



1

Profs. Martin Jaggi and Nicolas Flammarion  
Optimization for Machine Learning – CS-439 - IC  
08.07.2021 from 08h15 to 11h15  
Duration : 180 minutes

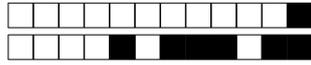
# Student One

SCIPER: 111111

Wait for the start of the exam before turning to the next page. This document is printed double sided, 16 pages. Do not unstaple.

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet (hand-written or 11pt min font size) if you have one; place all other personal items below your desk or on the side.
- You each have a different exam.
- For technical reasons, **do use black or blue pens for the MCQ part, no pencils!** Use white corrector if necessary.

Respectez les consignes suivantes   Observe this guidelines   Beachten Sie bitte die unten stehenden Richtlinien		
choisir une réponse   select an answer Antwort auswählen	ne PAS choisir une réponse   NOT select an answer NICHT Antwort auswählen	Corriger une réponse   Correct an answer Antwort korrigieren
  		 
ce qu'il ne faut <b>PAS</b> faire   what should <b>NOT</b> be done   was man <b>NICHT</b> tun sollte		
     		



## First part, multiple choice

There is **exactly one** correct answer per question.

### Smoothness and gradient descent

**Question 1** Let us define  $f : x \in \mathbb{R} \mapsto \cos(x)$ . We consider  $x_t \in \mathbb{R}$  and  $x_{t+1} = x_t - \nabla f(x_t)$ . Assume that  $x_t$  is not a critical point of  $f$ . Which one of the following statements is **true**:

- $f(x_{t+1}) = -1$
- There exists an  $x_t$  such that  $f(x_{t+1}) > f(x_t)$
- $\text{sign}(x_{t+1}) = \text{sign}(x_t)$
- $\|\nabla f(x_{t+1})\| < \|\nabla f(x_t)\|$
- $x_{t+1} < x_t$
- None of the above

**Question 2** Assume you want to minimize a function  $f : \mathbf{x} \in \mathbb{R}^d \mapsto \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}) \in \mathbb{R}$ , where for each  $i$ ,  $f_i$  is convex and  $L$ -smooth over  $\mathbb{R}^d$ . Which of the following statements is **false**:

- If I use a constant step-size  $\gamma < \frac{1}{L}$ , then GD will converge but not SGD.
- If  $n = 1$ , then SGD and GD correspond to the same recursion.
- If  $n$  is very big then gradient descent can be computationally infeasible.
- SGD corresponds to  $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f_{i_t}(\mathbf{x}_t)$  where  $i_t$  is the remainder of  $t$  divided by  $n$ :  $t = n \lfloor \frac{t}{n} \rfloor + i_t$ .

**Question 3** For  $a > 0$  and  $b \in \mathbb{R}$ , consider  $f(x) = a \cdot x^4 + b$ ,  $x \in \mathbb{R}$ . Assume you perform gradient descent on  $f$  with a constant step-size  $\gamma$ . Which one of the following statements is **true**:

- If  $|x_0| \leq 1$  and  $0 < \gamma \leq 1$  then the iterates converge to 0.
- Depending on my starting point  $x_0$  and my step size, either my iterates  $x_t$  converge to 0, or diverge  $|x_t| \xrightarrow[t \rightarrow \infty]{} +\infty$ .
- For the iterates to converge, my step size must depend on  $b$ .
- For a starting point  $x_0$ , if  $0 < \gamma < \frac{1}{2ax_0^3}$  then the iterates converge to 0.
- For a starting point  $x_0$ , whatever step size I pick, the iterates will never converge to 0.

### Newton's Method and Quasi-Newton

**Question 4** How many steps does the Newton's method require to reach an error smaller than  $\varepsilon > 0$  when minimizing a strictly convex quadratic function:

- It depends on the step size.
- $\mathcal{O}(\log(1/\varepsilon))$
- 1
- It depends on the condition number of the quadratic function.
- $\mathcal{O}(1/\varepsilon)$



**Question 5** We apply Newton's method to a function  $f$  with a critical point  $\mathbf{x}^*$  starting from iterate  $\mathbf{x}_0$ . Assume that  $f$  has bounded inverse Hessians and Lipschitz continuous Hessians. Among the following propositions, what is the extra assumption which allows to show that  $\|\mathbf{x}_T - \mathbf{x}^*\| < \varepsilon$  after  $T = \mathcal{O}(\log \log(1/\varepsilon))$  steps?

- Convexity.
- Smoothness.
- Taking the average iterate.
- Decreasing step size.
- Strong convexity.
- $\|\mathbf{x}_0 - \mathbf{x}^*\|$  should be small.

### Function properties

Consider the function  $d: \mathcal{D} \rightarrow \mathbb{R}$  with  $\mathcal{D} \subseteq \mathbb{R}^2$  defined as  $d(\mathbf{x}) = x_1^2 \cdot x_2^2$ , where  $x_1$  and  $x_2$  are the coordinates of  $\mathbf{x}$ . Let us consider three cases: **(A)** when  $\mathcal{D} = \mathbb{R}^2$ , **(B)** when  $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \leq 1\}$ , and **(C)** when  $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^2 : x_2 = 3\}$ .

**Question 6** In which cases is the function  $d$  convex ?

- C only.
- A, B and C.
- A and C only.
- A only.
- A and B only.
- B and C only.
- B only.
- None of them.

**Question 7** In which cases is the function  $d$   $L$ -smooth in the sense of the definition used in the course?

- C only.
- B and C only.
- A and B only.
- A and C only.
- A only.
- B only.
- A, B and C.
- None of them.



## Coordinate descent

**Question 8** Compared to gradient descent, coordinate descent with gradient-based updates can speed up optimization when coordinate-wise gradients are cheap to compute, and when the coordinates  $i$  ( $i = 1, \dots, d$ ) have varying smoothness constants  $L_i$ . We use coordinate-dependent step sizes  $\eta_i$ , and make a gradient step on coordinate  $i$  with probability  $p_i$ . To obtain a convergence rate that depends on  $\bar{L} = \frac{1}{d} \sum_{i=1}^d L_i$  instead of  $\max_i L_i$ , you would use

- $\eta_i < \eta_j$  and  $p_i < p_j$  if  $L_i > L_j$
- $\eta_i > \eta_j$  and  $p_i < p_j$  if  $L_i > L_j$
- $\eta_i < \eta_j$  and  $p_i > p_j$  if  $L_i > L_j$
- $\eta_i > \eta_j$  and  $p_i > p_j$  if  $L_i > L_j$

## Subgradient descent

**Question 9** The *Leaky ReLU* is an activation function defined as

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \lambda x & \text{if } x \leq 0 \end{cases},$$

where  $\lambda \in (0, 1)$  is a constant. Which of the following values is a subgradient of  $f$  at  $x = 0$ ?

- $\frac{1+\lambda}{2}$
- $-\frac{\lambda}{2}$
- 0
- $\frac{\lambda}{2}$
- 2

## Constrained optimization

Consider the Lasso regression  $\min_{\|\mathbf{x}\|_1 \leq 1} f(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$  where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

**Question 10** When using the Frank-Wolfe algorithm, which of the following points can be the output of the linear minimization oracle  $LMO(\nabla f(\mathbf{x}_0))$  where  $\mathbf{x}_0 = [\frac{1}{2}, \frac{1}{2}]^\top$ ?

- $[-1, 0]^\top$
- $[0, 0]^\top$
- $[1, 0]^\top$
- $[0, 1]^\top$

**Question 11** Which of the following points can be reached by applying 1 step of projected gradient descent, starting from  $\mathbf{x}_0 = [0, 1]^\top$ , with stepsize  $\gamma = 1$ ?

- $[-\frac{1}{2}, -\frac{1}{2}]^\top$
- $[-\frac{2}{3}, \frac{1}{3}]^\top$
- $[1, 0]^\top$
- $[0, 0]^\top$

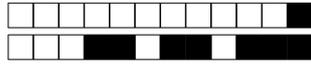


### Proximal gradient descent

**Question 12** For  $h(x) = |x|$ , the soft thresholding operator is defined by the proximal operator  $\mathbf{prox}_{h,t}(u)$ . Then for  $u \geq t > 0$ ,  $\mathbf{prox}_{h,t}(u)$  can be written as which of the following?

- $u + t$
- $0$
- $u$
- $u - t$

DRAFT



## Second part, true/false questions

**Question 13** (Convexity) A function  $f(x)$  is *convex* if and only if  $g(x) = -f(x)$  is *non-convex*.

TRUE       FALSE

**Question 14** (Convexity) Any critical point of a convex differentiable function on an open domain is a global minimizer of the function.

TRUE       FALSE

**Question 15** (Nesterov Accelerated Gradient) Nesterov's accelerated gradient method asymptotically requires fewer update steps than Gradient Descent on smooth convex functions to achieve the same sub-optimality  $\varepsilon$ . To achieve this, the method requires more memory of size  $\mathcal{O}(d^2)$ , where  $d$  is the dimensionality of the parameter vector to be optimized.

TRUE       FALSE

**Question 16** (Subgradient Descent) For strongly convex and non-differentiable function, subgradient descent achieves a  $\mathcal{O}(1/T)$  convergence rate with a small enough constant stepsize.

TRUE       FALSE

**Question 17** (Projected Gradient Descent) Applying projected gradient descent on an Euclidean ball  $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$  is equivalent to gradient descent with adaptive learning rate.

TRUE       FALSE

**Question 18** (Gradient Descent) Let  $f : \mathbf{x} \in \mathbb{R}^d \rightarrow \mathbb{R}$  be an  $L$ -smooth and convex function. We perform gradient descent with step-size  $0 < \gamma < \frac{1}{L}$ , from a starting point  $\mathbf{x}_0$ . Then the iterates will converge towards a point  $\mathbf{x}^*$  with  $f(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ .

TRUE       FALSE

**Question 19** (Frank-Wolfe) Consider  $\min_{(x_1, x_2) \in \mathbb{R}_+^2} |x_1 - 0.1|^2 + |x_2 - 0.1|^2$ , if we apply the Frank-Wolfe algorithm with stepsize  $\gamma := \frac{2}{t+2}$ , then it converge at a rate of  $\mathcal{O}(1/T)$  for any initial iterate.

TRUE       FALSE



Solution:

### Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Do not cross any checkboxes, they are reserved for correction.

#### Bregman Divergence

Let us consider a strictly convex and differentiable function  $h$  on  $\mathbb{R}^d$ . We define the Bregman divergence associated with the function  $h$  by:

$$D_h(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) - h(\mathbf{y}) - \nabla h(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

**Question 20:** 1 point. Show that the function  $\mathbf{x} \mapsto D_h(\mathbf{x}, \mathbf{y})$  is strictly convex, for any fixed  $\mathbf{y}$ .

0  1

**Solution:**  $D_h(\cdot, \mathbf{y})$  is strictly convex as the sum of a the strictly convex function  $h$  and of a linear function  $\nabla h(\mathbf{y})^\top (\cdot - \mathbf{y})$ .

**Question 21:** 1 point. Show that  $D_h(\mathbf{x}, \mathbf{y}) \geq 0$  for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  and that  $D_h(\mathbf{x}, \mathbf{y}) = 0$  if and only if  $\mathbf{x} = \mathbf{y}$ .

0  1

**Solution:** Since  $h$  is convex we have for all  $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ ,  $h(\mathbf{x}) \geq h(\mathbf{y}) + \nabla h(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$ . Since  $h$  is strictly convex  $h(\mathbf{x}) = h(\mathbf{y}) + \nabla h(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$  if and only if  $\mathbf{x} = \mathbf{y}$ .

**Question 22:** 1 point. Compute  $D_{1/2\|\cdot\|_2^2}$ .

0  1

**Solution:**  $D_{1/2\|\cdot\|_2^2}(\mathbf{x}, \mathbf{y}) = 1/2\|\mathbf{x} - \mathbf{y}\|_2^2$

**Question 23:** 2 points. Is  $D_h$  symmetric, i.e.,  $D_h(\mathbf{x}, \mathbf{y}) = D_h(\mathbf{y}, \mathbf{x})$ ? Prove your answer.

0  1  2

**Solution:** No, consider for example  $h(x) = x \log(x)$  and its associated Bregman divergence  $D_h(x, y) = x \log(x/y)$

**Question 24:** 2 point. Let  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$ . Simplify

$$D_h(\mathbf{x}, \mathbf{z}) - D_h(\mathbf{x}, \mathbf{y}) - D_h(\mathbf{y}, \mathbf{z}).$$

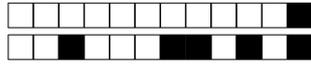
0  1  2

**Solution:**

$$D_h(\mathbf{x}, \mathbf{z}) - D_h(\mathbf{x}, \mathbf{y}) - D_h(\mathbf{y}, \mathbf{z}) = (\nabla h(\mathbf{y}) - \nabla h(\mathbf{z}))^\top (\mathbf{x} - \mathbf{y}).$$

Let us consider now a second convex function  $f$  also defined on  $\mathbb{R}^d$ . We assume that  $f$  is continuously differentiable on  $\mathbb{R}^d$ . We define the following key property

$$\exists L > 0 \text{ such that } L \cdot h - f \text{ is convex on } \mathbb{R}^d. \quad (\text{S})$$



**Question 25:** 2 points. Show that the condition (S) is equivalent to

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + L \cdot D_h(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

0  1  2

**Solution:** The function  $Lh - f$  is convex if and only if

$$(Lh - f)(\mathbf{x}) \geq (Lh - f)(\mathbf{y}) + \nabla(Lh - f)(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}).$$

Rearranging we obtain

$$\begin{aligned} f(\mathbf{x}) &\leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + L(h(\mathbf{x}) - h(\mathbf{y}) - \nabla h(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})) \\ &\leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + LD_h(\mathbf{x}, \mathbf{y}). \end{aligned}$$

**Question 26:** 3 points. Assume condition (S). Show that for any  $\mathbf{y}, \mathbf{x}, \mathbf{z} \in \mathbb{R}^d$  we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{y}) + LD_h(\mathbf{x}, \mathbf{z}).$$

0  1  2  3

**Solution:** Using Question 25 we have

$$f(\mathbf{x}) \leq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) + LD_h(\mathbf{x}, \mathbf{z}).$$

Using the convexity of  $f$  we also have

$$0 \leq f(\mathbf{y}) - f(\mathbf{z}) - \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z})$$

Summing both inequality yields to the result.

## The Mirror Descent Algorithm

We consider now the following update rule defined for a step size  $\gamma \geq 0$  by:

$$T_\gamma(\mathbf{x}) := \arg \min_{\mathbf{u} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{u} - \mathbf{x}) + \frac{1}{\gamma} D_h(\mathbf{u}, \mathbf{x}) \right\}.$$

**Question 27:** 2 points. We assume in this question that  $h = 1/2 \|\cdot\|_2^2$ . Show that  $T_\gamma(\mathbf{x})$  is well defined and compute it. Which algorithm do you recover if you iterate  $\mathbf{x}_{t+1} := T_\gamma(\mathbf{x}_t)$  ?

0  1  2

**Solution:** For this specific choice of  $h$ , the function in the bracket is strongly convex, and therefore has a unique global minimum. We obtain by setting the gradient to zero  $T_\gamma(\mathbf{x}) = \mathbf{x} - \gamma \nabla f(\mathbf{x})$ . We recover the gradient descent algorithm.



We consider that the function  $h$  satisfies additionally the following properties:

- The gradient of  $h$  takes all possible values, i.e.,  $\nabla h(\mathbb{R}^d) = \mathbb{R}^d$ .

We consider the optimization algorithm defined as  $\mathbf{x}_0 \in \mathbb{R}^d$  and which iterates:

$$\mathbf{x}_{t+1} := T_\gamma(\mathbf{x}_t) \text{ for } t \in \mathbb{N}. \quad (\text{MD})$$

This algorithm is called Mirror Descent.

**Question 28:** 3 points. Show that the operator  $T_\gamma$  is well-defined and that, for an appropriate function  $g$  you will give, the recursion can be rewritten as:

$$g(\mathbf{x}_{t+1}) = g(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t).$$

0  1  2  3

**Solution:** Since  $h$  is strictly convex, the objective can have at most one minimizer. Computing the gradient we obtain  $\nabla f(\mathbf{x}) + 1/\gamma(\nabla h(\mathbf{u}) - \nabla h(\mathbf{x}))$ . Using that  $\nabla h(\mathbb{R}^d) = \mathbb{R}^d$  grants the existence of  $u$  such that  $\nabla f(\mathbf{x}) + 1/\gamma(\nabla h(\mathbf{u}) - \nabla h(\mathbf{x})) = 0$ . Therefore such a  $u$  is a stationary point of a differentiable and convex function over an open set and therefore  $u$  is a global minimum of the function. And therefore

$$\nabla h(\mathbf{x}_{t+1}) = \nabla h(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t).$$

### Analysis of The Mirror Descent Algorithm

**Question 29:** 3 points. Let  $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$ . Define  $\mathbf{x}^+ := T_\gamma(\mathbf{x})$  and assume that  $\gamma < 1/L$  where  $L$  is defined in condition (S). Show that

$$\gamma(f(\mathbf{x}^+) - f(\mathbf{u})) \leq D_h(\mathbf{u}, \mathbf{x}) - D_h(\mathbf{u}, \mathbf{x}^+) - (1 - \gamma L)D_h(\mathbf{x}^+, \mathbf{x}).$$

0  1  2  3

**Solution:** Using the three point equality of Question 24 we directly obtain:

$$\begin{aligned} D_h(\mathbf{u}, \mathbf{x}) - D_h(\mathbf{u}, \mathbf{x}^+) - (1 - \gamma L)D_h(\mathbf{x}^+, \mathbf{x}) &= (\nabla h(\mathbf{x}^+) - \nabla h(\mathbf{x}))^\top (\mathbf{u} - \mathbf{x}^+) + \gamma L D_h(\mathbf{x}^+, \mathbf{x}) \\ &= -\gamma \nabla f(\mathbf{x})^\top (\mathbf{u} - \mathbf{x}^+) + \gamma L D_h(\mathbf{x}^+, \mathbf{x}) \\ &\geq \gamma(f(\mathbf{x}^+) - f(\mathbf{u})) \end{aligned}$$

where we have used Question 26 for the last inequality.

**Question 30:** 2 points. Let  $\mathbf{u} \in \mathbb{R}^d$  and consider the iterates defined in equation (MD). We denote the average of the iterates  $\mathbf{x}_t$  by  $\bar{\mathbf{x}}_t = \frac{1}{t} \sum_{i=1}^t \mathbf{x}_i$ . Show the following inequality:

$$f(\bar{\mathbf{x}}_t) - f(\mathbf{u}) \leq \frac{1}{\gamma t} D_h(\mathbf{u}, \mathbf{x}_0)$$

0  1  2

**Solution:** Using the previous question with  $\mathbf{u}, \mathbf{x}_i$  we have that

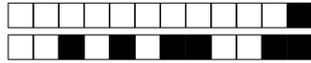
$$\gamma(f(\mathbf{x}_{i+1}) - f(\mathbf{u})) \leq D_h(\mathbf{u}, \mathbf{x}_i) - D_h(\mathbf{u}, \mathbf{x}_{i+1})$$

Summing these for  $i = 0, \dots, t-1$  yields to

$$\gamma \sum_{i=0}^{t-1} (f(\mathbf{x}_{i+1}) - f(\mathbf{u})) \leq D_h(\mathbf{u}, \mathbf{x}_0) - D_h(\mathbf{u}, \mathbf{x}_t),$$

and applying Jensen's inequality we obtain:

$$f\left(\frac{1}{t} \sum_{i=1}^t \mathbf{x}_i\right) - f(\mathbf{u}) \leq \frac{D_h(\mathbf{u}, \mathbf{x}_0) - D_h(\mathbf{u}, \mathbf{x}_t)}{\gamma t},$$



**Question 31:** 3 points. Show a similar inequality for the last iterate  $\mathbf{x}_t$ :

$$f(\mathbf{x}_t) - f(\mathbf{u}) \leq \frac{1}{\gamma t} D_h(\mathbf{u}, \mathbf{x}_0)$$

0  1  2  3

**Solution:** Using Question 30 with  $\mathbf{u} = \mathbf{x}_t$  we obtain that the function values are decreasing and therefore the result on the last iterate comes from the result on the averaged iterates.

**Question 32:** 2 points. Does the inequality proved in Question 32 imply convergence  $f(\mathbf{x}_t) \rightarrow f(\mathbf{u})$ ? Prove your answer.

0  1  2

**Solution:** No, it does not because we do not have that  $f(\mathbf{x}_t) - f(\mathbf{u}) \geq 0$

**Question 33:** 2 points. Let us assume that  $\arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \neq \emptyset$ . Show that for any solution  $\mathbf{x}_* \in \arg \min_{\mathbf{x} \in \mathbb{R}^d} f$ ,

$$f(\mathbf{x}_t) - \min_{\mathbb{R}^d} f \leq \frac{1}{\gamma t} D_h(\mathbf{x}_*, \mathbf{x}_0)$$

Does this inequality imply convergence  $f(\mathbf{x}_t) \rightarrow f(\mathbf{x}_*)$ ? Prove your answer.

0  1  2

**Solution:** ...



## Application to Poisson Linear Inverse Problems

Let us denote by  $\mathbb{R}_+ = \{x \in \mathbb{R}, x \geq 0\}$  and  $\mathbb{R}_{+*} = \{x \in \mathbb{R}, x > 0\}$ . Given a matrix  $A \in \mathbb{R}_{+*}^{m \times n}$  and a vector  $\mathbf{b} \in \mathbb{R}_{+*}^m$ , the goal is to reconstruct a signal  $\mathbf{x} \in \mathbb{R}_+^n$  such that

$$A\mathbf{x} \simeq \mathbf{b}.$$

A natural way of recovering  $\mathbf{x}$  is to minimize the Kullback-Leibler divergence

$$\min_{\mathbf{x} \in \mathbb{R}_+^n} f(\mathbf{x}) := \sum_{i=1}^m b_i \log \frac{b_i}{(\mathbf{A}\mathbf{x})_i} + (\mathbf{A}\mathbf{x})_i - b_i,$$

where  $b_i$  is the  $i$ -th coordinate of the vector  $\mathbf{b}$ .

**Question 34:** 2 points. Show that the function  $f$  is convex over  $\mathbb{R}_{+*}^n$ .

0  1  2

**Solution:** We note that the function  $f$  is the Bregman divergence associated with the Shannon entropy  $u \mapsto u \log u$  which is convex. You can also derive this with composition rules of convex functions, starting from  $-\log x$  being convex.

**Question 35:** 3 points. Is the function  $f$  smooth on  $\mathbb{R}_{+*}^n$ ? Justify your answer.

0  1  2  3

**Solution:** We can compute the Hessian of  $f$ :

$$\nabla^2 f(\mathbf{x}) = \sum_{i=1}^m b_i \frac{\mathbf{a}_i \mathbf{a}_i^\top}{(\mathbf{a}_i^\top \mathbf{x})^2},$$

where  $\mathbf{a}_i$  are the rows of the matrix  $\mathbf{A}$ . We note that  $\nabla^2 f(\mathbf{x})$  is diverging to  $\infty$  when  $\mathbf{x}$  converge to the boundary 0 and therefore the function is not globally smooth.

Let us denote by  $\mathbf{a}_i$  the  $i$ -th row of the matrix  $\mathbf{A}$ . We assume that  $\mathbf{a}_i \neq 0$  and  $r_j := \sum_{i=1}^m a_{ij} > 0$  for all  $j$ . Let us consider Burg's entropy defined by  $h(\mathbf{x}) := -\sum_{j=1}^n \log x_j$  on  $\mathbb{R}_{+*}^n$ .

**Question 36:** 5 points. Show that for any  $L$  satisfying  $L \geq \|\mathbf{b}\|_1$ , the function  $Lh - f$  is convex on  $\mathbb{R}_{+*}^n$  (HINT: you can compute the Hessian and show that it is positive semi-definite.)

0  1  2  3  4  5

**Solution:** Since  $f$  and  $h$  are  $C^2$  we can just compute the Hessian. First

$$\mathbf{d}^\top \nabla^2 h(\mathbf{x}) \mathbf{d} = \sum_{j=1}^n \frac{d_j^2}{x_j^2}. \quad (1)$$

Moreover using the definition of  $f$ ,  $\nabla f(\mathbf{x}) = \sum_{i=1}^m (1 - \frac{b_i}{\mathbf{a}_i^\top \mathbf{x}}) \mathbf{a}_i$  and thus

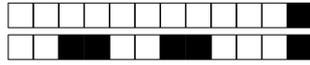
$$\mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d} = \sum_{i=1}^m b_i \frac{(\mathbf{a}_i^\top \mathbf{d})^2}{(\mathbf{a}_i^\top \mathbf{x})^2} \quad (2)$$

Now using the Jensen's inequality to  $t^2$  we have

$$\frac{(\mathbf{u}^\top \mathbf{d})^2}{(\mathbf{u}^\top \mathbf{x})^2} \leq \sum_j \frac{u_j x_j}{\mathbf{u}^\top \mathbf{x}} (d_j/x_j)^2 \leq \sum_j (d_j/x_j)^2 \quad (3)$$

Applying the previous inequality to  $u := \mathbf{a}_i \neq 0$  we obtain

$$\mathbf{d}^\top \nabla^2 g(\mathbf{x}) \mathbf{d} = \sum_{j=1}^n \frac{d_j^2}{x_j^2} \leq \sum_i b_i \sum_j (d_j/x_j)^2 \quad (4)$$



We will minimize the function  $f$  using the Mirror Descent algorithm and the potential  $h$  whose update rule is defined similarly by<sup>1</sup>

$$\mathbf{x}_{t+1} := T_\gamma(\mathbf{x}_t) \text{ for } t \in \mathbb{N}, \text{ where } T_\gamma(\mathbf{x}) := \arg \min_{\mathbf{u} \in \mathbb{R}_{+*}^d} \left\{ f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{u} - \mathbf{x}) + \frac{1}{\gamma} D_h(\mathbf{u}, \mathbf{x}) \right\}. \quad (\text{MDA})$$

**Question 37:** 4 points. Show that  $T_\gamma(\mathbf{x})$  is well defined for  $\gamma \leq \frac{1}{\|\mathbf{b}\|_1}$ . Find a closed form expression for the recursion defined in Eq. (MDA).

0  1  2  3  4

**Solution:** The function  $f$  is separable, and therefore the iterates defined in Eq.(MDA) reduces to solve the one dimensional convex problem:

$$x^+ = \arg \min_{u > 0} \left\{ gu + \frac{1}{\gamma} \left( \frac{u}{x} - \log \frac{u}{v} \right) \right\},$$

for  $u > 0$  and where  $g$  is one coordinate of the gradient  $\nabla f(\mathbf{x})$ . When  $u \rightarrow 0$ , the function is going to  $+\infty$ . Let us show that  $\gamma gx + 1 > 0$  such that the function is also going to  $+\infty$  when  $u \rightarrow +\infty$  and  $x^+ = \frac{x}{1+\gamma gx}$  is well defined. By definition of  $\nabla f$  and  $\gamma$ , the condition writes:

$$\begin{aligned} 1/\gamma + \sum_i a_{i,j} x_j - \sum_i \frac{b_i a_{i,j}}{\sum_j a_{i,j} x_j} x_j &\geq \sum_i b_i + \sum_i a_{i,j} x_j - \sum_j \frac{b_i a_{i,j}}{\sum_i a_{i,j} x_j} x_j \\ &\geq \sum_i b_i \left( 1 - \frac{a_{i,j} x_j}{\sum_j a_{i,j} x_j} \right) + \sum_i a_{i,j} x_j > 0 \end{aligned}$$

Also justify that it is  $> 0$ .

Define  $c_j(\mathbf{x}) := \sum_{i=1}^m b_i \frac{a_{i,j}}{\mathbf{a}_i^\top \mathbf{x}}$ . The  $j$ -th component of the gradient of  $\nabla f(\mathbf{x})$  can be written as

$$g_j = r_j - c_j(\mathbf{x}),$$

And the iterate writes:

$$x_j^+ = \frac{x_j}{1 + \gamma x_j (r_j - c_j(\mathbf{x}))},$$

for  $j = 1, \dots, n$ .

**Question 38:** 3 points. Assuming that you can apply the results derived in Question 34, which convergence rate do you obtain with this algorithm on this problem? Why is it surprising?

0  1  2  3

**Solution:** We obtain a  $1/t$  convergence for a non-smooth problem!

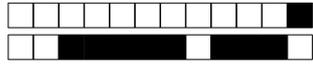
<sup>1</sup>Note that the update rule here is similar to the standard update rule but the minimum is now taken over  $\mathbb{R}_{+*}^d$ .



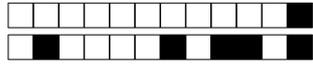
DRAFT



DRAFT



DRAFT



DRAFT