Modern Optimization Techniques

Second Take-home Exam

Prof. Dr. Dr. Lars Schmidt-Thieme
M.Sc. Eya Boumaiza

Dec 18^{th}, 2020
ISMLL Universität Hildesheim

Note:

- Time: 240 minutes
- Add your name and matriculation number on top of every page.
- Please provide clear and detailed answers to get full points.
- Please make sure the provided solution is clearly written and scanned.
- Solutions need to be submitted before the deadline. Please consider submitting their solutions 5 minutes before the deadline in order to ensure that the solution is uploaded and no internet problem interrupts it.
- Plagiarism, cheating and group submissions are not allowed. Any suspicious solution will be further investigated and the student will fail the course for this semester in case we prove it.

1. Newton and Quasi-Newton methods

1A. Convergence of Newton method (2 points)

Given the function $f(x) = ln(x)$. Discuss what happens when we use Newton method to optimize this function. When the method converges? and what is the disadvantage of using Newton method here when compared to Gradient Descent?

1B. Newton method (5 points)

In this question you have to use Newton method to perform a logistic regression task given the data below using the cross entropy loss. **Don't forget to add the bias!**

$$X = \begin{pmatrix} 5 & 2 \\ 3 & -1 \\ 4 & -1 \end{pmatrix} \quad Y = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Your tasks are:

a. Derive the update rule for a β^{t+1}.

b. Perform 2 iterations of the Newton method with an initial $\beta^0 = (1, 1, 1)^T$ and an update step $\mu = 0.1$ (you can use a calculator for the computation of the inverse).

c. At the end of each iteration, calculate the value of the cross entropy loss. Describe what happens with the loss.

1C. Quasi-Newton methods (3 points)

a. Explain in your own words, what is the most important criteria for Quasi-Newton methods to approximate the hessian. When can we use symmetric rank one update method?

b. Explain in your own words, what is the difference between BFSG and L-BFSG and discuss the memory needed and complexity of each of them in details.

2. Unconstrained Optimization: Coordinate Descent

2A. Coordinate Descent vs others (2 points)

a. Explain in your own words, what is the difference of coordinate descent as compared to gradient descent and stochastic gradient descent.

b. Explain in your own words, what are the advantages and disadvantages of using coordinate descent as compared to gradient descent and stochastic gradient descent and when they happen?

2B. Coordinate Descent Method (5 points)

In this question you have to use coordinate descent optimization method to find solution for a $L1$ regularized least squared objective function given the data below.

$$\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \end{pmatrix} \quad Y = \begin{pmatrix} 7 \\ 9 \\ 5 \end{pmatrix}$$

Your tasks are:

a. Derive the update rule for a β^{t+1} coordinate.

b. Perform 2 epochs (4 iterations) of the CD with an initial $\beta^0 = (2, 1)^T$ and a regularization weight $\lambda = 0.1$.

c. At the end of each iteration, calculate the value of the RMSE. Describe what happens with the error.

2C. Convergence of Coordinate Descent (3 points)

a. Assume the algorithm from the previous question converges to a fixed point β^*. Show that β^* is optimal, i.e. it minimizes our objective function. Hint: Use the sub-differential you have seen in the previous lecture and exercise sheet.

b. What happens when we use coordinate descent to optimize the following function:

$$g(x) = |x_1 x_2| + 0.1(x_1 + x_2)$$