# Math 164-1: Optimization

Instructor: Alpár R. Mészáros

Second Midterm, May 18, 2016

Name (use a pen):

Student ID (use a pen):

Signature (use a pen):

**Rules:**

- Duration of the exam: **50 minutes**.
- By writing your name and signature on this exam paper, you attest that you are the person indicated and will adhere to the UCLA Student Conduct Code.
- You may use either a pen or a pencil to write your solutions. However, if you use a pencil I will withheld your paper for **two** weeks after grading it.
- **No** calculators, computers, cell phones (all the cell phones should be turned off during the exam), notes, books or other outside material are permitted on this exam. If you want to use scratch paper, you should ask for it from one of the supervisors. Do not use your own scratch paper!
- Please justify all your answers with mathematical precision and write rigorous and clear proofs. You may loose points in the lack of justification of your answers.
- Theorems from the lectures and homework may be used in order to justify your solution. In this case state the theorem you are using.
- This exam has 3 problems and is worth **20 points**. Adding up the indicated points you can observe that there are **28 points**, which means that there are **8 "bonus" points**. This permits to obtain the highest score 20, even if you do not answer some of the questions. On the other hand nobody can be bored during the exam. All scores higher than 20 will be considered as 20 in the gradebook.
- I wish you success!

| Problem | Score |
|---|---|
| Exercise 1 | |
| Exercise 2 | |
| Exercise 3 | |
| Total | |

**Exercise 1** (11 points)**.**

Let $n \geq 1$ be an integer and let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix (non necessarily positive definite) for which all of its eigenvalues are non-zero. Let $a \in \mathbb{R}^n$ be a given vector and we consider the function $f : \mathbb{R}^n \to \mathbb{R}$, defined as

$$f(x) = \frac{1}{2}(x-a)^\top A^2(x-a),$$

where $A^2 = AA$.

(1) Using first and second order optimality conditions show that $f$ has a unique global minimizer on $\mathbb{R}^n$ and determine this optimizer. Denote it by $x^*$.

(2) Write the updates in the steepest descent algorithm (i.e. gradient descent with optimal step size) starting from a point $x^0 \in \mathbb{R}^n$ to approximate the optimizer $x^*$ of $f$ that has been determined in (1). Determine the step size $\alpha_k$ in each step.

(3) Imagine that one wants to use a fixed step gradient algorithm too, to approximate $x^*$. Which is maximal range for the step size $\alpha$ in terms of the eigenvalues of $A$ that ensures global convergence for the algorithm?

(4) Give an example of $A \in \mathbb{R}^{2\times 2}$ diagonal matrix that has a zero and a non-zero eigenvalue. Take $a \in \mathbb{R}^2$. Determine the global minimizers of $f$ in $\mathbb{R}^2$ in this case. What can we say about the uniqueness of them?

(5) Explain what will happen if we want to proceed with a fixed step size gradient algorithm for (4). Does an algorithm like this converge globally? If yes, for which values of the step size $\alpha$ and to which limit point $x^*$?

(6) Explain what is the major difference between the cases when $A$ has at least one zero eigenvalue and when it does not, from the point of view of the gradient descent algorithms.

**Solutions**

Notice first the since $A$ is symmetric, so is $A^2$. Moreover since $A$ has non-zero eigenvalues, $A^2$ has all its eigenvalues positive, hence it is a positive definite matrix. Let us define $Q := A^2$. Observe also that the function can be rewritten as

$$f(x) = \frac{1}{2}x^\top Qx - x^\top b + c,$$

where we set $b := Qa$ and $c := \frac{1}{2}a^\top Qa$. Mind that in the optimization problem the constant $c$ does not play any role.

(1) Since the optimization problem is without constraints, the first order necessary optimality condition for the minimizer reads as $\nabla f(x^*) = 0$, that is $Qx^* = b$, from where $x^* = Q^{-1}b = Q^{-1}Qa = a$. All these computations are meaningful, because $Q^{-1}$ exists. The second order sufficient condition of minimality (since $\nabla f(x^*) = 0$) reads as $D^2 f(x^*) = Q = A^2 > 0$, which is true, hence $x^* = a$ is the unique global minimizer of $f$ on $\mathbb{R}^n$.

(2) The updates in the steepest descent starting from $x^0$ are

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) = x^k - \alpha_k(Qx^k - b),$$

where $\alpha_k = \text{argmin}_{\alpha\in\mathbb{R}} f(x^k - \alpha\nabla f(x^k))$ and using the formula derived during the lecture, one has

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^\top Q \nabla f(x^k)}.$$

(3) For the fixed step size algorithm global convergence is equivalent (as we discussed during the lectures) to $0 < \alpha < \dfrac{2}{\lambda_{max}(Q)}$. The maximal eigenvalue of $Q$ actually can be written in terms of the

maximal (in absolute value) eigenvalue of $A$, i.e. $\lambda_{max}(Q) = \max\{\lambda_i^2 : i = 1, \ldots, n\}$, where the $\lambda_i$'s are the eigenvalues of $A$ counted with multiplicity.

(4) An example for such a matrix is

$$A = \begin{pmatrix} \gamma & 0 \\ 0 & 0 \end{pmatrix},$$

where $\gamma \neq 0$. The other option is, when the elements on the main diagonal are exchanged. In this case

$$Q = A^2 = \begin{pmatrix} \gamma^2 & 0 \\ 0 & 0 \end{pmatrix},$$

and the function can be written as $f(x_1, x_2) = \frac{1}{2}\gamma^2(x_1 - a_1)^2$, hence it is independent of the second variable. Setting $\nabla f(x) = 0$ one finds that the candidates for the optimizers are $x^* = (a_1, x_2)$, where $x_2 \in \mathbb{R}$ is arbitrary. Since the function is independent of the second variable and $f(a_1, x_2) = 0 \leq f(y_1, y_2)$ for any $(y_1, y_2) \in \mathbb{R}^2$, one has that all of them are global minimizers that have the same objective function value, hence they are not unique.

(5) In the case of (4) the problem is reduced to a 1D problem, hence a fixed step size gradient algorithm converges globally if and only if the step size $\alpha$ is in the range $0 < \alpha < \frac{2}{\gamma^2}$. From the 2D point of view what is happening is the following: choosing any initial guess$x^0 = (x_1^0, x_2^0)$, since $f$ is independent of the second variable (hence the second coordinate of its gradient is always zero), during each update in $x^{k+1} = (x_1^{k+1}, x_2^{k+1})$ the second coordinate $x_2^{k+1}$ remains unchanged. Hence the algorithm actually converges to *a global minimizer* namely the one $(a_1, x_2^0)$.

(6) If some of the eigenvalues of $A$ are zero, $Q = A^2$ will have also the corresponding eigenvalues 0. On the other hand, since $Q$ is symmetric, it is diagonalizable, so we can see it up to a change of coordinates as a diagonal matrix with the eigenvalues on the main diagonal. As we have seen in (5), the coordinates (in the new system of coordinates, if $Q$ was not diagonal at the first place) corresponding to the zero eigenvalues are unaffected by the gradient algorithms. And the dimension of the problem can be reduced by the number of zero eigenvalues. While for positive definite $Q$, i.e. if $A$ does not have zero eigenvalues, the problem is full dimensional. This is a major difference between the two cases.

**Exercise 2** (8 points)**.**

Let us consider the function $f : (0, +\infty) \to \mathbb{R}$ defined as

$$f(x) = x - \ln(x),$$

where ln denotes the natural logarithm of base $e$.

(1) Using eventually first and second order optimality conditions, show that $f$ has a unique minimizer on $(0, +\infty)$. Denote this by $x^*$.

In what follows, we are aiming to approximate $x^*$ from (1) using Newton's algorithm.

(2) Write the updates in Newton's algorithm used to approximate the minimizer of $f$ above. Denote the sequence of iterates by $(x^k)_{k\geq 0}$. Determine the biggest range for the initial guess $x^0 > 0$ for which one has after one iteration that $x^1 > 0$. Denote this range by $I$.

(3) Let $\varepsilon > 0$ be a given error term. Explain why is the condition $|1-x^k| \leq \varepsilon$ a good stopping condition for Newton's algorithm approximating $x^*$.

(4) Show that for all $x^0 \in I$ (where $I$ is determined in (2)) the sequence $(x^k)_{k\geq 0}$ is converging to $x^*$. *Hint:* compute for instance the error $1 - x^1$ in terms of $x^0$, then write this relation also for $x^{k+1}$ and $x^k$.

(5) Propose a modification of the above algorithm that will ensure that it is converging also if $x^0 \in (0, +\infty) \setminus I$. *Hint:* you may think to introduce a step size in the algorithm, which is exactly 1 in the usual Newton algorithm.

**Solution**

(1) Using the first order necessary optimality condition, if $x^*$ is a minimizer, then $1 - 1/x^* = 0$, hence the only candidate in the interior is $x^* = 1$. The second order sufficient condition (since $x^* = 1$ is an interior point and $f'(x^*) = 0$) $f''(1) = 1 > 0$ implies that $x^* = 1$ is a unique strict local minimizer. On the boundary one cannot have other local minimizers, since $\lim_{x\downarrow 0} f(x) = +\infty$ and $\lim_{x\to+\infty} f(x) = +\infty$ and $f$ is decreasing from 0 to 1, then it is increasing towards $+\infty$. Thus $x^*$ is actually a global minimizer as well.

(2) After choosing $x^0 > 0$, we construct the sequence with the recursive relation

$$x^{k+1} = x^k - f'(x^k)/f''(x^k) = 2x^k - (x^k)^2,$$

provided $f''(x^k) \neq 0$.

One aims to have $x^1 = 2x^0 - (x^0)^2 > 0$, which determines the range (since $x^0 > 0$) for $x^0 \in (0, 2) = I$.

(3) Since the unique minimizer is $x^* = 1$, the condition $|1 - x^k| \leq \varepsilon$ will result in an approximation of 1 by an error $\varepsilon > 0$, thus it is reasonable to stop the algorithm, once this approximation is achieved. On the other hand, since $f'(x^k) = \dfrac{x^k - 1}{x^k}$ and in general the condition $|f'(x^k)| \leq \varepsilon$ will give a point that is very close to the minimizer, the condition $|1 - x^k| \leq \varepsilon$ will imply that $|f'(x^k)| \leq \varepsilon/x^k$ and we expect for $x^k$ not to become very small of very large, thus this implies once more the good choice of this condition.

(4) By the formula of the update one has $1 - x^{k+1} = 1 - 2x^k + (x^k)^2 = (1 - x^k)^2$, and since this is true for every index $k > 0$, one has (also passing to absolute values)

$$|1 - x^{k+1}| = |1 - x^0|^{2^{k+1}},$$

and since $x^0 \in I$ implies $0 < x^0 < 2$, which means that $|1 - x^0| < 1$, passing to the limit in the above equality one obtains that the algorithm converges exponentially fast.

(5) Clearly, the problem with initial guesses $x^0$ outside of $I$ is that after one iteration $x^1$ becomes negative, for which values the function is not defined. To overcome this issue, we introduce a step size

$\alpha_k$ in the algorithm that ensures that we do not go outside of the domain of $f$. The modified algorithm reads as follows

$$x^{k+1} = x^k - \alpha_k f'(x^k)/f''(x^k) = x^k - \alpha_k((x^k)^2 - x^k),$$

where we chose $\alpha_k > 0$ to be small enough that prevents that $x^{k+1} \leq 0$ if $x^k \in (0, +\infty) \setminus I$ or $\alpha_k = 1$ if $x^k$ is already in $I$. The condition $x^{k+1} > 0$ implies that $\alpha_k$ has to be chosen such that $0 < \alpha_k < \dfrac{1}{x^k - 1}$ if $x^k > 2$ or $\alpha_k = 1$ if $x^k \in I$. This will result in a convergent algorithm, since in a finite number of steps we achieve that $0 < x^k < 2$, after which we know that the algorithm converges for all starting points in $I$.

**Exercise 3** (9 points)**.**

We aim to compute an approximation of $\sqrt{2}$. For this, we construct a sequence $(x^k)_{k\geq 0}$ that converges to one of the solutions of the equation $x^2 = 2$.

(1) Suppose we have two initial guesses $x^0, x^1 \in \mathbb{R}$. Write down the definition of the sequence $(x^k)_{k\geq 0}$ constructed by the *secant method*. Write the formula in a compact form.

(2) Setting $x^0 = 0$ and $x^1 = 1$, compute $x^2, x^3$ and $x^4$ using the algorithm given in (1). What do you observe?

(3) Give two initial guesses $x^0$ and $x^1$ for which the sequence constructed in (1) tends to converge to $-\sqrt{2}$ instead. Justify your choice.

(4) Explain analytically and geometrically the behavior of the algorithm described in (1), when one chooses $x^0 = a$ and $x^1 = -a$ for some $a \in \mathbb{R}$ as initial guesses.

(5) Give a sufficient condition for the initial guesses $x^0$ and $x^1$ (discuss also the geometrical intuition behind it) that guarantees that the algorithm described in (1) has a tendance to converge to $\sqrt{2}$.

*Notice:* by the notion of *tendance of convergence* we mean that we have a strong belief that the algorithm converges and this is supported by a couple of iterations and the geometrical intuition behind. You *do not* need to show actual convergences in this exercise!

**Solution**

(1) Introducing the function $f : \mathbb{R} \to \mathbb{R}$, defined as $f(x) = x^2 - 2$, the updates using the secant method for the roots of $f$ read as

$$x^{k+1} = x^k - (x^k - x^{k-1})\frac{f(x^k)}{f(x^k) - f(x^{k-1})},$$

provided $f(x^k) \neq f(x^{k-1})$. This can be written in a compact form as

$$x^{k+1} = x^k - \frac{(x^k)^2 - 2}{x^k + x^{k-1}} = \frac{2 + x^{k-1}x^k}{x^{k-1} + x^k}.$$

(2) Simple computations yield $x^2 = 2$, $x^3 = 4/3 \approx 1.33$ and $x^4 = 7/5 = 1.4$. The observation is that we start to get closer to $\sqrt{2} \approx 1.4142$.

(3) Since the problem is symmetric to the origin, one expects that the choice $x^0 = 0$ and $x^1 = -1$ produces a sequence that tends to converge to $-\sqrt{2}$. Indeed, computing the first few iterations in this case one obtains that $x^2 = -2$, $x^3 = -4/3 \approx -1.33$ and $x^4 = -7/5 = -1.4$

Notice also using the formula for the update that once we achieved two consecutive terms that are nonpositive, the rest of them will remain nonpositive as well (because $x^{k-1}x^k \geq 0$ and $x^{k-1} + x^k \leq 0$ if both terms are nonpositive). Thus for any two nonpositive initial guesses the algorithm would have the tendance to converge to $-\sqrt{2}$.

(4) Analytically it is clear that in this case $x^2$ cannot be well defined because we would divide by zero. Geometrically what is happening is the following: remember that $\dfrac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} = x^k + x^{k-1}$ is used to approximate $f'(x^k)$ in Newton's algorithm, that was the initial purpose to introduce the secant method. For $k = 1$, even if $x^0 = a$ could be far from $x^1 = -a$, still this approximation would be $x^1 + x^0 = 0$, meaning that the slope of the "approximated" tangent line would be 0. This would mean that it is parallel to the $Ox-$axes hence one cannot define the next term in the iteration (that is defined as the intersection of this tangent line with the $Ox-$axes).

(5) Once again, the problem is completely symmetric to the origin, meaning that the two solutions are $-\sqrt{2}$ and $\sqrt{2}$. Thus, as a first intuition, one should achieve after some iterations that two consecutive terms in the sequence to be nonnegative. Once this is achieved, using the formula for the update, it is sure that all the upcoming terms will be nonnegative, so one might hope for convergence to $\sqrt{2}$.

Actually assuming that the algorithm converges to a number $\ell$, passing to the limit is the formula for $x^{k+1}$, one obtains the equation for $\ell$ :

$$\ell = \frac{2 + \ell^2}{2\ell},$$

i.e. $\ell^2 = 2$ and because all the terms of the sequence after a certain point are nonnegative would imply that $\ell = \sqrt{2}$. Thus, indeed the intuition to have two consecutive terms that are nonnegative could result in a convergent algorithm that if converging cannot converge elsewhere but to $\sqrt{2}$. This reasoning implies that a good sufficient condition is to choose the two initial guesses nonnegative and distinct.

Other sufficient conditions could be considered as well (also involving initial guesses with opposite signs), however the precise conditions for these could be more technical to describe.