Homework 4
Trust Region, Cond. Gradient/Frank Wolfe
Lagrange Mult. Theory, Augmented Lagrangian, Barrier Methods

CMU 10-725/36-725: Convex Optimization (Fall 2017)

OUT: Oct 20
DUE: **Nov 3, 5:00 PM**

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3.4"). Second, write your solution independently: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submissions can be written in LaTeX. Upon submission, label each question using the template provided by Gradescope. Please refer to Piazza for detailed instruction for joining Gradescope and submitting your homework.
- **Programming**: All programming portions of the assignments should be submitted to Gradescope as well. We will not be using this for autograding, meaning you may use any language which you like to submit.

1 Trust Region [Hao; 25 pts]

Let

$$f(x) = \frac{1}{2}x_1^2 + x_2^2, \tag{1}$$

and let $x_0 = (1, 1)$, $g = \nabla f(x_0)$, $B = \nabla^2 f(x_0)$.

(a) **[10pt]** Explicitly compute the next step in the trust region method using values of $\triangle = 2$ and $\triangle = \frac{5}{6}$.

[Hint: The only non-negative solution for $\frac{1}{(1+\lambda)^2} + \frac{4}{(2+\lambda)^2} = \frac{25}{36}$ is $\lambda = 1$.]

(b) **[15pt]** Compute the next step in the dogleg method for all $\triangle > 0$.

2 Frank Wolfe [Hongyang; 25 pts]

In this problem, we will derive the convergence rate of Frank-Wolfe algorithm to address general constrained convex optimization problem

$$\min_{x \in \mathcal{D}} f(x), \tag{2}$$

where the objective function f is convex and continuously differentiable, and that the domain $\mathcal{D}$ is a compact convex subset of any vector space. Recall that the Frank-Wolfe algorithm is given by

1: Let $x^{(0)} \in \mathcal{D}$.
2: for $k = 0, 1, ..., K$ **do**
3: Compute $s := \text{argmin}_{s \in \mathcal{D}} \langle s, \nabla f(x^{(k)}) \rangle$
4: Update $x^{(k+1)} := (1 - \gamma)x^{(k)} + \gamma s$, for $\gamma := \frac{2}{k+2}$
5: end for

Algorithm 1: Frank-Wolfe Algorithm.

We define the curvature constant C_f of $f : \mathbb{R}^n \to \mathbb{R}$ with respect to domain $\mathcal{D}$ by

$$C_f := \sup_{x, s \in \mathcal{D} \gamma \in [0,1], y = x + \gamma(s-x)} \frac{2}{\gamma^2}(f(y) - f(x) - \langle y - x, \nabla f(x) \rangle).$$

Prove that for a step $x^{(k+1)} := x^{(k)} + \gamma(s - x^{(k)})$ with arbitrary step-size $\gamma \in [0, 1]$, it holds that

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \gamma g(x^{(k)}) + \frac{\gamma^2}{2} C_f,$$

if s is an approximate linear minimizer, i.e., $\langle s, \nabla f(x^{(k)}) \rangle = \min_{\hat{s}} \langle \hat{s}, \nabla f(x^{(k)}) \rangle$, where $g(x)$ is the duality gap defined by

$$g(x) = \max_{s \in \mathcal{D}} \langle x - s, \nabla f(x) \rangle.$$

3 Lagrange Methods and Augmented Lagrangian [Devendra; 25 pts]

(a) **[12pts]** A quantity of Q units of a certain product is available for sale in n outlets. For each $i = 1, .., n$ the quantity d_i, which is demanded at outlet i, and its sale price p_i are known. We wish to determine

the quantities s_i^* with $0 \le s_i^* \le d_i$ to be sold at each outlet that maximizes the revenue $\sum_{i=1}^n p_i s_i$ from the sale.

- Assuming that $d_i > 0$, $p_i > 0$, and $\sum_{i=1}^n d_i \ge Q$, show that there exists a cutoff price level y such that for each i, if $p_i > y$, then $s_i^* = d_i$, and if $p_i < y$, then $s_i^* = 0$.
- What happens if $p_i = y$?
- Describe a procedure for obtaining s_i^*.

(b) **[13pts]** Consider the problem of

$$\begin{aligned} \text{minimize} \quad & f(x) = \frac{1}{2}(x_1^2 - x_2^2) - 3x_2 \\ \text{subject to} \quad & x_2 = 0 \end{aligned}$$

- Calculate the optimal solution and the Lagrange multiplier.
- For $k = 0, 1, 2$ and $c^k = 10^{k+1}$, calculate and compare iterates of the quadratic penalty method with $\lambda^k = 0$ for all k and the method of multipliers with $\lambda^0 = 0$. Here λ is Lagrange multiplier and c is positive penalty parameter.
- If c is a constant in the method of Multipliers, for what values of c, would the method converge?

4 Barrier Methods for Support Vector Regression [Yichong; 25 pts]

In this problem we develop barrier methods for support vector regression(SVR). Suppose we have a set of data points $(X_1, Y_1), ..., (X_n, Y_n)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ for all i, and we want to fit a linear model of $y = w^T x + b$. For simplicity, we append a 1 to the end of each X_i so that we don't need to consider b. An ε-SVR solves the convex optimization problem

$$\begin{aligned} \min_{w \in \mathbb{R}^{d+1}, \xi \in \mathbb{R}^n} \tilde{f}(w, \xi) &= \frac{1}{2}\|w_{1:d}\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{s.t. } |Y_i - w^T X_i| &\le \varepsilon + \xi_i, \\ \xi_i &\ge 0, i = 1, 2, ..., n. \end{aligned}$$

Here ε and C are parameters of SVR, and $w_{1:d}$ is the first d dimension of w. Please refer to http://www.stat.cmu.edu/ ryantibs/convexopt/lectures/barr-method.pdf (page 12) for a description of barrier methods.

(a) **[3pts]** Derive the objective $f(w, \xi)$ of barrier method (with parameter t) for SVR problem.

(b) **[8pts]** Compute derivative and Hessian of f with respect to w and ξ.

(c) **[2pts]** Describe how to generate the initial point for barrier method.

(d) **[12pts]** Implement barrier method with backtracking line search on data in birthwt.zip. birthwt.zip contains data of features of newborn babies to predict their birth weight. Centralize the data such that $\sum_{i=1}^n X_{i,j} = 0$ for all $j \in \{1, 2, ..., d\}$, where $X_{i,j}$ is the j-th dimension of X_i. Use SVR with $C = 10, \varepsilon = 0.1$ to predict birth weights. For the barrier parameters t(the multiplier for the original objective of SVR) and μ(the constant by which t increases at each outer iteration of the barrier method), try two settings: $(t, \mu) = (5, 20)$ and $(t, \mu) = (10000, 5)$. A good number for m(the constant that, together with t, bounds the duality gap) is the number of constraints in the barrier problem. For backtracking during the Newton method steps, use $\alpha = 0.2$ and $\beta = 0.9$. You can use 1e-9 as the stopping threshold for both the Newton method and the barrier method.

Record the SVR objective and MSE $g(w) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - w^T X_i)^2$ for each outer step, i.e., each time you update values of w (the steps of backtracking line search is not considered). Give two plots comparing the performance for two settings of (t, μ), one on SVR objective f (excluding barrier terms) and one on MSE g. Plot the x-axis as the number of iterations, and y-axis as $\log\big(f^{(k)} - f^*\big)$ and $\log\big(g^{(k)} - g^*\big)$, with $f^* = 729.652$ and $g^* = 0.365$.