Homework 4

Convex Optimization 10-725/36-725

Due Friday November 4 at 5:30pm
submitted to Christoph Dann in Gates 8013
(Remember to a submit separate writeup for each problem, with your name at the top)

Total: 75 points
v1.4

1 Newton's Method [Christoph] (16pts)

(a) Invariance Under Affine Transformation

(i, 4pts) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and twice differentiable, $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ be invertible. Define g as $g(x) = f(Ax + b)$ for all x and let $u_0 \in \mathbb{R}^n$ be arbitrary but fix. A step of Newton's method applied to f at u_0 results in

$$u_1 = u_0 - \left(\nabla^2 f(u_0)\right)^{-1} \nabla f(u_0). \tag{1}$$

Show that a step of the Newton's method applied to g at $x_0 = A^{-1}(u_0 - b)$ results in $x_1 = A^{-1}(u_1 - b)$.

This will imply that $g(x_1) = f(u_1)$, that is, the criterion values match after a Newton step. This will continue to be true at all iterations, and thus we say that Newton's method is affine invariant.

(ii, 1pt) Show that gradient descent is not invariant under affine transformation by providing a concise counterexample. Be specific, that is, define a function f, an affine transformation A, b and an intial u_0 at least.

(b) Newton Decrement Characterization (5pts)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and twice differentiable, $b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times n}$ be a matrix with rank p. Assume $\hat{x}$ satisfies $A\hat{x} = b$. Recall that the Newton step Δx at $\hat{x}$ for problem $\min f(x)$ sb.t. $Ax = b$ is the solution of the linear equations

$$\begin{bmatrix} \nabla^2 f(\hat{x}) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ u \end{bmatrix} = \begin{bmatrix} -\nabla f(\hat{x}) \\ 0 \end{bmatrix}, \tag{2}$$

and the Newton decrement $\lambda(\hat{x})$ is given by

$$\lambda(\hat{x}) = \sqrt{-\nabla f(\hat{x})^\top \Delta x} = \sqrt{\Delta x^\top \nabla^2 f(\hat{x}) \Delta x}. \tag{3}$$

Assume the coefficient matrix in the linear equations above is nonsingular and that $\lambda(\hat{x}) > 0$. Express the solution y of the optimization problem

$$\min_y \ \nabla f(\hat{x})^\top y \tag{4}$$

$$\text{subject to } \ Ay = 0 \tag{5}$$

$$y^\top \nabla^2 f(\hat{x}) y \leq 1 \tag{6}$$

in terms of the Newton step Δx and the Newton decrement $\lambda(\hat{x})$.

(c) Quadratic Convergence of Newton's Method in $\mathbb{R}$ (6pts)

Let $f : \mathbb{R} \to \mathbb{R}$ be convex and three times continuously differentiable with $|f'''(x)| \leq C_1$ bounded and $f''(x) \geq C_2 > 0$ for all x. Let x^* be a local minimum. Show that the Newton method iterates $x_1, x_2, \ldots$ converge quadratically towards the optimum x^*, that is,

$$|x_{k+1} - x^*| = O(|x_k - x^*|^2). \tag{7}$$

2 Nuclear norm, duality, and matrix completion (17 pts) [Alnur]

In this problem, we will take a look at convex optimization problems involving *matrix* variables. We touched on these a little bit in class, and while they might seem somewhat abstract, these sorts of problems are very much connected to real-world applications like Netflix's movie recommendation engine (you are welcome to ask us for further details here).
Let $X \in \mathbb{R}^{m \times n}$ be a matrix. The *trace norm* (also known as the *nuclear norm*) of a matrix X, which we write as $\|X\|_{\text{tr}}$, can be defined as the sum of the singular values of X.

(a, 5pts) Show that computing the trace norm of a matrix, i.e., computing $\|X\|_{\text{tr}}$, can be expressed as the following (convex) optimization problem:

$$\begin{array}{ll} \underset{Y \in \mathbb{R}^{m \times n}}{\text{maximize}} & \text{tr}(X^T Y) \\ \text{subject to} & \begin{bmatrix} I_m & Y \\ Y^T & I_n \end{bmatrix} \succeq 0, \end{array} \tag{8}$$

where I_p is the $p \times p$ identity matrix. (By the way, problem (8) is a semidefinite program; more on this in part (d) below.)

Hint: think about using the "Schur complement" somewhere here. A good reference for this might be Section A.5.5 in the "Convex Optimization" book, by Stephen Boyd and Lieven Vandenberghe.

(b, 5pts) Show that the dual problem associated with (8) can be expressed as

$$\begin{array}{ll} \underset{\substack{W_1 \in \mathbb{S}^m, \\ W_2 \in \mathbb{S}^n}}{\text{minimize}} & \text{tr}(W_1) + \text{tr}(W_2) \\ \text{subject to} & \begin{bmatrix} W_1 & (1/2)X \\ (1/2)X^T & W_2 \end{bmatrix} \succeq 0, \end{array} \tag{9}$$

where, just to remind you, $\mathbb{S}^p$ is the space of $p \times p$ real, symmetric matrices.

(c, 2pts) Show that the optimal values for problems (8) and (9) are equal to each other, and that both optimal values are attained.

(d, 5pts) In the *matrix completion problem*, we want to find a matrix $X \in \mathbb{R}^{m\times n}$ of low rank that is close, in a squared error sense, to some observed matrix $Z \in \mathbb{R}^{m\times n}$. We do not assume that all of the entries of Z are observed, so we will look at the squared error over Z's observed entries only, which we store in a set Ω of (observed) row and column indices. Putting all this together leads us to the following (convex) optimization problem:

$$\underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega}(X_{ij}-Z_{ij})^2 + \lambda\|X\|_{\text{tr}}, \tag{10}$$

with tuning parameter $\lambda > 0$.

Show that problem (10) can be expressed as a semidefinite program of the form

$$\begin{array}{ll}\underset{x\in\mathbb{R}^p}{\text{minimize}} & c^T x \\ \text{subject to} & x_1A_1 + \cdots + x_pA_p \preceq B,\end{array}$$

for some fixed $c, B, A_i,\ i = 1,\ldots,p$.

Hint: you will probably need to use each of the above parts (in different ways) here.

3 Second Order Methods for Logistic Regression (20 pts) [Mariya]

In this problem, we'll revisit the goal of classifying a viewer's age group from his movie ratings. We again formulate the problem as a binary classification with output label $y \in \{0,1\}^n$, corresponding to whether a person's age is under 40, and input features $X \in \mathbb{R}^{n\times(p+1)}$. Similarly to the second homework, the first column of X is taken to be 1_n to account for the intercept term. We will apply second order methods to solve the logistic regression problem.
We model each $y_i|x_i$ with the probabilistic model

$$\log\left(\frac{p_\beta(y_i=1|x_i)}{1-p_\beta(y=1|x_i)}\right) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = (X\beta)_i,$$

for $i = 1,\ldots,n$, $\beta \in \mathbb{R}^{p+1}$. We use β_0 to denote the intercept and $\beta_{1:p}$ for the rest of the weights. As a reminder, the negative log likelihood (NLL) under the logistic probability model can be expressed as:

$$f(\beta) = -\sum_{i=1}^n y_i(X\beta)_i + \sum_{i=1}^n \log(1+\exp\{(X\beta)_i\}),$$

(a) In this part of the problem, we will implement Newton's method for the nonpenalized logistic regression problem and see that in this context, Newton's method is known as Iteratively Reweighted Least Squares (IRLS).

(i, 1pt) Express the gradient and the Hessian of the NLL in terms of X, y, and μ.

(ii, 1pt) Write the Newton update for β using the above gradient and Hessian, using t to denote step-size.

(iii, 2pts) Show that the Newton update has the form of a weighted least-squares estimation problem. This is why in this context, Newton's method is known as IRLS.

(iv, 2pts) Given X and y, write out the steps for performing IRLS to estimate $\hat{\beta}$. Assume a fixed step size t in the steps.

(v, 3pts) Now, implement IRLS with backtracking using the movie data set on the website (in `hwk4_gs3.zip`, for this homework, not the second one). Initialize your weights with zeros. Use $\alpha = 0.01$ and shrink parameter $\beta = 0.9$ for backtracking. You can stop IRLS when the change between consecutive objective values is less than 1e-6. Report both the train and test errors of classifying whether a person is under 40. Plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at outer iterations k of IRLS, and the optimal objective value is $f^\star = 186.637$ on a semi-log scale (i.e. where the y-axis is in log scale).

(b) Now, we introduce regularization to the NLL to improve our test error. The problem becomes:

$$\min_{\beta \in \mathbb{R}^{p+1}} \quad f(\beta) + \lambda \|\beta_{1:p}\|_1 \tag{11}$$

(i, 2pts) Rewrite (11) as a problem that has a smooth criterion, aims to preserve the sparsity of the original problem, and can be solved by an interior point method. (Hint: consider introducing a new variable and corresponding inequality constraints).

(ii, 3pts) Describe the iterations of the barrier method for the problem in (i), explicitly deriving the Newton updates.

(iii, 6pts) Implement the barrier method described above. For the barrier parameters t(the multiplier for the original criterion in (11)) and μ(the constant by which t increases at each outer iteration of the barrier method), use $\mu = 20$ and start with $t = 5$. 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. Remember to initialize the centering Newton method step with a strictly feasible point (hint: in MATLAB, one simple way to find such a point is with *linprog*).

Use the barrier method with $\lambda = 15$ to classify whether a viewer is under 40. Report the train and test classification errors. Report the number of zeros at the solution $\beta^\star$. Here we consider any number with absolute value under 1e-10 to be zero.

Now, revisit your code from homework 2 for proximal gradient descent with backtracking for the lasso. Using shrinkage parameter $\beta = 0.5$ and $\lambda = 15$, run proximal gradient descent with backtracking on the training data.

To compare the convergence of both the barrier method and the proximal gradient descent, plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at outer iterations k of the algorithms, and the optimal objective value is $f^\star = 306.476$ on a semi-log scale (i.e. where the y-axis is in log scale). Plot these in the same figure.

Attach all relevant code to the end of this problem.

4 Interior Point Methods for SVMs [Han & Justin] (22 pts)

Overview In this question we will develop the dual of the primal optimization problem in kernel support vector machine, and solve it using the barrier method and the primal-dual interior point method, on the dataset `Q4c_movies.zip` on the class website. Recall from homework 2 that we used group lasso to classify a persons age group (above/under 40) from his movie ratings. Here, we will ignore the movie groupings and try classification on the features. Note, the labels are coded as 0's and 1's in this dataset.

Kernel SVM Suppose we have a sample set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of labeled examples in $\mathbb{R}^d$ with labels $y_i \in \{1, -1\}$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^D$ be a feature map that transform each input example

to a feature vector in $\mathbb{R}^D$. The primal optimization of kernel SVM is given by

$$\begin{aligned} \underset{\beta,\xi_i}{\text{minimize}} \quad & \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\beta^T\phi(x_i)+\beta_0) \geq 1-\xi_i \quad \forall i = 1,\ldots,n \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad \forall i = 1,\ldots,n \end{aligned}$$

which is equivalent to the following dual optimization

$$\begin{aligned} \underset{w}{\text{maximize}} \quad & 1^T w - \frac{1}{2}w^T\tilde{K}w \\ \text{subject to} \quad & 0 \leq w \leq C1, \quad w^T y = 0 \end{aligned} \tag{12}$$

where we define $K_{ij} = \langle\phi(x_i),\phi(x_j)\rangle$, and $\tilde{K}_{ij} = y_i y_j K_{ij}$. $\xi_1,\ldots,\xi_n$ are called slack variables. As we saw in a previous homework, the optimal slack variables have intuitive geometric interpretations. Basically, when $\xi_i = 0$, the corresponding feature vector $\phi(x_i)$ is correctly classified and it will either lie on the margin of the separator or on the correct side of the margin. Feature vectors with $0 < \xi_i \leq 1$ lie within the margin but are still correctly classified. When $\xi_i > 1$, the corresponding feature vector is misclassified. Support vectors correspond to the instances with the dual variable $w_i > 0$, or alternatively, these correspond to instances that lie on the margin or instances with $\xi_i > 0$. The optimal primal vector β^* can be represented in terms of the dual optimal $w_i^*, i = 1,\ldots,n$ as $\beta^* = \sum_{i=1}^n w_i^* y_i \phi(x_i)$. To solve for β_0^* from the dual, we can pick any instance j that lies on the margin, and compute β_0^* as

$$\beta_0^* = y_j - (\beta^*)^T\phi(x_j) = y_j - \sum_{i=1}^{n} w_i^* y_i K_{ij}$$

In practice, in order to reduce the variance, we can take the average as

$$\beta_0^* = \frac{1}{|J|}\sum_{j\in J}\left(y_j - \sum_{i=1}^{n} w_i^* y_i K_{ij}\right)$$

where J is the set of instances on the margin.
Lastly, the prediction for a given vector $x \in \mathbb{R}^d$ can be calculated as follows:

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{n} w_i^* y_i \langle\phi(x),\phi(x_i)\rangle + \beta_0^*\right) = \text{sign}\left(\sum_{i=1}^{n} w_i^* y_i K(x,x_i) + \beta_0^*\right) \tag{13}$$

In this problem we will use the feature map ϕ induced by the RBF kernel as:

$$\langle\phi(x_i),\phi(x_j)\rangle = K_{ij} = \exp\left(-\frac{||x_i - x_j||_2^2}{2\sigma^2}\right)$$

where we fix $\sigma = 1000$, i.e., $\sigma^2 = 10^6$.

(a) **Barrier Method**

(i, 5pts) Derive the gradient and the Hessian of the dual criterion $f(w)$, and of the log barrier objective of the form $tf(w) + g(w)$, where $g(w)$ is the log barrier term which takes the role of the inequality constraint. (Note, the equality constraint remains unchanged in the log barrier problem!) Confirm that the Hessian only changes from that of the original version in the diagonal entries. (Hint: This may be helpful in understanding the update directions in the interior point implementation in part (b).)

(ii, Bonus question, 5pts) Implement the barrier method as described in class to solve the SVM dual. Note, you will be using an equality-constrained Newton method with backtracking, in the barrier method updates. Using initial $t = 1000$ and $C = 1000$, run your barrier method until $m/t < 10^{-8}$. Use $\mu = 1.5$. Several steps you should address:

i. Describe how to obtain an initial feasible point in our problem.

ii. Describe how to obtain the equality-constrained Newton update direction.

iii. Describe the stopping rule for in terms of the Newton decrement, using the equality-constrained Newton direction update. Note, your stopping criterion should be sufficiently small, for your inner Newton to have good precision.

iv. Describe how to find an initial step size prior backtracking step. (Hint: the key is to choose a step size that would allow the upcoming update to be feasible; there are several ways to do this.)

v. For backtracking line search, you can use $\alpha = 0.01$ and $\beta = 0.5$, per notation in the lecture notes.

Report the optimal objective value at the optimal training weights w^*: $f^* = \frac{1}{2}(w^*)^T \tilde{K} w^* - 1^T w^*$. Report the number of support vectors, considering all $w^* < 10^{-6}$ to be zero. Use the optimal w^* to do prediction on both the training set and the test set, report both the training set and the test set classification accuracies. A good check is to see if your answer agrees with (b) (or with CVX).

(b) **Primal-Dual Interior Point Method**

(i, 2pts) For a given $t > 0$,List the perturbed KKT condition for this problem.

(ii, 2pts) Let $u_i \geq 0$ be the dual variable correspond to the inequality constraint $-w_i \leq 0, \forall i$ and $v_i \geq 0$ be the dual variable correspond to the inequality constraint $w_i \leq C, \forall i$. Let λ be the dual variable correspond to the equality constraint $y^T w = 0$. Write down the expression for the primal residual r_{prim}, the dual residual r_{dual} and the centrality residual r_{cent}.

(iii, 2pts) Define the residual vector $r(w, u, v, \lambda) = (r_{\text{dual}}, r_{\text{cent}}, r_{\text{prim}})$ and $z = (w, u, v, \lambda)$. The primal-dual interior point method is trying to find an update direction $\Delta z = (\Delta w, \Delta u, \Delta v, \Delta\lambda)$ such that $r(z + \Delta z) = 0$. However, in general $r(z + \Delta z) = 0$ corresponds to a system of nonlinear equations that does not admit closed form solution, so instead we make a first-order approximation to $r(z + \Delta z)$ and solves the corresponding linear system: $r(z + \Delta z) \approx r(z) + Dr(z)\Delta z = 0$. Write down the linear system to be solved in order to get Δz. Δz is also known as the Newton direction.

(iv, 2pts) Primal-dual interior point method requires a strictly feasible point of (12) as a start point. Form a linear program that can be used to find such a strictly feasible initial point.

(v, 9pts) Implement the primal-dual interior point method to solve the minimization problem you formed in (i). We fix $C = 1000$. You should stop your algorithm when both the following two conditions are met:

- $(||r_{\text{prim}}||_2^2 + ||r_{\text{dual}}||_2^2)^{1/2} \leq 10^{-6}$.
- The surrogate duality gap is less than or equal to 2×10^{-6}.

Here is a list of tips that might be helpful to you:

- Use the LP formed in the last question to find a strictly feasible initial point.
- After obtaining each Newton direction, it is important to compute a corresponding step size such that both the primal and the dual variables are feasible.
- For backtracking line search, you can use $\alpha = 0.01$ and $\beta = 0.5$ (Note this β corresponds to the parameter to shrink the step size, not the β in the primal SVM).

- It is important to tune to barrier parameter μ such that your algorithm stops within a fair amount of iterations. As a reference, the TA's implementation stops in 52 iterations.

Report the function value at the optimal w^*: $f^* = \frac{1}{2}(w^*)^T \tilde{K} w^* - 1^T w^*$. Report the number of support vectors. Here we consider any w^* that is less than 10^{-6} to be 0. Use the optimal w^* to do prediction on both the training set and the test set, report both the training set and the test set classification accuracies.

Attach all relevant code to the end of this problem.