Exam Continuous Optimization

18 January 2021, 14.00–17.00

The exam consists of 4 questions. In total you can obtain 90 points. The final grade is $1 + \#\text{points}/10$ rounded to the nearest integer.

This is an open-book exam. It is NOT allowed to discuss with anyone else. If you have any questions regarding the exam, or technical questions regarding uploading of your answer, please contact David de Laat at d.delaat@tudelft.nl.

Please review the instructions posted on the announcement page for the course. The most important points are repeated below:

- Write your answers **by hand** and start each exercise on a new sheet.
- On your first answer sheet, you should write the following statement: "This exam will be solely undertaken by myself, without any assistance from others, and without use of sources other than those allowed."
- When scanning your work place your student ID on the first page. If you do not have a student ID please use some other form of identification but in that case make sure only your name and photo are visible.
- Scan your work and submit it as **one single pdf-file** at 17.00.
- You should keep an eye on your email from 17.00-18.00 because you can be asked to join the zoom call for a random check.

Good luck!

1. Consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \text{ for } i = 1, \ldots, m, \\ & Ax = b. \end{aligned}$$

where $f_0, \ldots, f_m$ are convex and twice-continuously differentiable on $\mathbb{R}^n$.

(a) (6 points) Show Slater's condition holds if there exist feasible points $x_1, \ldots, x_m \in \mathbb{R}^n$ with $f_i(x_i) < 0$ for $i = 1, \ldots, m$.

Solution: Let $x = (x_1 + \ldots + x_n)/n$. Then,

$$Ax = A(x_1 + \ldots + x_n)/n = (Ax_1 + \ldots + Ax_n)/n = b$$

and

$$f_i(x) = f_i((x_1 + \ldots + x_n)/n) \leq (f_i(x_1) + \ldots + f_i(x_n))/n < 0$$

for all $i = 1, \ldots, m$. This shows x is a strictly feasible point. Since the problem is convex and admits a strictly feasible point, Slater's condition holds.

(b) (6 points) Use the second-order condition for convexity to show that the barrier functional

$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$$

is convex.

Solution: The domain of ϕ is the set of all x with $f_i(x) < 0$ for all $i = 1, \ldots, m$. Since the functions $f_1, \ldots, f_m$ are convex, this open set is convex. For x in the domain of ϕ we have

$$\nabla^2 \phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x).$$

The first part

$$\sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T$$

is positive semidefinite because it is a conic combination of outer products (which are positive semidefinite). The second part

$$\sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x)$$

is positive semidefinite because it is a conic combination of positive semidefinite matrices, since the Hessians $\nabla^2 f_i$ are positive semidefinite because the functions $f_1, \ldots, f_m$ are convex.

(c) (6 points) The barrier (or centralizer) problem for a given t is defined as

$$\begin{aligned} \text{minimize} \quad & t f_0(x) + \phi(x) \\ \text{subject to} \quad & Ax = b. \end{aligned}$$

Write down the Lagrangian and the KKT conditions for this problem.

Solution: The Lagrangian:

$$L(x, \nu) = t f_0(x) + \phi(x) + \nu^T (Ax - b).$$

The KKT conditions:

- $Ax = b$
- $t\nabla f_0(x) + \sum_{i=1}^m \frac{1}{-f_i(x)} \nabla f_i(x) + A^\mathsf{T}\nu = 0.$
- (Optionally one can explicitly state x has to lie in the domain of the objective.)

(d) (6 points) The optimal solution to the barrier (or centralizer) problem is denoted by $x^*(t)$ and for $t > 0$ these solutions form a path called the central path. Explain how the tangent vector $dx^*(t)/dt$ to the central path can be computed. (Hint: use the KKT conditions form (c).)

Solution: The KKT conditions give

$$Ax^*(t) = b$$

and

$$t\nabla f_0(x^*(t)) + \nabla\phi(x^*(t)) + A^T\hat{\nu}(t) = 0.$$

Implicit differentiation with respect to t gives

$$A\frac{dx^*(t)}{dt} = 0$$

and

$$\nabla f_0(x^*(t)) + (t\nabla^2 f_0(x^*(t)) + \nabla^2\phi(x^*(t)))\frac{dx^*(t)}{dt} + A^T\frac{d\hat{\nu}(t)}{dt} = 0.$$

This can be written as the linear system

$$\begin{pmatrix} t\nabla^2 f_0(x^*(t)) + \nabla^2\phi(x^*(t)) & A^\mathsf{T} \\ A & 0 \end{pmatrix} \begin{pmatrix} dx^*(t)/dt \\ d\hat{\nu}(t)/dt \end{pmatrix} = \begin{pmatrix} -\nabla f_0(x^*(t)) \\ 0 \end{pmatrix},$$

so we can compute the tangent vector by solving this system.

2. Consider the unconstrained optimization problem

$$\text{minimize} \quad f(x),$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex and continuously differentiable on $\mathbb{R}^n$.

(a) (6 points) Give an example of a strongly convex function f with $n = 2$ for which gradient descent performs badly but Newton's method works well. Why does gradient descent perform badly for this example? Why does Newton's method work well for this example?

Solution: Let $f(x_1, x_2) = \frac{10^5}{2}x_1^2 + \frac{1}{2}x_2^2$. Then the Hessian is $\nabla^2 f(x) = \left(\begin{smallmatrix} 10^5 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, which has condition number 10^5. Gradient descent does not work well because the level sets are higly eccentric ellipses, so that the gradient vectors at most points are almost orthogonal to the vector pointing in the direction of the minimizer $(0, 0)$. For most initial points this will result in a zig-zag behaviour. The convergence rate is almost equal to 1 is this case. Newton's method converges in one step for a quadratic function.

(b) (6 points) Consider the norm $\|\cdot\|$ defined by

$$\|x\| = 2\|x\|_2.$$

Express the steepest descent direction Δx_{sd} and the normalized steepest descent direction Δx_{nsd} explicitly in terms of the gradient $\nabla f(x)$.

Solution: This is a quadratic norm with matrix $P = 4I$. Then,

$$\begin{aligned}
\Delta x_{\text{nsd}} &= -(\nabla f(x)^T P^{-1} \nabla f(x))^{-1/2} P^{-1} \nabla f(x) \\
&= -(\nabla f(x)^T \nabla f(x))^{-1/2} P^{-1/2} \nabla f(x) \\
&= -\frac{\nabla f(x)}{2\|\nabla f(x)\|_2}
\end{aligned}$$

and

$$\Delta x_{\text{sd}} = -P^{-1}\nabla f(x) = -\frac{1}{4}\nabla f(x).$$

Alternatively, the definitions can be used directly. We have

$$\begin{aligned}
\Delta x_{\text{nsd}} &= \operatorname{argmin}\{\nabla f(x)^{\mathsf{T}} v : \|v\| \le 1\} \\
&= \operatorname{argmin}\{\nabla f(x)^{\mathsf{T}} v : 2\|v\|_2 \le 1\} \\
&= \frac{1}{2}\operatorname{argmin}\{\nabla f(x)^{\mathsf{T}} v : \|v\|_2 \le 1\} \\
&= -\frac{\nabla f(x)}{2\|\nabla f(x)\|_2},
\end{aligned}$$

and since

$$\begin{aligned}
\|\nabla f(x)\|_* &= \sup\{\nabla f(x)^{\mathsf{T}} x : \|x\| \le 1\} \\
&= \sup\{\nabla f(x)^{\mathsf{T}} x : 2\|x\|_2 \le 1\} \\
&= \sup\{\frac{1}{2}\nabla f(x)^{\mathsf{T}} x : \|x\|_2 \le 1\} \\
&= \frac{1}{2}\|\nabla f(x)\|_2
\end{aligned}$$

we get

$$\Delta x_{\mathrm{sd}} = \|f(x)\|_* \Delta x_{\mathrm{nsd}} = -\frac{1}{4}\nabla f(x).$$

(c) (6 points) Explain whether or not steepest descent using the above norm is the same as gradient descent when exact line search is used. And what about when backtracking line search is used?

Solution: Since the descent direction for the steepest descent method with norm $\|.\|$ is a positive multiple of the gradient descent direction, both methods are exactly the same when exact line search is used. With backtracking line search the two methods can be different (backtracking line search starts at a certain multiple of the descent direction and tracks backwards. If the descent direction had a different magnitude this can be different).

(d) (6 points) Suppose $n = 2$ and $f(x) = x_1^2 + x_2^2 - \cos(x_1)$. Find the gradient, Hessian, and Newton step Δx_{nt} at the point $(0, 1)$.

Solution: The gradient is

$$\nabla f(x) = \begin{pmatrix} 2x_1 + \sin(x_1) \\ 2x_2 \end{pmatrix}$$

The Hessian is

$$\nabla^2 f(x) = \begin{pmatrix} 2 + \cos(x_1) & 0 \\ 0 & 2 \end{pmatrix}$$

The Newton direction is

$$\Delta x_{\mathrm{nt}} = -(\nabla^2 f(0,1))^{-1} \nabla f(0,1) = -\begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

3. Consider the 1-dimensional optimization problem

$$\begin{aligned} \text{minimize} \quad & (x-2)^2 \\ \text{subject to} \quad & 0 \leq x \leq 5. \end{aligned}$$

(a) (6 points) Explain why the objective function is unimodal (according to the definition of unimodal we used in the lecture).

Solution: It is strictly decreasing for $x < 2$ and strictly increasing for $x > 2$.

(b) (6 points) Suppose we apply Fibonacci line search with initial bracket $[0, 5]$. What is the bracket after 4 function evaluations? Make a sketch to support your answer.

Solution: (skip) The final bracket is something like $[1, 2.000001]$ or $[1.999999, 3]$

(c) (3 points) How many iterations does quadratic fit search need to find the minimum? Explain your answer.

Solution: Only 1 since the function is quadratic itself. (The answer 2 is also correct, to detect a minimizer has been found.)

(d) (9 points) Find the Lagrangian, Lagrange dual function, and Lagrange dual problem.

Solution: The Lagrangian is

$$L(x, \lambda) = (x-2)^2 - \lambda_1 x + \lambda_2 (x-5) = x^2 + (\lambda_2 - 4 - \lambda_1)x - 5\lambda_2 + 4.$$

We have

$$\frac{\partial}{\partial x} L(x, \lambda) = 2x + \lambda_2 - 4 - \lambda_1.$$

So the infimum over x is attained for $x = (\lambda_1 + 4 - \lambda_2)/2$. The dual function is

$$\begin{aligned} g(\lambda) &= \inf_x L(x, \lambda) = (\lambda_1 + 4 - \lambda_2)^2/4 - (\lambda_1 + 4 - \lambda_2)^2/2 - 5\lambda_2 + 4 \\ &= 4 - 5\lambda_2 - (\lambda_1 + 4 - \lambda_2)^2/4 \end{aligned}$$

So the dual problem is

$$\begin{aligned} \text{maximize} \quad & 4 - 5\lambda_2 - (\lambda_1 + 4 - \lambda_2)^2/4 \\ \text{subject to} \quad & \lambda \geq 0. \end{aligned}$$

4. Let $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \{-1, 1\}$ be a training set and $\gamma > 0$ a parameter. Consider the support vector problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|a\|_2^2 + \gamma 1^\mathsf{T} u \\ \text{subject to} \quad & y_i(a^\mathsf{T} x_i + b) \geq 1 - u_i \text{ for } i = 1, \ldots, N, \\ & u \geq 0 \end{aligned}$$

with optimal solution (a^*, b^*, u^*).

(a) (3 points) Does strong duality hold? Explain why or why not.

Solution: Yes, the problem is convex and the solution (a, b, u) with $a = 0$, $b = 0$, and $u_i = 2$ for all i, is strictly feasible, so strong duality holds by Slater's condition. (In fact, since the constraints are affine and the domain is everything, any feasible solution is strictly feasible.)

(b) (3 points) How do we use the solution (a^*, b^*, u^*) to decide to which class ($+1$ or -1) a new point $z \in \mathbb{R}^n$ belongs?

Solution: If $(a^*)^\mathsf{T} x_i + b^* > 0$ we assign it to class $+1$, and otherwise to class -1.

(c) (6 points) Explain why we have the terms $\frac{1}{2}\|a\|_2^2$ and $\gamma 1^\mathsf{T} u$ in the objective. What are these terms achieving in relation to the hyperplane and slab around the hyperplane defined by a and b? What happens when the parameter γ is very large?

Solution: We can think of a and b defining a hyperplane $\{x : a^\mathsf{T} x = b\}$ with a slab around it of width $2/\|a\|_2$. The term $\frac{1}{2}\|a\|_2^2$ means we are trying to find a hyperplane with a wide slab around it, and the term $\gamma 1^\mathsf{T} u$ penalizes a point lying in the slab or on the wrong side of the slab. When γ is very large (almost) no point will lie in the slab or on the wrong side of the slab; that is, when γ is very large we are trying to find a hyperplane that separates the two point sets with an as wide as possible slab around it containing no points.

(d) (6 points) Suppose that for a given i and given dual optimal solution, the dual variables corresponding to the constraints $y_i(a^\mathsf{T} x_i + b) \geq 1 - u_i$ and $u_i \geq 0$ are both nonzero. What does this say about x_i in relation to the hyperplane and slab around the hyperplane defined by (a^*, b^*)?

Solution: If these dual variables are nonzero, then by complementary slackness (note that strong duality holds) $y_i(a^\mathsf{T} x_i + b) = 1 - u_i$ and $u_i = 0$. This means the point x_i lies on the boundary of the slab.

End of test