Homework 5

Due: 10am Thursday 11/16/17

1. Proximal Gradient Method
   The proximal operator of a function $r : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as
   \[
   \text{prox}_r(z) = \arg\min_w \left( r(w) + \frac{1}{2}\|w - z\|^2_2 \right).
   \]
   In class, we saw how to use the $\ell_1$ regularizer to encourage sparsity. In this problem, we will see a different regularizer that enforces sparsity.

   (a) Define the $k$-sparse indicator $1_k : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\},$
   \[
   1_k(w) = \begin{cases} 
   0 & \text{nnz}(w) \leq k \\
   \infty & \text{otherwise}
   \end{cases}
   \]
   where $\text{nnz}(w) =$ the number of non-zero entries of $w.$
   Compute the proximal operator of the $k$-sparse indicator $1_k.$

   (b) A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if the line between any two points on the function lies (weakly) above the graph of the function. Formally, $\forall x, y \in \mathbb{R}^d$ and $\forall t \in [0, 1],$
   \[
   f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).
   \]
   Find a counterexample in two dimensions that shows the 1-sparse indicator is not convex. (That is, $d = 2$ and $k = 1.$)

   (c) Our goal now is to solve the Sparse Least Squares (SLS) problem,
   \[
   \min \|Xw - y\|^2 + 1_k(w).
   \]
   Write pseudocode describing how the proximal gradient method could be used to solve this problem. (That is, write the prox and gradient steps explicitly for this loss function and regularizer.)

   (d) Code the proximal gradient method for the SLS problem and run it on the instance in
You may find the Julia function `sortperm` useful. Recall that the Lipschitz constant of the gradient of the least squares objective is \( L = 2\|X\|^2 \), where \( \|X\| \) is the maximum singular value of \( X \). Make sure to use an appropriate step size to ensure convergence.

Plot the objective value as a function of the number of iterations. You may want to plot the \( y \) axis of the plot on a log scale using `semilogy()` instead of the `plot()` command in the package `PyPlot`. (This is called a convergence plot; it helps us understand how quickly the method finds a solution, and the quality of that solution.)

(e) Run the algorithm starting at multiple locations and create a histogram of the objective value. Use 100 iterations for each run. What do you observe?

(f) Solve the LASSO problem (\( \ell_1 \) regularized least squares regression) using the proximal gradient method on this problem. You may use the code from the demo in class, found at `https://github.com/ORIE4741/demos/blob/master/ProximalGradient.ipynb`.

(g) Does LASSO converge to the same place starting from different initial vectors \( w^0 \)?

(h) Compare the SLS solution with the LASSO solution. Which is more sparse? Which achieves a better objective value? Which method is more reliable?

2. *Stochastic proximal gradient method.*

(a) Write pseudocode for the stochastic proximal gradient method applied to the Sparse Least Squares problem above.

(b) Code the stochastic proximal gradient method for the Sparse Least Squares problem.

(c) Using the same data used in `ProxGradHomework.ipynb`, plot the objective value as a function of the number of iterations.

(d) How long does the stochastic proximal gradient method take compared to the standard proximal gradient method? Compare both the number of iterations and the time required for convergence. You may find Julia’s `@time` macro useful: place it in front of line of code to evaluate the running time of that line.

(e) Run the stochastic algorithm starting at multiple locations and create a histogram of the final objective values. What do you observe?

3. *Loss functions.*

In our first look at regression in this course, we showed how to predict \( y \in \mathbb{R} \) given \( x \in \mathbb{R}^d \) by finding a vector \( w \) minimizing the least squares loss

\[
\|y - Xw\|^2.
\]
This problem is also called $\ell_2$ regression, and the loss is sometimes also called a quadratic loss. However, now that we have grown more sophisticated both in modeling and in optimization, we understand that the quadratic loss is not always the best choice, and that it can be beneficial to use regularization to ensure model interpretability or to improve generalization.

Please list at least two cases where we should use a loss function that is not quadratic. For each, state the input space $\mathcal{X}$, the output space $\mathcal{Y}$, describe the loss function and regularizer you would use for this problem (and, optionally, any feature transformations), and explain why your choice of loss function and regularizer make sense for this problem. Feel free to use a problem you’ve encountered in your class project.

4. Quantile Regression. You can find data and starter code for this problem in the Jupyter notebook quantileRegression.ipynb available at

http://github.com/orie4741/homework/blob/master/quantileRegression.ipynb

We will be using a random sample of datapoints taken from the 2015 Natality Data of the National Center for Health Statistics. We are interested in investigating the effect of gender, mother’s marital status, and prenatal care in the first trimester on the baby’s birth weight.

(a) Fit an ordinary least squares regression to the data. Interpret the coefficients that you find.

(b) Fit a quantile regression on the data for the 5th quantile $q = 0.05$ and for the 95th quantile $q = 0.95$. What do these models predict, and how does it differ from the prediction of the least squares regression? Compare these coefficients to those you found in part a).

*Hint:* if you’re having trouble fitting this using the proximal gradient method, try starting with a larger stepsize like 10 or 100.

(c) Fit quantile regressions for $q = 0.05, 0.10, \ldots, 0.95$.

(d) Create an intercept plot that plots quantiles against the intercept coefficient from that quantile regression. Create coefficient plots for MaritalStatus, Male, and PrenatalCare coefficients. How do the coefficients change as the quantile increases?

(e) What is the meaning of the intercepts of the quantile regressions?

(f) What does the coefficient plot tell you about the effect of prenatal care for infants with low birth weight compared to those with average birth weights?