Logic and Theory of Discrete Systems
RWTH AACHEN UNIVERSITY

Prof. Dr. M. Grohe E. Fluck, P. Lindner

Exercise Sheet 2

Due date: Monday, May 18 until 15:00

- **There are exercises marked with 0 points (Exercise 2 and 5). These exercises will not be corrected or graded. We advise you to work on them anyway and we will upload solutions to these exercises.**
- Please upload your solutions to Moodle.
- Hand in your solutions in groups of **two to three students**.
- Please hand the solutions of your group in as a single PDF file.
- You will not be able to change your upload.
- The solutions for this exercise sheet will be published on **Monday, May 18 15:00**.
- A discussion regarding this exercise sheet will take place on **Friday, May 22 13:00** via Zoom.

Exercise 1 (Expectation, Variance and Concentration Bounds) 2+2+3=7 points

Compute the following properties:

a) Suppose we roll a fair dice twice. Let X_2 be the random variable that is the product of the two values obtained from rolling the dice twice. So formally $\Omega = \{1, \dots, 6\}^2$ and $X_2(i, j) = i \cdot j$. Compute $E(X_2)$ and $\mathrm{Var}(X_2)$.

b) We generalize the previous exercise to rolling the dice n times. Let X_n be the product of the results. Again, compute $E(X_n)$ and $\mathrm{Var}(X_n)$.

c) Suppose we rool a fair dice 200 times and count the number of 1's. Give an upper bound for the probability that the count of 1's stays below 8.

Towards this end, compare the bounds given by the Chebychev, Chernoff and Hoeffding inequalities. Wich one gives the best bound?

Exercise 2 (Joint Entropy) 0 points

Let X and Y be random variables over the same probability space with finite range. The *joint entropy* of X and Y is defined as

$$H(X, Y) = \sum_{(x,y)} P(X = x, Y = y) \log \frac{1}{P(X = x, Y = y)}.$$

Prove that if X and Y are independent random variables then

$$H(X, Y) = H(X) + H(Y).$$

Exercise 3 (Sample Sizes for Decision Trees) — 2 points

Consider again the Decision Tree example from the lecture given on Slide 1.29. Suppose the target function can be represented by a Decision Tree with 16 nodes.
Assume we want to compute a Decision Tree with an error rate of at most 5% with probability at least 80%. Give an upper bound on the number of training examples required for this. Justify your answer.

Exercise 4 (VC Dimension) — 4+3=7 points

What is the VC dimension of the following hypothesis classes? Justify your answers.

a) The class of all 12-element subsets of $\mathbb{R}$, i.e. the class of all functions $h\colon \mathbb{R} \to \{0,1\}$ such that $|h^{-1}(1)| = 12$.

b) The class of all circles in the plane, i.e. the class of all functions $h_{a,b,r}\colon \mathbb{R}^2 \to \{0,1\}$ defined by

$$h_{a,b,r}(x,y) = \begin{cases} 1 & \text{if } (x-a)^2 + (y-b)^2 \leq r \\ 0 & \text{otherwise} \end{cases}$$

for all $a, b, r \in \mathbb{R}$.

Note: Often a hypothesis class for a Boolean Classification Problem is specified by a class of sets rather than a class of functions. This is equivalent since 0-1-functions $f\colon \mathbb{X} \to \{0,1\}$ can be viewed as sets $S_f = \{x \in \mathbb{X} \mid f(x) = 1\}$ and vice-versa.

Exercise 5 (Linear Separators) — 0 points

We say that a set S of points is shattered by linear separators of margin γ if every labeling of the points in S is achievable by a homogeneous linear separator of margin at least γ. Decide whether there is a set of $1/\gamma^2 + 1$ points in the unit ball (i.e., each point $x \in S$ satisfies $\|x\| \leq 1$) which is shattered by linear separators of margin γ. Prove your answer.

Exercise 6 (Multiplicative Weight Update Algorithm) — 4 points

In the lecture you have seen two different MWU algorithms (first deterministic, then randomized). In this exercise you are supposed to execute the randomized algorithm on a given input.
We consider an MWU problem with 3 experts and 4 events.
Let $\alpha = \frac{1}{2}$, the event sequence 121234 (e.g. $j^{(3)} = 1$ and so on) and $L = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & \frac{1}{2} \end{pmatrix}$.
Give every vector $w^{(t)}$ for $t \leq 7$ and additionally $p^{(6)}$. Would changing the order of events change the result of $w^{(7)}$? Justify your answer.
Note: the vector $w^{(t)}$ contains all the weights at time step t, i.e. $w^{(t)} = (w_1^{(t)}, w_2^{(t)}, w_3^{(t)})$ in this example. The vector $p^{(t)}$ is defined analogously.