Prof. Michael Schaub
Florian Frantzen, Michael Scholkemper

Algorithmic Foundations of Data Science

Sample Exam With Solutions , solve whenever you want

Name: _____

Student ID: _____

Study Program: _____

Remarks

- Write your name and student ID on **every** sheet of paper.
- Write your solution in the space provided on the problem sheets. You may use the back of the pages if you need extra space. Please mark it below the problem if you do so.
- Do **not** use your **own paper**.
- You may answer in either English or German. Please do not mix languages within the answer to a problem.
- Only use **black** or **blue** document-proof pens. Do **not** use pencils.
- Clearly mark your solutions and results as such. If you provide **multiple** solutions to a question, the **worst** of those counts.
- You have **120 minutes** to work on the exam.
- With **60 points** you have passed this exam.

I hereby declare that I have read the above guidelines and that I am healthy enough to take the exam.

_____

(Signature)

Do not write below this line.

	1	2	3	4	5	6
Points	/ 20	/ 20	/ 21	/ 19	/ 19	/ 21
Sign.						

$\sum$	/ 120

Problem 1 (General Questions) 4+4+4+4+4 = 20 points

a) Briefly explain the concepts of batch and online learning in the context of supervised learning. What is the difference?

Solution:

For batch learning all training data is received initially in one batch. On the other hand, in online learning, training data arrives over time and we have to gradually improve the classifier.

b) What is the property of the hypothesis returned by an Empirical Risk Minimization algorithm?

Solution:

An ERM algorithm will choose a hypothesis that minimizes the training error.

c) Where is the volume of high-dimensional balls concentrated? Name the intuitive meaning of both bounds we considered in the lecture.

Solution:

The volume of an high-dimensional object is concentrated near the surface ($\frac{\text{vol}(1-\varepsilon)X}{\text{vol}(X)} \leq e^{-\varepsilon\ell}$). Moreover, the volume of an high-dimensional ball is also concentrated near the equator ($\frac{\text{vol}\left(\left\{x \in B^\ell \middle| |x_1| > \frac{c}{\sqrt{\ell-1}}\right\}\right)}{\text{vol}(B^\ell)} \leq \frac{2}{c}e^{-c^2/2}$)

d) Name two applications of the Multiplicative Weight Update Algorithm discussed in the lecture.

Solution:

- Boosting

- Bandit Learning
- (Stock Market Predictions)

---

e) Sketch the Page Rank algorithm.

Solution:

Page Rank simulates a random walk on the web graph with random restarts.

Ignoring the random restart every outgoing edge of a vertex has the same weight $(1/d_i^+)$.

Since the web graph is neither ergodic nor connected we introduce random restarts. Vertices without outgoing edges have outgoing probabilities $1/n$, for all other vertices we add r/n and subtract $rq_{i,j}$ from the original edges.

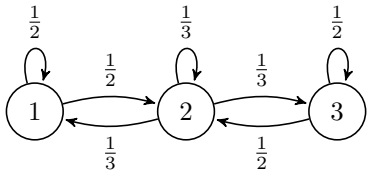
Problem 2 (Markov Chains) **5+5+5+5 = 20 points**

a) Let $\mathcal{Q}$ be a Markov chain and let G_Q denote its graph. Give the (formal) definitions of (i) *connectedness* of $\mathcal{Q}$, (ii) *aperiodicity* of $\mathcal{Q}$, and (iii) *ergodicity* of $\mathcal{Q}$. Additionally, give an example of an ergodic Markov chain with at least three states.

Solution:

- $\mathcal{Q}$ is connected if G_Q is strongly connected.
- $\mathcal{Q}$ is aperiodic if the greatest common divisor of all cycles in G_Q is 1.
- $\mathcal{Q}$ is ergodic if $\mathcal{Q}$ is both connected and aperiodic.

The following Markov chain (for example) is ergodic:



b) Give an example of a *connected* Markov chain $\mathcal{Q}$ with transition matrix Q such that there exists an initial distribution $\boldsymbol{p}_0$ for which the sequence $\left(\boldsymbol{p}_0 Q^i\right)_{i\in\mathbb{N}}$ does *not* converge to the the stationary distribution of $\mathcal{Q}$. For this, determine the stationary distribution $\boldsymbol{\pi}$ of your chain $\mathcal{Q}$ and briefly argue why $\lim_{i\to\infty} \boldsymbol{p}_0 Q^i \neq \boldsymbol{\pi}$.

Solution:

Let $\mathcal{Q}$ be the Markov chain with transition matrix $Q = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Then $\mathcal{Q}$ is connected and has stationary distribution $\boldsymbol{\pi} = \left(\frac{1}{2}, \frac{1}{2}\right)$ but for $\boldsymbol{p}_0 = (1, 0)$, the limit of $\boldsymbol{p}_0 Q^i$ does not exist, because

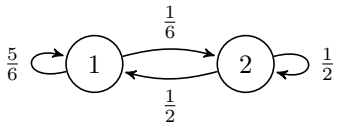
$$\boldsymbol{p}_0 Q^i = \begin{cases} (1,0) & \text{if } i \text{ is even, and} \\ (0,1) & \text{if } i \text{ is odd.} \end{cases}$$

c) You are observing traffic on the A4 motorway at the border from Germany to the Netherlands. There are only cars and trucks on the road. You find that on average

- for five in six cars crossing the border, the next vehicle crossing the border is again a car;
- for every second truck crossing the border, the next vehicle crossing the border is again a truck.

Model this situation with a two state Markov chain (state 1 $\widehat{=}$ car, state 2 $\widehat{=}$ truck). Compute the stationary distribution of your Markov chain to find the total percentage of cars and trucks on the road.

Solution:



The stationary distribution is $\left(\frac{3}{4}, \frac{1}{4}\right)$. That is, 75% of the vehicles are cars, and 25% are trucks.

d) Let $Q = (q_{ij}) \in \mathbb{R}^{n \times n}$ be the transition matrix of a connected Markov chain $\mathcal{Q}$. We construct a new matrix $Q' = (q'_{ij}) \in \mathbb{R}^{n \times n}$ with

$$q'_{ij} := \begin{cases} \frac{1}{2} q_{ij} & \text{if } i \neq j \\ \frac{1}{2}(q_{ij} + 1) & \text{if } i = j. \end{cases}$$

Prove that Q' is the transition matrix of an *ergodic* Markov chain $\mathcal{Q}'$. Prove that $\mathcal{Q}'$ has the same stationary distribution as $\mathcal{Q}$.

Solution:

It holds that $Q' = \frac{1}{2}Q + \frac{1}{2}I$ where I is the n-by-n identity matrix. Because Q and I are square matrices, Q' is a square matrix. The sum of the ith row in both Q and I is 1, respectively, so the sum of the ith row in Q' is $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1$. Thus, Q' is a stochastic square matrix, i. e. the transition matrix of a Markov chain $\mathcal{Q}'$.

We let $G_{\mathcal{Q}}$, $G_{\mathcal{I}}$ and $G_{\mathcal{Q}'}$ denote the graphs underlying the chains belonging to Q, I and Q'. They all have vertex set $\{1, \ldots, n\}$. The edge set of $G_{\mathcal{Q}'}$ is exactly the union of the edge sets of $G_{\mathcal{Q}}$ and $G_{\mathcal{I}}$. Because $G_{\mathcal{Q}}$ is already

strongly connected, $G_{\mathcal{Q}'}$ is strongly connected as well, i. e. $\mathcal{Q}'$ is connected. Moreover, $G_{\mathcal{Q}'}$ contains a loop at every vertex from the edges of $G_{\mathcal{I}}$. Thus, $\mathcal{Q}'$ is aperiodic. Together, $\mathcal{Q}'$ is ergodic.

Let $\boldsymbol{\pi}$ be the stationary distribution of $\mathcal{Q}$. Note that $\boldsymbol{\pi}$ is unique because $\mathcal{Q}$ is connected. Then

$$\boldsymbol{\pi}Q' = \boldsymbol{\pi}\left(\tfrac{1}{2}Q + \tfrac{1}{2}I\right) = \tfrac{1}{2}\boldsymbol{\pi}Q + \tfrac{1}{2}\boldsymbol{\pi}I = \tfrac{1}{2}\boldsymbol{\pi} + \tfrac{1}{2}\boldsymbol{\pi} = \boldsymbol{\pi}.$$

Thus, $\boldsymbol{\pi}$ is the stationary distribution of $\mathcal{Q}'$. It is also unique, because $\mathcal{Q}'$ is connected.

Problem 3 (Streaming) 4+2+3+6+6 = 21 points

a) Let $p \geq 0$. Define the pth frequency moment $F_p(\boldsymbol{a})$ of a data stream $\boldsymbol{a} = a_1, \ldots, a_n$ of elements from a universe U. Which special restriction is needed for $p = 0$?

Solution:

Let $\boldsymbol{a} = a_1, \ldots, a_n$ be a data stream with elements from the universe U. Then the pth frequency moment of $\boldsymbol{a}$ is

$$F_p(\boldsymbol{a}) := \sum_{u \in U} (|\{i \in [n] | a_i = u\}|)^p .$$

For $p = 0$ we only sum over the existing elements.

b) What is the intuitive meaning of $F_0(\boldsymbol{a})$ and $F_1(\boldsymbol{a})$?

Solution:

$F_0(\boldsymbol{a})$ is the number of distinct elements in $\boldsymbol{a}$.

$F_1(\boldsymbol{a}) = n$, the length of the stream $\boldsymbol{a}$.

c) Compute $F_0(\boldsymbol{a})$, $F_1(\boldsymbol{a})$, and $F_2(\boldsymbol{a})$ for the stream $\boldsymbol{a} = 2, 3, 1, 6, 1, 2, 6, 3, 2, 7, 1$ from $U = [8]$.

Solution:

$\boldsymbol{a} = 2, 3, 1, 6, 1, 2, 6, 3, 2, 7, 1$

$F_0(\boldsymbol{a}) = 3^0 + 3^0 + 2^0 + 2^0 + 1^0 = 5$

$F_1(\boldsymbol{a}) = 3^1 + 3^1 + 2^1 + 2^1 + 1^1 = 11$

$F_2(\boldsymbol{a}) = 3^2 + 3^2 + 2^2 + 2^2 + 1^2 = 27$

d) Describe the Flajolet-Martin Algorithm for estimating the number of distinct elements of the stream. Which properties of the family of hash functions used by the Algorithm are assumed?

Solution:

Let $\boldsymbol{a} = a_1, \ldots, a_n$ be our stream. We draw a hash function h uniformly at random from a family of hash functions (see below). For each stream element a, we calculate the number of trailing zeros of $h(a)$ (in binary representation). We only remember the largest value z of these trailing zeros. At the end we return $2^{z+1/2}$.

The hash family has the properties of being strongly 2-universal and from U to $[M]$, where M is the first power of 2 greater than or equal to $|U|$.

e) What is the result of applying the Flajolet-Martin Algorithm to the stream in (c) if the randomly chosen hash function $h\colon U \to [8]$ is given below. Give the result as well as some intermediate steps showing how the estimator is computed from $\boldsymbol{a}$

a	1	2	3	4	5	6	7	8
$h(a)$	6	5	4	8	7	2	1	3

Solution:

$\boldsymbol{a} = 2, 3, 1, 6, 1, 2, 6, 3, 2, 7, 1$

$h(1) = 6 = 110b \Rightarrow \text{zeros}(h(1)) = 1$
$h(2) = 5 = 101b \Rightarrow \text{zeros}(h(2)) = 0$
$h(3) = 4 = 100b \Rightarrow \text{zeros}(h(3)) = 2$
$h(6) = 2 = 10b \Rightarrow \text{zeros}(h(6)) = 1$
$h(7) = 1 = 1b \Rightarrow \text{zeros}(h(7)) = 0$

$z = 2 \Rightarrow 2^{z+1/2} = 4\sqrt{2} \approx 5,6$

Problem 4 (Power Iteration Algorithm) **6+7+6 = 19 points**

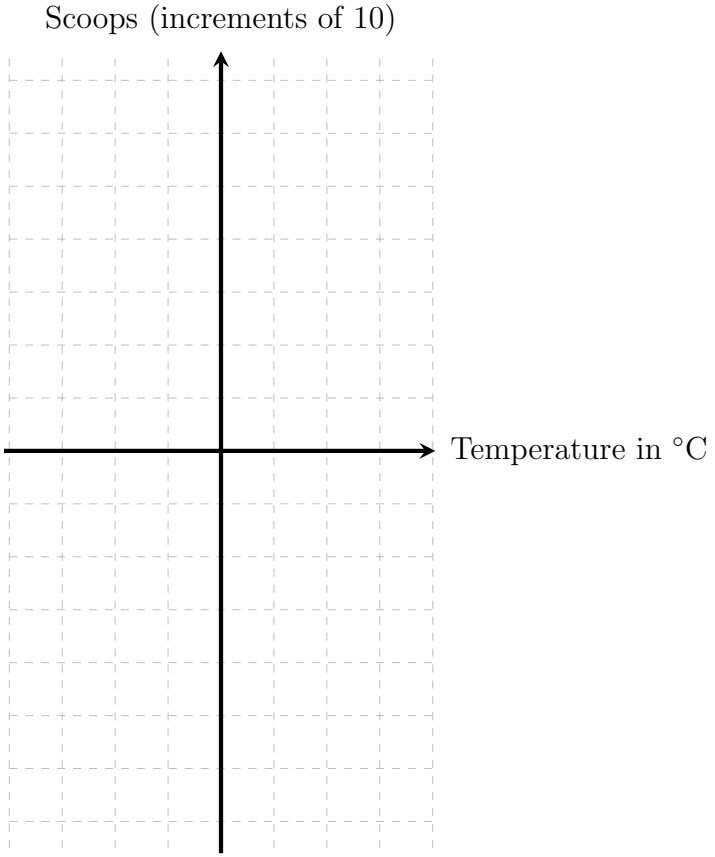
The local ice cream shop wants to find out how their sells depend on the weather. Towards this goal they measure the temperature on four days and count how many scoops of ice cream they sell. They find that the mean temperature was 30°C and the mean of sold scoops was 700 per day.

The following table notes the deviation from the mean on each day.

	Temperature (in °C)	Scoops of ice cream (in increments of 10)
day 1	-2	-3
day 2	3	7
day 3	2	3
day 4	-3	-7

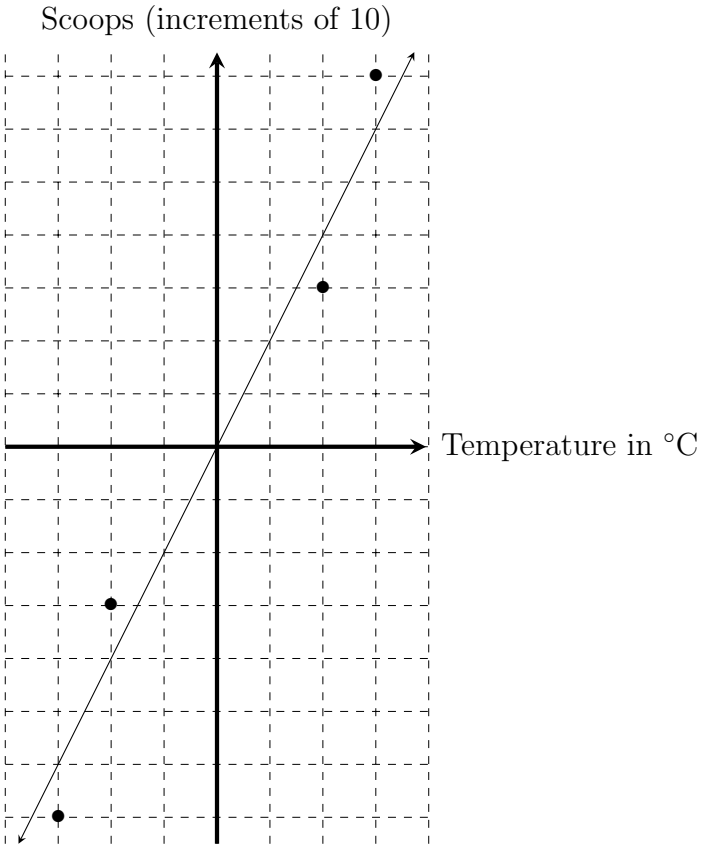
Let A be the 4×2 matrix representing these data-points.

a) Plot the 4 data-points above on the coordinate system drawn below. Using this plot, graphically *estimate* the first principal component of the above data-set. State this estimate as a linear equation of the form $ax + by = c$.



Solution:

Pick a line (through origin) which fits the data well enough. The easiest example is $0.5y - x = 0$.



b) Recall that the co-variance matrix C of the data-set A is defined as $C = A^\mathsf{T}A$. Using $(1 \quad 1)^\mathsf{T}$ as the starting vector, apply 2 rounds of the Power Iteration algorithm to estimate the eigenvector of C corresponding to the largest eigenvalue. To make the computations easier do the following:

- Use the L^∞-norm (i.e., the absolute value of the largest entry in the vector) for normalization.
- Round the vectors to one digit after the decimal point, at *every* step of Power Iteration.

Solution:

We first compute

$$C = A^{\mathsf{T}}A = \begin{pmatrix} 26 & 54 \\ 54 & 116 \end{pmatrix}$$

First iteration: C $(1 \quad 1)^{\mathsf{T}} = (80 \quad 170)^{\mathsf{T}} \approx (0.5 \quad 1)^{\mathsf{T}}$ Second iteration: C $(0.5 \quad 1)^{\mathsf{T}} = (67 \quad 143)^{\mathsf{T}} \approx (0.5 \quad 1)^{\mathsf{T}}$. Hence, the estimate of the eigenvector is $(0.5 \quad 1)^{T}\mathsf{T}$.

---

c) Let v be the eigenvector estimated by you in Part b). Let $\langle v \rangle$ denote the line in the x-y plane which passes through $(0, 0)^{\mathsf{T}}$ and v. Compare $\langle v \rangle$ with the estimate obtained in Part a) for the first principal component. Are they similar? If yes, briefly explain why it is reasonable to expect this. If no, give a reason why they are far from each other in this example.

Solution:

The two estimates should be very similar. Why? The first principal component of A is the span of the dominant eigenvector vector of $A^{\mathsf{T}}A$: this was the main theorem of PCA lectures. Hence, the two should match (modulo some error).

Problem 5 (Map-Reduce) **3+10+6 = 19 points**

a) Name and describe three cost measures that are used in the analysis of Map-Reduce algorithms.

Solution:

- **Wall-clock time:** total time for the MR-process to finish.
- **Number of rounds:** total number of MR-rounds.
- **Communication cost:** sum of the input sizes of all tasks.
- **Replication rate:** total number of key-value produced by all map tasks divided by the input size.
- **Maximum load / reducer size:** maximum input length for a single reducer / reduce task.

b) We consider meteorological data, given in key-value pairs of the shape $(c, (s, t, d))$ where

- c is a country,
- s is a weather station,
- t is a temperature measurement (in °C) and
- d is the date of the recording.

Specify *single-round* Map-Reduce algorithms for the following problems in pseudocode.

(i) *Average temperature per country.* Output all key-value pairs (c, t) where c is a country and t is the average temperature in country c (taken over all measurements ever recorded in country c).

(ii) *Stations with extreme temperature differences.* Output all key-value pairs (c, s) where c is a country and s is a station in country c with the property that the difference between the lowest and the highest temperature ever recorded at station s is at least 30 °C.

Additional notes:

- Use the "on input ..., [do some computation,] emit ..." format for specifying your pseudocode.
- Make sure your algorithm can benefit from parallelisation.

Solution:

(i) Map: On input $(c, (s, t, d))$, emit (c, t).

Reduce: On input (c, values), emit (c, t) where t is the average of the entries in values.

(ii) Map: On input $(c, (s, t, d))$, emit $((c, s), t)$.

Reduce: On input $((c, s), \mathit{values})$, emit (c, s) if the difference of the maximum and minimum entry of values is ≥ 30.

c) The following shows a Map-Reduce algorithm for computing the difference of two relations $\mathcal{R}$ and $\mathcal{S}$ in Relational Algebra.

MAP:	On input (R, t), emit $(Q, (R, t))$. On input (S, t), emit $(Q, (S, t))$.
REDUCE:	On input (Q, values), emit (Q, t) if $(R, t) \in \mathit{values}$ and $(S, t) \notin \mathit{values}$.

Although this algorithm is technically correct, it is a very bad example of a Map-Reduce algorithm. Why?

Propose a new algorithm for computing the difference. Explain why your algorithm is better by comparing it to the algorithm above.

Solution:

Possible explanations:

- The algorithm loads all tuples of $\mathcal{R}$ and $\mathcal{S}$ into a single reduce task and thus prevents any kind of parallel execution during the reduce phase.
- Bad load balancing. The load of the reducer may exceed its available space.

The following is much better:

MAP:	On input (R, t), emit (t, R). On input (S, t), emit (t, S).
REDUCE:	On input (t, values), emit (Q, t) if $R \in \mathit{values}$ but $S \notin \mathit{values}$.

The proposed algorithm creates a single task per tuple in $\mathcal{R} \cup \mathcal{S}$. Potentially, all of these tasks can be worked on in parallel.

Problem 6 (Linear Separators) 5+6+10 = 21 points

In this problem we study linear separators and Boolean functions $f\colon \{0,1\}^n \to \{-1,1\}$. Recall that we say f can be *represented by a linear separator* if there is a *weight* vector $\boldsymbol{w} \in \mathbb{R}^n$ and a *bias* $b \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \{0,1\}^n$,

$$f(\boldsymbol{x}) = \operatorname{sgn}\big(\langle \boldsymbol{w} \rangle \boldsymbol{x} - b\big).$$

where $\operatorname{sgn}(0) = 0$, $\operatorname{sgn}(x) = 1$ if $x > 0$, and $\operatorname{sgn}(x) = -1$ if $x < 0$. In this case, we say $(\boldsymbol{w}, b)$ *represents* f.

a) Consider the function $f\colon \{0,1\}^2 \to \{-1,1\}$ that is defined as follows:

$\boldsymbol{x}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
$f(\boldsymbol{x})$	1	-1	1	-1

Find a pair $(\boldsymbol{w}, b)$ representing f.

Solution:

Observe the following for $\boldsymbol{w} = (w_1, w_2)^\intercal$:

$\boldsymbol{x}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
$f(\boldsymbol{x})$	1	-1	1	-1
$\operatorname{sgn}\big(\langle \boldsymbol{w} \rangle \boldsymbol{x} - b\big)$	$\operatorname{sgn}(-b)$	$\operatorname{sgn}(w_2 - b)$	$\operatorname{sgn}(w_1 - b)$	$\operatorname{sgn}(w_1 + w_2 - b)$

Thus, any pair with $w_2 < b < 0$ and $b < w_1 < b - w_2$ will work, for example, $\boldsymbol{w} = \begin{pmatrix}0\\-1\end{pmatrix}$ and $b = -\frac{1}{2}$.

b) We consider a generalisation of the function in part a): For all $n \in \mathbb{N}_{>0}$ and every $I \subseteq \{1, \dots, n\}$, we let $f_{n,I}\colon \{0,1\}^n \to \{-1,1\}$ with

$$f_{n,I}(\boldsymbol{x}) := \begin{cases} -1 & \text{if } x_i = 1 \text{ for all } i \in I \text{ and} \\ 1 & \text{otherwise} \end{cases}$$

for all $\boldsymbol{x} = (x_1, \dots, x_n)^\intercal \in \{0,1\}^n$.

Depending on n and I, find a pair $(\boldsymbol{w}_{n,I}, b_{n,I})$ representing $f_{n,I}$. Justify your solution by discussing the value of $\langle \boldsymbol{w}_{n,I} \rangle \boldsymbol{x} - b_{n,I}$ for $\boldsymbol{x} \in \{0,1\}^n$.

Solution:

Fix arbitrary $n \in \mathbb{N}_{>0}$ and $I \subseteq \{1, \dots, n\}$. Define $\boldsymbol{w}_{n,I} = \big(w_1, \dots, w_n\big) \in \mathbb{R}^n$ with

$$w_i = \begin{cases} -1 & \text{if } i \in I, \text{ and} \\ 0 & \text{if } i \in \{1, \dots, n\} \setminus I. \end{cases}$$

Define $b_{n,I} = -\big(|I| - \frac{1}{2}\big) = \frac{1}{2} - |I|$. Then $(\boldsymbol{w}_{n,I}, b_{n,I})$ represents $f_{n,I}$, because

$$\langle \boldsymbol{w}_{n,I} \rangle \boldsymbol{x} - b_{n,I} = \sum_{i \in [n]} w_i x_i + |I| - \tfrac{1}{2} = \sum_{i \in I} (1 - x_i) + \tfrac{1}{2}.$$

This equals $-\frac{1}{2}$ if $x_i \in I$ for all $i \in I$ and is $\geq \frac{1}{2}$ otherwise. Thus, $(\boldsymbol{w}_{n,I}, b_{n,I})$ represents $f_{n,I}$.

---

c) Let $n \in \mathbb{N}_{>0}$ and let $f_n \colon \{0,1\}^n \to \{-1, 1\}$ be the Boolean function defined by

$$f_n(\boldsymbol{x}) := \begin{cases} -1 & \text{if } 2x_1 + \sum_{i=2}^{n} 3x_i \equiv 0 \pmod 4 \\ 1 & \text{otherwise.} \end{cases}$$

(i) Find pairs $(\boldsymbol{w}, b)$ and $(\boldsymbol{w}', b')$ representing f_1 and f_2, respectively.

(ii) Show that for $n \geq 3$, f_n can not be represented by a linear separator.

Hint:. First show that f_3 has no linear separator and then generalise the proof to $n > 3$.

Solution:

It is $f_1(0) = -1$ and $f_1(1) = 1$. Note that for all $w, b \in \mathbb{R}$ it holds that

$$\operatorname{sgn}(wx - b) = \begin{cases} \operatorname{sgn}(-b) & \text{if } x = 0 \\ \operatorname{sgn}(w - b) & \text{if } x = 1. \end{cases}$$

Thus, (w, b) represents f_1 if and only if $w > b > 0$. For example, $(1, \frac{1}{2})$ represents f_1.

For f_2 observe the following (the last row is the value of $\operatorname{sgn}(\langle \boldsymbol{w} \rangle \boldsymbol{x} - b)$).

$\boldsymbol{x}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
$f_2(\boldsymbol{x})$	-1	1	1	1
	$\operatorname{sgn}(-b)$	$\operatorname{sgn}(w_2 - b)$	$\operatorname{sgn}(w_1 - b)$	$\operatorname{sgn}(w_1 + w_2 - b)$

Thus, $(\boldsymbol{w}, b)$ represents f_2 if and only if $w_1, w_2 > b > 0$. For example, $(\begin{pmatrix}1\\1\end{pmatrix}, \frac{1}{2})$ represents f_2.

Suppose $(\boldsymbol{w}, b)$ represents f_n with $n \geq 3$, $\boldsymbol{w} = (w_1, \ldots, w_n)^\intercal \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then

$$
\begin{aligned}
1 &= f_n\big((1,0,0,0,\ldots,0)^\intercal\big) = \operatorname{sgn}(w_1 - b) && \Rightarrow w_1 > b,\\
1 &= f_n\big((0,1,0,0,\ldots,0)^\intercal\big) = \operatorname{sgn}(w_2 - b) && \Rightarrow w_2 > b,\\
1 &= f_n\big((0,0,1,0,\ldots,0)^\intercal\big) = \operatorname{sgn}(w_3 - b) && \Rightarrow w_3 > b, \text{ and}\\
-1 &= f_n\big((1,1,1,0,\ldots,0)^\intercal\big) = \operatorname{sgn}(w_1 + w_2 + w_3 - b) && \Rightarrow w_1 + w_2 + w_3 < b.
\end{aligned}
$$

Thus, $b < w_1 + w_2 + w_3 < b$, a contradiction.

Empty Page for Notes