# Foundations of data science, summer 2020

JONATHAN LENNARTZ, MICHAEL NÜSKEN, ANNIKA TARNOWSKI

## 11. Exercise sheet
## Hand in solutions until Thursday, 2 July 2020, 12:00

**Exercise 11.1** (Mixture of densities). (8 points)

Suppose you are given some random variables $X^{(i)} \xleftarrow{\text{RND}} \mathbb{R}$ with density $p_i$. For the computer scientist: some routine `Xi` produces samples of $X^{(i)}$. How do you construct a routine `X` that samples acc. to the overlayed density $\sum_{i<k} w_i p_i$? Prove correctness: 8

**Theorem.** *Consider $X^{(i)} \xleftarrow{\text{RND}} p_i$ for $i < k$ and $\hat{\imath} \xleftarrow{\text{RND}} w$, reading $w$ as a distribution on $\mathbb{N}_{<k}$. Finally, let $X \leftarrow X^{(\hat{\imath})}$. Then $X \sim p$.*

*Hint*: $X \sim p$ means that $p$ is the density of $X$, ie. $\text{prob}\,(X \in [a,b]) = \int_a^b p(x)\,\mathrm{d}x$ for all $a < b$.

Remark: This generalizes to random variables with other outputs instead of values in $\mathbb{R}$.

**Exercise 11.2** (Application of the SVD). (0+13 points)

In this exercise you shall play with the example from

Alex Thomo (2009). Latent Semantic Analysis (Tutorial).

(i) Reprogram it, denote by $k$ the used dimension. +3

(ii) Examine the resulting ranking if... +5

(a) ...you modify $k \in \{2,3,4,5\}$.

(b) ...you omit the scaling step.

(c) ...you change the selection of words by omitting words that only occur in a single document or by adding more words.

(d) ...you use the Euclidean metric instead of the angle metric.

That's a total of at least 24 cases. You need a careful analysis to isolate important insights.

(iii) Redo similar analysis with a larger dataset: You will find documents `11-document*.txt` in the exercises folder, which contain (parts of) the short overviews of some Wikipedia articles. +5

*Hint*: We expect you to present an analysis with insights, explanations and arguments. So, no large tables or thelike.