Homework 4

Convex Optimization 10-725/36-725

Due Thursday March 19 at 4:00pm
submitted to Mallory Deptola in GHC 8001
(make sure to submit each problem separately)

1 Convergence Analysis of Univariate Newton Method [Nicole]

Under certain restrictions on the function being minimized, we can provide a quadratic convergence guarantee for pure Newton's method (i.e., without backtracking).

Assume that we are minimizing a triply smooth function $f : [a, b] \to \mathbb{R}$, and that the following is true for all $x \in [a, b]$:

$$|f''(x)| > C_1 > 0$$
$$|f'''(x)| < C_2$$

Show that these conditions are sufficient for proving a quadratic convergence rate. I.e. show that:

$$|x^{(k+1)} - x^\star| \leq \frac{C_2}{2C_1}|x^{(k)} - x^\star|^2,$$

where $x^\star \in (a, b)$ is the global minimizer of f, and $x^{(k)} \in (a, b)$ is the estimate after k Newton iterations.

(Hint: you may want to Taylor expand.)

2 Iteratively Reweighted Least Squares [Veeru]

How to fit generalized linear models (GLMs)? Given samples $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, consider an exponential family model, wherein $y_i|x_i$ follows a density

$$f(y_i; \theta_i) = \exp\Big(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big).$$

Here θ_i is called the natural parameter, ϕ is called the dispersion parameter, and a, b, c are functions. In a generalized linear model, we assume that $\theta_i = x_i^T\beta$ for each i, a linear function of the predictor variables.

(a) Assuming independent samples, and ϕ known, write down the (conditional) likelihood as a function of β. Prove that maximizing the likelihood over β is equivalent to the following problem

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \big(- y_i x_i^T\beta + b(x_i^T\beta)\big). \tag{1}$$

What is a sufficient condition for this to be a convex problem? Be as general as possible.

(b) Write down the Newton's method for this problem assuming that the second derivative of b exists and is positive everywhere. Assume pure Newton's method, so the step size is always $t = 1$. Show that it takes the form of iteratively reweighted least squares.

(c) Assume that we further have a constraint on the absolute value of θ_i and we want to solve the following problem instead:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \sum_{i=1}^{n} \big(- y_i x_i^T \beta + b(x_i^T \beta) \big) \\ \text{subject to} \quad & |x_i^T \beta| \leq u_i, \ i = 1, \dots n \end{aligned}$$

where u_i are fixed positive numbers. Write down the log barrier function, and write down the steps of the barrier method here. Note that we require backtracking here, i.e., the step-size is no longer 1. Does this still take the form of some kind of least squares?

3 Poisson Interior Point Method [Mattia & Junier]

In this exercise you will fit a Poisson regression model with linear constraints on the mean. Remember that in a Poisson regression model, we are modeling the mean of the Poisson distributed $y_i|x_i$ as $\mu_i = e^{x_i^T \beta}$. One only needs to set $b(w) = e^w$ (so that $b(x_i^T \beta) = e^{x_i^T \beta}$) in equation (1) of Question 2. You will model the expected number of customers shopping in a store as a function of the following covariates: the time of the day, the weather on that day and a variable indicating whether or not the store was offering a special deal.

The `poisLogBarrier` data files contains a dataset where the response variable (`count`) corresponds to the number of people that are shopping in the store at a given time of the day (from 8 AM to 8 PM). In the dataset, time is recorded using the 24 hour format in 13 dummy variables (one for each hour from 8AM to 8PM). Two additional covariates are available: the binary variable `weather` (1: sunny, 0: not sunny) and the variable `offer` (1: the store offered a special discount/deal on the day of the observation, 0: there were no special deals on the day of the observation). The covariates are stored in the matrix `X`. You are asked to fit a Poisson regression model on these data with

$$\log \mu_i = x_i^T \beta = \beta_{\texttt{weather}} * \texttt{weather}_i + \beta_{\texttt{offer}} * \texttt{offer}_i + \sum_{t=8}^{20} \beta_t * \mathbb{1}(t_i = t),$$

where $\mathbb{1}(t_i = t) = 1$ if the time t_i corresponding to the $i-th$ observation is equal to t and $\mathbb{1}(t_i = t) = 0$ otherwise. The vector of coefficients β is thus a vector of length 15 (1 coefficients for `weather`, 1 coefficients for `offer`, and 1 coefficient for each hour of the day between 8 AM and 8 PM).

Based on her long experience in the store, the store manager has a belief on how the expected count of customers varies as a function of time during the day. The belief of the store manager can be expressed in terms of the band

$$\log \mu_i \in [l_i, u_i],$$

where the bounds l_i and u_i are provided as vectors in the `poisLogBarrier` dataset (`belief.lower` and `belief.upper`, respectively).

Fit the Poisson regression model under the above linear (with respect to β) inequality constraints using either a barrier method or a primal dual interior point method. **Your final solution must contain your code, the optimal value of the objective, and the final vector of fitted coefficients.**

Here are some suggestions regarding the tuning parameters if you are using the log-barrier method:

- $t_0 = 5$ – initial t value for the log-barrier penalty parameter
- $\mu = 2$ – coefficient for the update of the log-barrier penalty parameter

- $\epsilon_{\text{in}} = 10^{-6}$ – relative tolerance for the objective within the Newton's method loop (i.e. check convergence in Newton's method using $|f(x_i) - f(x_{i-1})|/|f(x_{i-1})| \leq \epsilon_{\text{in}}$)
- $\epsilon_{\text{out}} = 10^{-6}$ – threshold for the log-barrier algorithm stopping rule (i.e. stop when $m/t \leq \epsilon_{\text{out}}$, where m is the number of constraints)
- we recommend setting the damping parameter controlling the step size in Newton's method to 0.9 (see the lecture notes on Newton's method, page 11)
- $\alpha = 0.5$ – Newton's method backtracking parameter α

Notice that

1. you will need to find an initial feasible vector β_0: that requires solving a linear program as discussed in the lecture notes on the Barrier Method;
2. the implementation of the backtracking for interior point methods can be a little tricky: before computing the value of the objective at a new candidate point, make sure that that point is feasible! If the candidate point in the backtracking loop is not feasible, that means that the current stepsize in the is presumably too large and should be shrunk.

4 Binary Image Denoising Revisited [Yu-Xiang]

Recall from Q4 of Homework 3 the binary image denoising problem:

$$\min_{\beta} \sum_{i=1}^{n} \log(1 + e^{-y_i \beta_i}) + \lambda \|D\beta\|_1,$$

where y is an image to be denoised, each component corresponding to a pixel, i.e., $y \in \mathbb{R}^{n_1 n_2}$ is an unraveled version of an $n_1 \times n_2$ image, and each component of β corresponds to the probability that a given pixel in the image is equal to 1. The matrix D chosen so that

$$\|D\beta\|_1 = \sum_{i \sim j} |\beta_i - \beta_j|$$

where $i \sim j$ means that pixels i and j are either horizontally or vertically, adjacent.

Since it is hard to solve the primal problem directly (with our current first- and second-order toolset), Q4 of Homework 3 explored solving the dual problem:

$$\begin{aligned} \min_{u} \quad & \sum_{i=1}^{n} y_i (D^T u)_i \log(y_i (D^T u)_i) + (1 - y_i (D^T u)_i) \log(1 - y_i (D^T u)_i) \\ \text{subject to} \quad & 0 \leq y_i (D^T u)_i \leq 1, \quad i = 1, \dots n, \quad \|u\|_\infty \leq \lambda. \end{aligned}$$

From a dual solution $\hat{u}$, a primal solution $\hat{\beta}$ is given by

$$\hat{\beta}_i = -y_i \log\left(y_i (D^T \hat{u})_i\right) + y_i \log\left(1 - y_i (D^T \hat{u})_i\right), \quad i = 1, \dots n.$$

As the dual constraints were complicated to deal with, we "lifted" them into the dual criterion in what may have looked like a somewhat arbitrary manner. Now we can make this rigorous with barrier functions and interior point methods.

(a) Let $g(u)$ be the dual criterion function, and let $\phi(u)$ be the log barrier function. For the latter, you should expand the constraint $\|u\|_\infty \leq \lambda$ into componentwise constraints on u. Show that

- $\nabla g(u) = Dc(u)$ for some vector c;
- $\nabla^2 g(u) = DW(u)D^T$ for some diagonal matrix W;
- $\nabla \phi(u) = a(u) + Db(u)$ for some vectors a, b;
- $\nabla^2 \phi(u) = U(u) + DV(u)D^T$ for some diagonal matrices U, V.

Hence for some barrier constant τ, when minimizing $\tau g(u) + \phi(u)$, write down the Newton step direction.

(b) Implement the barrier method using Newton's method for the inner loops. Your function for the barrier method should take as inputs (besides the obvious inputs y, D, λ): an initial barrier parameter $\tau^{(0)} > 0$, an update parameter $\mu > 1$ for the barrier parameter, parameters $\gamma_1, \gamma_2 > 0$ for the backtracking in Newton's method, a tolerance $\epsilon_{\text{inner}} > 0$ for the inner loop (we stop when the change in objective values is less than ϵ_{inner}), and a tolerance $\epsilon_{\text{outer}} > 0$ for the outer loop (we stop when the duality gap is less than ϵ_{outer}).

Apply your algorithm on the image data from Q4 of Homework 3, at a tuning parameter value of $\lambda = 0.25$. You can set $\tau^{(0)} = 5$, $\mu = 10$, $\gamma_1 = 0.1$, $\gamma_2 = 0.8$, $\epsilon_{\text{inner}} = 10^{-6}$, $\epsilon_{\text{outer}} = 10^{-6}$. How many outer iterations did it take to converge? How many total inner iterations (Newton steps)? Roughly speaking, how does this compare in terms of the number of iterations, and the computational cost, to your gradient descent implementation from Q4 of Homework 3?

Use the computed dual solution $\hat{u}$ to recover the primal solution $\hat{\beta}$, and then the estimated probabilities $\hat{p}_i = 1/(1 + e^{-\hat{\beta}_i})$ at each pixel $i = 1, \ldots n_1 n_2$. Plot the corresponding binary image, by classifying according to whether or not the predicted probabilities are larger than 0.5.

(Hint 1: for the barrier method, you must begin with a strictly feasible point for the dual problem. But from Q4 of Homework 3, you already have a strictly feasible point for the dual problem at $\lambda = 0.5$. There is an easy way to use such a point to get a strictly feasible point for the dual at $\lambda = 0.25$; this is much easier than recomputing a new strictly feasible point at $\lambda = 0.25$...)

(Hint 2: to make your implementation efficient, you should take advantage of the sparsity of D, when you solve the linear systems at each iteration of Newton's method. In Matlab or R, this is just done by making sure that the Hessian is stored as a sparse matrix.)

(c) If the ith component of the dual solution $\hat{u}$ satisfies $|\hat{u}_i| < \lambda$, then what does this mean about $(D\hat{\beta})_i$ in the primal? Verify your answer empirically, on the computed solution from part (b).

(Hint: recall the construction of the dual problem, by introducing the auxiliary variables $z = D\beta$ in the primal problem; inspect the KKT stationarity condition for z, to a derive a relationship here.)

(d) At the end of each outer iteration in the barrier method application in part (b), transform the dual estimate u into a primal estimate β, and plot this as a grayscale image. Note: this will not be binary, because the components of β are real-valued, but you can still plot it in grayscale over its dynamic range. (If the number of outer iterations required until convergence is large, then just pick estimates from 6 or so outer iterations to display.) What do you notice, as the barrier method proceeds, about the primal estimates?

Bonus: Can you explain the phenomenon from part (d)?