Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Institute of Theoretical Computer Science
Bernd Gärtner, David Steurer

Optimization for Data Science Final Exam (8 August 2020) FS20

Candidate

First name: ....

Last name: ....

Student ID (Legi) Nr.: ....

I attest with my signature that I was able to take the exam under regular conditions and that I have read and understood the general remarks below.

Signature: ....

General remarks and instructions

1. Check your exam documents for completeness (pages numbered from 1 to 13).
2. You can solve the exercises in any order. They are not ordered by difficulty.
3. Immediately inform an assistant in case you are not able to take the exam under regular conditions. Later complaints will not be accepted.
4. Pencils are not allowed. Pencil-written solutions will not be reviewed.
5. No auxiliary material allowed. All electronic devices must be turned off and are not allowed to be on your desk or carried with you to the toilet. We will write the current time on the blackboard every 15 minutes.
6. Attempts to cheat/defraud lead to immediate exclusion from the exam and can have judicial consequences.
7. Provide only one solution to each exercise. Cancel invalid solutions clearly.
8. **All solutions must be understandable and well-founded. Write down the important thoughts in clear sentences and keywords. Unless stated otherwise, no points will be awarded for unfounded or incomprehensible solutions. Please write your solutions in English.**
9. You may use anything that has been introduced and proved in the lecture or in the exercise sessions without reproving it. However, if you need something *different* than what we have shown, you must write a new proof or at least list all necessary changes.

Good luck!

	achieved points (maximum)
1	(15)
2	(9)
3	(11)
4	(18)
5	(8)
6	(9)
Σ	(70)

Useful Facts

In your solutions, you may use the following facts:

Fact 1. *Let* $h : \mathbf{dom}(h) \to \mathbb{R}$ *be convex and differentiable, and let* $X \subseteq \mathbf{dom}(h)$ *be a convex set. Then* x^* *is a minimizer of* h *over* X *if and only if*

$$\nabla h(x^*)^\top (x - x^*) \geq 0 \qquad \forall x \in X.$$

Fact 2. *Let* $G \sim N(0,1)^{m \times m}$ *for* $m \geq 1$. *Then*

$$\mathbb{P}\left(\|G\| \geq 10\sqrt{m} \right) \leq 0.001\,.$$

Fact 3. *Let* $M, m \in \mathbb{N}$ *and* $G \sim N(0,1)^{M \times m}$. *Then*

$$\mathbb{P}\left(\left\| \frac{1}{M} G^\top G - \mathrm{Id}_m \right\| \leq 0.001 \right) \geq 1 - 2\exp\left(-\frac{M}{10^6 m} \right)\,.$$

Assignment 1. **(15 points)** *Define*

$$\Delta_n := \{x \in \mathbb{R}^n : x > \mathbf{0},\ \mathbf{1}^\top x = 1\}.$$

where $\mathbf{1}$ *is the* n*-dimensional all-one vector.*

Let a *be a given* n*-dimensional vector and* $f : \mathbb{R}^n_+ \to \mathbb{R}$ *be the function given by*

$$f(x) = \sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} a_i x_i.$$

(a) **(2 points)** *Prove that* Δ_n *is convex.*

(b) **(2 points)** *Let* $g : \mathbf{dom}(g) \to \mathbb{R}$ *be convex and differentiable. Prove that if there exists* $c \in \mathbb{R}$ *such that* $\nabla g(x^*) = c\mathbf{1}$*, then* x^* *is a minimizer of* g *over* Δ_n*.*

(c) **(4 points)** *Prove that the reverse direction of Part (b) is also true. That is, if* x^* *is a minimizer of* g *over* Δ_n*, then there exists* $c \in \mathbb{R}$ *such that* $\nabla g(x^*) = c\mathbf{1}$*. Hint: If two coordinates of* $\nabla g(x^*)$ *are not the same then you may find a contradiction to one of the useful facts.*

(d) **(4 points)** *Prove that* $f(x)$ *is convex.*

(e) **(3 points)** *Prove that there exists an optimal solution to the problem* $\min\{f(x) : x \in \Delta_n\}$ *and compute that optimal solution.*

Solution:

(a) For $x, y \in \Delta_n$ and $\lambda \in [0,1]$, we have $z = \lambda x + (1-\lambda)y > \mathbf{0}$ since $x, y > \mathbf{0}$, and

$$\mathbf{1}^\top z = \lambda \mathbf{1}^\top x + (1-\lambda)\mathbf{1}^\top y = 1.$$

Therefore, Δ_n is convex.

(b) We have

$$\nabla g(x^*)^\top (x - x^*) = c\mathbf{1}^\top (x - x^*) = c\mathbf{1}^\top x - c\mathbf{1}x^* = c - c = 0.$$

Hence, x^* is a minimizer of g over Δ_n, by Fact 1.

(c) Let y and z be the vectors which agree with x^* on all coordinates except for two coordinates i, j where $y_i = x_i + t$, $y_j = x_j - t$, $z_i = x_i - t$ and $z_j = x_j + t$ for some $t > 0$ such that $t < \min\{x_i, x_j\}$. It is easy to see that $y, z \in \Delta_n$.

By Fact 1, we have for all $x \in \Delta_n$,

$$\nabla g(x^*)^\top (x - x^*) \geq 0.$$

Consider $x = y$, we have

$$\nabla g(x^*)^\top (y - x^*) = t\nabla g_i(x^*) - t\nabla g_j(x^*) \geq 0.$$

Similarly, consider $x = z$, we have

$$\nabla g(x^*)^\top (z - x^*) = -t\nabla g_i(x^*) + t\nabla g_j(x^*) \geq 0.$$

From the above two equations, we have $\nabla g_i(x^*) = \nabla g_j(x^*)$.

As the above holds for any $i, j \in [n]$, it follows that all coordinates of $\nabla g(x^*)$ have the same value. Hence, $\nabla g(x^*) = c\mathbf{1}$ for some $c \in \mathbb{R}$.

(d) Firstly, observe that $\mathbf{dom}(f) = \mathbb{R}_+$ is convex. Next, we have

$$\frac{\partial f}{\partial x_i}(x) = \log x_i + 1 - a_i.$$

Hence, for $i \in [n]$,

$$\frac{\partial^2 f}{\partial x_i \partial x_i}(x) = \frac{1}{x_i}, \text{ and}$$
$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = 0 \text{ for } j \neq i.$$

Therefore, the eigenvalues of $\nabla^2 f(x)$ are $\frac{1}{x_i} > 0$. This implies that $\nabla^2 f(x) \succ \mathbf{0}$. By the second-order characterization of convexity, we can conclude that $f(x)$ is convex.

(e) Suppose there is an optimal solution x^*. Applying (c), we have for some $c \in \mathbb{R}$ and any $i \in [n]$,

$$\log x_i^* + 1 - a_i = c.$$

Therefore, for any $i \in [n]$,

$$x_i^* = e^{c-1+a_i} = e^{c-1}e^{a_i}.$$

Since $\mathbf{1}^\top x^* = 1$, we have

$$e^{c-1} = \frac{1}{\sum_{j=1}^n e^{a_j}}.$$

It follows that for any $i \in [n]$,

$$x_i^* = \frac{e^{a_i}}{\sum_{j=1}^n e^{a_j}}.$$

On the other hand, the value of x, where

$$x_i = \frac{e^{a_i}}{\sum_{j=1}^n e^{a_j}},$$

satisfies that for any $i \in [n]$, $\log x_i + 1 - a_i = c$, where c is defined above. Hence, applying part (b), we have that value x is an optimal solution. Therefore, there exists an optimal solution of f over Δ_n.

Assignment 2. **(9 points)** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and let X be a closed, convex set. Consider the following optimization problem*

$$\begin{array}{ll} \textit{minimize} & f(x) \\ \textit{subject to} & x \in X. \end{array} \tag{1}$$

To solve this problem, we introduced Projected Gradient Descent with update rule for $\gamma > 0$,

$$\begin{aligned} y_{t+1} &:= x_t - \gamma \nabla f(x_t) \\ x_{t+1} &:= \Pi_X(y_{t+1}) := \operatorname*{argmin}_{x \in X} \|x - y_{t+1}\|^2 \end{aligned}$$

as an optimization technique to solve this problem.

A potential issue is that we may get "stuck", i.e., for some t, we have $x_{t+1} = x_t$. In this assignment, you will show that this is actually not an issue, as when it occurs, we have found the optimal solution.

More specifically, prove that x^* is an optimal solution to (1) if and only if $x^* = \Pi_X(x^* - \gamma \nabla f(x^*))$.

Hint: There are actually two constrained optimization problems.

Solution: By Fact 1, x^* is the optimal solution of Problem (1), if and only if $\nabla f(x^*)^\top (y - x^*) \geq 0$ for all $y \in X$.

Let $g(x) = \|x - (x^* - \gamma \nabla f(x^*))\|^2$. Observe that $\Pi_X(x^* - \gamma \nabla f(x^*))$ is the minimizer of the problem

$$\min\{g(x) : x \in X\}. \tag{2}$$

Again, by Fact 1, a vector w is the optimal solution to (??), if and only if $\nabla g(w)^\top (y - w) \geq 0$ for all $y \in X$.

Observe that $\nabla g(x^*)^\top (y - x^*) = 2\gamma \nabla f(x^*)^\top (y - x^*)$, for all $y \in X$. Since $\gamma > 0$, $\nabla g(x^*)^\top (y - x^*) \geq 0$ if and only if $\nabla f(x^*)^\top (y - x^*) \geq 0$.

Putting all the above together, we obtain that the following statements are equivalent:

- x^* is the optimal solution of Problem (1);
- $\nabla f(x^*)^\top (y - x^*) \geq 0$;
- $\nabla g(x^*)^\top (y - x^*) \geq 0$; and
- x^* is the optimal solution of Problem (??), that is $x^* = \Pi_X(x^* - \gamma \nabla f(x^*))$.

The result then follows.

Assignment 3. **(11 points)** *In class, we learned that the Newton's method with step size 1 can find the minimizer of a quadratic function in one step. In this assignment, we will investigate what happens when we vary the step size.*

Consider the Newton's method with step size γ, where the update step is as follows:

$$x_{t+1} := x_t - \gamma \nabla^2 f(x_t)^{-1} \nabla f(x_t), \qquad t \geq 0.$$

Suppose $M \in \mathbb{R}^{n \times n}$ is positive definite and $x^* \in \mathbb{R}^n$ is a given vector. Consider the function

$$f(x) = \frac{1}{2}(x - x^*)^\top M (x - x^*).$$

Answer the following questions and provide justifications for your answers.

(a) **(2 points)** *Prove that $f(x)$ has only one minimizer and that minimizer is x^*.*

(b) **(3 points)** *If we run the Newton's method with some initial vector $x_0 \in \mathbb{R}^n$, for which values of γ does the sequence $(x_t)_{t \geq 0}$ converge to x^*?*

(c) **(3 points)** *For the values of γ in Part (b), how many iterations are required to obtain an $x \in \mathbb{R}^n$ such that $\|x - x^*\| < \epsilon$ for some $\epsilon > 0$?*

(d) **(3 points)** *For the values of γ in Part (b), how many iterations are required to obtain an $x \in \mathbb{R}^n$ such that $f(x) - f(x^*) < \epsilon$ for some $\epsilon > 0$?*

Solution:

(a) Since M is positive definite, by the definition of positive definiteness, we have for $x \neq x^*$, $f(x) > 0$. Further, $f(x^*) = 0$. Hence, f only has one minimizer x^*.

(b) Calculating the gradient and the Hessian, we obtain

$$\nabla f(x) = M(x - x^*),$$
$$\nabla^2 f(x) = M.$$

Therefore, the update step becomes

$$\begin{aligned} x_{t+1} &= x_t - \gamma M^{-1} M (x - x^*) \\ &= x_t - \gamma (x - x^*). \end{aligned}$$

Hence, $x_{t+1} - x^* = (x_t - x^*)(1 - \gamma)$. Applying this many times, we obtain

$$x_{t+1} - x^* = (x_0 - x^*)(1 - \gamma)^{t+1}.$$

Since x_0 and x^* are given, the sequence $\{x_t\}_{t \geq 0}$ converges to x^* if and only if $|1 - \gamma| < 1$ or $0 < \gamma < 2$.

(c) We have

$$||x_t - x^*|| = (1-\gamma)^t ||x_0 - x^*||.$$

For $\gamma = 1$, we note that $x_1 = x^*$, so we only need one step.

For $\gamma \neq 1$, in order to achieve $||x_t - x^*|| < \epsilon$ for some t, we should have

$$(1-\gamma)^t ||x_0 - x^*|| < \epsilon.$$

This is equivalent to

$$t \geq \frac{1}{-\log(1-\gamma)} \log \frac{||x_0 - x^*||}{\epsilon}.$$

(d)

$$f(x_t) - f(x^*) = (1-\gamma)^{2t} (x_0 - x^*)^\top M (x_0 - x^*).$$

Again, for $\gamma = 1$, we note that $x_1 = x^*$, so we only need one step.

For $\gamma \neq 1$, in order to achieve $f(x_t) - f(x^*) < \epsilon$, noting that $f(x^*) = 0$, we should have

$$\frac{1}{2}(1-\gamma)^{2t} (x_0 - x^*)^\top M (x_0 - x^*) \leq \epsilon.$$

This is equivalent to

$$t \geq \frac{1}{-\log((1-\gamma)^2)} \log \frac{(x_0 - x^*)^\top M (x_0 - x^*)}{2\epsilon}.$$

Assignment 4. (18 points) *Let* $n \geq 10^{10} \cdot d$ *and suppose that* $n/d \in \mathbb{N}$. *Let* $W \in \mathbb{R}^{n\times d}$ *be a matrix with i.i.d. rows* $w_1^\top, \ldots, w_n^\top$, *where* $w_1, \ldots, w_n \sim N(0, \Sigma)$ *for some positive definite matrix* $\Sigma \in \mathbb{R}^{d\times d}$ *such that* $\|\Sigma\| \geq 1$.

(a) **(7 points)** *Show that with probability at least* 0.999,

$$\left\| \frac{1}{n} \sum_{i\in[n]} w_i w_i^\top - \Sigma \right\| \leq 0.001 \cdot \|\Sigma\| \,.$$

(b) **(5 points)** *Let* $X \in \mathbb{R}^{n\times d}$ *be a matrix with rows* $x_1^\top, \ldots, x_n^\top$ *obtained as follows: let* $S_1, \ldots, S_d \subset [n]$ *be disjoint sets, each of size* n/d. *For all* $i \in S_j \subset [n]$, *set row* $x_i^\top = e_j^\top$ *where* e_j *is the* j*-th vector of the standard basis.*

Show that with probability at least 0.999,

$$\left\| \frac{1}{n} \sum_{i\in[n]} w_i x_i^\top \right\| \leq 0.001 \cdot \sqrt{\|\Sigma\|} \,,$$

where $w_1, \ldots, w_n \sim N(0, \Sigma)$ *as above.*

(c) **(6 points)** *Suppose* $d \geq 1000$, $n \geq 10^{10} \cdot d$. *Consider a matrix of the form* $Y = X + W$, *where* X, W *are defined as above. Assuming Part (a) and Part (b), show that with probability at least* 0.99,

$$\left\| \frac{1}{n} \sum_{i\in[n]} y_i y_i^\top - \Sigma \right\| \leq 0.01 \cdot \|\Sigma\| \,,$$

where $y_1^\top, \ldots, y_n^\top$ *are the rows of* Y.

Solution:

(a) Notice that

$$\frac{1}{n} \sum_{i\in[n]} w_i {w_i}^\top - \Sigma = \Sigma^{1/2} \left(\frac{1}{n} \sum_{i\in[n]} g_i {g_i}^\top - \mathrm{Id} \right) \Sigma^{1/2} \,,$$

where $g_1, \ldots, g_n \sim N(0, \mathrm{Id})$. Thus by Fact 3 with probability at least 0.999,

$$\left\| \Sigma^{1/2} \left(\frac{1}{n} \sum_{i\in[n]} g_i {g_i}^\top - \mathrm{Id} \right) \Sigma^{1/2} \right\| \leq \|\Sigma\| \left\| \frac{1}{n} \sum_{i\in[n]} g_i {g_i}^\top - \mathrm{Id} \right\| \leq 0.001 \cdot \|\Sigma\| \,.$$

(b) Notice that

$$\frac{1}{n} \sum_{i\in[n]} w_i {x_i}^\top = \frac{1}{n} \Sigma^{1/2} \sum_{i\in[n]} g_i {x_i}^\top \,,$$

where $g_1, \dots, g_n \sim N(0, \mathrm{Id})$. A matrix $\sum\limits_{i\in[n]} g_i {x_i}^\top$ has iid Gaussian entries with mean 0 and variance n/d. Hence by Fact 2 with probability at least 0.999,

$$\left\|\frac{1}{n}\sum_{i\in[n]} w_i {x_i}^\top\right\| \le \frac{1}{n}\cdot\sqrt{\|\Sigma\|}\cdot 10\sqrt{d}\cdot\sqrt{\frac{n}{d}} \le \frac{10}{\sqrt{n}}\cdot\sqrt{\|\Sigma\|} \le 0.001\cdot\sqrt{\|\Sigma\|}\,.$$

(c) Opening the product,

$$\frac{1}{n}\sum_{i\in[n]} y_i {y_i}^\top = \frac{1}{n}\sum_{i\in[n]} \left(x_i {x_i}^\top + w_i {w_i}^\top + x_i {w_i}^\top + w_i {x_i}^\top\right)\,.$$

So by triangle inequality,

$$\left\|\frac{1}{n}\sum_{i\in[n]} y_i {y_i}^\top - \Sigma\right\| \le \left\|\frac{1}{n}\sum_{i\in[n]} w_i {w_i}^\top - \Sigma\right\| + \left\|\frac{1}{n}\sum_{i\in[n]} x_i {x_i}^\top\right\| + \left\|\frac{1}{n}\sum_{i\in[n]} x_i {w_i}^\top\right\| + \left\|\frac{1}{n}\sum_{i\in[n]} w_i {x_i}^\top\right\|\,.$$

We bound each term separately. Observe that by construction,

$$\frac{1}{n}\sum_{i\in[n]} x_i {x_i}^\top = \frac{1}{d}\mathrm{Id}\,.$$

Combining with the bounds from Parts (a) and (b) and the fact that $\|\Sigma\| \ge 1$, we get the desired result.

Assignment 5. (**8 points**) *Let* $Z \in \mathbb{R}^{n\times d}$ *be a matrix with linearly independent columns (known to the data analyst).*

Consider the following family of an n*-by-*d *matrix-valued random variables*

$$\left\{ \mathbf{Y} = ZX + \mathbf{W} \mid X \in \mathbb{R}^{d\times d},\ \mathbf{W}_{ij} \overset{i.i.d.}{\sim} N(0,1) \right\}.$$

Given a realization of a random variable $\mathbf{Y} = ZX + \mathbf{W}$ *in this family for unknown* X*, the goal is to estimate the matrix* ZX *in Frobenius norm.*

Let $\hat{X} : \mathbb{R}^{n\times d} \to \mathbb{R}^{d\times d}$ *be an arbitrary estimator for this problem. Show that for every* $\epsilon > 0$*, there exists a random variable* $\mathbf{Y} = ZX + \mathbf{W}$ *in the above family such that*

$$\frac{1}{nd}\mathbb{E}\left\| ZX - Z\hat{X}(\mathbf{Y}) \right\|_F^2 \geq (1-\epsilon)\cdot\frac{d}{n}.$$

Remark: We have discussed statistical lower bounds for general estimators in the lectures and in the exercises. The lecture notes also discussed lower bounds for the special case of unbiased estimators. You can also get full points if you only prove the above lower bound for unbiased estimators.

Solution: This problem is a special case of linear regression: if we consider $\mathbf{Y}$, X and $\mathbf{W}$ as vectors $y' \in \mathbb{R}^{nd}$, $x' \in \mathbb{R}^{d^2}$ and $w' \in \mathbb{R}^{nd}$, then we can write $y' = Z'x' + w'$, where Z' is an $nd \times d^2$ block-diagonal matrix with blocks equal to Z. Notice that Z' has linearly independent columns, hence using the lower bound for linear regression, we get the desired bound.

Assignment 6. (9 points) *Consider the following variant of the matrix completion problem in which we see* $n \geq d \log d$ *independent samples of the form*

$$y_i = \langle X_i, \Theta^* + E \rangle + w_i \,,$$

where $X_i = e_a e_b^\top$ *for some* a, b *sampled independently and uniformly at random from* $\{1, \ldots, d\}$*,* $\Theta^* \in \mathbb{R}^{d \times d}$ *satisfies* $\operatorname{rank}(\Theta^*) \leq r$*,* $\|\Theta^*\|_\infty \leq R$*,* $w_1, \ldots, w_n \sim N(0, \mathrm{Id}_d)$ *and* $E \in \mathbb{R}^{d \times d}$ *is an arbitrary matrix (possibly dependent on* $\Theta^*, X_1, \ldots, X_n, w_1, \ldots, w_n$*) satisfying* $\|E\|_\infty \leq \sqrt{\frac{R \cdot \log d}{n \cdot d}}$*. Note that the difference with the problem discussed in class is the additional error matrix* $E \in \mathbb{R}^{d \times d}$*.*

Let

$$Y := \frac{1}{n} \sum_{i \in [n]} y_i \cdot d^2 \cdot X_i \,.$$

Consider the estimator

$$\hat{\Theta} := \arg\min \left\{ \|\Theta - Y\|_F^2 \;\middle|\; \|\Theta\|_* \leq r, \|\Theta\|_\infty \leq R \right\}$$

Show that with probability at least 0.99*, despite the presence of* E*,*

$$\left\| \hat{\Theta} - \Theta^* \right\|_F^2 \leq O\left(\frac{d^3 \cdot \log d}{n} \cdot R \cdot r \right) \,.$$

As shown in the lecture notes, you may directly assume the following facts

F.1 It holds that $\left\| \hat{\Theta} - \Theta^* \right\|_F^2 \leq 10r \|Y - \Theta^*\|^2$*.*

F.2 With probability at least 0.99*, it holds that*

$$\left\| \frac{1}{n} \sum_{i \in [n]} \left(\langle X_i, \Theta^* \rangle + w_i \right) \cdot d^2 \cdot X_i - \Theta^* \right\|^2 \leq O\left(\frac{d^3 \log d}{n} \cdot R \right) \,.$$

Solution: We may rewrite

$$\begin{aligned} Y - \Theta^* &= \frac{1}{n} \sum_{i \in [n]} y_i \cdot d^2 \cdot X_i - \Theta^* \\ &= \frac{1}{n} \sum_{i \in [n]} \Big(\left(\langle X_i, \Theta^* \rangle + w_i \right) \cdot d^2 \cdot X_i - \Theta^* \Big) + \frac{1}{n} \sum_{i \in [n]} \langle X_i, E \rangle \cdot d^2 \cdot X_i \,. \end{aligned}$$

Using the hint, it remains to bound the last term.

$$\left\| \frac{1}{n} \sum_{i \in [n]} \langle X_i, E \rangle \cdot d^2 \cdot X_i \right\| \leq \frac{1}{n} \sum_{i \in [n]} d^2 \cdot |\langle X_i, E \rangle| \leq d^2 \cdot \|E\|_\infty \leq \sqrt{\frac{d^3 \cdot \log d}{n} \cdot R} \,.$$

Putting everything together, the result follows.