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 **Special Assignment 1** **FS20**

- The solution is due on **Friday, April 3, 2020** by **11:59 pm**. Please submit your solution as PDF on Moodle. The name of the file should follow the format `SA1-{Legi number}`, e.g., `SA1-19-123-456`. Please press "Submit assignment", after uploading your solution and make sure that the status is "Submitted" and not "Draft". You should then receive an automatic email that confirms the submission.

- If you submit your solution within six hours before the deadline and a technical problem prevents you from submitting on Moodle, you can send your solution as PDF to `hung.hoang@inf.ethz.ch`. The cutoff time still needs to be observed. If there is any problem with the submission, complain timely.

- Please solve the exercises carefully and then write a nice and complete exposition of your solution using a computer, where we strongly recommend to use LaTeX. A tutorial can be found at `http://www.cadmo.ethz.ch/education/thesis/latex`. Handwritten solutions will not be graded!

- For geometric drawings that can easily be integrated into LaTeX documents, we recommend the drawing editor IPE, retrievable at `http://ipe7.sourceforge.net/` in source code and as an executable for Windows.

- Keep in mind the following premises:
  - When writing in English, write short and simple sentences.
  - When writing a proof, write precise statements.

  The conclusion is, of course, that your solution should consist of sentences that are short, simple, and precise!

- This is a theory course, which means: if an exercise does not explicitly say "you do not need to prove your answer" or "justify intuitively", then a formal proof is **always** required. You can of course refer in your solutions to the lecture notes and to the exercises, if a result you need has already been proved there.

- We would like to stress that the ETH Disciplinary Code applies to this special assignment as it constitutes part of your final grade. The only exception we make to the Code is that we encourage you to verbally discuss the tasks with your colleagues. It is strictly prohibited to share any (hand)written or electronic (partial) solutions with any of your colleagues. We are obligated to inform the Rector of any violations of the Code.

- There will be two special assignments and an exam. We will assign percentages to each of them (i.e., the ratio of the points you obtain over the maximum points). The weighted average will then be converted to a grade on the usual scale from 1 to 6. That is, if S_1 and S_2 are the percentages from your respective special assignments and E is the percentage from your exam, then your weighted percentage will be $P = 0.1 \cdot S_1 + 0.1 \cdot S_2 + 0.8 \cdot E$. P will then be converted to your final grade. If you do not hand in one of the special assignments, it will be awarded 0%.
- As with all exercises, the material of the special assignments is relevant for the exam.

Separating Hyperplane of Data Points

Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$ for $i \in [m]$. We say a vector $\mathbf{w}$ in $\mathbb{R}^d$ *separates* the elements of S, if $y_i \mathbf{w}^\top x_i > 0$ for all $i \in [m]$. In other words, an open halfspace defined by the hyperplane $\mathbf{w}^\top x = 0$ contains all the points in the set $\{x_i \mid i \in [m], y_i = 1\}$, while the other open halfspace contains all points in the set $\{x_i \mid i \in [m], y_i = -1\}$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be the function such that

$$f(\mathbf{w}) = \max_{i \in [m]} (-y_i \mathbf{w}^\top x_i).$$

Suppose that there exists $\mathbf{w}$ that separates the elements of S. The purpose of this exercise is to use subgradient descent to find such a vector.

For this, we will use the lemma below.

Lemma 1. *Let $g : \mathbb{R}^d \to R$ be the function such that $g(x) = \max_{i \in [m]} g_i(x)$, for m convex differentiable functions $g_1, \ldots, g_m$. At any point $x \in \mathbb{R}^d$, let $k(x) \in [m]$ be such that $g_{k(x)}(x) = \max_{i \in [m]} g_i(x)$. Then $\nabla g_{k(x)}(x)$ is a subgradient of $g(x)$.*

Assignment 1. (10 points) *Prove Lemma 1 and apply the lemma to calculate a subgradient of f.*

Assignment 2. (7 points) *Let $B := \max_{i \in [m]} \|x_i\|$. Prove that f is convex and B-Lipschitz.*

Assignment 3. (10 points) *Let $\mathbf{w}^{(1)}$ be the vector that has the minimum norm among the vectors $\mathbf{w} \in \mathbb{R}^d$ that satisfies $y_i \mathbf{w}^\top x_i \geq 1$ for all $i \in [m]$.*

Prove that

$$\min_{\mathbf{w} : \|\mathbf{w}\| \leq \|\mathbf{w}^{(1)}\|} f(\mathbf{w}) = -1.$$

Note: *In general, $\mathbf{w}^{(1)}$ only depends on the set S, so we will consider it as a parameter. However, we do not know the exact value of it.*

Assignment 4. (20 points) *Describe a subgradient descent algorithm to find $\mathbf{w}$ that separates the elements of S. (You should write clearly and sufficiently to show that the algorithm indeed produces such $\mathbf{w}$.)*

Assignment 5. (10 points) *For the algorithm in Assignment 4, analyse the number of iterations and the time taken to perform each iteration.*

Accelerated Gradient Descent for Strongly Convex Functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum x^*. Suppose f is smooth with parameter L and strongly convex with parameter $\mu > 0$. Further suppose $\|x_0 - x^*\| \leq R$, for some R.

From the lectures, we have learnt that applying gradient descent with the starting point x_0 and the step-size $\gamma = 1/L$, in order to obtain an ϵ-optimal solution (i.e., solution for which the function value is at most $f(x^*) + \epsilon$), the required number of iterations is $\mathcal{O}\left(\frac{L}{\mu} \ln\left(\frac{R^2 L}{\epsilon}\right)\right)$.

Here, we will bootstrap the Accelerated Gradient Descent algorithm to bring down the number of iterations to $\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \ln\left(\frac{R^2 L}{\epsilon}\right)\right)$.

Assignment 6. (10 points) *Prove that after running the Accelerated Gradient Descent algorithm for* $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\right)$ *iterations, we can find a point* x *such that*

$$\|x - x^*\|^2 \leq \frac{R^2}{2}.$$

Assignment 7. (13 points) *Prove that by repetitively restarting Accelerated Gradient Descent, we can obtain an* ϵ*-optimal solution after the total number of* $\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \ln\left(\frac{R^2 L}{\epsilon}\right)\right)$ *iterations.*