10-725 Optimization, Spring 2010: Homework 5

Due: Wednesday, April 14, beginning of class

Instructions There are 6 questions on this assignment. The last question involves coding. Do not attach your code to the writeup. Instead, copy your implementation to /afs/andrew.cmu.edu/course/10/725/Submit/your_andrew_id/HW5 for andrew, or you can log into, for example, unix.andrew.cmu.edu. Please submit your homework with your name and userid on it. Refer to the webpage for policies regarding collaboration, due dates, and extensions.

1 Conjugate functions [Sivaraman, 15 points]

In class, Geoff discussed the dual \( f^*(y) = \sup_x (y^T x - f(x)) \) of a function \( f(x) \), which is also called the conjugate \( f^* \) of the function \( f \). Derive the conjugates of the following functions (3 points each):

1. Max function. \( f(x) = \max_{i=1,...,n} x_i \) on \( \mathbb{R}^n \).
2. Sum of largest elements. \( f(x) = \sum_{i=1}^r x[i] \) on \( \mathbb{R}^n \). (where \( x[i] \) denotes the \( i \)-th largest element in \( x \)).
3. Power function. \( f(x) = x^p \) on \( \mathbb{R}^{++} \), where \( p > 1 \).
4. Geometric mean. \( f(x) = -(\prod_{i=1}^n x_i)^{1/n} \) on \( \mathbb{R}^{++} \).
5. Negative generalized logarithm for second-order cone. \( f(x,t) = -\log(t^2 - x^T x) \) on \( \{(x,t) \in \mathbb{R}^n \times \mathbb{R}^n | \|x\|_2 < t\} \).

2 The Lagrange dual function and conjugate functions [Yi, 5 points]

The Lagrange dual (of an optimization problem) and the conjugate (of the objective function \( f \)) are closely related. This question will show a simple example.

Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{s.t.} & \quad Ax \leq b \\
& \quad Cx = d
\end{align*}
\]

with variables \( x \in \mathbb{R}^n \). Derive the Lagrange dual of this problem, and express it using the conjugate \( f^* \) of function \( f \).

3 Convexity, Strong Duality and KKT Conditions [Yi, 25 points]

3.1 A Convex Problem with Strong Duality Failed

Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad e^{-x} \\
\text{s.t.} & \quad x^2/y \leq 0
\end{align*}
\]

with variables \( x \) and \( y \), and domain \( D = \{(x,y) | y > 0\} \).
1. [3 pts] Verify that this is a convex optimization problem and check whether Slater’s condition holds or not.

2. [4 pts] Give the Lagrange dual problem, and find the optimal solution $\lambda^*$ and optimal value $d^*$ of the dual problem. What is the optimal duality gap?

### 3.2 KKT Conditions for Non-convex Problems

*This question is Problem 5.29 from Boyd & Vandenberghe.*

Consider this non-convex optimization problem:

\[
\begin{align*}
\min_{x_1, x_2, x_3} & \quad -3x_1^2 + x_2^2 + 2x_3^2 + 2(x_1 + x_2 + x_3) \\
\text{such that} & \quad x_1^2 + x_2^2 + x_3^2 = 1
\end{align*}
\]  \hfill (1)

We will see that although it is non-convex, strong duality still holds.

1. [3 pts] Derive and state the KKT conditions.

2. [3 pts] Find all solutions $x, \lambda$ (where $\lambda$ is the dual variable) which satisfy the KKT conditions. You can use any tool for this question.

3. [3 pts] Which pair $(x, \lambda)$ corresponds to the optimal solution, and what is the optimal value of the objective?

**Note** that multiple solutions satisfy the KKT conditions, but only one is optimal. For non-convex problems, a solution satisfying the KKT conditions is not necessarily optimal. However, if the objective and constraint functions are differentiable, strong duality implies that any optimal point must satisfy the KKT conditions.

### 3.3 KKT & Supporting Hyperplanes for Convex Problems

*This question is based on Problem 5.31 from Boyd & Vandenberghe.*

Consider this problem where all $f_i, i = 0, \ldots, m,$ are convex and differentiable:

\[
\begin{align*}
\min_x & \quad f_0(x) \\
\text{such that} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m
\end{align*}
\]  \hfill (3)

Assume $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^m$ satisfy the KKT conditions:

\[
\begin{align*}
f_i(x^*) & \leq 0, \quad i = 1, \ldots, m \\
\lambda_i^* & \geq 0, \quad i = 1, \ldots, m \\
\lambda_i^* f_i(x^*) & = 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^\top \nabla f_i(x^*) & = 0
\end{align*}
\]  \hfill (4)

1. [5 pts] Show that $\nabla f_0(x^\top)(x - x^*) \geq 0$ for all feasible $x$. (Note that this has a nice geometric interpretation: if $\nabla f_0(x^*) \neq 0$, then $\nabla f_0(x^\top)$ defines a supporting hyperplane to the feasible set at $x^*$.)

Since $f_0$ is convex, we know that for all $x, y$ in the domain of $f_0$,

\[
f_0(y) \geq f_0(x) + \nabla f_0(x\top)(y - x)
\]

So, if $\nabla f_0(x\top)(y - x) \geq 0$, then $f_0(y) \geq f_0(x)$. Thus, we have shown that if $x^*$ satisfies the KKT conditions, $f_0(y) \geq f_0(x^*)$ for all feasible $y$; i.e. $x^*$ is an optimal solution.
To get a better idea of this geometric interpretation, you’ll plot the supporting hyperplane to the feasible set at \( x^* \) for the following simple problem:

\[
\min_{x} x_1^2 + x_2^2 \quad \text{such that} \quad x_2 \geq 2x_1^2 - 7x_1 + 6
\]

2. [2 pts] Find the optimal solution \( x^* \) for this problem. (You may do this by hand or using e.g. CVX, but either way, show your work or your code.)

3. [2 pts] Plot the feasible region for this problem (with \( x_1 \) on the horizontal axis and \( x_2 \) on the vertical axis), as well as the supporting hyperplane for the feasible region defined by \( \nabla f_0(x^*)^T (x - x^*) \geq 0 \).

4 Optimality conditions for QCQP [Yi, 10 points]

[Ex. 5.26, B&V]. Consider the following Quadratic Constraints Quadratic Programming (QCQP) problem

\[
\begin{align*}
\text{minimize} & \quad x_1^2 + x_2^2 \\
\text{s.t.} & \quad (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1 \\
& \quad (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1
\end{align*}
\]

with variables \( x \in \mathbb{R}^2 \).

1. [3 pts] Sketch the feasible set and level sets of the objective. Find the optimal point \( x^* \) and optimal value \( p^* \).

2. [3 pts] Give the KKT conditions. Do there exist Lagrange multipliers \( \lambda_1^* \) and \( \lambda_2^* \) that prove that \( x^* \) is optimal?

3. [4 pts] Derive and solve the Lagrange dual problem. Does strong duality hold?

5 Dual of SOCP [Yi, 10 points]

Show that the dual of the SOCP

\[
\begin{align*}
\text{minimize} & \quad f^T x \\
\text{s.t.} & \quad \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m
\end{align*}
\]

with variables \( x \in \mathbb{R}^n \) can be expressed as

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{m} (b_i^T u_i + d_i v_i) \\
\text{s.t.} & \quad \sum_{i=1}^{m} (A_i^T u_i + c_i v_i) + f = 0 \\
& \quad \|u_i\|_2 \leq -v_i, \quad i = 1, \ldots, m
\end{align*}
\]

with variables \( u_i \in \mathbb{R}^n, v_i \in \mathbb{R}, \ i = 1, \ldots, m \). The problem data are \( f \in \mathbb{R}^n, A_i \in \mathbb{R}^{n_i \times n}, b_i \in \mathbb{R}^{n_i}, c_i \in \mathbb{R}^n \) and \( d_i \in \mathbb{R}, \ i = 1, \ldots, m \). Derive the dual in the following two ways.

(a) [5 points] Introduce new variables \( y_i \in \mathbb{R}^{n_i} \) and \( t_i \in \mathbb{R} \) and equalities \( y_i = A_i x + b_i, \ t_i = c_i^T x + d_i, \) and derive the Lagrange dual.

(b) [5 points] Start from the conic formulation of the SOCP and use the conic dual. Use the fact that the second-order cone is self-dual.
In this question you will implement Newton’s method for multi-class logistic regression.

Consider, the expression for K-class logistic regression,

\[ P(y = k | X = x) = \frac{\exp(w_k^T x)}{\sum_{k=1}^{K} \exp(w_k^T x)} \]  

where each \( w_i \) is a p-dimensional vector (where \( p \) is the number of features). More generally you would consider a bias term for each class, but you can ignore the bias term for this problem.

6.1 Theory

1. [2 pts] Consider you are given a training set \( (X, Y) = \{(X_1, y_1), \ldots, (X_n, y_n)\} \). Write an expression for the log-likelihood of the training set, and cast the problem of maximizing this log-likelihood with a ridge penalty of the form \( \frac{1}{2} \sum_{k=1}^{K} ||w_i||^2 \) on the parameters as an optimization problem.

2. [5 pts] Derive analytic expressions for the gradient, and the hessian of the problem.

3. [3 pts] Show analytically that the hessian you derived is negative semidefinite. Also, show that if the weight on the ridge term \( \lambda > 0 \), then the objective is “strictly” concave. A function \( f \) is strictly concave if \( f(x) < f(x_0) + g(x_0)^T(x - x_0) \) for all \( x \) in \( \text{dom}(f) \) where \( g \) is the gradient (or an element of the sub-gradient) of \( f \).

   \text{Hint: A diagonally dominant symmetric matrix with real entries with positive diagonal entries is positive semidefinite. You can additionally use the fact that the Kronecker product of two, PSD matrices is PSD. Show that the Hessian can be written as the Kronecker product of a PSD matrix and a diagonally dominant one. Are the diagonal entries of the diagonally dominant matrix positive or negative? What can you conclude?}

6.2 Implementation

In this part of the question you will work with two data sets, the first is a synthetic dataset and the second is the UCI Iris dataset.

1. [10 pts] Implement Newton’s method for this problem. Your implementation should accept a value for \( \lambda \), however don’t use values of \( \lambda \) that are too small, or your hessian might not be invertible. Declare convergence whenever the L2-norm of the parameter vector (or matrix) changes by less than \( 10^{-6} \).

2. [7 pts] You should now download the synthetic dataset (“data_large.mat”) from the website. Train your logistic regressor on the training set, and use the learned weights to predict class labels for the test set. In your report you should show the weights you learned (you can print these out since this is a big vector or matrix but clearly indicate the class and feature for the weights). Report your accuracy on the test set. Use \( \lambda = 0.01 \) for this part.

3. [6 pts] Download “data_small.mat” from the website. Plot regularization paths for the weights (you should have only 4 weights), by varying \( \lambda \) (values of \( \{10^{-2}, 1, 10^2, 10^4, 10^6, 10^8, 10^{10}\} \) work well for me but you can use any values as long as you get a reasonable curve).

4. [7 pts] Download the UCI Iris dataset from the website (we have split the data for you into train, validation and test sets). Use the validation set to pick the best value of \( \lambda \). In your report, clearly write down the value of \( \lambda \) and your accuracy on the training, test and validation sets.