CMSC764 Homework 6 – Gradient methods and duality: Due Wed April 25

You must turn in two things. (1) Your code, including a script called `hmwk6` that runs all the experiments (and generates the plots) asked for below. (2) A short writeup (pdf) containing the plots asked for below, and your answers to the questions.

1. (a) Write a method with signature

```
function x_sol, res = grad_descent(f,grad,x0)
```

The inputs `f` and `grad` are function handles. The function `f`: $\mathbb{R}^N \to \mathbb{R}$ is an arbitrary objective function, and `grad`: $\mathbb{R}^N \to \mathbb{R}^N$ is its gradient. The method should minimize $f$ using gradient descent, and terminate when the gradient of $f$ is small. I suggest stopping when

$$\|\nabla f(x^k)\| < \|\nabla f(x^0)\| * tol$$

where $x^0$ is an initial guess and $tol$ is a small tolerance parameter (a typical value would be $10^{-4}$).

Use a backtracking line search to guarantee convergence. The stepsize should be monotonically decreasing. Each iteration should begin by trying the stepsize that was used on the previous iteration, and then backtrack until the Armijo condition holds:

$$f(x^{k+1}) \le f(x^k) + \alpha\langle x^{k+1} - x^k, \nabla f(x^k)\rangle,$$

where $\alpha \in (0,1)$, and $\alpha = 0.1$ is suggested.

The function returns the solution vector `x_sol`, and also a vector `res` containing the norm of the residual (i.e., the norm of the gradient) at each iteration.

You can obtain the initial stepsize by estimating a Lipschitz constant for `f` using the formula:

$$L \approx \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|}$$

where $x = x^0$ and $y$ is obtained by adding a small random perturbation to $x$. The initial stepsize is then $\tau = \frac{2}{L}$. Use a vector of zeros as the initial guess $x^0$.

(b) Test your solver using the logistic regression classification problem from homework 4. Remember, you already have code for generating this objective function and gradient. You handed these functions to your gradient checker.

Your test should use a classification problem with 200 feature vectors, 20 features per vector, and condition number $\kappa = 1$. Then use your solvers to minimize `logreg_objective` using the gradient function `logreg_grad`. Use a stopping condition with $tol = 10^{-6}$ to verify that the methods converge to a high degree of precision. Plot the residuals using a *logarithmic* y-axis. For example, in Matlab you could call

```
[D, c] = create_classification_problem(200, 20, 1);
[x_sol, res] = grad_descent( @(x) logreg_objective(x,D,c),
                             @(x) logreg_grad(x,D,c), x0);
semilogy(res)
```

Now, re-run the experiment with condition number $\kappa = 100$.

**Create one plot with residual curves for $\kappa = 1$ and $\kappa = 100$. Make sure the plot has labeled axes and legends.**

(c) Answer the following questions. Keep your answers short (but complete).

a Why is the stopping condition in problem 1 based on the residual? Why don't we just terminate when the objective function gets small?

b Why is the stopping condition suggested in problem 1 better than just terminating when the residual is small, i.e. when $\|\nabla f(x^k)\| < tol$.

d What effect does the condition number have on the convergence speed?

2. Consider the "monotropic" program

$$\begin{aligned} \text{minimize} \quad & \|x\|_\infty \\ \text{subject to} \quad & Ax = b. \end{aligned} \tag{1}$$

(a) Write this as an unconstrained (or implicitly constrained) problem using the characteristic function of the zero vector $\chi_0(z)$. This function is zero if it's argument is zero, and infinite otherwise.

(b) What is the conjugate of $f(z) = \|z\|_\infty$?

(c) What is the conjugate of $g(z) = \chi_0(z)$?

(d) Using the conjugate functions, write down the dual of (1).

3. Consider the linear program

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0. \end{aligned}$$

(a) Write the optimality conditions for this problem (i.e., the KKT system).

(b) Write the Lagrangian for this problem.

(c) Minimize out the primal variables in the Lagrangian, and write the dual formulation of this linear program.

4. (a) Let $\text{prox}_f(x,t) = \arg\min_z f(z) + \frac{1}{2t}\|z - x\|^2$. Prove the "Moreau decomposition" identity

$$x = \text{prox}_f(x,t) + t\,\text{prox}_{f^*}(x/t, 1/t).$$

(b) Using your result from part (a), prove that

$$\text{prox}_{|x|}(x,t) = x - \max\{\min\{x,t\}, -t\}$$

where $|x|$ denotes the 1-norm of $x$, and "min" and "max" are applied element-wise.