ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Institute of Theoretical Computer Science
Bernd Gärtner and Niao He

Optimization for Data Science — Graded Assignment 1 — FS23

- The solution is due on **March 28, 2023** by **11:59 pm**. Please submit your solution as a PDF on Moodle. The name of the file should follow the format `GA4-{Legi number}`, e.g., `GA2-19-123-456`. After uploading your solution, please make sure that the status is "Submitted for grading". You should receive an automatic email that confirms your submission. Please notify us if you don't receive this.
- If you want to submit your solution within six hours before the deadline and a technical problem prevents you from submitting it on Moodle, you can send your solution as PDF to `saeed.ilchi@inf.ethz.ch`. The same submission deadline still applies. If you encounter any trouble with the submission process, complain timely.
- Please solve the exercises carefully and typeset your solution using LaTeX or a similar typesetting program. A tutorial can be found at `http://www.cadmo.ethz.ch/education/thesis/latex`. Handwritten solutions will not be graded! The same applies to solutions written with any kind of tablet device and stylus, etc.
- For geometric drawings that can easily be integrated into LaTeX documents, we recommend the drawing editor IPE, which you can find at `http://ipe7.sourceforge.net/`.
- Keep in mind the following premises:
  - When writing in English, write short and simple sentences.
  - When writing a proof, write precise statements.
- 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 Graded 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 obliged to inform the Rector of any violations of the Code.
- As with all exercises, the material of the graded assignments is relevant for the exam.

Separating Points on the Unit Interval (20 points)

Consider a learning problem where the data source $\mathcal{X} = [0, 1]$ is the unit interval and each sample point $X \in \mathcal{X}$ is drawn uniformly from $\mathcal{X}$ and is labeled as zero if $X < p^*$ and labeled as 1 otherwise, where p^* is an unknown parameter. Suppose we want to model finding p^* with 0-1-loss and the class of hypotheses is $\mathcal{H} = [0, 1]$. Provide a function $f(\cdot, \cdot)$ such that for any $0 < \varepsilon, \delta < 1$ and given $n \geqslant f(\varepsilon, \delta)$ many samples, any hypothesis $H \in \mathcal{H}$ with zero empirical risk[1] has low expected risk with probability at least $1 - \delta$. That is:

$$\ell(H) \leqslant \varepsilon.$$

In other words, if $n \geqslant f(\varepsilon, \delta)$, then the probability of existence of a hypothesis with zero empirical risk but with expected risk more than ε is at most δ.

Solution: The expected risk of an arbitrary H is $d = |p^* - H|$ and H has zero empirical risk if no sample point lies between H and p^*. So there is high expected risk hypothesis $H < p^*$ with zero empirical risk if $p^* > \varepsilon$ and there is no sample point in the interval $(p^* - \varepsilon, p^*]$. This happens with probability at most $(1 - \varepsilon)^n$. So the probability of error is:

$$(1 - \varepsilon)^n \leqslant e^{-\varepsilon n}$$

which means that for $n \geqslant f(\varepsilon, \delta) = \ln(2/\delta)/\varepsilon$, this happens with probability at most $\frac{\delta}{2}$. We can use the same argument for $H > p^*$ and concludes the proof for $f(\varepsilon, \delta) = \ln(2/\delta)/\varepsilon$. □

Continuous Convex Functions (35 points)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. Show that the following are equivalent:

(a) f is a convex function.

(b) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following inequality holds:

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leqslant \frac{f(\mathbf{x}) + f(\mathbf{y})}{2}.$$

Solution: $(a) \implies (b)$: By convexity of f (Definition 2.11), we have

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leqslant \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}),$$

which is equivalent to

$$f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) \leqslant f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})),$$

进行了变形，主要原因是要和b进行吻合

[1] Observe that there is always at least one such hypothesis.

for $\lambda \in [0,1]$. Integrating both sides from 0 to 1, we get

这里利用了定积分的性质

$$\int_0^1 f(\mathbf{y}+\lambda(\mathbf{x}-\mathbf{y}))d\lambda \leqslant \int_0^1 f(\mathbf{y})+\lambda(f(\mathbf{x})-f(\mathbf{y}))d\lambda = \frac{f(\mathbf{x})+f(\mathbf{y})}{2}.$$

(a) $\Longleftarrow$ (b): We now prove the converse. In order to do that, we show that if f is not convex, then there exist $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ and $\lambda \in (0,1)$, such that

$$\int_0^1 f(\mathbf{y}+\lambda(\mathbf{x}-\mathbf{y}))d\lambda > \frac{f(\mathbf{x})+f(\mathbf{y})}{2}.$$

One can easily check that this is equivalent to showing that if for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and $\lambda \in (0,1)$,

$$\int_0^1 f(\mathbf{y}+\lambda(\mathbf{x}-\mathbf{y}))d\lambda \leqslant \frac{f(\mathbf{x})+f(\mathbf{y})}{2},$$

then f is convex. (Simply check that given two propositions p, q, the truth table for $p \implies q$ is the same as that for $\neg q \implies \neg p$.)

Suppose that f is not convex. Then there are $\mathbf{x}$ and $\mathbf{y}$ and $\lambda' \in (0,1)$ such that

$$f(\lambda'\mathbf{x}+(1-\lambda')\mathbf{y}) > \lambda' f(\mathbf{x})+(1-\lambda')f(\mathbf{y}).$$

这个地方不是很清楚，但是是反证法

Consider the function of λ given by

$$F(\lambda) = f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) - \lambda f(\mathbf{x}) - (1-\lambda)f(\mathbf{y}),$$

which is continuous since f is. Note that F is zero for $\lambda = 0$ and $\lambda = 1$, and positive at $\lambda = \lambda'$. Let α be the largest zero crossing of F below λ' and let β be the smallest zero crossing of F above λ'. Define $\mathbf{u} = \alpha\mathbf{x}+(1-\alpha)\mathbf{y}$ and $\mathbf{v} = \beta\mathbf{x}+(1-\beta)\mathbf{y}$. On the interval (α,β), we have

$$F(\lambda) = f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) > \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}),$$

so for $\lambda \in (0,1)$,

$$f(\lambda\mathbf{u}+(1-\lambda)\mathbf{v}) > \lambda f(\mathbf{u})+(1-\lambda)f(\mathbf{v}).$$

Integrating this expression from 0 to 1 yields

$$\int_0^1 f(\lambda\mathbf{u}+(1-\lambda)\mathbf{v})d\lambda > \int_0^1 \lambda f(\mathbf{u})+(1-\lambda)f(\mathbf{v})d\lambda = \frac{f(\mathbf{u})+f(\mathbf{v})}{2},$$

which proves the converse.

□

Gradient Descent with Inexact Gradient Oracle (45 points)

Consider an unconstrained optimization problem $\min_{x\in\mathbb{R}^d} f(x)$. Assume f is μ-strongly convex and L-Lipschitz smooth. Now we only have access to an inexact gradient $g(x)$ at each point x such that $\|g(x) - \nabla f(x)\| \leqslant \delta$ with $\delta > 0$. Consider gradient descent with this inexact gradient:

$$x_{t+1} = x_t - \gamma g(x_t),$$

where $\gamma > 0$ is the step-size. Define $x^* \triangleq \arg\min_{x\in\mathbb{R}^d} f(x)$ and $f^* = f(x^*)$.

(a) Show that for any $x, y \in \mathbb{R}^d$, we have

$$f(x) \geqslant f(y) + \langle g(y), x - y\rangle + \frac{\mu}{4}\|x - y\|^2 - \frac{\delta^2}{\mu},$$

and moreover,

$$\frac{1}{\mu}\|g(y)\|^2 \geqslant f(y) - f^* - \frac{\delta^2}{\mu}.$$

(b) Show that for any $x, y \in \mathbb{R}^d$ we have

$$f(x) \leqslant f(y) + \langle g(y), x - y\rangle + L\|x - y\|^2 + \frac{\delta^2}{2L}.$$

(c) Show that by running gradient descent with inexact gradient and setting $\gamma = \frac{1}{2L}$, we have

$$f(x_{t+1}) - f^* \leqslant \left(1 - \frac{\mu}{4L}\right)(f(x_t) - f^*) + \frac{3\delta^2}{4L}.$$

This directly implies

$$f(x_T) - f^* \leqslant \left(1 - \frac{\mu}{4L}\right)^T (f(x_0) - f^*) + \frac{3\delta^2}{\mu}.$$

(d) Find a function that is μ-strongly-convex and show that the algorithm above can not guarantee to find a point x such that $f(x) - f^* < \frac{\delta^2}{2\mu}$.

Solution:

(a) Because f is μ-strongly convex,

$$\begin{aligned}
f(x) &\geqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \\
&= f(y) + \langle g(y), x - y\rangle + \langle \nabla f(y) - g(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \\
&\geqslant f(y) + \langle g(y), x - y\rangle - \|\nabla f(y) - g(y)\|\|x - y\| + \frac{\mu}{2}\|x - y\|^2 \\
&\geqslant f(y) + \langle g(y), x - y\rangle - \frac{\|\nabla f(y) - g(y)\|^2}{\mu} - \frac{\mu}{4}\|x - y\|^2 + \frac{\mu}{2}\|x - y\|^2
\end{aligned}$$

$$\geqslant f(y) + \langle g(y), x - y\rangle - \frac{\delta^2}{\mu} + \frac{\mu}{4}\|x - y\|^2.$$

In the fourth inequality we use the fact that $a^2 + b^2 \geqslant 2ab$ for all $a, b \in \mathbb{R}$. Minimizing the right hand side over y,

$$f(x) \geqslant f(y) - \frac{1}{\mu}\|g(y)\|^2 - \frac{\delta^2}{\mu}.$$

Since this inequality holds for all x, we reach our conclusion.

(b) Because f is L-smooth,

$$\begin{aligned}
f(x) &\leqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \\
&= f(y) + \langle g(y), x - y\rangle + \langle \nabla f(y) - g(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \\
&\leqslant f(y) + \langle g(y), x - y\rangle + \|\nabla f(y) - g(y)\|\|x - y\| + \frac{L}{2}\|x - y\|^2 \\
&\leqslant f(y) + \langle g(y), x - y\rangle + \frac{\|\nabla f(y) - g(y)\|^2}{2L} + \frac{L}{2}\|x - y\|^2 + \frac{L}{2}\|x - y\|^2 \\
&\leqslant f(y) + \langle g(y), x - y\rangle + \frac{\delta^2}{2L} + L\|x - y\|^2.
\end{aligned}$$

(c) By part (b),

$$\begin{aligned}
f(x_{t+1}) - f^* &\leqslant f(x_t) - f^* + \langle g(x_t), x_{t+1} - x_t\rangle + L\|x_{t+1} - x_t\|^2 + \frac{\delta^2}{2L} \\
&= f(x_t) - f^* - (\gamma - L\gamma^2)\|g(x_t)\|^2 + \frac{\delta^2}{2L} \\
&\leqslant \left[1 - \mu(\gamma - L\gamma^2)\right](f(x_t) - f^*) + (\gamma - L\gamma^2)\delta^2 + \frac{\delta^2}{2L},
\end{aligned}$$

where in the equality we use the update rule and in the last inequality we use part (b) and $\gamma - L\gamma^2 > 0$ with our choice of γ. Setting $\gamma = \frac{1}{2L}$,

$$f(x_{t+1}) - f^* \leqslant \left(1 - \frac{\mu}{4L}\right)(f(x_t) - f^*) + \frac{3\delta^2}{4L}.$$

Recursing this, we have:

$$\begin{aligned}
f(x_T) - f^* &\leqslant \left(1 - \frac{\mu}{4L}\right)^T (f(x_0) - f^*) + \sum_{i=0}^{T} \left(1 - \frac{\mu}{4L}\right)^i \left(\frac{3\delta^2}{4L}\right) \\
&\leqslant \left(1 - \frac{\mu}{4L}\right)^T (f(x_0) - f^*) + \sum_{i=0}^{\infty} \left(1 - \frac{\mu}{4L}\right)^i \left(\frac{3\delta^2}{4L}\right) \\
&\leqslant \left(1 - \frac{\mu}{4L}\right)^T (f(x_0) - f^*) + \frac{3\delta^2}{\mu}
\end{aligned}$$

(d) Consider function $f(x) = \frac{\mu}{2}x^2$. Assume the algorithm starts from point $x_0 = \frac{\delta}{\mu}$ and observe that $\|\nabla f(x_0)\| = \delta$. We can set the inexact gradient $g(x_0) = 0$. Then the algorithm will not move and $f(x_0) - f^* = \frac{\delta^2}{2\mu}$.

□