Modern Optimization Techniques

Retake Exam

Prof. Dr. Dr. Lars Schmidt-Thieme
M.Sc. Daniela Thyssens

Mar 25^{th}, 2021
ISMLL Universität Hildesheim

1. Unconstrained Optimization and Convexity

1A. Convexity of Functions (3 points)

Explain in your own words convexity of functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ by using the initial definition and drawing a small sketch! Why is convexity such an important property for minimizing objective functions in optimization?

1B. Unconstrained Optimization (5 (+ 2 BONUS) points)

Given the following matrices:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ -1 & -2 & -3 \\ 7 & 11 & 9 \end{pmatrix}, \quad Y = \begin{pmatrix} 21.5 \\ 51.5 \\ -27.5 \\ 83.0 \end{pmatrix}, \quad \beta^0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

For an un-regularized linear regression problem, **compute the following**:

a. **Two** iterations of Gradient Descent (GD). Use a constant step length $\mu = 0.005$. Show the initial value of the loss and the loss value after each iteration.

b. **One** iteration of Stochastic Gradient Descent (SGD). Use a constant step length $\mu = 0.005$, and for simplicity, use each instance in order. Show the initial value of the loss and the loss value after each iteration.

- **BONUS** [**2 points**]:

  - **One** iteration of Stochastic Gradient Descent (SGD) using an adaptive gradient algorithm (AdaGrad). Find a suitable starting step length μ and for simplicity, use each instance in order. Show the initial value of the loss function and its value after each iteration.
  - **One** iteration of Coordinate Descent (CD). Please use each instance in order. Show the initial value of the loss and the loss value after each iteration.

1C. Convex Function on and below secant line segment (2 points)

Suppose $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a convex function. Show that the following holds for all $x \in [a, b]$.

$$f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

2. Newton's Method and Subdifferentials

2A. Newton's Method for Unconstrained Optimization (3 points)

Derive the Newton's Step (formula) starting from the quadratic Taylor expansion of a function $f(x)$ around x_t. Comment each step of the derivation. And explain why the Newton's Step can in some cases be faster than gradient methods.

2B. Newton's Method for Constrained Optimization (5 points)

Consider the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & x_1^2 + x_2^2 \\ \text{subject to} \quad & x_2 - 2x_1 = \frac{1}{2} \end{aligned}$$

Solve **one iteration** of the optimization problem using Newton's Method with $x^{(0)} = (1, 1)$ by calculating $x^{(1)}$. Did the algorithm converge after this one iteration if we assume a convergence threshold of $\epsilon = 0.1$?

2C. Subdifferentials (2 points)

Given the following function

$$f(x) = \begin{cases} x^2 - x - 1, & \text{if } x < -2 \\ 5, & \text{if } x \in [-2, 2] \\ x^2 + x - 1 & \text{if } x > 2 \end{cases}$$

Draw a sketch of this function. Compute its subdifferential!

3. Equality Constraint Optimization and the Dual Problem

3A. KKT Conditions (3 points)

Consider the following problem:

$$\begin{aligned} \text{minimize} \quad & f(x_1, x_2) = x_1^3 + x_2^2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 9 \\ & x_2 \geq 0 \end{aligned}$$

Write down the KKT conditions and argue whether the conditions are necessary for guaranteeing local optimality for this problem. Are they sufficient?

3B. Computing the Dual Problem (5 points)

Consider the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & f(x_1, x_2) = 25 - a_1 x_1^2 - a_2 x_2^2 \\ \text{subject to} \quad & x_1 - x_2 = 1 \end{aligned}$$

a. Write down the Primal (Lagrangian) problem.

b. Compute the Dual problem $g(\lambda)$.

c. Solve the Dual problem in order to find expressions for x_1^* and x_2^*.

3C. Duality (2 points)

The SVM primal problem can be written as follows:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\beta\|^2 \\ \text{s.t.} \quad & y_i(\beta_0 + \beta^\top x_i) \geq 0, \quad i = 1, \ldots, n \\ & \beta \in \mathbb{R}^p, \quad \beta_0 \in \mathbb{R} \end{aligned}$$

Show that the dual problem can be written as follows:

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \end{aligned}$$

4. Inequality Constrained Optimization

4A. Active Set Method (2 points)

Explain in your own words how the Active Set method works. What are the limitations of the method and potential disadvantages? And what are the situations that would cause an empty active set?

4B. Inequality Constrained: Active Set Method (6 points)

Solve the following optimization problem using the Active Set method with **Quadratic Programming** to find the **search direction**. Use the backtracking line search to find μ to ensure feasibility of a given solution. Start with the initial point $x^{(0)} = (-1, 0)^T$.

$$
\begin{aligned}
\text{minimize} \quad & f(x_1, x_2) = x_1^2 + x_2^2 - 4x_1 - 4x_2 \\
\text{subject to} \quad & 2x_1 + x_2 \leq 2 \\
& -2x_1 + x_2 \leq 2 \\
& x_1 - x_2 \leq 1 \\
& -x_1 - x_2 \leq 1
\end{aligned}
$$

Hints:

- To find the optimal **search direction** Δx the QP to solve is $\begin{pmatrix} P & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \nu \end{pmatrix} = \begin{pmatrix} -\tilde{g}^{(t)} \\ 0 \end{pmatrix}$, where ν is the vector of slack variables and $\tilde{g}^{(t)} = Px^{(t)} + q$
- To stay in the feasible region, you may need to find the step size μ that projects an infeasible temporary solution, for example $x^{(1)}$, back to the feasible region by solving: $x^{(0)} + \mu \Delta x^{(1)} \geq 0$.

4C. Quadratic Penalty Function (2 points)

For quadratic penalty functions, the penalty is equal to the sum of square of all constraints, such that the penalty function is defined as below, where ρ is the penalty weight:

$$P(x, \rho) = \rho g(x)^\top g(x)$$

a. Can this formulation of the penalty function $P(x, \rho)$ handle inequality constraints?
   If yes, explain how. If not, reformulate the function $P(x, \rho)$ by defining a function $\phi(x, \rho)$ that can handle inequality constraints.

b. Show how penalty functions can suffer from ill conditioning, when considering inequality constraints. **Hint**: Consider that the penalty weight increases over time during the optimization procedure.