Homework 4

Convex Optimization 10-725/36-725

Due Friday November 13 at 4:00pm
submitted to Mallory Deptola in GHC 8001
(Remember to submit each problem on a separate sheet of paper, with your name on at the top)

1 Newton's method [Dallas]

For the questions below, assume f is strongly convex with parameter m, ∇f is Lipschitz with parameter M, and $\nabla^2 f$ is Lipschitz with parameter L.

1. Assume that we are in the quadratic convergence phase of Newton's method (such that $\|\nabla f(x^{(k)})\|_2 \leq \eta$). Starting with $\|\nabla^2 f(x + t\Delta x) - \nabla^2 f(x)\|_2 \leq tL\|\Delta x\|_2$ (which we can obtain from the Lipschitz condition on $\nabla^2 f(x)$), show that

$$\tilde{f}(t) \leq \tilde{f}(0) - t\lambda(x)^2 + t^2\frac{1}{2}\lambda(x)^2 + t^3\frac{L}{6m^{3/2}}\lambda(x)^3$$

   where $\tilde{f}(t) = f(x + t\Delta x)$.

   You may use the following:

$$\begin{aligned}\tilde{f}''(t) &= \Delta x^T \nabla^2 f(x + t\Delta x)\Delta x \\ \tilde{f}'(0) &= -\lambda(x)^2 \\ \tilde{f}''(0) &= \lambda(x)^2 \geq m\|\Delta x\|_2^2\end{aligned}$$

2. Now take $t = 1$, and assume $\eta \leq 3(1 - 2\alpha)m^2/L$, and show that

$$f(x + \Delta x) \leq f(x) + \alpha\nabla f(x)^T\Delta x$$

   (i.e. a step size of $t = 1$ will always be accepted).

3. Still in the quadratic convergence phase, for $\eta \leq m^2/L$, and step size $t = 1$, we can show that

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$$

   Use this to prove that

$$f(x^{(l)}) - f^* \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{l-k+1}}$$

   for $l > k$, where k is the iteration at which we begin quadratic convergence.

2 Iteratively reweighted least squares for logistic regression [Dallas, but you can blame Ryan if this homework seems long]

Consider binary logistic regression. We assume that $y_i \sim \text{Bernoulli}(\mu_i)$, independently, where

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta,$$

for $i = 1, \ldots n$. Thus $p(y_i = 1|x_i) = \mathbb{E}(y_i) = g^{-1}(x_i^T\beta)$. We can think of this as doing linear regression of $g(y)$ onto x, where the variance of y depends on x.

1. Write out the negative log likelihood (NLL) for the logistic regression problem with y_i drawn i.i.d. from a Bernoulli distribution with unknown parameter μ_i.
2. Write out the gradient and the Hessian of the NLL as functions of X, y and μ.
3. Write out the Newton update using the above gradient and Hessian. (You can assume a fixed step size of 1).
4. Show that this takes the form of an weighted least-squares estimation problem.
5. Newton's method in this context is known as Iteratively Reweighted Least Squares (IRLS). Given X and y, write out a step-by-step algorithm to perform IRLS to estimate $\hat{\beta}$.

As a side note, since each iteration of IRLS is a weighted linear regression problem, we can use the weighted regression connection to compute useful quantities, such as standard errors. Many software packages will use IRLS to solve logistic regression problems (e.g. `glm()` in R), and will obtain standard errors from the final iteration.

3 Sparse eigenvectors via the barrier method [Matt]

In this problem we consider finding a sparse "eigenvector" of a matrix $A \succeq 0$.

(a) Consider solving the optimization problem

$$\begin{aligned} \max_x \ & x^T A x \\ \text{subject to } & \|x\|_2 = 1 \\ & \|x\|_1 \leq C. \end{aligned}$$

Is this problem convex? Why or why not?

(b) Now, consider the semidefinite problem

$$\begin{aligned} \max_{X \in \mathbb{S}^n} \ & \text{tr} AX \\ \text{subject to } & X \succeq 0 \\ & \text{tr} X = 1 \\ & \sum_{i,j} |X_{ij}| \leq C^2. \end{aligned}$$

What is the relationship between this problem and the problem from part (a)? (Hint: what is the relationship between x and X that makes the objectives of the two problems equal?)

(c) Form the logarithmic barrier for the constraint $\sum_{i,j} |X_{ij}| \leq C^2$ and use this along with the logarithmic barrier for $X \succeq 0$ ($\phi(X) = \log\det X$) to modify the problem in part (b) to have a smooth objective and equality constraint.

(d) Setup and describe the iterations of the barrier method for the problem in part (c), explicitly deriving the Newton update for $X^{(k)}$.

4 Barrier method for ℓ_1 SVM [Shashank]

In Question 4 of the previous assignment, we solved the standard SVMs using a QP solver. The standard SVM solves the following optimization:

$$\min_{\beta,\beta_0} \quad \sum_{i=1}^{n} \Big[1 - y_i\big(x_i^T\beta + \beta_0\big)\Big]_+ + \lambda\|\beta\|_2^2$$

In this question, we will be solving variant of this for the same (newsgroup) data as you used in the previous assignment. This version of the SVM also uses a hinge loss, but replaces the ridge regularizer for the weight-vector with a ℓ_1 penalty:

$$\min_{\beta,\beta_0} \quad \sum_{i=1}^{n} \Big[1 - y_i\big(x_i^T\beta + \beta_0\big)\Big]_+ + \lambda\|\beta\|_1 \tag{1}$$

1. Formulate the ℓ_1 SVM problem as a linear program. (Hint: you will need to introduce additional variables for this.)

2. Implement the barrier method to solve the LP you formulated in the last part, on the newsgroup data (`news.mat`). You must use Newton's method for the centering step (inner loops). Your procedure for the barrier method should take as parameters the initial value of the barrier parameter t_0, the update factor μ for the barrier parameter, α and β for back-tracking in the Newton step. Use $\lambda = 0.25$.

   Here are some notes and suggestions for the barrier method implementation.

   - The barrier method requires an initial strictly feasible point to start from. You can do this by solving a very simple feasibility program, as discussed in the lecture. Note that this feasibility program is indeed itself an LP; for this, you can either run the barrier method until you have a feasible point, or simply solve the feasibility LP directly using e.g., `linprog()` in MATLAB or the `lpsolve()` in R.
   - Suggested parameter values for the log-barrier method: $t_0 = 5$, $\mu = 30$.
   - Suggested parameter values of Newton method backtracking: $\alpha = 0.1$, $\beta = 0.8$.
   - Log-barrier stopping threshold $\epsilon_{outer} = 10^{-6}$ (stop when duality gap $m/t < \epsilon_{outer}$), and inner iterations stopping tolerance: $\epsilon_{inner} = 10^{-6}$ (stop when relative change in barrier objective is less than ϵ_{outer}).

   (Hint: For the log-barrier method, you must ensure that all prospective iterates in the procedure are strictly feasible. Before computing the objective at any point in Newton method during backtracking, check for feasibility. If a candidate point is infeasible, it means that the current step-size would take you outside the feasible region, and must be shrunk until you are in the feasible region.)

3. What is the optimal value for the actual objective (not the log-barrier objective) that you get from the barrier method? Show a plot of the actual objective value against cumulative number of inner (Newton) iterations in your procedure.

4. For the optimal weight vector, what are the classification accuracies on the training set and the test set? How does this compare to the performance of the standard SVM, from the last homework?

5. How many support vectors in the training set does the optimal solution have?

6. How many nonzero components does the optimal weight vector have (considering weights with absolute values below 10^{-6} to be zero)?

5 Primal-dual interior-point method for ℓ_1 SVM [Hanzhang]

In Question 3, you have derived an equivalent linear program for the ℓ_1 SVM problem, and solved it using the barrier method. Here, we will solve ℓ_1 SVM with the primal-dual interior-point method.

1. Rewrite the ℓ_1 SVM problem as a standard form linear program. What is the relationship of the soluion in this problem with β, the solution in the original ℓ_1 SVM problem?

2. Solve this standard form LP with a primal-dual interior-point method, and use this to compute the solution in the original problem, β. Your final solution should contain your code, the optimal value of the objective (of the standard form LP version of ℓ_1 SVM), the number of the nonzero components in β, and the indices and values of the first 30 components that have absolute value of greater than 0.01, not including the intercept β_0.

   Here are some suggestions and notes for the implementation of the primal-dual interior-point method.

   - Similar to the barrier method, you will need to find a strictly feasible initial solution.
   - During backtracking line search for seeking update step size, it is important to check feasibility. If the solution is no longer feasible then the step size is too large. Standard form makes checking feasibility easy.
   - A suggested value of the barrier parameter μ is 2.
   - Since the feature matrix is sparse, you should use sparse representation of matrices in your impementation to reduce to computational complexity.