coll(P_{lang})$	0.066	0.070	0.077	0.072	0.071

Table 5: Collision Probabilities

□

(1 pt) **c.** Why is it called collision probability?

Solution: It shows the probability of finding two distinct pieces of text in the given language that collide —by having the same character frequency distributions. It is similar to finding two different pieces of data represented with the same hash-function. □

(2 pt) **d.** You are given the file `permuted_cipher.txt` that has been encrypted by (first removing spaces and then) shuffling around the characters (i.e. by applying a permutation cipher). Note that this kind of encryption preserves the letter frequencies. Compute the frequency distribution P_{cipher} and figure out which language the original text was by picking the one that minimizes by the variational distance $\|P_{cipher} - P_{lang}\|$ with $lang \in \{$eng, ger, esp, ita, fin$\}$ as above.
Note: You do not have to submit your code.

Solution:

As we can see from Table 6, finding an $argmin$ of Variational Distance tells us that the closest language to our cipher is Spanish. □

	eng	esp	fin	ger	ita
cipher	0.212	0.040	0.177	0.243	0.099

Table 6: Variational Distances to the encrypted version

(1 pt) **e.** Would you have picked the same language when comparing the collision probability $Coll(P_{\mathit{cipher}})$ to the ones above?

Solution: $Coll(P_{cipher})$ = 0.070083.

Yes, since Spanish language once again has the closest collision probability. □