SUGGESTED EXERCISES: MODERN OPTIMIZATION

JEAN-LUC BOUCHOT

1. Introduction

2. Calculus-free optimization

Homework 1. Prove the following proposition:

Proposition 2.1 (Homework)**.** *The convexity of a function can equivalently be written as: If f is convex, then for any $\alpha_1, \cdots, \alpha_n$ positive numbers such that $\alpha_1 + \cdots + \alpha_n = 1$, if holds*

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i).$$

Homework 2. Prove the following result:

Proposition 2.2 (Generalized AGM Inequality – Homework)**.** *Let $x_1, \cdots, x_n$ be positive numbers and let $\alpha_1, \cdots, \alpha_n$ be positive numbers such that $\alpha_1 + \cdots + \alpha_n = 1$. Then the GAGM inequality reads*

$$\prod_{i=1}^n x_i^{\alpha_i} \leq \sum_{i=1}^n \alpha_i x_i.$$

Homework 3. Use the generalized AGM to prove Hölder's inequality:

Proposition 2.3 (Höder's inequality – Homework)**.** *For two sequence of numbers $\{a_k, 1 \leq k \leq n\}$ and $\{b_k, 1 \leq k \leq n\}$, Hölder's inequality reads, for p, q such that $\frac{1}{p} + \frac{1}{q} = 1$ (we say that p and q are Hölder conjugates)*

$$\sum_{k=1}^n |c_k d_k| \leq \|c\|_p \|d\|_q$$

with equality if and only if

$$\left(\frac{|c_k|}{\|c\|_p}\right)^p = \left(\frac{|d_k|}{\|d\|_q}\right)^q, \quad \text{for all } k.$$

And use this result to prove the triangle inequality for general p norms (so called Minkowksi's inequality).

3. Geometric programming

Homework 4. Minimize the function $f_0(x, y) = \frac{1}{xy} + xy + x + y$ for $x > 0$ and $y > 0$.

(You might need some help of a computer.)

4. Convex functions and analysis– review

Homework 5. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function defined on the whole of $\mathbb{R}^d$. Show that if f is bounded above, then f is constant.

Homework 6. Show that if $\mathrm{dom}(f)$ is closed, than it is not necessarily continuous.

Homework 7. Show that the ℓ_1 norm defined as $\|x\|_1 = \sum_{i=1}^d |x_i|$ is a convex function.

Homework 8. Show that the following function is convex:

$$f : \left\{ \begin{array}{rcl} \mathbb{R}^n & \to & \mathbb{R} \\ x = (x_1, \cdots, x_n) & \mapsto & f(x) = \log\left(\sum_{i=1}^n e^{x_i}\right). \end{array} \right.$$

Date: May 19, 2021.

Homework 9. Let f be defined as

$$f(x,y) = x^2(1-y^2) + y^2(1-x^2)$$

for $(x,y) \in [-1,1]^2$.

(1) Show that the function f is convex along the canonical x axis. (Make sure to look at all directions parallel to the x axis, not just the one going through the point $(0,0)$)
(2) Show that the function f is convex along the canonical y axis. (same remark)
(3) Show that the function f is not convex.

This example shows that canonical along all directions is not sufficient to show convexity globally.

Homework 10. Let $S^n_{\geq 0} := \{A \in \mathbb{R}^{n\times n} : A^T = A \text{ and } A \succcurlyeq 0\}$ be the set of symmetric positive semidefinite matrices. Show that $S^n_{\geq 0}$ is a convex cone. For the case $n = 2$, characterize its boundary as a surface in dimension 3.

Homework 11. Let $K := \{\alpha \in \mathbb{R}^{n+1} : \sum_{i=0}^n \alpha_i x^i \geq 0, \forall x \in [0,1]\}$ be the set of (coefficients of) nonnegative polynomials on $[0,1]$. Show that K is a convex cone.

Homework 12. Let C and D be two disjoint subset of $\mathbb{R}^d$. Consider the set of separating hyperplanes $(a,b) \in \mathbb{R}^{d+1}$ such that $a^T x - b \leq 0$ for all $x \in C$ and $a^T x - b \geq 0$ for all $x \in D$. Show that this set is convex.

Homework 13. Compute the Fenchel conjugate functions of

(1) Negative entropy: $f(x) = x\ln(x)$.
(2) ℓ_1 norm: $f(x) = \|x\|_1$.
(3) $f(x) = x^p$ for some $p > 1$.
(4) $f(x) = \max_{1\leq i\leq d} |x_i|$.

Homework 14. Show that a geometric programming problem can be expressed as a convex problem in standard form.

Homework 15. Express the first order condition when the inequality constraints are just positivity of the coordinates.

5. Gradient descent algorithms

Homework 16. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as

$$f(x) = \frac{2}{3}|x|^{3/2}.$$

Define a step size $\gamma > 0$.

(1) Show that f is convex and admits a unique minimizer.
(2) Define $x^* = \left(\frac{\gamma}{2}\right)^2$. Detail the sequence of iterates given by the vanilla gradient descent starting with $x^0 = x^*$.
(3) Consider now $x^0 \in (0, x^*)$. What can be said about the sequence of iterates?
(4) Conclude with regards to the use of Vanilla gradient descent without care.

Homework 17. Let $f(\mathbf{x}) := \mathbf{x}^T Q\mathbf{x} + c^T\mathbf{x}$ with

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \kappa & 0 \\ 0 & 0 & \kappa^2 \end{pmatrix} \quad \text{and} \quad c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Moreover, assume given the starting point $x^0 = (0,0,0)^T$.

(1) Solve theoretically the unconstrained minimization problem.
(2) Compute the condition number of the matrix Q.
(3) Perform 3 iterations of the gradient descent for each $\kappa \in \{10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$. Use the exact line search at each iteration for setting the step size.
(4) Compare the results obtained after these iterations.
(5) If you have access to a computer and wish to program something (or re-use the file shared with you), compare the number of iterations required for convergence with the condition number of Q.
(6) Can you prove the observed behaviour?

Homework 18. Let Q be an $d \times d$ symmetric matrix, $b \in \mathbb{R}^d$ and c a scalar. Show that the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(x) = x^T Q x + b^T x + c$ is smooth with parameter $2\|Q\|$.

Homework 19 (Difficult). Prove the following theorem

Theorem 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L smooth convex differentiable function with a global minimum x^*. Then the iterates obtained by gradient descent with step size*

$$\gamma = \frac{1}{L}$$

satisfy

$$f(x^K) - f(x^*) \leq \frac{2L}{K+4}\|x^0 - x^*\|^2.$$

Homework 20. We have seen a proof based on the mean value theorem to prove the equivalence between monotonicity of the derivative and the convexity of a univariate function. Prove the multivariate extension:

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if $\mathrm{dom}(()f)$ is convex and

$$\left(\nabla f(x) - \nabla f(y)\right)^T (x - y) \geq 0, \quad \forall x, y \in \mathrm{dom}(()f).$$

Homework 21. Show that Nesterov's second accelerated gradient descent algorithm is equivalent to the first accelerated gradient descent (AGM1).

Proposition 5.1. *Nesterov's accelerated gradient descent updates are equivalent to the following iterations.*

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = \left(1 - \frac{1-\lambda_k}{\lambda_{k+1}}\right) x^{k+1} + \frac{1-\lambda_k}{\lambda_{k+1}} x^k,$$

where λ_k are defined recursively as

$$\lambda_0 = 0$$
$$\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}.$$

Homework 22 (Programming assignment – obviously optional). Based on the given notebook, implement the second approach to Nesterov's acceleration and compare with the one given.

6. Constrained optimization: Projected methods

Homework 23. Let $\Omega \subset \mathbb{R}^d$ be a closed conex set. Show that the projection defined as

$$P_\Omega(x) := \operatorname*{argmin}_{y \in \Omega} N(x - y)$$

where N denotes a certain given norm is not necessarily unique.

Homework 24. Assume Ω is a convex body. Show that the projection

$$P_\Omega(x) := \operatorname*{argmin}_{y \in \Omega} \|x - y\|_2$$

might not exist at all.

Homework 25. Show the following result:

Lemma 1. *Let x^k be the sequence of iterates generated by the projected gradient descent of an L-smooth convex differentiable function f over a closed convex domain Ω. Then, using a fixed gradient step*

$$\gamma = \frac{1}{L}$$

we have

$$f(x^{k+1}) \leq f(x^k) - \frac{L}{2}\|x^{k+1} - x^k\|^2$$

Homework 26. Prove the following result.

Proposition 6.1. *Let $u \in \mathbb{R}^d$ be such that $u_i \geq 0$, for all $1 \leq i \leq d$ and $\sum_{i=1}^d u_i > 1$. For $p \in \{1, \cdots, d\}$, define $y(p)$ as*

$$y(p)_i = u_i - \theta_p, \quad 1 \leq i \leq p, \qquad \text{and } 0 \text{ for } i > p.$$

Then

$$y(p^*) = \operatorname{argmin}_{x \in \Delta_d} \|x - u\|_2$$

where

$$p^* = \max\{p : u_p - \frac{1}{p}\left(\sum_{i=1}^p u_i - 1\right) > 0\}$$

and Δ_d defines the d dimensional unit simplex:

$$\Delta_d = \left\{ x \in \mathbb{R}^d : x_i \geq 0 \text{ and } \sum_{i=1}^d x_i = 1 \right\}.$$

7. Proximal methods

Homework 27. Let f be an L-smooth convex function and define $\{x^k\}$ the sequence of iterates obtained from the gradient descent with step size $\frac{1}{L}$. Show that

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{1}{2\gamma} \left\| x - \left(x^k - \gamma \nabla f(x^k)\right) \right\|_2.$$

Homework 28. Let $f(x) = c$ be a constant function. Compute its proximal operator.

Homework 29. This exercise is here to show that, although the examples we will face in our class will not create any trouble, some care should be taken before using proximal operators.

Compute the proximal operators (with $\gamma = 1$) of the following functions

(1) $f_1(x) = -\lambda$ if $x = 0$ and $f_1(x) = 0$ elsewhere (for a positive $\lambda > 0$).
(2) $f_2(x) = \lambda$ if $x = 0$ and $f_1(x) = 0$ elsewhere (for a positive $\lambda > 0$).

Homework 30. Find the proximal operators (for general $\gamma > 0$) for of the following functions

- $f(x) = |x|$ for $x \in \mathbb{R}$.
- $f(x) = x^3$ if $x \geq 0$ and $f(x) = \infty$ elsewhere.
- $f(x) = x$ for $x \geq 0$ and $f(x) = 0$ elsewhere.

Homework 31. Prove the following result

Lemma 2. *Let f be a (proper closed) convex function and $\gamma > 0$. For any x in the domain*

$$u = \operatorname{prox}_{f,\gamma}(x) \Rightarrow \frac{1}{\gamma}\langle x - u, y - u\rangle \leq f(y) - f(u), \quad \forall y.$$

Homework 32. Let $g : \mathbb{R}^d \to \mathbb{R}$ be defined as

$$g(x) = \begin{cases} -\sum_{i=1}^d \log(x_i), & x > 0 \\ \infty & \text{elsewhere.} \end{cases}$$

Compute its proximal operator.

Homework 33. Let $\|x\|_0 = \#\{i : x_i \neq 0\}$ denotes the number of non zero entries in x (so-called sparsity of x). Compute its proximal operator.

Homework 34. Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. Assume given a regularization parameter $\lambda > 0$. We are interested in solving the following optimization problem

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|_1.$$

Derive the proximal gradient methods updates as well as the accelerated proximal gradient iterations.

Remarks:

(1) The *classical* proximal gradient method is called ISTA: Iterative Shrinkage-Thresholding Algorithm.
(2) The *accelerated* version is known as FISTA: Fast ISTA.

By extension, (F)ISTA is used to denote any ℓ_1 regularized minimization problem, albeit the original version is an algorithm for solving ℓ_1 regularized least squares problem.

8. Stochastic gradient descent

Homework 35. We want to compare gradient descent and stochastic gradient descent for least squares problems. We will, however, introduce a new algo in which the sampling part of stochastic gradient descent is done by importance sampling instead of uniform sampling.

We consider the linear system $Ax = y$ for some $A \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$. Assume that 1) $n > d$, 2) A has full column rank, and 3) there exists a solution. Let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A)$ denote the singular values of A. Finding a solution may be solved by finding a solution to the following problem:

$$x^* = \operatorname*{argmin}_{x\in\mathbb{R}^d} \|Ax - y\|_2^2.$$

(1) We start with the analysis of classical gradient descent with step size $\gamma = \frac{1}{\sigma_1(A)^2}$.
   (a) Show that
   $$\sigma_1(I - \gamma A^T A)^2 = 1 - \gamma\sigma_d(A)^2 = 1 - \frac{\sigma_d(A)^2}{\sigma_1(A)^2}.$$
   (b) Detail the gradient descent iterations.
   (c) Show the following convergence rate
   $$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\sigma_d(A)^2}{\sigma_1(A)^2}\right)\|x^k - x^*\|_2^2.$$

(2) We look now at the stochastic gradient descent with importance sampling. In importance sampling, we use the stepsize $\gamma_i = \frac{1}{\|A_i\|_2^2}$ where A_i denotes the i^{th} row of A. An index i is selected with probability $\frac{\|A_j\|_2^2}{\|A\|_F^2}$.
   (a) Define $P_i = \gamma_i A_i^T A_i$. Show that it is the orthogonal operator onto the range of A_i.
   (b) Show that $\mathbb{E}[P_i] = \frac{A^T A}{\|A\|_F^2}$.
   (c) Show that
   $$\mathbb{E}\left[\|x^{k+1} - x^k\|_2^2\right] \leq \left(1 - \frac{\sigma_d(A)^2}{\|A\|_F^2}\right)\mathbb{E}\left[\|x^k - x^*\|_2^2\right].$$

Homework 36. Prove the following result

Proposition 8.1. *The variance of the mini-batch stochastic gradient descent decreases linearly with the mini-batch size. More precisely:*

Let $S \subset \{1, \cdots, n\}$ be a subset sampled uniformly at random and let g_k denotes the stochastic mini batch gradient over S. Then

$$\mathbb{E}\left[\left\|g_k - \nabla f(x^k)\right\|^2\right] \leq \frac{B^2}{|S|}.$$

School of Mathematics and Statistics, Beijing Institute of Technology
Email address: jlbouchot@bit.edu.cn